Functions Preserving Matrix Groups and Iterations for the Matrix Square Root

Higham, Nicholas J. and Mackey,
D. Steven and Mackey, Niloufer and Tisseur, Françoise

2005

MIMS EPrint: **2005.8**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: http://eprints.maths.manchester.ac.uk/
And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

FUNCTIONS PRESERVING MATRIX GROUPS AND ITERATIONS FOR THE MATRIX SQUARE ROOT*

NICHOLAS J. HIGHAM†, D. STEVEN MACKEY‡, NILOUFER MACKEY§, AND FRANÇOISE TISSEUR¶

Abstract. For which functions f does $A \in \mathbb{G} \Rightarrow f(A) \in \mathbb{G}$ when $\mathbb{G}$ is the matrix automorphism group associated with a bilinear or sesquilinear form? For example, if A is symplectic when is $f(A)$ symplectic? We show that group structure is preserved precisely when $f(A^{-1}) = f(A)^{-1}$ for bilinear forms and when $f(A^{-*}) = f(A)^{-*}$ for sesquilinear forms. Meromorphic functions that satisfy each of these conditions are characterized. Related to structure preservation is the condition $f(\overline{A}) = \overline{f(A)}$, and analytic functions and rational functions satisfying this condition are also characterized. These results enable us to characterize all meromorphic functions that map every $\mathbb{G}$ into itself as the ratio of a polynomial and its "reversal", up to a monomial factor and conjugation.

The principal square root is an important example of a function that preserves every automorphism group $\mathbb{G}$. By exploiting the matrix sign function, a new family of coupled iterations for the matrix square root is derived. Some of these iterations preserve every $\mathbb{G}$; all of them are shown, via a novel Fréchet derivative-based analysis, to be numerically stable.

A rewritten form of Newton's method for the square root of $A \in \mathbb{G}$ is also derived. Unlike the original method, this new form has good numerical stability properties, and we argue that it is the iterative method of choice for computing $A^{1/2}$ when $A \in \mathbb{G}$. Our tools include a formula for the sign of a certain block 2×2 matrix, the generalized polar decomposition along with a wide class of iterations for computing it, and a connection between the generalized polar decomposition of $I + A$ and the square root of $A \in \mathbb{G}$.

Key words. automorphism group, bilinear form, sesquilinear form, scalar product, adjoint, Fréchet derivative, stability analysis, perplectic matrix, pseudo-orthogonal matrix, Lorentz matrix, generalized polar decomposition, matrix sign function, matrix pth root, matrix square root, structure preservation, matrix iteration, Newton iteration

AMS subject classifications. 65F30, 15A18

1. Introduction.

Theory and algorithms for structured matrices are of growing interest because of the many applications that generate structure and the potential benefits to be gained by exploiting it. The benefits include faster and more accurate algorithms as well as more physically meaningful solutions. Structure comes in many forms, including Hamiltonian, Toeplitz or Vandermonde structure and total positivity. Here we study a nonlinear structure that arises in a variety of important applications and has an elegant mathematical formulation: that of a matrix automorphism group $\mathbb{G}$ associated with a bilinear or sesquilinear form.

Our particular interest is in functions that preserve matrix automorphism group structure. We show in Section 3 that $A \in \mathbb{G} \Rightarrow f(A) \in \mathbb{G}$ precisely when $f(A^{-1}) =$

*Numerical Analysis Report 446, Manchester Centre for Computational Mathematics, June 2004.

†Department of Mathematics, University of Manchester, Manchester, M13 9PL, England (higham@ma.man.ac.uk, http://www.ma.man.ac.uk/~higham/). This work was supported by Engineering and Physical Sciences Research Council grant GR/R22612 and by a Royal Society-Wolfson Research Merit Award.

‡Department of Mathematics, University of Manchester, Manchester, M13 9PL, England (smackey@ma.man.ac.uk). This work was supported by Engineering and Physical Sciences Research Council grant GR/S31693.

§Department of Mathematics, Western Michigan University, Kalamazoo, MI 49008, USA (nil.mackey@wmich.edu, http://homepages.wmich.edu/~mackey/).

¶Department of Mathematics, University of Manchester, Manchester, M13 9PL, England (ftisseur@ma.man.ac.uk, http://www.ma.man.ac.uk/~ftisseur/). This work was supported by Engineering and Physical Sciences Research Council grant GR/R45079.

$f(A)^{-1}$ for bilinear forms or $f(A^{-*}) = f(A)^{-*}$ for sesquilinear forms; in other words, f has to commute with the inverse function or the conjugate inverse function at A. We characterize meromorphic functions satisfying each of these conditions. For sesquilinear forms, the condition $f(\overline{A}) = \overline{f(A)}$, that is, f commutes with conjugation, also plays a role in structure preservation. 0We characterize analytic functions and rational functions satisfying this conjugation condition. We show further that any meromorphic function that is structure preserving for all automorphism groups is rational and, up to a monomial factor and conjugation, the ratio of a polynomial and its "reversal".

The matrix sign function and the matrix principal pth root are important examples of functions that preserve all automorphism groups. Iterations for computing the sign function in a matrix group were studied by us in [15]. We concentrate here on the square root, aiming to derive iterations that exploit the group structure. Connections between the matrix sign function, the matrix square root, and the generalized polar decomposition are developed in Section 4. A new identity for the matrix sign function (Lemma 4.3) establishes a link with the generalized polar decomposition (Corollary 4.4). For $A \in \mathbb{G}$ we show that the generalized polar decomposition of $I + A$ has $A^{1/2}$ as the factor in $\mathbb{G}$, thereby reducing computation of the square root to computation of the generalized polar decomposition (Theorem 4.7).

A great deal is known about iterations for the matrix sign function. Our results in Section 4 show that each matrix sign function iteration of a general form leads to two further iterations:

- A coupled iteration for the principal square root of any matrix A. The iteration is structure preserving, in the sense that $A \in \mathbb{G}$ implies all the iterates lie in $\mathbb{G}$, as long as the underlying sign iteration is also structure preserving.
- An iteration for the generalized polar decomposition, and hence for the square root of $A \in \mathbb{G}$.

Iterations for matrix roots are notorious for their tendency to be numerically unstable. In Section 5 Fréchet derivatives are used to develop a stability analysis of the coupled square root iterations that arise from superlinearly convergent sign iterations. We find that all such iterations are stable, but that a seemingly innocuous rewriting of the iterations can make them unstable. The technique developed in this section should prove to be of wider use in analyzing matrix iterations.

In Section 6 two instances of the connections identified in Section 4 between the sign function and the square root are examined in detail. We obtain a family of coupled structure-preserving iterations for the square root whose members have order of convergence $2m + 1$ for $m = 1, 2, \ldots$. We also derive a variant for $A \in \mathbb{G}$ of the well-known but numerically unstable Newton iteration for $A^{1/2}$ by using the connection with the generalized polar decomposition. Our numerical experiments and analysis in Section 7 confirm the numerical stability of both the structure-preserving iterations and the Newton variant, showing both to be useful in practice. Because the Newton variant has a lower cost per iteration and shows better numerical preservation of structure, it is our preferred method in general.

2. Preliminaries. We give a very brief summary of the required definitions and notation. For more details, see Mackey, Mackey, and Tisseur [25].

Consider a scalar product on $\mathbb{K}^n$, that is, a bilinear or sesquilinear form $\langle \cdot, \cdot \rangle_M$ defined by any nonsingular matrix M: for $x, y \in \mathbb{K}^n$,

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

Here $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$ and the superscript $*$ denotes conjugate transpose. The associated automorphism group is defined by

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

The adjoint $A^\star$ of $A \in \mathbb{K}^{n\times n}$ with respect to $\langle\cdot,\cdot\rangle_{\mathrm{M}}$ is the unique matrix satisfying

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

It can be shown that the adjoint is given explicitly by

$$A^\star = \begin{cases} M^{-1}A^TM, & \text{for bilinear forms,} \\ M^{-1}A^*M, & \text{for sesquilinear forms} \end{cases} \tag{2.1}$$

and has the following basic properties:

$$(A+B)^\star = A^\star + B^\star, \quad (AB)^\star = B^\star A^\star, \quad (A^{-1})^\star = (A^\star)^{-1},$$
$$(\alpha A)^\star = \begin{cases} \alpha A^\star, & \text{for bilinear forms,} \\ \overline{\alpha} A^\star, & \text{for sesquilinear forms.} \end{cases}$$

The automorphism group can be characterized in terms of the adjoint by

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

Table 2.1 lists some of the "classical" matrix groups. Observe that M, the matrix of the form, is real orthogonal with $M = \pm M^T$ in all these examples. Our results, however, place no restrictions on M other than nonsingularity; they therefore apply to all scalar products on $\mathbb{R}^n$ or $\mathbb{C}^n$ and their associated automorphism groups.

We note for later use that

$$A \in \mathbb{G} \text{ and } M \text{ unitary} \quad \Rightarrow \quad \|A\|_2 = \|A^{-1}\|_2. \tag{2.2}$$

We recall one of several equivalent ways of defining $f(A)$ for $A \in \mathbb{C}^{n\times n}$, where f is an underlying scalar function. Let A have distinct eigenvalues $\lambda_1, \dots, \lambda_s$ occurring in Jordan blocks of maximum sizes $n_1, \dots, n_s$, respectively. Thus if A is diagonalizable, $n_i \equiv 1$. Then $f(A) = q(A)$, where q is the unique Hermite interpolating polynomial of degree less than $\sum_{i=1}^s n_i$ that satisfies the interpolation conditions

$$q^{(j)}(\lambda_i) = f^{(j)}(\lambda_i), \qquad j = 0\colon n_i - 1, \quad i = 1\colon s. \tag{2.3}$$

Stated another way, q is the Hermite interpolating polynomial of minimal degree that interpolates f at the roots of the minimal polynomial of A. We use the phrase *f is defined on the spectrum of A* or, for short, *f is defined at A* or *A is in the domain of f*, to mean that the derivatives in (2.3) exist.

At various points in this work the properties $f\big(\mathrm{diag}(X_1, X_2)\big) = \mathrm{diag}(f(X_1), f(X_2))$ and $f(P^{-1}AP) = P^{-1}f(A)P$, which hold for any matrix function [24, Thms. 9.4.1, 9.4.2], will be used. We will also need the following three results.

LEMMA 2.1. *Let $A, B \in \mathbb{C}^{n\times n}$ and let f be defined on the spectrum of both A and B. Then there is a single polynomial p such that $f(A) = p(A)$ and $f(B) = p(B)$.*

Proof. It suffices to let p be the polynomial that interpolates f and its derivatives at the roots of the least common multiple of the minimal polynomials of A and B. See the discussion in [16, p. 415]. ∎

TABLE 2.1
A sampling of automorphism groups.

Here, $R = \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix}$, $J = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix}$, $\Sigma_{p,q} = \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix} \in \mathbb{R}^{n\times n}$.

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

[a] Also known as Lorentz matrices.

COROLLARY 2.2. *Let $A, B \in \mathbb{C}^{n\times n}$ and let f be defined on the spectra of both AB and BA. Then*

$$Af(BA) = f(AB)A.$$

Proof. By Lemma 2.1 there is a single polynomial p such that $f(AB) = p(AB)$ and $f(BA) = p(BA)$. Hence

$$Af(BA) = Ap(BA) = p(AB)A = f(AB)A. \qquad \square$$

LEMMA 2.3. *Any rational function r can be uniquely represented in the form $r(z) = z^n p(z)/q(z)$, where p is monic, n is an integer, p and q are relatively prime, and $p(0)$ and $q(0)$ are both nonzero.*

Proof. Straightforward. $\square$

We denote the closed negative real axis by $\mathbb{R}^-$. For $A \in \mathbb{C}^{n\times n}$ with no eigenvalues on $\mathbb{R}^-$, the principal matrix pth root $A^{1/p}$ is defined by the property that the eigenvalues of $A^{1/p}$ lie in the segment $\{\, z : -\pi/p < \arg(z) < \pi/p \,\}$. We will most often use the principal square root, $A^{1/2}$, whose eigenvalues lie in the open right half-plane.

Finally, we introduce some notation connected with a polynomial p. The polynomial obtained by replacing the coefficients of p by their conjugates is denoted by $\overline{p}$. The polynomial obtained by reversing the order of the coefficients of p is denoted by $\mathrm{rev}p$; thus if p has degree m then

$$\mathrm{rev}p(x) = x^m p(1/x). \tag{2.4}$$

3. Structure-preserving functions. Our aim in this section is to characterize functions f that preserve automorphism group structure. For a given $\mathbb{G}$, if $f(A) \in \mathbb{G}$ for all $A \in \mathbb{G}$ for which $f(A)$ is defined, we will say that f *is structure preserving for* $\mathbb{G}$. As well as determining f that preserve structure for a particular $\mathbb{G}$, we wish to determine f that preserve structure *for all* $\mathbb{G}$.

The (principal) square root is an important example of a function that preserves all groups. To see this for $\mathbb{G}$ associated with a bilinear form, recall that $A \in \mathbb{G}$ is equivalent to $M^{-1}A^TM = A^{-1}$. Assuming that A has no eigenvalues on $\mathbb{R}^-$, taking the (principal) square root in this relation gives

$$A^{-1/2} = (M^{-1}A^TM)^{1/2} = M^{-1}\big(A^T\big)^{1/2}M = M^{-1}\big(A^{1/2}\big)^TM, \tag{3.1}$$

which shows that $A^{1/2} \in \mathbb{G}$.

In order to understand structure preservation we need first to characterize when $f(A) \in \mathbb{G}$ for a fixed f and a fixed $A \in \mathbb{G}$. The next result relates this property to various other relevant properties of matrix functions.

THEOREM 3.1. *Let $\mathbb{G}$ be the automorphism group of a scalar product. Consider the following eight properties of a matrix function f at a (fixed) matrix $A \in \mathbb{K}^{n\times n}$, where f is assumed to be defined at the indicated arguments:*

(a)	$f(A^T) = f(A)^T$,	(e)	$f(A^{-1}) = f(A)^{-1}$,
(b)	$f(A^*) = f(A)^*$,	(g)	$f(A^{-*}) = f(A)^{-*}$,
(c)	$f(\overline{A}) = \overline{f(A)}$,	(h)	$f(A^{-\star}) = f(A)^{-\star}$,
(d)	$f(A^\star) = f(A)^\star$,	(i)	*when* $A \in \mathbb{G}$, $f(A) \in \mathbb{G}$.

(a) *always holds.* (b) *is equivalent to* (c). (c) *is equivalent to the existence of a single* real *polynomial p such that $f(A) = p(A)$ and $f(\overline{A}) = p(\overline{A})$. Moreover,*

- for bilinear forms: (d) *always holds.* (e), (h) *and* (i) *are equivalent;*
- for sesquilinear forms: (d) *is equivalent to* (b) *and to* (c). (g), (h) *and* (i) *are equivalent. Any two of* (d) *for A^{-1} and* (e) *and* (h) *for A imply the third*[1].

Proof. (a) follows because the same polynomial can be used to evaluate $f(A)$ and $f(A^T)$, by Lemma 2.1. Property (b) is equivalent to $f(\overline{A}^T) = \overline{f(A)}^T$, which on applying (a) becomes (c). So (b) is equivalent to (c).

Next, we consider the characterization of (c). Suppose p is a real polynomial such that $f(A) = p(A)$ and $f(\overline{A}) = p(\overline{A})$. Then $f(\overline{A}) = p(\overline{A}) = \overline{p(A)} = \overline{f(A)}$, which is (c). Conversely, assume (c) holds and let q be any complex polynomial that simultaneously evaluates f at A and $\overline{A}$, so that $f(A) = q(A)$ and $f(\overline{A}) = q(\overline{A})$; the existence of such a q is assured by Lemma 2.1. Then

$$q(A) = f(A) = \overline{f(\overline{A})} = \overline{q(\overline{A})} = \overline{q}(A),$$

and hence $p(x) := \frac{1}{2}(q(x) + \overline{q}(x))$ is a *real* polynomial such that $f(A) = p(A)$. Since

$$q(\overline{A}) = f(\overline{A}) = \overline{f(A)} = \overline{q(A)} = \overline{q}(\overline{A}),$$

we also have $f(\overline{A}) = p(\overline{A})$.

We now consider (d). From the characterization (2.1) of the adjoint, for bilinear forms we have

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

so (d) always holds. For sesquilinear forms,

$$f(A^\star) = f(M^{-1}A^*M) = M^{-1}f(A^*)M,$$

[1] We will see at the end of Section 3.4 that for sesquilinear forms neither (d) for A^{-1} (equivalently (c) for A^{-1}) nor (e) is a necessary condition for (h) (or, equivalently, for the structure-preservation property (i)).

which equals $f(A)^\star = M^{-1}f(A)^*M$ if and only if (b) holds.

To see that (h) and (i) are equivalent, consider the following cycle of potential equalities:

$$\begin{array}{ccc} f(A^{-\star}) & \overset{\text{(h)}}{=\!=} & f(A)^{-\star} \\ {\scriptstyle \text{(x)}}\,\backslash\backslash & & //\,{\scriptstyle \text{(y)}} \\ & f(A) & \end{array}$$

Clearly (x) holds if $A \in \mathbb{G}$, and (y) holds when $f(A) \in \mathbb{G}$. Hence (h) and (i) are equivalent.

For bilinear forms,

$$\begin{aligned} f(A^{-\star}) = f(A)^{-\star} &\iff f(M^{-1}A^{-T}M) = M^{-1}f(A)^{-T}M \\ &\iff f(A^{-T}) = f(A)^{-T} \\ &\iff f(A^{-1}) = f(A)^{-1} \quad \text{by (a).} \end{aligned}$$

Thus (h) is equivalent to (e). For sesquilinear forms a similar argument shows that (h) is equivalent to (g).

Finally, it is straightforward to show for sesquilinear forms that any two of (d) for A^{-1} and (e) and (h) for A imply the third. □

The main conclusion of Theorem 3.1 is that f is structure preserving for $\mathbb{G}$ precisely when $f(A^{-1}) = f(A)^{-1}$ for all $A \in \mathbb{G}$ for bilinear forms, or $f(A^{-*}) = f(A)^{-*}$ for all $A \in \mathbb{G}$ for sesquilinear forms. We can readily identify two important functions that satisfy both these conditions more generally for all $A \in \mathbb{K}^{n\times n}$ in their domains, and hence are structure preserving *for all* $\mathbb{G}$.

- The matrix sign function. Recall that for a matrix $A \in \mathbb{C}^{n\times n}$ with no pure imaginary eigenvalues the sign function can be defined by $\operatorname{sign}(A) = A(A^2)^{-1/2}$ [12], [22]. That the sign function is structure preserving is known: proofs specific to the sign function are given in [15] and [26].
- Any matrix power A^α, subject for fractional α to suitable choice of the branches of the power at each eigenvalue; in particular, the principal matrix pth root $A^{1/p}$. The structure-preserving property of the principal *square* root is also shown by Mackey, Mackey, and Tisseur [26].

In the following three subsections we investigate three of the properties in Theorem 3.1 in detail, for general matrices $A \in \mathbb{C}^{n\times n}$. Then in the final two subsections we characterize meromorphic structure-preserving functions and conclude with a brief consideration of M-normal matrices.

3.1. Property (c): $f(\overline{A}) = \overline{f(A)}$. Theorem 3.1 shows that this property for A^{-1}, together with property (e), namely $f(A^{-1}) = f(A)^{-1}$, is sufficient for structure preservation in the sesquilinear case. While property (c) is not necessary for structure preservation, it plays an important role in our understanding of the preservation of realness, and so is of independent interest.

We first give a characterization of analytic functions satisfying property (c) for all A in their domain, followed by an explicit description of all rational functions with the property. We denote by $\Lambda(A)$ the set of eigenvalues of A.

THEOREM 3.2. *Let f be analytic on an open subset $\Omega \subseteq \mathbb{C}$ such that each connected component of Ω is closed under conjugation. Consider the corresponding*

matrix function f on its natural domain in $\mathbb{C}^{n\times n}$, the set $\mathcal{D} = \{\, A \in \mathbb{C}^{n\times n} : \Lambda(A) \subseteq \Omega \,\}$. Then the following are equivalent:

(a) $f(\overline{A}) = \overline{f(A)}$ *for all* $A \in \mathcal{D}$.

(b) $f(\mathbb{R}^{n\times n} \cap \mathcal{D}) \subseteq \mathbb{R}^{n\times n}$.

(c) $f(\mathbb{R} \cap \Omega) \subseteq \mathbb{R}$.

Proof. Our strategy is to show that (a) $\Rightarrow$ (b) $\Rightarrow$ (c) $\Rightarrow$ (a).

(a) $\Rightarrow$ (b): If $A \in \mathbb{R}^{n\times n} \cap \mathcal{D}$ then

$$\begin{aligned} f(A) &= f(\overline{A}) \quad (\text{since } A \in \mathbb{R}^{n\times n}) \\ &= \overline{f(A)} \quad (\text{given}), \end{aligned}$$

so $f(A) \in \mathbb{R}^{n\times n}$, as required.

(b) $\Rightarrow$ (c): If $\lambda \in \mathbb{R} \cap \Omega$ then $\lambda I \in \mathcal{D}$. But $f(\lambda I) \in \mathbb{R}^{n\times n}$ by (b), and hence, since $f(\lambda I) = f(\lambda)I$, $f(\lambda) \in \mathbb{R}$.

(c) $\Rightarrow$ (a): Let $\widetilde{\Omega}$ be any connected component of Ω. Since $\widetilde{\Omega}$ is open and connected it is path-connected, and since it is also closed under conjugation it must contain some $\lambda \in \mathbb{R}$ by the intermediate value theorem. The openness of $\widetilde{\Omega}$ in $\mathbb{C}$ then implies that $U = \widetilde{\Omega} \cap \mathbb{R}$ is a nonempty open subset of $\mathbb{R}$, with $f(U) \subseteq \mathbb{R}$ by hypothesis. Now since f is analytic on $\widetilde{\Omega}$, it follows from the "identity theorem" [27, pp. 227–236 and Ex. 4, p. 236] that $f(\overline{z}) = \overline{f(z)}$ for all $z \in \widetilde{\Omega}$. The same argument applies to all the other connected components of Ω, so $f(\overline{z}) = \overline{f(z)}$ for all $z \in \Omega$. Thus $f(\overline{A}) = \overline{f(A)}$ holds for all diagonal matrices in $\mathcal{D}$, and hence for all diagonalizable matrices in $\mathcal{D}$. Since the scalar function f is analytic on Ω, the matrix function f is continuous on $\mathcal{D}$ [16, Thm. 6.2.27][2] and therefore the identity holds for all matrices in $\mathcal{D}$, since diagonalizable matrices are dense in any open subset of $\mathbb{C}^{n\times n}$. ∎

Turning to the case when f is rational, we need a preliminary lemma.

LEMMA 3.3. *Suppose r is a complex rational function that maps all reals (in its domain) to reals. Then r can be expressed as the ratio of two real polynomials. In particular, in the canonical form for r given by Lemma* 2.3 *the polynomials p and q are both real.*

Proof. Let $r(z) = z^n p(z)/q(z)$ be the canonical form of Lemma 2.3 and consider the rational function

$$h(z) := \overline{r(\overline{z})} = z^n \frac{\overline{p}(z)}{\overline{q}(z)}.$$

Clearly, $h(z) = r(z)$ for all real z in the domain of r, and hence $p(z)/q(z) = \overline{p}(z)/\overline{q}(z)$ for this infinitude of z. It is then straightforward to show (cf. the proof of Lemma 3.6 below) that $p = \alpha\overline{p}$ and $q = \alpha\overline{q}$ for some nonzero $\alpha \in \mathbb{C}$. But the monicity of p implies that $\alpha = 1$, so p and q are real polynomials. ∎

Combining Lemma 3.3 with Theorem 3.2 gives a characterization of all rational matrix functions with property (c) in Theorem 3.1.

THEOREM 3.4. *A rational matrix function $r(A)$ has the property $r(\overline{A}) = \overline{r(A)}$ for all $A \in \mathbb{C}^{n\times n}$ such that A and $\overline{A}$ are in the domain of r if and only if the scalar function r can be expressed as the ratio of two real polynomials. In particular, the polynomials p satisfying $p(\overline{A}) = \overline{p(A)}$ for all $A \in \mathbb{C}^{n\times n}$ are precisely those with real coefficients.*

[2] Horn and Johnson require that Ω should be a *simply-connected* open subset of $\mathbb{C}$. However, it is not difficult to show that just the openness of Ω is sufficient to conclude that the matrix function f is continuous on $\mathcal{D}$.

3.2. Property (e): $f(A^{-1}) = f(A)^{-1}$. We now investigate further the property $f(A^{-1}) = f(A)^{-1}$ for matrix functions f. We would like to know when this property holds for all A such that A and A^{-1} are in the domain of f. Since a function of a diagonal matrix is diagonal, a necessary condition on f is that $f(z)f(1/z) = 1$ whenever z and $1/z$ are in the domain of f. The following result characterizes meromorphic functions satisfying this identity. Recall that a function is said to be meromorphic on an open subset $U \subseteq \mathbb{C}$ if it is analytic on U except for poles. In this paper we only consider meromorphic functions on $\mathbb{C}$, so the phrase "f is meromorphic" will mean f is meromorphic on $\mathbb{C}$.

LEMMA 3.5. *Suppose f is a meromorphic function on $\mathbb{C}$ such that $f(z)f(1/z) = 1$ holds for all z in some infinite compact subset of $\mathbb{C}$. Then*

(a) *The identity $f(z)f(1/z) = 1$ holds for all nonzero $z \in \mathbb{C} \setminus S$, where S is the discrete set consisting of the zeros and poles of f together with their reciprocals.*

(b) *The zeros and poles of f come in reciprocal pairs $\{a, 1/a\}$ with matching orders. That is,*

$$(3.2)\quad z = a \text{ is a zero (pole) of order } k \iff z = 1/a \text{ is a pole (zero) of order } k.$$

Consequently, the set S is finite and consists of just the zeros and poles of f. Note that $\{0, \infty\}$ is also to be regarded as a reciprocal pair for the purpose of statement (3.2).

(c) *The function f is meromorphic at ∞.*

(d) *The function f is rational.*

Proof. (a) The function $g(z) := f(z)f(1/z)$ is analytic on the open connected set $\mathbb{C} \setminus \{S \cup \{0\}\}$, so the result follows by the identity theorem.

(b) Consider first the case where $a \neq 0$ is a zero or a pole of f. Because f is meromorphic the set S is discrete, so by (a) there is some open neighborhood U of $z = a$ such that $f(z)f(1/z) = 1$ holds for all $z \in U \setminus \{a\}$ *and* such that f can be expressed as $f(z) = (z-a)^k g(z)$ for some nonzero $k \in \mathbb{Z}$ ($k > 0$ for a zero, $k < 0$ for a pole) and some function g that is analytic and nonzero on all of U. Then for all $z \in U \setminus \{a\}$ we have

$$f(1/z) = \frac{1}{f(z)} = (z-a)^{-k}\frac{1}{g(z)}.$$

Letting $w = 1/z$, we see that there is an open neighborhood $\widetilde{U}$ of $w = 1/a$ in which

$$f(w) = \left(\frac{1}{w} - a\right)^{-k} \frac{1}{g(1/w)} = \left(w - \frac{1}{a}\right)^{-k} \frac{(-1)^k w^k}{a^k g(1/w)} =: \left(w - \frac{1}{a}\right)^{-k} h(w)$$

holds for all $w \in \widetilde{U} \setminus \{\frac{1}{a}\}$, where $h(w)$ is analytic and nonzero for all $w \in \widetilde{U}$. This establishes (3.2), and hence that the set S consists of just the zeros and poles of f.

Next we turn to the case of the "reciprocal" pair $\{0, \infty\}$. First note that the zeros and poles of any nonzero meromorphic function can never accumulate at any finite point z, so in particular $z = 0$ cannot be a limit point of S. In our situation the set S also cannot have $z = \infty$ as an accumulation point; if it did, then the reciprocal pairing of the nonzero poles and zeros of f just established would force $z = 0$ to be a limit point of S. Thus if $z = \infty$ is a zero or singularity of f then it must be an isolated zero or singularity, which implies that S is a finite set.

Now suppose $z = 0$ is a zero or pole of f. In some open neighborhood U of $z = 0$ we can write $f(z) = z^k g(z)$ for some nonzero $k \in \mathbb{Z}$ and some g that is analytic and

nonzero on U. Then for all $z \in U \setminus \{0\}$ we have

$$f(1/z) = \frac{1}{f(z)} = z^{-k}\frac{1}{g(z)} =: z^{-k}h(z),$$

where h is analytic and nonzero in U. Thus $z = 0$ being a zero (pole) of f implies that $z = \infty$ is a pole (zero) of f. The converse is established by the same kind of argument.

(c) That f is meromorphic at ∞ follows from (3.2), the finiteness of S, and the identity $f(z)f(1/z) = 1$, together with the fact that f (being meromorphic on $\mathbb{C}$) can have only a pole, a zero, or a finite value at $z = 0$.

(d) By [9, Thm. 4.7.7] a function is meromorphic on $\mathbb{C}$ and at ∞ if and only if it is rational. ∎

Since Lemma 3.5 focuses attention on rational functions, we next give a complete description of all rational functions satisfying the identity $f(z)f(1/z) = 1$. Recall that $\mathrm{rev}p$ is defined by (2.4).

LEMMA 3.6. *A complex rational function $r(z)$ satisfies the identity $r(z)r(1/z) = 1$ for infinitely many $z \in \mathbb{C}$ if and only if it can be expressed in the form*

$$r(z) = \pm z^k \frac{p(z)}{\mathrm{rev}p(z)} \tag{3.3}$$

for some $k \in \mathbb{Z}$ and some polynomial p. For any r of the form (3.3) the identity $r(z)r(1/z) = 1$ holds for all nonzero $z \in \mathbb{C}$ except for the zeros of p and their reciprocals. Furthermore, there is always a unique choice of p in (3.3) so that p is monic, p and $\mathrm{rev}p$ are relatively prime, and $p(0) \neq 0$; in this case the sign is also uniquely determined. In addition, $r(z)$ is real whenever z is real if and only if this unique p is real.

Proof. For any r of the form (3.3) it is easy to check that $r(1/z) = \pm z^{-k}(\mathrm{rev}p(z))/p(z)$, so that the identity $r(z)r(1/z) = 1$ clearly holds for all nonzero $z \in \mathbb{C}$ except for the zeros of p and their reciprocals (which are the zeros of $\mathrm{rev}p$).

Conversely, suppose that $r(z)$ satisfies $r(z)r(1/z) = 1$ for infinitely many $z \in \mathbb{C}$. By Lemma 2.3, we can uniquely write r as $r(z) = z^k p(z)/q(z)$, where p and q are relatively prime, p is monic, and $p(0)$ and $q(0)$ are both nonzero. For this unique representation of r, we will show that $q(z) = \pm\mathrm{rev}p(z)$, giving us the form (3.3). Begin by rewriting the condition $r(z)r(1/z) = 1$ as

$$p(z)p(1/z) = q(z)q(1/z). \tag{3.4}$$

Letting n be any integer larger than $\deg p$ and $\deg q$ (where $\deg p$ denotes the degree of p), multiplying both sides of (3.4) by z^n and using the definition of rev results in

$$z^{n-\deg p}p(z)\,\mathrm{rev}p(z) = z^{n-\deg q}q(z)\,\mathrm{rev}q(z).$$

Since this equality of polynomials holds for infinitely many z, it must be an identity. Thus $\deg p = \deg q$, and

$$p(z)\,\mathrm{rev}p(z) = q(z)\,\mathrm{rev}q(z) \tag{3.5}$$

holds for all $z \in \mathbb{C}$. Since p has no factors in common with q, p must divide $\mathrm{rev}q$. Therefore $\mathrm{rev}q(z) = \alpha p(z)$ for some $\alpha \in \mathbb{C}$, which implies $q(z) = \alpha\,\mathrm{rev}p(z)$ since

$q(0) \neq 0$. Substituting into (3.5) gives $\alpha^2 = 1$, so that $q(z) = \pm \mathrm{rev}p(z)$, as desired. The final claim follows from Lemma 3.3. ∎

Lemmas 3.5 and 3.6 now lead us to the following characterization of meromorphic matrix functions with the property $f(A^{-1}) = f(A)^{-1}$. Here and in the rest of this paper we use the phrase "meromorphic matrix function $f(A)$" to mean a matrix function whose underlying scalar function f is meromorphic on all of $\mathbb{C}$.

THEOREM 3.7. *A meromorphic matrix function $f(A)$ has the property $f(A^{-1}) = f(A)^{-1}$ for all $A \in \mathbb{C}^{n\times n}$ such that A and A^{-1} are in the domain of f if and only if the scalar function f is rational and can be expressed in the form*

$$f(z) = \pm z^k \frac{p(z)}{\mathrm{rev}p(z)} \tag{3.6}$$

for some $k \in \mathbb{Z}$ and some polynomial p. If desired, the polynomial p may be chosen (uniquely) so that p is monic, p and $\mathrm{rev}p$ are relatively prime, and $p(0) \neq 0$. The matrix function $f(A)$ maps real matrices to real matrices if and only if this unique p is real.

Proof. As noted at the start of this subsection, $f(z)f(1/z) = 1$ for all z such that z and $1/z$ are in the domain of f is a necessary condition for having the property $f(A^{-1}) = f(A)^{-1}$. That f is rational then follows from Lemma 3.5, and from Lemma 3.6 we see that f must be of the form (3.6). To prove sufficiency, consider any f of the form (3.6) with $\deg p = n$. Then we have

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

The final claim follows from Theorem 3.2 and Lemma 3.6. ∎

3.3. Property (g): $f(A^{-*}) = f(A)^{-*}$. The results in this section provide a characterization of meromorphic matrix functions satisfying $f(A^{-*}) = f(A)^{-*}$. Consideration of the action of f on diagonal matrices leads to the identity $f(z)\overline{f(1/\overline{z})} = 1$ as a necessary condition on f. Thus the analysis of this identity is a prerequisite for understanding the corresponding matrix function property.

The following analogue of Lemma 3.5 can be derived, with a similar proof.

LEMMA 3.8. *Suppose f is a meromorphic function on $\mathbb{C}$ such that $f(z)\overline{f(1/\overline{z})} = 1$ holds for all z in some infinite compact subset of $\mathbb{C}$. Then f is a rational function with its zeros and poles matched in conjugate reciprocal pairs $\{a, 1/\overline{a}\}$. That is,*

$$z = a \text{ is a zero (pole) of order } k \iff z = 1/\overline{a} \text{ is a pole (zero) of order } k, \tag{3.7}$$

where $\{0, \infty\}$ is also to be regarded as a conjugate reciprocal pair.

In view of this result we can restrict our attention to rational functions.

LEMMA 3.9. *A complex rational function $r(z)$ satisfies the identity $r(z)\overline{r(1/\overline{z})} = 1$ for infinitely many $z \in \mathbb{C}$ if and only if it can be expressed in the form*

$$r(z) = \alpha z^k \frac{p(z)}{\mathrm{rev}\overline{p}(z)} \tag{3.8}$$

for some $k \in \mathbb{Z}$, some $|\alpha| = 1$, and some polynomial p. For any r of the form (3.8) *the identity $r(z)r(1/\overline{z}) = 1$ holds for all nonzero $z \in \mathbb{C}$ except for the zeros of p and their conjugate reciprocals. Furthermore, there is always a unique choice of p in* (3.8) *so that p is monic, p and $\mathrm{rev}\overline{p}$ are relatively prime, and $p(0) \neq 0$; in this case the scalar α is also unique. In addition, $r(z)$ is real whenever z is real if and only if this unique p is real and $\alpha = \pm 1$.*

Proof. The proof is entirely analogous to that of Lemma 3.6 and so is omitted. □

With Lemmas 3.8 and 3.9 we can now establish the following characterization of meromorphic functions with the property $f(A^{-*}) = f(A)^{-*}$.

THEOREM 3.10. *A meromorphic matrix function $f(A)$ has the property $f(A^{-*}) = f(A)^{-*}$ for all $A \in \mathbb{C}^{n\times n}$ such that A and A^{-*} are in the domain of f if and only if the scalar function f is rational and can be expressed in the form*

$$f(z) = \alpha z^k \frac{p(z)}{\mathrm{rev}\overline{p}(z)} \tag{3.9}$$

for some $k \in \mathbb{Z}$, some $|\alpha| = 1$, and some polynomial p. If desired, the polynomial p may be chosen (uniquely) so that p is monic, p and $\mathrm{rev}\overline{p}$ are relatively prime, and $p(0) \neq 0$; in this case the scalar α is also unique. The matrix function $f(A)$ maps real matrices to real matrices if and only if this unique p is real and $\alpha = \pm 1$.

Proof. As noted at the start of this subsection, $f(z)\overline{f(1/\overline{z})} = 1$ for all z such that z and $1/\overline{z}$ are in the domain of f is a necessary condition for having the property $f(A^{-*}) = f(A)^{-*}$. That f is rational then follows from Lemma 3.8, and from Lemma 3.9 we see that f must be of the form (3.9). To prove sufficiency, consider any f of the form (3.9) with $\deg p = n$. Then we have

$$\begin{aligned}
f(A)^* f(A^{-*}) &= \left[\alpha A^k p(A)[\mathrm{rev}\overline{p}(A)]^{-1}\right]^* \cdot \alpha(A^{-*})^k p(A^{-*})\left[\mathrm{rev}\overline{p}(A^{-*})\right]^{-1} \\
&= \left[\alpha A^k p(A)[A^n \overline{p}(A^{-1})]^{-1}\right]^* \cdot \alpha(A^*)^{-k} p(A^{-*})\left[(A^{-*})^n \overline{p}(A^*)\right]^{-1} \\
&= (A^*)^{-n}\overline{p}(A^{-1})^{-*} p(A)^* (A^*)^k \overline{\alpha}\alpha (A^*)^{-k} p(A^{-*})\overline{p}(A^*)^{-1}(A^*)^n \\
&= (A^*)^{-n} p(A^{-*})^{-1}\overline{p}(A^*) p(A^{-*})\overline{p}(A^*)^{-1}(A^*)^n \\
&= I.
\end{aligned}$$

The final claim follows from Lemma 3.9. □

Perhaps surprisingly, one can also characterize *general* analytic functions f satisfying $f(A^{-*}) = f(A)^{-*}$. The next result has a proof very similar to that of Theorem 3.2.

THEOREM 3.11. *Let f be analytic on an open subset $\Omega \subseteq \mathbb{C}$ such that each connected component of Ω is closed under reciprocal conjugation (i.e., under the map $z \mapsto 1/\overline{z}$). Consider the corresponding matrix function f on its natural domain in $\mathbb{C}^{n\times n}$, the set $\mathcal{D} = \{ A \in \mathbb{C}^{n\times n} : \Lambda(A) \subseteq \Omega \}$. Then the following are equivalent:*

(a) *$f(A^{-*}) = f(A)^{-*}$ for all $A \in \mathcal{D}$.*

(b) *$f(U(n) \cap \mathcal{D}) \subseteq U(n)$, where $U(n)$ denotes the group of $n \times n$ unitary matrices.*

(c) *$f(C \cap \Omega) \subseteq C$, where C denotes the unit circle $\{ z : |z| = 1 \}$.*

This theorem has the striking corollary that if a function is structure preserving for the unitary group then it is automatically structure preserving for *any other* automorphism group associated with a sesquilinear form.

COROLLARY 3.12. *Consider any function f satisfying the conditions of Theorem* 3.11. *Then f is structure preserving for all $\mathbb{G}$ associated with a sesquilinear form if and only if f is structure preserving for the unitary group $U(n)$.*

In view of the connection between the identity $f(z)\overline{f(1/\overline{z})} = 1$ and the property $f(C) \subseteq C$ established by Theorem 3.11, we can now see Lemma 3.9 as a natural generalization of the well known classification of all Möbius transformations mapping the open unit disc bijectively to itself, and hence mapping the unit circle to itself. These transformations are given by [9, Thm. 6.2.3], [27, Sec. 2.3.3]

$$f(z) = \alpha \frac{z - \beta}{1 - \overline{\beta} z},$$

where α and β are any complex constants satisfying $|\alpha| = 1$ and $|\beta| < 1$. This formula is easily seen to be a special case of Lemma 3.9.

3.4. Structure-preserving meromorphic functions. We can now give a complete characterization of structure-preserving meromorphic functions. This result extends [15, Thm. 2.1], which covers the "if" case in part (e).

THEOREM 3.13. *Consider the following two types of rational function, where $k \in \mathbb{Z}$, $|\alpha| = 1$, and p is a polynomial:*

$$\text{(I)}: \ \pm z^k \frac{p(z)}{\mathrm{rev}p(z)}, \qquad \text{(II)}: \ \alpha z^k \frac{p(z)}{\mathrm{rev}\overline{p}(z)},$$

and let $\mathbb{G}$ denote the automorphism group of a scalar product. A meromorphic matrix function f is structure preserving for all groups $\mathbb{G}$ associated with

(a) *a bilinear form on $\mathbb{C}^n$ iff f can be expressed in Type* I *form;*
(b) *a bilinear form on $\mathbb{R}^n$ iff f can be expressed in Type* I *form with a* real *p;*
(c) *a sesquilinear form on $\mathbb{C}^n$ iff f can be expressed in Type* II *form;*
(d) *a scalar product on $\mathbb{C}^n$ iff f can be expressed in Type* I *form with a* real *p;*
(e) *any scalar product iff f can be expressed in Type* I *form with a* real *p.*

Any such structure-preserving function can be uniquely expressed with a monic polynomial p such that p and rev$p$ *(or p and* rev$\overline{p}$ *for Type II) are relatively prime and $p(0) \neq 0$.*

Proof. (a) Theorem 3.1 shows that structure preservation is equivalent to the condition $f(A^{-1}) = f(A)^{-1}$ for all $A \in \mathbb{G}$ (in the domain of f), although not necessarily for all $A \in \mathbb{C}^{n\times n}$ (in the domain of f). Thus we cannot directly invoke Theorem 3.7 to reach the desired conclusion. However, note that the complex symplectic group contains diagonal matrices with arbitrary nonzero complex numbers z in the (1,1) entry. Thus $f(z)f(1/z) = 1$ for all nonzero complex numbers in the domain of f is a necessary condition for $f(A^{-1}) = f(A)^{-1}$ to hold for all $\mathbb{G}$. Hence f must be rational by Lemma 3.5, and a Type I rational by Lemma 3.6. That being of Type I is sufficient for structure preservation follows from Theorem 3.7.

(b) The argument used in part (a) also proves (b), simply by replacing the word "complex" throughout by "real", and noting that Lemma 3.6 implies that p may be chosen to be real.

(c) The argument of part (a) can be adapted to the sesquilinear case. By Theorem 3.1, structure preservation is in this case equivalent to the condition $f(A^{-*}) = f(A)^{-*}$ for all $A \in \mathbb{G}$ (in the domain of f). Again we cannot directly invoke Theorem 3.10 to complete the argument, but a short detour through Lemma 3.8 and Lemma 3.9 will yield the desired conclusion. Observe that the *conjugate* symplectic

group contains diagonal matrices D with arbitrary nonzero complex numbers z in the (1,1) entry. The condition $f(D^{-*}) = f(D)^{-*}$ then implies that $f(1/\overline{z}) = 1/\overline{f(z)}$ for all nonzero z in the domain of f, or equivalently $f(z)\overline{f(1/\overline{z})} = 1$. Lemma 3.8 now implies that f must be rational, Lemma 3.9 implies that f must be of Type II, and Theorem 3.10 then shows that any Type II rational function is indeed structure preserving.

(d) The groups considered here are the union of those in (a) and (c), so any structure-preserving f can be expressed in both Type I and Type II forms. But Lemmas 3.6 and 3.9 show that when f is expressed in the Lemma 2.3 canonical form $z^n p(z)/q(z)$, with p monic, $p(0) \neq 0$, and p relatively prime to q, then this particular expression for f is simultaneously of Type I and Type II. Thus $q = \pm \mathrm{rev} p = \beta \mathrm{rev} \overline{p}$ for some $|\beta| = 1$, and hence $p = \gamma \overline{p}$ ($\gamma = \pm\beta$). The monicity of p then implies $\gamma = 1$, so that p must be real. Conversely, it is clear that any Type I rational with real p is also of Type II, and hence is structure preserving for automorphism groups of both bilinear and sesquilinear forms on $\mathbb{C}^n$.

(e) Finally, (b) and (d) together immediately imply (e). ∎

We note a subtle feature of the sesquilinear case. That the conditions $f(A^{-1}) = f(A)^{-1}$ and $f(\overline{A}) = \overline{f(A)}$ hold for all $A \in \mathbb{G}$ is sufficient for f to be structure preserving for $\mathbb{G}$ (as shown by Theorem 3.1). However neither condition is necessary, as a simple example shows. Consider the function $f(z) = iz$. Since f is a Type II rational, it is structure preserving by Theorem 3.13 (c). But it is easy to see that neither $f(A^{-1}) = f(A)^{-1}$ nor $f(\overline{A}) = \overline{f(A)}$ holds for any nonzero matrix A.

3.5. M-normal matrices. We conclude this section with a structure-preservation result of a different flavor. It is well known that if A is normal ($A^*A = AA^*$) then $f(A)$ is normal. This result can be generalized to an arbitrary scalar product space. Following Gohberg, Lancaster and Rodman [8, Sec. I.4.6], define $A \in \mathbb{K}^{n\times n}$ to be M-normal, that is, normal with respect to the scalar product defined by a matrix M, if $A^\star A = AA^\star$. If A belongs to the automorphism group $\mathbb{G}$ of the scalar product then A is certainly M-normal. The next result shows that any function f preserves M-normality. In particular, for $A \in \mathbb{G}$, even though $f(A)$ may not belong to $\mathbb{G}$, $f(A)$ is M-normal and so has some structure.

THEOREM 3.14. *Consider a scalar product defined by a matrix M. Let $A \in \mathbb{K}^{n\times n}$ be M-normal and let f be any function defined on the spectrum of A. Then $f(A)$ is M-normal.*

Proof. Let p be any polynomial that evaluates f at A. Then

$$f(A)^\star = p(A)^\star = \begin{cases} p(A^\star), & \text{for bilinear forms,} \\ \overline{p}(A^\star), & \text{for sesquilinear forms.} \end{cases}$$

Continuing with the two cases,

$$\begin{aligned} f(A)f(A)^\star &= \begin{cases} p(A)p(A^\star) = p(A^\star)p(A) \\ p(A)\overline{p}(A^\star) = \overline{p}(A^\star)p(A) \end{cases} \quad (\text{since } AA^\star = A^\star A) \\ &= f(A)^\star f(A). \end{aligned}$$ ∎

Theorem 3.14 generalizes a result of Gohberg, Lancaster, and Rodman [8, Thm. I.6.3], which states that if $A \in \mathbb{G}$ or $A = A^\star$ with respect to a Hermitian sesquilinear form then $f(A)$ is M-normal.

4. Connections between the matrix sign function, the generalized polar decomposition, and the matrix square root. Having identified the matrix sign function and square root as structure preserving for all groups, we now consider computational matters. We show in this section that the matrix sign function and square root are intimately connected with each other and also with the generalized polar decomposition. Given a scalar product on $\mathbb{K}^n$ with adjoint $(\cdot)^\star$, a generalized polar decomposition of a matrix $A \in \mathbb{K}^{n\times n}$ is a decomposition $A = WS$, where W is an automorphism and S is self-adjoint with spectrum contained in the open right half-plane, that is, $W^\star = W^{-1}$, $S^\star = S$, and $\operatorname{sign}(S) = I$. The existence and uniqueness of a generalized polar decomposition is described in the next result, which extends [15, Thm. 4.1].

THEOREM 4.1 (generalized polar decomposition). *With respect to an arbitrary scalar product on $\mathbb{K}^n$, a matrix $A \in \mathbb{K}^{n\times n}$ has a generalized polar decomposition $A = WS$ if and only if $(A^\star)^\star = A$ and $A^\star A$ has no eigenvalues on $\mathbb{R}^-$. When such a factorization exists it is unique.*

Proof. ($\Rightarrow$) Note first that if the factorization exists then

$$(A^\star)^\star = (S^\star W^\star)^\star = (SW^{-1})^\star = W^{-\star}S^\star = WS = A.$$

Also we must have

$$A^\star A = S^\star W^\star W S = S^\star S = S^2. \tag{4.1}$$

But if $\operatorname{sign}(S) = I$ is to hold then the only possible choice for S is $S = (A^\star A)^{1/2}$, and this square root exists only if $A^\star A$ has no eigenvalues on $\mathbb{R}^-$.

($\Leftarrow$) Letting $S = (A^\star A)^{1/2}$, the condition $\operatorname{sign}(S) = I$ is automatically satisfied, but we need also to show that $S^\star = S$. First, note that for any B with no eigenvalues on $\mathbb{R}^-$ we have $(B^\star)^{1/2} = (B^{1/2})^\star$. Indeed $(B^{1/2})^\star$ is a square root of $B^\star$, because $(B^{1/2})^\star(B^{1/2})^\star = (B^{1/2}\cdot B^{1/2})^\star = B^\star$, and the fact that $(B^{1/2})^\star$ is similar to $(B^{1/2})^T$ (for bilinear forms) or $(B^{1/2})^*$ (for sesquilinear forms) implies that $(B^{1/2})^\star$ must be the principal square root. Then, using the assumption that $(A^\star)^\star = A$, we have

$$S^\star = \left((A^\star A)^{1/2}\right)^\star = \left((A^\star A)^\star\right)^{1/2} = (A^\star A)^{1/2} = S.$$

Finally, the uniquely defined matrix $W = AS^{-1}$ satisfies

$$W^\star W = (AS^{-1})^\star(AS^{-1}) = S^{-\star}(A^\star A)S^{-1} = S^{-1}(S^2)S^{-1} = I,$$

using (4.1), and so $W \in \mathbb{G}$. ∎

For many scalar products, including all those in Table 2.1, $(A^\star)^\star = A$ holds *for all* $A \in \mathbb{K}^{n\times n}$, in which case we say that the adjoint is involutory. It can be shown that the adjoint is involutory if and only if $M^T = \pm M$ for bilinear forms and $M^* = \alpha M$ with $|\alpha| = 1$ for sesquilinear forms [26]. But even for scalar products for which the adjoint is not involutory, there are always many matrices A for which $(A^\star)^\star = A$, as the next result shows. We omit the straightforward proof.

LEMMA 4.2. *Let $\mathbb{G}$ be the automorphism group of a scalar product. The condition*

$$(A^\star)^\star = A \tag{4.2}$$

is satisfied if $A \in \mathbb{G}$, $A = A^\star$, or $A = -A^\star$. Moreover, arbitrary products and linear combinations of matrices satisfying (4.2) *also satisfy* (4.2).

The generalized polar decomposition as we have defined it is closely related to the polar decompositions corresponding to Hermitian sesquilinear forms on $\mathbb{C}^n$ studied by Bolshakov et al. [2], [3], the symplectic polar decomposition introduced by Ikramov [17], and the polar decompositions corresponding to symmetric bilinear forms on $\mathbb{C}^n$ considered by Kaplansky [19]. In these papers the self-adjoint factor S may or may not be required to satisfy additional conditions, but $\operatorname{sign}(S) = I$ is not one of those considered. The connections established below between the matrix sign function, the principal matrix square root, and the generalized polar decomposition as we have defined it, suggest that $\operatorname{sign}(S) = I$ is the appropriate extra condition for a generalized polar decomposition of computational use.

The following result, which we have not found in the literature, is the basis for the connections to be established.

LEMMA 4.3. *Let $A, B \in \mathbb{C}^{n\times n}$ and suppose that AB (and hence also BA) has no eigenvalues on $\mathbb{R}^-$. 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. The matrix $P = \left[\begin{smallmatrix} 0 & A \\ B & 0 \end{smallmatrix}\right]$ cannot have any eigenvalues on the imaginary axis, because if it did then $P^2 = \left[\begin{smallmatrix} AB & 0 \\ 0 & BA \end{smallmatrix}\right]$ would have an eigenvalue on $\mathbb{R}^-$. Hence $\operatorname{sign}(P)$ is defined and

$$\begin{aligned}
\operatorname{sign}(P) = P(P^2)^{-1/2} &= \begin{bmatrix} 0 & A \\ B & 0 \end{bmatrix} \begin{bmatrix} AB & 0 \\ 0 & BA \end{bmatrix}^{-1/2} \\
&= \begin{bmatrix} 0 & A \\ B & 0 \end{bmatrix} \begin{bmatrix} (AB)^{-1/2} & 0 \\ 0 & (BA)^{-1/2} \end{bmatrix} \\
&= \begin{bmatrix} 0 & A(BA)^{-1/2} \\ B(AB)^{-1/2} & 0 \end{bmatrix} =: \begin{bmatrix} 0 & C \\ D & 0 \end{bmatrix}.
\end{aligned}$$

Since the square of the matrix sign of any matrix is the identity,

$$I = (\operatorname{sign}(P))^2 = \begin{bmatrix} 0 & C \\ D & 0 \end{bmatrix}^2 = \begin{bmatrix} CD & 0 \\ 0 & DC \end{bmatrix},$$

so $D = C^{-1}$. Alternatively, Corollary 2.2 may be used to see more directly that $CD = A(BA)^{-1/2}B(AB)^{-1/2}$ is equal to I. ∎

Two important special cases of Lemma 4.3 are, for $A \in \mathbb{C}^{n\times n}$ with no eigenvalues on $\mathbb{R}^-$ [13],

$$\operatorname{sign}\left(\begin{bmatrix} 0 & A \\ I & 0 \end{bmatrix}\right) = \begin{bmatrix} 0 & A^{1/2} \\ A^{-1/2} & 0 \end{bmatrix}, \tag{4.3}$$

and, for nonsingular $A \in \mathbb{C}^{n\times n}$ [12],

$$\operatorname{sign}\left(\begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix}\right) = \begin{bmatrix} 0 & U \\ U^* & 0 \end{bmatrix}, \tag{4.4}$$

where $A = UH$ is the polar decomposition. A further special case, which generalizes (4.4), is given in the next result.

COROLLARY 4.4. *If $A \in \mathbb{K}^{n\times n}$ has a generalized polar decomposition $A = WS$ then*

$$\operatorname{sign}\left(\begin{bmatrix} 0 & A \\ A^\star & 0 \end{bmatrix}\right) = \begin{bmatrix} 0 & W \\ W^\star & 0 \end{bmatrix}. \tag{4.5}$$

Proof. Lemma 4.3 gives

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

where $C = A(A^\star A)^{-1/2}$. Using the given generalized polar decomposition, $C = WS \cdot S^{-1} = W$, and so

$$\operatorname{sign}\left(\begin{bmatrix} 0 & A \\ A^\star & 0 \end{bmatrix}\right) = \begin{bmatrix} 0 & W \\ W^{-1} & 0 \end{bmatrix} = \begin{bmatrix} 0 & W \\ W^\star & 0 \end{bmatrix}. \qquad \square$$

The significance of (4.3)–(4.5) is that they enable results and iterations for the sign function to be translated into results and iterations for the square root and generalized polar decomposition. For example, Roberts' integral formula [28], $\operatorname{sign}(A) = (2/\pi)A\int_0^\infty (t^2 I + A^2)^{-1}\,dt$ translates, via (4.5), into an integral representation for the generalized polar factor W:

$$W = \frac{2}{\pi} A \int_0^\infty (t^2 I + A^\star A)^{-1}\,dt.$$

Our interest in the rest of this section is in deriving iterations, beginning with a family of iterations for the matrix square root[3].

THEOREM 4.5. *Suppose the matrix A has no eigenvalues on $\mathbb{R}^-$, so that $A^{1/2}$ exists. Let g be any matrix function of the form $g(X) = Xh(X^2)$ such that the iteration $X_{k+1} = g(X_k)$ converges to $\operatorname{sign}(X_0)$ with order of convergence m whenever $\operatorname{sign}(X_0)$ is defined. Then in the coupled iteration*

$$\begin{aligned} Y_{k+1} &= Y_k h(Z_k Y_k), & Y_0 &= A, \\ Z_{k+1} &= h(Z_k Y_k) Z_k, & Z_0 &= I, \end{aligned} \tag{4.6}$$

$Y_k \to A^{1/2}$ and $Z_k \to A^{-1/2}$ as $k \to \infty$, both with order of convergence m, Y_k commutes with Z_k, and $Y_k = AZ_k$ for all k. Moreover, if g is structure preserving for an automorphism group $\mathbb{G}$, then iteration (4.6) is also structure preserving for $\mathbb{G}$, that is, $A \in \mathbb{G}$ implies $Y_k, Z_k \in \mathbb{G}$ for all k.

Proof. Observe that

$$\begin{aligned} g\left(\begin{bmatrix} 0 & Y_k \\ Z_k & 0 \end{bmatrix}\right) &= \begin{bmatrix} 0 & Y_k \\ Z_k & 0 \end{bmatrix} h\left(\begin{bmatrix} Y_k Z_k & 0 \\ 0 & Z_k Y_k \end{bmatrix}\right) \\ &= \begin{bmatrix} 0 & Y_k \\ Z_k & 0 \end{bmatrix} \begin{bmatrix} h(Y_k Z_k) & 0 \\ 0 & h(Z_k Y_k) \end{bmatrix} \\ &= \begin{bmatrix} 0 & Y_k h(Z_k Y_k) \\ Z_k h(Y_k Z_k) & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & Y_k h(Z_k Y_k) \\ h(Z_k Y_k) Z_k & 0 \end{bmatrix} = \begin{bmatrix} 0 & Y_{k+1} \\ Z_{k+1} & 0 \end{bmatrix}, \end{aligned}$$

[3]We note that if we generalize to $Z_0 = B$ in (4.6), where BA has no eigenvalues on $\mathbb{R}^-$, then $Y_k \to A(BA)^{-1/2}$, which is a solution of the special Riccati equation $XBX = A$, while $Z_k \to B(AB)^{-1/2}$, which solves $XAX = B$.

where the penultimate equality follows from Corollary 2.2. The initial conditions $Y_0 = A$ and $Z_0 = I$ together with (4.3) now imply that Y_k and Z_k converge to $A^{1/2}$ and $A^{-1/2}$, respectively. It is easy to see that Y_k and Z_k are polynomials in A for all k, and hence Y_k commutes with Z_k. Then $Y_k = AZ_k$ follows by induction. The order of convergence of the coupled iteration (4.6) is clearly the same as that of the sign iteration from which it arises.

Finally, if g is structure preserving for $\mathbb{G}$ and $A \in \mathbb{G}$, then we can show inductively that $Y_k, Z_k \in \mathbb{G}$ for all k. Clearly $Y_0, Z_0 \in \mathbb{G}$. Assuming that $Y_k, Z_k \in \mathbb{G}$, then $Z_k Y_k \in \mathbb{G}$. Since $\operatorname{sign}\left[\begin{smallmatrix} 0 & Y_k \\ Z_k & 0 \end{smallmatrix}\right] = \operatorname{sign}\left[\begin{smallmatrix} 0 & A \\ I & 0 \end{smallmatrix}\right]$, we know that $P = \left[\begin{smallmatrix} 0 & Y_k \\ Z_k & 0 \end{smallmatrix}\right]$ has no imaginary eigenvalues, and hence that $P^2 = \left[\begin{smallmatrix} Y_k Z_k & 0 \\ 0 & Z_k Y_k \end{smallmatrix}\right]$ has no eigenvalues on $\mathbb{R}^-$. Thus $(Z_k Y_k)^{1/2}$ exists, and from Section 3 we know that $(Z_k Y_k)^{1/2} \in \mathbb{G}$. But for any $X \in \mathbb{G}$, $g(X) \in \mathbb{G}$, and hence $h(X^2) = X^{-1} g(X) \in \mathbb{G}$. Thus with $X = (Z_k Y_k)^{1/2}$, we see that $h(X^2) = h(Z_k Y_k) \in \mathbb{G}$, and therefore $Y_{k+1}, Z_{k+1} \in \mathbb{G}$. ∎

The connection between sign iterations and square root iterations has been used previously [13], but only for some particular g. By contrast, Theorem 4.5 is very general, since all commonly used sign iteration functions have the form $g(X) = X\,h(X^2)$ considered here. Note that the commutativity of Y_k and Z_k allows several variations of (4.6); the one we have chosen has the advantage that it requires only one evaluation of h per iteration. We have deliberately avoided using commutativity properties in deriving the iteration within the proof above (instead, we invoked Corollary 2.2). In particular, we did not rewrite the second part of the iteration in the form $Z_{k+1} = Z_k h(Z_k Y_k)$, which is arguably more symmetric with the first part. The reason is that experience suggests that exploiting commutativity when deriving matrix iterations can lead to numerical instability (see, e.g., [11]). Indeed we will show in Section 5 that while (4.6) is numerically stable, the variant just mentioned is not.

We now exploit the connection in Corollary 4.4 between the sign function and the generalized polar decomposition. The corollary suggests that we apply iterations for the matrix sign function to

$$X_0 = \begin{bmatrix} 0 & A \\ A^\star & 0 \end{bmatrix},$$

so just as in Theorem 4.5 we consider iteration functions of the form $g(X) = X\,h(X^2)$. It is possible, though nontrivial, to prove by induction that all the iterates X_k of such a g have the form

$$X_k = \begin{bmatrix} 0 & Y_k \\ Y_k^\star & 0 \end{bmatrix} \tag{4.7}$$

with $(Y_k^\star)^\star = Y_k$, and that $Y_{k+1} = Y_k h(Y_k^\star Y_k)$—under an extra assumption on g in the sesquilinear case. Corollary 4.4 then implies that Y_k converges to the generalized polar factor W of A. While this approach is a useful way to derive the iteration for W, a shorter and more direct demonstration of the claimed properties is possible, as we now show.

THEOREM 4.6. *Suppose the matrix A has a generalized polar decomposition $A = WS$ with respect to a given scalar product. Let g be any matrix function of the form $g(X) = X\,h(X^2)$ such that the iteration $X_{k+1} = g(X_k)$ converges to* $\operatorname{sign}(X_0)$ *with order of convergence m whenever* $\operatorname{sign}(X_0)$ *is defined. For sesquilinear forms assume*

that g also satisfies (d) *in Theorem* 3.1 *for all matrices in its domain. Then the iteration*

$$Y_{k+1} = Y_k h(Y_k^\star Y_k), \qquad Y_0 = A \tag{4.8}$$

converges to W with order of convergence m.

Proof. Let $X_{k+1} = g(X_k)$ with $X_0 = S$, so that $\lim_{k\to\infty} X_k = \operatorname{sign}(S) = I$. We claim that $X_k^\star = X_k$ and $Y_k = WX_k$ for all k. These equalities are trivially true for $k = 0$. Assuming that they are true for k, we have

$$X_{k+1}^\star = g(X_k)^\star = g(X_k^\star) = g(X_k) = X_{k+1}$$

and

$$Y_{k+1} = WX_k h(X_k^\star W^\star W X_k) = WX_k h(X_k^2) = WX_{k+1}.$$

The claim follows by induction. Hence $\lim_{k\to\infty} Y_k = W \lim_{k\to\infty} X_k = W$. The order of convergence is readily seen to be m. ∎

Theorem 4.6 shows that iterations for the matrix sign function automatically yield iterations for the generalized polar factor W. The next result reveals that the square root of a matrix in an automorphism group is the generalized polar factor W of a related matrix. Consequently, iterations for W also lead to iterations for the matrix square root, although only for matrices in automorphism groups. We will take up this topic again in Section 6.

THEOREM 4.7. *Let $\mathbb{G}$ be the automorphism group of a scalar product and $A \in \mathbb{G}$. If A has no eigenvalues on $\mathbb{R}^-$, then $I+A = WS$ with $W = A^{1/2}$ and $S = A^{-1/2}+A^{1/2}$ is the generalized polar decomposition of $I+A$ with respect to the given scalar product.*

Proof. Clearly, $I + A = WS$ and $W^\star = A^{\star/2} = A^{-1/2} = W^{-1}$. It remains to show that $S^\star = S$ and $\operatorname{sign}(S) = I$. We have

$$S^\star = A^{-\star/2} + A^{\star/2} = A^{1/2} + A^{-1/2} = S.$$

Moreover, the eigenvalues of S are of the form $\mu = \lambda^{-1} + \lambda$, where $\lambda \in \Lambda(A^{1/2})$ is in the open right half-plane. Clearly μ is also in the open right half-plane, and hence $\operatorname{sign}(S) = I$. ∎

Note that Theorem 4.7 does not make any assumption on the scalar product or its associated adjoint. The condition $(B^\star)^\star = B$ that is required to apply Theorem 4.1 is automatically satisfied for $B = I + A$, since $A \in \mathbb{G}$ implies that $I + A$ is one of the matrices in Lemma 4.2.

Theorem 4.7 appears in Cardoso, Kenney, and Silva Leite [5, Thm. 6.3] for real bilinear forms only and with the additional assumption that the matrix M of the scalar product is symmetric positive definite.

5. Stability analysis of coupled square root iterations. Before investigating any specific iterations from among the families obtained in the previous section, we carry out a stability analysis of the general iteration (4.6) of Theorem 4.5. A whole section is devoted to this analysis for two reasons. First, as is well known, minor rewriting of matrix iterations can completely change their stability properties [11], [13]. As already noted, (4.6) can be rewritten in various ways using commutativity and/or Corollary 2.2, and it is important to know that a choice of form motivated by computational cost considerations does not sacrifice stability. Second, we are able

to give a stability analysis of (4.6) in its full generality, and in doing so introduce a technique that is novel in this context and should be of wider use in analyzing the stability of matrix iterations.

We begin by slightly changing the notation of Theorem 4.5. Consider matrix functions of the form $g(X) = Xh(X^2)$ that compute the matrix sign by iteration, and the related function

$$G(Y,Z) = \begin{bmatrix} g_1(Y,Z) \\ g_2(Y,Z) \end{bmatrix} = \begin{bmatrix} Yh(ZY) \\ h(ZY)Z \end{bmatrix}. \tag{5.1}$$

Iterating G starting with $(Y,Z) = (A,I)$ produces the coupled iteration (4.6), which we know converges to $(A^{1/2}, A^{-1/2})$. Recall that the Fréchet derivative of a map $F : \mathbb{C}^{m\times n} \to \mathbb{C}^{m\times n}$ at a point $X \in \mathbb{C}^{m\times n}$ is a linear mapping $L_X : \mathbb{C}^{m\times n} \to \mathbb{C}^{m\times n}$ such that for all $E \in \mathbb{C}^{m\times n}$ [6], [29],

$$F(X+E) - F(X) - L_X(E) = o(\|E\|).$$

For our purposes it will not matter whether L_X is $\mathbb{C}$-linear or only $\mathbb{R}$-linear.

Our aim is to find the Fréchet derivative of the map G at the point $(Y,Z) = (A^{1/2}, A^{-1/2})$, or more generally at any point of the form (B, B^{-1}); later, these points will all be seen to be fixed points of the map G. We denote the Fréchet derivative of G by dG, the derivative at a particular point (A,B) by $dG_{(A,B)}$, and the matrix inputs to dG by dY and dZ. With this notation, we have

$$dG_{(Y,Z)}(dY,dZ) = \begin{bmatrix} dg_1(dY,dZ) \\ dg_2(dY,dZ) \end{bmatrix} = \begin{bmatrix} Y dh_{ZY}(ZdY + dZ\cdot Y) + dY\cdot h(ZY) \\ dh_{ZY}(ZdY + dZ\cdot Y)\cdot Z + h(ZY)dZ \end{bmatrix}.$$

At the point $(Y,Z) = (B, B^{-1})$ this simplifies to

$$dG_{(B,B^{-1})}(dY,dZ) = \begin{bmatrix} B\, dh_I(B^{-1}dY + dZ\cdot B) + dY\cdot h(I) \\ dh_I(B^{-1}dY + dZ\cdot B)\cdot B^{-1} + h(I)dZ \end{bmatrix}. \tag{5.2}$$

In order to further simplify this expression we need to know more about $h(I)$ and dh_I. We give a preliminary lemma and then exploit the fact that h is part of a function that computes the matrix sign.

LEMMA 5.1. *For any matrix function $F(X)$ with underlying scalar function f that is analytic at $z = 1$, the Fréchet derivative of F at the matrix I is just scalar multiplication by $f'(1)$, that is, $dF_I(E) = f'(1)E$.*

Proof. Expand the scalar function f as a convergent power series about $z = 1$: $f(z) = \sum_{k=0}^{\infty} b_k(z-1)^k$, where $b_k = f^{(k)}(1)/k!$. Then

$$F(I+E) - F(I) = \sum_{k=0}^{\infty} b_k E^k - b_0 I = b_1 E + O(\|E\|^2).$$

Thus $dF_I(E) = b_1 E = f'(1)E$. ∎

LEMMA 5.2. *Suppose h is part of a matrix function of the form $g(X) = Xh(X^2)$ such that the iteration $X_{k+1} = g(X_k)$ converges superlinearly to* $\operatorname{sign}(X_0)$ *whenever* $\operatorname{sign}(X_0)$ *exists. If the scalar function h is analytic at $z = 1$ then $h(I) = I$ and $dh_I(E) = -\frac{1}{2}E$.*

Proof. Since $\operatorname{sign}(I) = I$, I is a fixed point of the iteration, so $g(I) = I$ and hence $h(I) = I$.

At the scalar level, $g(x) = xh(x^2)$ and $g'(x) = 2x^2h'(x^2) + h(x^2)$, so $g'(1) = 2h'(1) + h(1)$. But $h(I) = I$ implies $h(1) = 1$, so $g'(1) = 2h'(1) + 1$. Now we are assuming that the iterates of g converge superlinearly to $\mathrm{sign}(X_0)$, so in particular we know that a neighborhood of 1 contracts superlinearly to 1 under iteration by g. From fixed point iteration theory this means that $g'(1) = 0$. Hence $h'(1) = -\frac{1}{2}$ and, using Lemma 5.1, $dh_I(E) = h'(1)E = -\frac{1}{2}E$. ∎

Because $h(I) = I$, it is now clear that any point (B, B^{-1}) is a fixed point for G. Furthermore, our knowledge of $h(I)$ and dh_I allows us to complete the simplification of dG, continuing from (5.2):

$$\begin{aligned} dG_{(B,B^{-1})}(dY, dZ) &= \begin{bmatrix} -\frac{1}{2}dY - \frac{1}{2}B\,dZ\,B + dY \\ -\frac{1}{2}B^{-1}dY\,B^{-1} - \frac{1}{2}dZ + dZ \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{2}dY - \frac{1}{2}B\,dZ\,B \\ \frac{1}{2}dZ - \frac{1}{2}B^{-1}dY\,B^{-1} \end{bmatrix}. \end{aligned}$$

A straightforward computation shows that $dG_{(B,B^{-1})}$ is idempotent, and hence is a projection. We summarize our findings in a theorem.

THEOREM 5.3. *Consider any iteration of the form* (4.6) *and its associated mapping*

$$G(Y, Z) = \begin{bmatrix} Yh(ZY) \\ h(ZY)Z \end{bmatrix},$$

where $X_{k+1} = g(X_k) = X_k h(X_k^2)$ is any superlinearly convergent iteration for the matrix sign such that the scalar function h is analytic at $z = 1$. Then any matrix pair of the form $P = (B, B^{-1})$ is a fixed point for G, and the Fréchet derivative of G at P is given by

$$dG_P(E, F) = \frac{1}{2}\begin{bmatrix} E - BFB \\ F - B^{-1}EB^{-1} \end{bmatrix}.$$

The derivative map dG_P is idempotent, that is, $dG_P \circ dG_P = dG_P$.

Following Cheng, Higham, Kenney, and Laub [7], we define an iteration $X_{k+1} = g(X_k)$ to be stable in a neighborhood of a fixed point $X = g(X)$ if for $X_0 := X + H_0$, with arbitrary H_0, the errors $H_k := X_k - X$ satisfy

$$H_{k+1} = L_X(H_k) + O(\|H_k\|^2), \tag{5.3}$$

where L_X is a linear operator (necessarily the Fréchet derivative of g at X) with bounded powers, that is, there exists a constant c such that for all $s > 0$ and arbitrary H of unit norm, $\|L_X^s(H)\| \le c$. Note that the iterations we are considering have a specified X_0 and so the convergence analysis in Section 4 says nothing about the effect of arbitrary errors H_k in the X_k. In practice, such errors are of course introduced by the effects of roundoff. The significance of Theorem 5.3 is that it shows that any iteration belonging to the broad class (4.6) is stable, for L_X is here idempotent and hence trivially has bounded powers.

A further use of our analysis is to predict the limiting accuracy of the iteration in floating point arithmetic, that is, the smallest error we can expect. Consider $X_0 = X + H_0$ with $\|H_0\| \le u\|X\|$, where u is the unit roundoff, so that X_0 can be thought of as X rounded to floating point arithmetic. Then from (5.3) we have $\|H_1\| \lesssim \|L_X(H_0)\|$, and so an estimate of the absolute limiting accuracy is any bound

for $\|L_X(H_0)\|$. In the case of iteration (4.6), a suitable bound is, from Theorem 5.3 with $B = A^{1/2}$,

$$\max\{\, \|E_0\| + \|A^{1/2}\|^2\|F_0\|, \|F_0\| + \|A^{-1/2}\|^2\|E_0\| \,\},$$

where $\|E_0\| \le \|A^{1/2}\|u$ and $\|F_0\| \le \|A^{-1/2}\|u$. For any of the classical groups in Table 2.1, M is unitary and so $A \in \mathbb{G}$ implies $\|A^{1/2}\|_2 = \|A^{-1/2}\|_2$, by (2.2) (since $A^{1/2} \in \mathbb{G}$). Hence this bound is just $\|A^{1/2}\|_2(1 + \|A^{1/2}\|_2^2)u$, giving an estimate for the *relative* limiting accuracy of $(1 + \|A^{1/2}\|_2^2)u$.

The Fréchet derivative-based analysis of this section would be even more useful if it also allowed us to identify otherwise plausible iterations that are unstable. To see that it does, consider the mathematically equivalent variant of (4.6)

$$\begin{aligned} Y_{k+1} &= Y_k h(Z_k Y_k), & Y_0 &= A, \\ Z_{k+1} &= Z_k h(Z_k Y_k), & Z_0 &= I, \end{aligned} \tag{5.4}$$

mentioned earlier as being arguably more symmetric, but of questionable stability since its derivation relies on commutativity properties. For this iteration we define the map $\widetilde{G}(Y, Z) = \left[\begin{smallmatrix} Yh(ZY) \\ Zh(ZY) \end{smallmatrix}\right]$, analogous to the map G for iteration (4.6), and see by a calculation similar to the one above that

$$d\widetilde{G}_P(E, F) = \frac{1}{2}\begin{bmatrix} E - BFB \\ 2F - B^{-1}FB - B^{-2}E \end{bmatrix}. \tag{5.5}$$

The following lemma, whose proof we omit, shows that for many B the map $d\widetilde{G}_P$ has an eigenvalue of modulus exceeding 1 and hence does not have bounded powers; the iteration is then unstable according to our definition.

LEMMA 5.4. *If α and β are any two eigenvalues of B then $\gamma = \frac{1}{2}(1 - \frac{\alpha}{\beta})$ is an eigenvalue for $d\widetilde{G}_P$ in* (5.5)*, where $P = (B, B^{-1})$.*

The stability and instability, respectively, of particular instances of iterations (4.6) and (5.4) are confirmed in the numerical experiments of Section 7.

Finally, we note that the following analog of Theorem 5.3 can be proved for the iterations computing the generalized polar factor W described in Theorem 4.6.

THEOREM 5.5. *Consider any iteration of the form* (4.8) *and the associated mapping $f(Y) = Yh(Y^\star Y)$, where $X_{k+1} = g(X_k) = X_k h(X_k^2)$ is any superlinearly convergent iteration for the matrix sign such that the scalar function h is analytic at $z = 1$. Then any $B \in \mathbb{G}$ is a fixed point for f, and the Fréchet derivative of f at B is given by $df_B(E) = \frac{1}{2}(E - BE^\star B)$. If the underlying scalar product has an involutory adjoint, then the derivative map df_B is idempotent.*

As an immediate consequence we see that any iteration of the form (4.8) is stable[4], at least when the adjoint is involutory (see the remarks preceding Lemma 4.2 for details of when this condition holds). Special cases of this include the unitary polar factor iterations developed in [15] and the iteration (6.7) for the square root of matrices in $\mathbb{G}$ derived in the next section.

6. Iterations for the matrix square root. We now use the theory developed above to derive some specific new iterations for computing the square root of a matrix

[4]Because of the presence of the adjoint in iteration (4.8), the map L_X in (5.3) is no longer complex linear in the sesquilinear case, but it is a real linear map and hence we can still deduce stability.

in an automorphism group. We assume throughout that A has no eigenvalues on $\mathbb{R}^-$, so that $A^{1/2}$ is defined. First, we recall the well-known Newton iteration

$$X_{k+1} = \frac{1}{2}\big(X_k + X_k^{-1}A\big), \qquad X_0 = A, \tag{6.1}$$

which can be thought of as a generalization to matrices of Heron's iteration for the square root of a scalar. This iteration converges quadratically to $A^{1/2}$, but it is numerically unstable and therefore not of practical use [11], [23]. There has consequently been much interest in deriving numerically stable alternatives.

We first derive a structure-preserving iteration. We apply Theorem 4.5 to the family of structure-preserving matrix sign function iterations identified by Higham, Mackey, Mackey, and Tisseur [15], which comprises the main diagonal of a table of Padé-based iterations discovered by Kenney and Laub [20].

THEOREM 6.1. *Let $A \in \mathbb{K}^{n\times n}$ and consider the iterations*

$$Y_{k+1} = Y_k\, p_m(I - Z_kY_k)\,[\mathrm{rev}p_m(I - Z_kY_k)]^{-1}, \qquad Y_0 = A, \tag{6.2a}$$

$$Z_{k+1} = p_m(I - Z_kY_k)\,[\mathrm{rev}p_m(I - Z_kY_k)]^{-1}Z_k, \qquad Z_0 = I, \tag{6.2b}$$

where $p_m(t)$ is the numerator in the $[m/m]$ Padé approximant to $(1-t)^{-1/2}$ and $m \geq 1$. Assume that A has no eigenvalues on $\mathbb{R}^-$ and $A \in \mathbb{G}$, where $\mathbb{G}$ is any automorphism group. 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 of convergence $2m+1$.

Proof. It was shown in [15] that the iteration $X_{k+1} = X_k\, p_m(I - X_k^2)\,[\mathrm{rev}p_m(I - X_k^2)]^{-1}$, with $X_0 = A$, is on the main diagonal of the Padé table in [20] and so converges to $\mathrm{sign}(A)$ with order of convergence $2m+1$. This iteration was shown in [15] to be structure preserving, a property that can also be seen from Theorem 3.13 (e). The theorem therefore follows immediately from Theorem 4.5. ∎

The polynomial $p_m(1-x^2)$ in Theorem 6.1 can be obtained by taking the odd part of $(1+x)^{2m+1}$ and dividing through by x [20]. The first two polynomials are $p_1(1-x^2) = x^2+3$ and $p_2(1-x^2) = x^4+10x^2+5$. The cubically converging iteration ($m=1$) is therefore

$$Y_{k+1} = Y_k(3I + Z_kY_k)(I + 3Z_kY_k)^{-1}, \qquad Y_0 = A, \tag{6.3a}$$

$$Z_{k+1} = (3I + Z_kY_k)(I + 3Z_kY_k)^{-1}Z_k, \qquad Z_0 = I. \tag{6.3b}$$

A rearrangement of these formulae that can be evaluated in fewer flops is the continued fraction form, adapted from [15],

$$Y_{k+1} = \frac{1}{3}Y_k\big[I + 8\big(I + 3Z_kY_k\big)^{-1}\big], \qquad Y_0 = A, \tag{6.4a}$$

$$Z_{k+1} = \frac{1}{3}\big[I + 8\big(I + 3Z_kY_k\big)^{-1}\big]Z_k, \qquad Z_0 = I. \tag{6.4b}$$

This iteration can be implemented in two ways: using 3 matrix multiplications and 1 (explicit) matrix inversion per iteration, or with 1 matrix multiplication and 2 solutions of matrix equations involving coefficient matrices that are transposes of each other. The latter approach has the smaller operation count, but the former could be faster in practice as it is richer in matrix multiplication, which is a particularly efficient operation on modern computers.

A related family[5] of coupled iterations for the square root was derived by Higham [13] from the first superdiagonal of Kenney and Laub's Padé table. However, unlike (6.2), that family is not structure preserving: when $A \in \mathbb{G}$ the iterates do not stay in the group.

With the aid of Theorem 4.7 we can derive iterations that, while not structure preserving, are specifically designed for matrices in automorphism groups. Theorem 4.7 says that computing the square root of $A \in \mathbb{G}$ is equivalent to computing the generalized polar factor W of $I + A$. Theorem 4.6 says that any of a wide class of iterations for the sign of a matrix yields a corresponding iteration for the generalized polar factor W of the matrix. The simplest application of this result is to the Newton iteration for the sign function,

$$X_{k+1} = \frac{1}{2}(X_k + X_k^{-1}), \qquad X_0 = A. \tag{6.5}$$

Applying Theorem 4.6 we deduce that for any A having a generalized polar decomposition $A = WS$, the iteration

$$Y_{k+1} = \frac{1}{2}\bigl(Y_k + Y_k^{-\star}\bigr), \qquad Y_0 = A \tag{6.6}$$

is well-defined and Y_k converges quadratically to W. This iteration is also analyzed by Cardoso, Kenney, and Silva Leite [5, Sec. 4], who treat real bilinear forms only and assume that the matrix M underlying the bilinear form is orthogonal and either symmetric or skew-symmetric. Higham [14] analyzes (6.6) in the special case of the pseudo-orthogonal group. In the special case of the real orthogonals, $M = I$, and (6.6) reduces to the well known Newton iteration for the orthogonal polar factor [10].

On invoking Theorem 4.7 we obtain the matrix square root iteration in the next result.

THEOREM 6.2. *Let $\mathbb{G}$ be any automorphism group and $A \in \mathbb{G}$. If A has no eigenvalues on $\mathbb{R}^-$ then the iteration*

$$\begin{aligned} Y_{k+1} &= \frac{1}{2}\bigl(Y_k + Y_k^{-\star}\bigr) \\ &= \begin{cases} \frac{1}{2}\bigl(Y_k + M^{-1}Y_k^{-T}M\bigr), & \textit{for bilinear forms,} \\ \frac{1}{2}\bigl(Y_k + M^{-1}Y_k^{-*}M\bigr), & \textit{for sesquilinear forms,} \end{cases} \end{aligned} \tag{6.7}$$

with starting matrix $Y_1 = \frac{1}{2}(I + A)$, is well-defined and Y_k converges quadratically to $A^{1/2}$. The iterates Y_k are identical to the X_k ($k \ge 1$) in (6.1) *generated by Newton's method.*

Proof. Only the last part remains to be explained. It is easy to show by induction that $X_k^{\star} = A^{-1}X_k$ ($k \ge 1$), from which $X_k = Y_k$ ($k \ge 1$) follows by a second induction. □

Note that the factor $\frac{1}{2}$ in Y_1 is chosen to ensure that $Y_k \equiv X_k$ for $k \ge 1$; since $\frac{1}{2}(I + A) = W(\frac{1}{2}S)$, W is unaffected by this factor.

Theorem 6.2 shows that for A in an automorphism group the Newton iteration (6.1) can be rewritten in an alternative form—one that has much better numerical stability properties, as we will show below.

[5] In [13], iteration (2.8) therein was rewritten using commutativity to obtain a more efficient form (2.10), which was found to be unstable. This form is (essentially) a particular case of (5.4). If instead (2.8) is rewritten using Corollary 2.2, as we did in deriving (4.6) in Section 4, efficiency is gained without the loss of stability.

The iteration in Theorem 6.2 is also investigated by Cardoso, Kenney, and Silva Leite [5, Sec. 6], with the same assumptions on $\mathbb{G}$ as mentioned above for their treatment of (6.6).

If M is a general matrix then the operation count for (6.7) is higher than that for the Newton iteration (6.1). However, for all the classical groups M is a permutation of $\mathrm{diag}(\pm 1)$ (see Table 2.1) and multiplication by M^{-1} and M is therefore of trivial cost; for these groups the cost of iteration (6.7) is one matrix inversion per iteration, which operation counts show is about 75% the cost per iteration of (6.1) and 30% of that for (6.4).

Matrix Newton iterations benefit from scaling when the starting matrix A is far from the limit. Much is known about scalings for the sign function iteration (6.5) of the form

$$X_{k+1} = \frac{1}{2}\bigl(\alpha_k X_k + \alpha_k^{-1} X_k^{-1}\bigr), \qquad X_0 = A; \tag{6.8}$$

see Kenney and Laub [21]. The corresponding scaled version of (6.7) is

$$Y_{k+1} = \frac{1}{2}\bigl(\gamma_k Y_k + (\gamma_k Y_k)^{-\star}\bigr), \qquad Y_1 = \frac{1}{2}(I + A). \tag{6.9}$$

By considering the discussion just before the proof of Theorem 4.6 we can see how to map α_k into γ_k. In particular, the determinantal scaling of Byers [4], which for $A \in \mathbb{C}^{n\times n}$ takes $\alpha_k = |\det(X_k)^{-1/n}|$ in (6.8), yields

$$\gamma_k = |\det(Y_k)^{-1/n}| \tag{6.10}$$

in (6.9), while the spectral scaling $\alpha_k = (\rho(X_k^{-1})/\rho(X_k))^{1/2}$ of Kenney and Laub [21] yields $\gamma_k = (\rho(Y_k^{-1}Y_k^{-\star})/\rho(Y_k^{\star}Y_k))^{1/4}$. The latter acceleration parameter is suggested in [5]; it has the disadvantage of significantly increasing the cost of each iteration.

Finally, we give another example of the utility of Theorem 4.6. The Schulz iteration

$$X_{k+1} = \frac{1}{2} X_k(3I - X_k^2), \qquad X_0 = A, \tag{6.11}$$

is a member of Kenney and Laub's Padé table of iterations for $\mathrm{sign}(A)$. Applying Theorem 4.6 (or, strictly, a slightly modified version, since (6.11) is not globally convergent), we obtain the iteration

$$Y_{k+1} = \frac{1}{2} Y_k(3I - Y_k^{\star} Y_k), \qquad Y_0 = A \tag{6.12}$$

for computing W, assuming that the generalized polar decomposition $A = WS$ exists. Using a known recurrence for the residuals $I - X_k^2$ of (6.11) [1, Prop. 6.1] we find that

$$R_{k+1} = \frac{3}{4} R_k^2 + \frac{1}{4} R_k^3, \qquad \text{for either } R_k = I - Y_k^{\star} Y_k \text{ or } R_k = I - Y_k Y_k^{\star}.$$

Hence a sufficient condition for the convergence of (6.12) is that the spectral radius $\rho(R_0) = \rho(I - A^{\star}A) < 1$. Iteration (6.12) was stated in [14] for the pseudo-orthogonal group, but the derivation there was ad hoc. Our derivation here reveals the full generality of the iteration.

TABLE 7.1
Results for a perplectic matrix $A \in \mathbb{R}^{7\times 7}$ with $\kappa_2(A) = 10$. Here, $\mathrm{err}(X)$ and $\mu_{\mathbb{G}}(X)$ are defined in (7.1) and (7.2).

k	Newton, (6.1)	(6.9) with $\gamma_k \equiv 1$		(6.9) with γ_k of (6.10)			Cubic, (6.4)	
	$\mathrm{err}(X_k)$	$\mathrm{err}(Y_k)$	$\mu_{\mathbb{G}}(Y_k)$	$\mathrm{err}(Y_k)$	$\mu_{\mathbb{G}}(Y_k)$	γ_k	$\mathrm{err}(Y_k)$	$\mu_{\mathbb{G}}(Y_k)$
0	1.0e+0						1.0e+0	2.5e-15
1	6.1e-1	6.1e-1	4.1e-1	6.1e-1	4.1e-1	1.4e+0	5.1e-1	8.9e-16
2	3.6e-1	3.6e-1	3.7e-1	2.5e-1	2.3e-1	1.1e+0	4.7e-2	4.4e-16
3	8.1e-2	8.1e-2	5.1e-2	2.0e-2	1.6e-2	1.0e+0	4.0e-5	4.7e-16
4	3.5e-3	3.5e-3	2.1e-3	2.3e-4	2.0e-4	1.0e+0	1.7e-14	5.3e-16
5	5.7e-6	5.7e-6	4.0e-6	1.9e-8	1.5e-8	1.0e+0	2.1e-15	4.2e-16
6	1.4e-11	1.4e-11	1.3e-11	2.0e-15	2.1e-16	1.0e+0		
7	2.2e-15	1.9e-15	1.2e-16					

7. Numerical properties. Key to the practical utility of the iterations we have described is their behaviour in floating point arithmetic. We begin by presenting two numerical experiments in which we compute the square root of

- a random perplectic matrix $A \in \mathbb{R}^{7\times 7}$, with $\|A\|_2 = \sqrt{10} = \|A^{-1}\|_2$, generated using an algorithm of D. S. Mackey described in [18],
- a random pseudo-orthogonal matrix $A \in \mathbb{R}^{10\times 10}$, with $p = 6$, $q = 4$ and $\|A\|_2 = 10^5 = \|A^{-1}\|_2$, generated using the algorithm of Higham [14]. The matrix A is also chosen to be symmetric positive definite, to aid comparison with the theory, as we will see later.

For definitions of the perplectic and pseudo-orthogonal groups see Table 2.1. All our experiments were performed in MATLAB, for which $u \approx 1.1 \times 10^{-16}$.

Tables 7.1 and 7.2 display the behavior of the Newton iteration (6.1), the cubic iteration (6.4), iteration (6.9) without scaling, and iteration (6.9) with determinantal scaling (6.10). We report iterations up to the last one for which there was a significant decrease in the error

$$\mathrm{err}(X) = \frac{\|X - A^{1/2}\|_2}{\|A^{1/2}\|_2}. \tag{7.1}$$

We also track the departure from $\mathbb{G}$-structure of the iterates, as measured by

$$\mu_{\mathbb{G}}(X) = \frac{\|X^\star X - I\|_2}{\|X\|_2^2}; \tag{7.2}$$

see Section 7.1 for justification of this measure. The next lemma gives a connection between these two quantities that applies to all the classical groups in Table 2.1.

LEMMA 7.1. *Let $A \in \mathbb{G}$, where $\mathbb{G}$ is the automorphism group of any scalar product for which M is unitary. Then for $X \in \mathbb{K}^{n\times n}$ close to $A^{1/2} \in \mathbb{G}$,*

$$\mu_{\mathbb{G}}(X) \le 2\mathrm{err}(X) + O(\mathrm{err}(X)^2). \tag{7.3}$$

Proof. Let $A \in \mathbb{G}$ and $X = A^{1/2} + E$. Then

$$\begin{aligned} X^\star X - I &= (A^{1/2})^\star (A^{1/2} + E) + E^\star A^{1/2} + E^\star E - I \\ &= A^{-1/2}E + E^\star A^{1/2} + E^\star E. \end{aligned}$$

TABLE 7.2
Results for a pseudo-orthogonal matrix $A \in \mathbb{R}^{10\times 10}$ with $\kappa_2(A) = 10^{10}$. Here, $\mathrm{err}(X)$ *and* $\mu_{\mathbb{G}}(X)$ *are defined in* (7.1) *and* (7.2).

k	Newton, (6.1)	(6.9) with $\gamma_k \equiv 1$		(6.9) with γ_k of (6.10)			Cubic, (6.4)	
	$\mathrm{err}(X_k)$	$\mathrm{err}(Y_k)$	$\mu_{\mathbb{G}}(Y_k)$	$\mathrm{err}(Y_k)$	$\mu_{\mathbb{G}}(Y_k)$	γ_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	1.6e+2	1.0e-5	2.0e-2	1.0e+2	7.2e-15
2	7.8e+1	7.8e+1	1.0e-5	7.4e-1	2.1e-3	3.7e-1	3.4e+1	6.0e-14
3	3.9e+1	3.9e+1	1.0e-5	1.9e-1	1.8e-4	6.5e-1	1.1e+1	5.1e-13
4	1.9e+1	1.9e+1	1.0e-5	6.0e-2	1.7e-5	8.7e-1	3.0e+0	2.9e-12
5	8.9e+0	8.9e+0	9.9e-6	4.9e-3	1.6e-6	9.8e-1	5.5e-1	4.4e-12
6	4.0e+0	4.0e+0	9.6e-6	1.2e-4	3.1e-8	1.0e+0	2.0e-2	4.1e-12
7	3.2e+1	1.6e+0	8.5e-6	3.6e-8	1.4e-11	1.0e+0	2.0e-6	4.1e-12
8	2.3e+5	4.9e-1	5.5e-6	2.1e-11	1.3e-16		2.1e-11	4.1e-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					

Taking 2-norms and using (2.1) and (2.2) gives

$$\begin{aligned}\|X^\star X - I\|_2 &\le \|E\|_2(\|A^{-1/2}\|_2 + \|A^{1/2}\|_2) + \|E\|_2^2\\ &= 2\|E\|_2\,\|A^{1/2}\|_2 + \|E\|_2^2.\end{aligned}$$

The result follows on multiplying throughout by $\|X\|_2^{-2}$ and noting that $\|X\|_2^{-2} = \|A^{1/2}\|_2^{-2} + O(\|E\|_2)$. □

The analysis in Section 6 shows that for $A \in \mathbb{G}$ the Newton iteration (6.1) and iteration (6.9) without scaling generate precisely the same sequence, and this explains the equality of the errors in the first two columns of Tables 7.1 and 7.2 for $1 \le k \le 6$. But for $k > 6$ the computed Newton sequence diverges for the pseudo-orthogonal matrix, manifesting the well known instability of the iteration (even for symmetric positive definite matrices). Table 7.2 shows that scaling brings a clear reduction in the number of iterations for the pseudo-orthogonal matrix and makes the scaled iteration (6.9) more efficient than the cubic iteration in this example.

The analysis of Section 5 shows that the cubic structure-preserving iteration is stable, and for the classical groups it provides an estimate $(1 + \|A^{1/2}\|_2^2)u$ of the relative limiting accuracy. This fits well with the observed errors in Table 7.2, since in this example $\|A^{1/2}\|_2^2 = \|A\|_2 = 10^5$ (which follows from the fact that A is symmetric positive definite). We know from Theorem 5.5 that the unscaled iteration (6.7) is stable if the adjoint is involutory, and the same estimate of the relative limiting accuracy as for the cubic iteration is obtained for the classical groups. These findings again match the numerical results very well.

The original Newton iteration (6.1) has a Fréchet derivative map whose powers are bounded if the eigenvalues λ_i of A satisfy $\frac{1}{2}|1 - \lambda_i^{1/2}\lambda_j^{-1/2}| < 1$ for all i and j [11]. This condition is satisfied for our first test matrix, but not for the second. The term on the left of this inequality also arises in Lemma 5.4 with $B = A^{1/2}$. Hence our theory predicts that the variant of (6.4) that corresponds to (5.4), in which (6.4b) is replaced by $Z_{k+1} = \frac{1}{3}Z_k[I + 8(I + 3Z_kY_k)^{-1}]$, will be unstable for the second matrix. Indeed it is, with minimum error 7.5e-3 occurring at $k = 7$, after which the errors

increase; it is stable for the first matrix.

Turning to the preservation of structure, the values for $\mu_{\mathbb{G}}(Y_k)$ in the tables confirm that the cubic iteration is structure preserving. But Table 7.2 also reveals that for the pseudo-orthogonal matrix, iteration (6.9), with or without scaling, is numerically better at preserving group structure at convergence than the cubic structure-preserving iteration, by a factor 10^4. The same behavior has been observed in other examples. Partial explanation is provided by the following lemma.

LEMMA 7.2. *Assume that* $(A^\star)^\star = A$ *for all* $A \in \mathbb{K}^{n\times n}$. *If*

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

then

$$Y_{k+1}^\star Y_{k+1} - I = \frac{1}{4}\big(Y_k^\star Y_k\big)^{-1}\big(Y_k^\star Y_k - I\big)^2.$$

Proof.

$$\begin{aligned} Y_{k+1}^\star Y_{k+1} - I &= \frac{1}{4}\big(Y_k^\star Y_k + Y_k^\star Y_k^{-\star} + (Y_k^{-\star})^\star Y_k + (Y_k^{-\star})^\star Y_k^{-\star} - 4I\big) \\ &= \frac{1}{4}\big(Y_k^\star Y_k + I + I + Y_k^{-1}Y_k^{-\star} - 4I\big) \\ &= \frac{1}{4}\big(Y_k^\star Y_k\big)^{-1}\big((Y_k^\star Y_k)^2 - 2Y_k^\star Y_k + I\big), \end{aligned}$$

which gives the result. □

Since Lemma 7.2 makes no assumptions about Y_k, we can think of Y_k as being an exact iterate perturbed by errors. The lemma shows that the iteration enforces quadratic convergence to the structure: an arbitrary error introduced at a particular stage can be expected to have rapidly decreasing effect on the departure from structure (though not necessarily on the error). The structure-preserving cubic iteration does not satisfy such a relation: while it automatically preserves structure, it has no mechanism for reducing a loss of structure caused by arbitrary perturbations in the iterates. However, as Lemma 7.1 shows, for any method the loss of structure is approximately bounded by the relative error, so severe loss of structure in the cubic iteration can occur only for ill conditioned problems.

7.1. Justification of measure $\mu_{\mathbb{G}}(A)$. The measure of structure $\mu_{\mathbb{G}}$ in (7.2) was used in [15] and justified by Lemma 4.2 therein, which shows that if A has a generalized polar decomposition $A = WS$, the matrix M of the scalar product is unitary, and $\|S - I\|_2 < 1$, then $W \in \mathbb{G}$ is within relative distance approximately $\mu_{\mathbb{G}}(A)$ of A. In Theorem 7.4 below we simplify this result to assume only that $\|A^\star A - I\| < 1$, and strengthen it to apply to any consistent norm and any scalar product.

LEMMA 7.3. *Suppose that* $\operatorname{sign}(S) = I$ *and* $S^2 = I + E$ *where* $\|E\| < 1$, *for any consistent norm. Then*

$$\|S - I\| \le \frac{\|E\|}{1 + \sqrt{1 - \|E\|}} < \|E\|.$$

Proof. We will make use of the observation that if $|x| < 1$ then $(1 + x)^{1/2}$ has a convergent Maclaurin series $1 + \sum_{k=1}^\infty a_k x^k$ such that $\sum_{k=1}^\infty |a_k||x|^k = 1 - \sqrt{1 - x}$.

Since $\operatorname{sign}(S) = I$ we have $S = (S^2)^{1/2}$ and hence $S = (I+E)^{1/2} = I + \sum_{k=1}^{\infty} a_k E^k$, since $\|E\| < 1$. Then

$$\begin{aligned}\|S - I\| &= \left\| \sum_{k=1}^{\infty} a_k E^k \right\| \le \sum_{k=1}^{\infty} |a_k| \|E\|^k \\ &= 1 - \sqrt{1 - \|E\|} = \frac{\|E\|}{1 + \sqrt{1 - \|E\|}} < \|E\|. \qquad \square\end{aligned}$$

The following theorem generalizes [12, Lem. 5.1], [14, Lem. 5.3] and [15, Lem. 4.2].

THEOREM 7.4. *Let $\mathbb{G}$ be the automorphism group of a scalar product. Suppose that $A \in \mathbb{K}^{n\times n}$ satisfies $(A^\star)^\star = A$ and $\|A^\star A - I\| < 1$. Then A has a generalized polar decomposition $A = WS$ and, for any consistent norm, the factors W and S satisfy*

$$\frac{\|A^\star A - I\|}{\|A\|(\|A^\star\| + \|W^\star\|)} \le \frac{\|A - W\|}{\|A\|} \le \frac{\|A^\star A - I\|}{\|A\|^2} \|A\| \, \|W\|, \tag{7.4}$$

$$\frac{\|A^\star A - I\|}{\|S\| + \|I\|} \le \|S - I\| \le \|A^\star A - I\|. \tag{7.5}$$

The inequalities (7.4) *can be rewritten as*

$$\frac{\mu_{\mathbb{G}}(A)\|A\|}{\|A^\star\| + \|W^\star\|} \le \frac{\|A - W\|}{\|A\|} \le \mu_{\mathbb{G}}(A) \, \|A\| \, \|W\|.$$

Proof. The condition $\|A^\star A - I\| < 1$ implies that the spectral radius of $A^\star A - I$ is less than 1, and hence that $A^\star A$ has no eigenvalues on $\mathbb{R}^-$. Since $(A^\star)^\star = A$, Theorem 4.1 implies that A has a (unique) generalized polar decomposition $A = WS$. Using $W^\star = W^{-1}$ and $S^\star = S$ we have

$$\begin{aligned}(A+W)^\star(A-W) &= A^\star A - A^\star W + W^\star A - W^\star W \\ &= A^\star A - S^\star W^\star W + W^\star W S - I = A^\star A - I.\end{aligned}$$

The lower bound in (7.4) follows on taking norms and using $\|(A+W)^\star\| = \|A^\star + W^\star\| \le \|A^\star\| + \|W^\star\|$.

The upper bound in (7.5) follows from Lemma 7.3, since

$$A^\star A - I = S^\star W^\star W S - I = S^2 - I. \tag{7.6}$$

The upper bound in (7.4) then follows by taking norms in $A - W = WS - W = W(S - I)$. Finally, the lower bound in (7.5) follows by writing (7.6) as $A^\star A - I = (S-I)(S+I)$ and taking norms. $\square$

Note that the term $\|A^\star\|$ in the denominator of (7.4) can be replaced by $\kappa(M)\|A^T\|$ or $\kappa(M)\|A^*\|$ for bilinear forms and sesquilinear forms, respectively, and for a unitarily invariant norm both expressions are just $\|A\|$ for all the groups in Table 2.1; likewise for $\|W^\star\|$.

7.2. Conclusions on choice of method for $A^{1/2}$ when $A \in \mathbb{G}$. Our overall conclusion is that the rewritten form (6.9) of Newton's iteration, with the scaling (6.10) or perhaps some alternative, is the best iteration method for computing the square root of a matrix A in an automorphism group. This iteration

- Overcomes the instability in the standard Newton iteration (6.1) and is less costly per iteration than (6.1) for the classical groups.
- Is generally more efficient than the cubic structure-preserving iteration (6.4): it costs significantly less per iteration than (6.4), and (6.4) typically requires approximately the same number of iterations.
- When iterated to convergence to machine precision, is likely to produce a computed result lying closer to the group than the cubic iteration (6.4) when A is ill conditioned.
- For the classical groups has half the cost per iteration of the mathematically equivalent Denman–Beavers iteration recommended in [13]. In fact, another way to derive (6.7) is to exploit the structure in the Denman–Beavers iteration that results when $A \in \mathbb{G}$.

If a structure-preserving iteration is required then an iteration from the family (6.2) can be recommended, such as the cubically convergent iteration (6.4). These iterations have the advantage that even if they are terminated well before convergence to machine precision, the result will lie in the group to approximately machine precision, though some loss of structure (no worse than that described by (7.3)) may occur for ill conditioned problems.

8. Acknowledgements. We thank a referee for helpful comments and suggestions.

REFERENCES

[1] Dario A. Bini, Nicholas J. Higham, and Beatrice Meini, *Computing the matrix pth root*, numerical analysis report, Manchester Centre for Computational Mathematics, Manchester, England, 2004. In preparation.
[2] Yuri Bolshakov, Cornelis V. M. van der Mee, André C. M. Ran, Boris Reichstein, and Leiba Rodman, *Extension of isometries in finite-dimensional indefinite scalar product spaces and polar decompositions*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 752–774.
[3] Yuri Bolshakov, Cornelis V. M. van der Mee, André C. M. Ran, Boris Reichstein, and Leiba Rodman, *Polar decompositions in finite dimensional indefinite scalar product spaces: General theory*, Linear Algebra Appl., 261 (1997), pp. 91–141.
[4] Ralph Byers, *Solving the algebraic Riccati equation with the matrix sign function*, Linear Algebra Appl., 85 (1987), pp. 267–279.
[5] João R. Cardoso, Charles S. Kenney, and F. Silva Leite, *Computing the square root and logarithm of a real P-orthogonal matrix*, Appl. Numer. Math., 46 (2003), pp. 173–196.
[6] Henri Cartan, *Differential Calculus*, Hermann, Paris, 1971.
[7] Sheung Hun Cheng, Nicholas J. Higham, Charles S. Kenney, and Alan J. Laub, *Approximating the logarithm of a matrix to specified accuracy*, SIAM J. Matrix Anal. Appl., 22 (2001), pp. 1112–1125.
[8] Israel Gohberg, Peter Lancaster, and Leiba Rodman, *Matrices and Indefinite Scalar Products*, Birkhäuser, Basel, Switzerland, 1983.
[9] Robert E. Greene and Steven G. Krantz, *Function Theory of One Complex Variable*, Wiley, New York, 1997.
[10] Nicholas J. Higham, *Computing the polar decomposition—with applications*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1160–1174.
[11] ———, *Newton's method for the matrix square root*, Math. Comp., 46 (1986), pp. 537–549.
[12] ———, *The matrix sign decomposition and its relation to the polar decomposition*, Linear Algebra Appl., 212/213 (1994), pp. 3–20.
[13] ———, *Stable iterations for the matrix square root*, Numerical Algorithms, 15 (1997), pp. 227–242.
[14] ———, *J-orthogonal matrices: Properties and generation*, SIAM Rev., 45 (2003), pp. 504–519.
[15] Nicholas J. Higham, D. Steven Mackey, Niloufer Mackey, and Françoise Tisseur, *Computing the polar decomposition and the matrix sign decomposition in matrix groups*, SIAM J. Matrix Anal. Appl., (2004). To appear.

[16] Roger A. Horn and Charles R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, 1991.
[17] Khakim D. Ikramov, *Hamiltonian square roots of skew-Hamiltonian matrices revisited*, Linear Algebra Appl., 325 (2001), pp. 101–107.
[18] David P. Jagger, *MATLAB toolbox for classical matrix groups*, M.Sc. Thesis, University of Manchester, Manchester, England, Sept. 2003.
[19] Irving Kaplansky, *Algebraic polar decomposition*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 213–217.
[20] Charles S. Kenney and Alan J. Laub, *Rational iterative methods for the matrix sign function*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 273–291.
[21] ———, *On scaling Newton's method for polar decomposition and the matrix sign function*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 688–706.
[22] ———, *The matrix sign function*, IEEE Trans. Automat. Control, 40 (1995), pp. 1330–1348.
[23] Pentti Laasonen, *On the iterative solution of the matrix equation* $AX^2 - I = 0$, M.T.A.C., 12 (1958), pp. 109–116.
[24] Peter Lancaster and Miron Tismenetsky, *The Theory of Matrices*, Academic Press, London, second ed., 1985.
[25] D. Steven Mackey, Niloufer Mackey, and Françoise Tisseur, *Structured tools for structured matrices*, The Electronic Journal of Linear Algebra, 10 (2003), pp. 106–145.
[26] ———, *Structured factorizations in scalar product spaces*, Numerical Analysis Report No. 432, Manchester Centre for Computational Mathematics, Manchester, England, 2004. In preparation.
[27] Reinhold Remmert, *Theory of Complex Functions*, Springer-Verlag, Berlin, 1991.
[28] J. D. Roberts, *Linear model reduction and solution of the algebraic Riccati equation by use of the sign function*, Internat. J. Control, 32 (1980), pp. 677–687. First issued as report CUED/B-Control/TR13, Department of Engineering, University of Cambridge, 1971.
[29] Jacob T. Schwartz, *Nonlinear Functional Analysis*, Gordon and Breach, New York, 1969.