The University of Manchester

Computing the Polar Decomposition and the Matrix Sign Decomposition in Matrix Groups

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

2004

MIMS EPrint: **2005.17**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: `http://www.manchester.ac.uk/mims/eprints`

And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 4, pp. 1178–1192

COMPUTING THE POLAR DECOMPOSITION AND THE MATRIX SIGN DECOMPOSITION IN MATRIX GROUPS*

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

Abstract. For any matrix automorphism group $\mathbb{G}$ associated with a bilinear or sesquilinear form, Mackey, Mackey, and Tisseur have recently shown that the matrix sign decomposition factors of $A \in \mathbb{G}$ also lie in $\mathbb{G}$; moreover, the polar factors of A lie in $\mathbb{G}$ if the matrix of the underlying form is unitary. Groups satisfying the latter condition include the complex orthogonal, real and complex symplectic, and pseudo-orthogonal groups. This work is concerned with exploiting the structure of $\mathbb{G}$ when computing the polar and matrix sign decompositions of matrices in $\mathbb{G}$. We give sufficient conditions for a matrix iteration to preserve the group structure and show that a family of globally convergent rational Padé-based iterations of Kenney and Laub satisfy these conditions. The well-known scaled Newton iteration for computing the unitary polar factor does not preserve group structure, but we show that the approach of the iterates to the group is precisely tethered to the approach to unitarity, and that this forces a different and exploitable structure in the iterates. A similar relation holds for the Newton iteration for the matrix sign function. We also prove that the number of iterations needed for convergence of the structure-preserving methods can be precisely predicted by running an associated scalar iteration. Numerical experiments are given to compare the cubically and quintically converging iterations with Newton's method and to test stopping criteria. The overall conclusion is that the structure-preserving iterations and the scaled Newton iteration are all of practical interest, and which iteration is to be preferred is problem-dependent.

Key words. automorphism group, bilinear form, sesquilinear form, adjoint, complex orthogonal matrix, symplectic matrix, perplectic matrix, pseudo-orthogonal matrix, polar decomposition, matrix sign decomposition, structure preservation, matrix iteration, Newton iteration, convergence tests

AMS subject classifications. 65F30, 15A18

DOI. 10.1137/S0895479803426644

1. Introduction.

The polar decomposition of $A \in \mathbb{C}^{n\times n}$ factors A as the product $A = UH$, where U is unitary and H is Hermitian positive semidefinite. The Hermitian factor H is always unique and can be expressed as $(A^*A)^{1/2}$, and the unitary factor is unique if A is nonsingular [13]. Here, the exponent $1/2$ denotes the principal square root: the one whose eigenvalues lie in the right half-plane. The polar decomposition is an important theoretical and computational tool, and much is known about its approximation properties, its sensitivity to perturbations, and its computation.

Closely related to the polar decomposition is the matrix sign decomposition, which is defined for $A \in \mathbb{C}^{n\times n}$ having no pure imaginary eigenvalues. The most concise definition of the decomposition is

$$A = SN \equiv A(A^2)^{-1/2} \cdot (A^2)^{1/2}.$$

*Received by the editors April 23, 2003; accepted for publication (in revised form) by D. Bini October 8, 2003; published electronically August 4, 2004. This work was supported by Engineering and Physical Sciences Research Council Visiting Fellowship GR/S15563/01.

http://www.siam.org/journals/simax/25-4/42664.html

†Department of Mathematics, University of Manchester, Manchester, M13 9PL, England (higham@ma.man.ac.uk, http://www.ma.man.ac.uk/~higham/; ftisseur@ma.man.ac.uk, http://www.ma.man.ac.uk/~ftisseur/). The work of the first author was supported by Engineering and Physical Sciences Research Council grant GR/R22612. The work of the fourth author was supported by Engineering and Physical Sciences Research Council grant GR/R45079.

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

Here, $S = \mathrm{sign}(A)$ is the matrix sign function, introduced by Roberts [23]. Note that for scalar $z \in \mathbb{C}$ lying off the imaginary axis, $\mathrm{sign}(z) = 1$ or -1 according as z is in the right half-plane or left half-plane, respectively. An alternative definition is via the Jordan canonical form

$$A = ZJZ^{-1} = Z\mathrm{diag}(J_1, J_2)Z^{-1},$$

where the eigenvalues of J_1 are assumed to lie in the open left half-plane and those of J_2 in the open right half-plane. With this notation,

$$A = SN \equiv Z\mathrm{diag}(-I, I)Z^{-1} \cdot Z\mathrm{diag}(-J_1, J_2)Z^{-1},$$

from which it is clear that S is involutory and the eigenvalues of N lie in the right half-plane.

The polar and matrix sign decompositions are intimately connected [9]. For example, Roberts' integral formula [23],

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

has an analogue in

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

This example illustrates the rule of thumb that any property or iteration involving the matrix sign function can be converted into one for the polar decomposition by replacing A^2 by A^*A, and vice versa.

Practical interest in the polar decomposition stems mainly from the fact that the unitary polar factor of A is the nearest unitary matrix to A in any unitarily invariant norm [6]. The polar decomposition is therefore of interest whenever it is required to orthogonalize a matrix [8]. The matrix sign function was originally developed as a tool to solve algebraic Riccati equations [23] and it is also used more generally in determining invariant subspaces corresponding to eigenvalues lying in particular regions of the complex plane [1].

Almost all existing work on the polar decomposition and the matrix sign decomposition assumes no special properties of A. However, in an investigation of factorizations in structured classes of matrices, Mackey, Mackey, and Tisseur [22] consider, among other things, the structure of the polar and sign factors. The structures they work with include the automorphism group

$$\mathbb{G} = \{\, A \in \mathbb{K}^{n\times n} : \langle Ax, Ay\rangle_{\mathrm{M}} = \langle x, y\rangle_{\mathrm{M}} \ \forall x, y \in \mathbb{K}^n \,\} \tag{1.1}$$

associated with a bilinear or sesquilinear form defined by any nonsingular matrix M:

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

Here $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$ and the superscript $*$ denotes conjugate transpose. It is easy to see that $\mathbb{G}$ is indeed a group under matrix multiplication. Recall that the adjoint $A^\star$ of $A \in \mathbb{K}^{n\times n}$ with respect to $\langle\cdot,\cdot\rangle_{\mathrm{M}}$ is defined by

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

TABLE 1.1
A sampling of automorphism groups $\mathbb{G} \in \mathfrak{U}$.

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
$\mathbb{R}^n$	R	$A^\star = R A^T R$	Real perplectics
$\mathbb{R}^{2n}$	J	$A^\star = -J A^T J$	Real symplectics
$\mathbb{C}^{2n}$	J	$A^\star = -J A^T J$	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	$A^\star = -J A^* J$	Conjugate symplectics

It can be shown that the adjoint is given explicitly by

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

The adjoint provides a useful alternative characterization of the automorphism group:

$$\mathbb{G} = \{\, A \in \mathbb{K}^{n\times n} : A^\star = A^{-1} \,\}. \tag{1.2}$$

For further details of this background algebra see, for example, [14], [19], or [24].

Automorphism groups for which the matrix M defining the underlying form is unitary ($M^{-1} = M^*$) play an important role in this paper. We use $\mathfrak{U}$ to denote this set of groups. Table 1.1 lists some examples of automorphism groups in $\mathfrak{U}$; here, the matrix M is one of I,

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

We need the following results, all from [22].

THEOREM 1.1. *Let $\mathbb{G} \in \mathfrak{U}$. Then any matrix in $\mathbb{G}$ has singular values that occur in reciprocal pairs σ and $1/\sigma$, with the same multiplicity.*

THEOREM 1.2. *Let $\mathbb{G} \in \mathfrak{U}$ and A be any matrix in $\mathbb{G}$. Then in the polar decomposition $A = UH$ the factors U and H also belong to $\mathbb{G}$.*

The following result places no restrictions on $\mathbb{G}$.

THEOREM 1.3. *Let $\mathbb{G}$ be any automorphism group and A be any matrix in $\mathbb{G}$ having a matrix sign decomposition $A = SN$. Then the factors S and N also belong to $\mathbb{G}$.*

We give proofs of Theorems 1.2 and 1.3 at the end of section 2 that provide alternatives to the proofs in [22].

For the real orthogonal and unitary groups, Theorems 1.2 and 1.3 are trivial. For the other groups the results are nontrivial, and indeed in three recent papers devoted to some of these groups the structured nature of the polar and sign factors of matrices in the groups is not noted [3], [4], [11].

This work is concerned with the exploitation of structure when computing the polar or sign factors of matrices from an automorphism group. In section 2 we identify

a general family of rational iterations that are structure-preserving, and we show that certain globally convergent Padé-based iterations of Kenney and Laub belong to this family. In sections 3–6 we concentrate on the polar decomposition, for $A \in \mathbb{G}$ and $\mathbb{G} \in \mathfrak{U}$. In section 3 we identify the most efficient implementations of the cubically and quintically convergent iterations and compare them for efficiency with the scaled Newton iteration. In section 4 we show that although the Newton iteration does not preserve the group structure, under a suitable condition on the scaling parameter it has the remarkable property that its iterates X_k satisfy $X_k^\star = X_k^*$. This relation implies that the approach of the iterates to the group is precisely tethered to the approach to unitarity, and also that for certain automorphism groups the iterates have a different structure that can be exploited.

Numerical stability of the iterations is discussed in section 5. In section 6 we show that the number of iterations needed by one of our structure-preserving methods can be predicted by running the corresponding scalar iteration starting with the largest singular value of A. Corresponding results for the matrix sign decomposition are summarized in section 7. Numerical experiments are presented in section 8 that compare the Newton and quintic iterations and test different stopping criteria. Finally, conclusions are given in section 9.

2. Structure-preserving iterations. A great deal is known about matrix iterations of the form

$$X_{k+1} = f(X_k), \qquad X_0 = A,$$

for computing the unitary polar factor or the matrix sign function; see, for example, [8], [9], [15], [16]. Motivated by Theorems 1.2 and 1.3, we ask the question, "If A belongs to an automorphism group $\mathbb{G}$, when do all the iterates X_k also belong to $\mathbb{G}$?" When this property holds, we say the iteration is structure-preserving for $\mathbb{G}$. Sufficient conditions for such an iteration are given in the next theorem. For a polynomial p of degree m we introduce the notation $\mathrm{rev}p(x) = x^m p(1/x)$; thus $\mathrm{rev}p$ is p with its coefficients reversed. (Note that $\mathrm{rev}(\mathrm{rev}p)$ is not necessarily p, as the example $p(x) = x^2 + x$ shows.)

THEOREM 2.1. *Let p be any polynomial with real coefficients and let f be a matrix function having the form*

$$f(X) = Yp(Z)[\mathrm{rev}p(Z)]^{-1}. \tag{2.1}$$

Assume that the appropriate inverses exist, so that f is well defined.

(a) *If Y and Z are integer powers of $X \in \mathbb{G}$, then $f(X) \in \mathbb{G}$ for any automorphism group $\mathbb{G}$.*

(b) *If Y and Z are finite products of X, X^{-1} and X^*, in any combination, where $X \in \mathbb{G}$, then $f(X) \in \mathbb{G}$, for any automorphism group $\mathbb{G} \in \mathfrak{U}$.*

Proof. We note first the properties that $(ST)^\star = T^\star S^\star$ for all S and T and $(S^{-1})^\star = (S^\star)^{-1}$ for all nonsingular S, the latter equality implying that we can write $S^{-\star}$ without ambiguity. Observe also that since p has real coefficients, $p(T)^\star = p(T^\star)$ for all T.

For part (a), Y and Z are readily seen to belong to $\mathbb{G}$, since $\mathbb{G}$ is a group under multiplication. For part (b), $\mathbb{G} \in \mathfrak{U}$ implies $M^{-1} = M^*$, and so $(T^\star)^* = (T^*)^\star$ for all T. Consequently $X \in \mathbb{G}$ implies $X^* \in \mathbb{G}$; hence Y and Z belong to $\mathbb{G}$.

Marshalling these facts, and denoting by m the degree of p, we obtain

$$f(X)^\star f(X) = [\mathrm{rev}p(Z)]^{-\star} \cdot p(Z)^\star \cdot \underbrace{Y^\star \cdot Y}_{I} \cdot p(Z) \cdot [\mathrm{rev}p(Z)]^{-1}$$

$$
\begin{aligned}
&= \big([\mathrm{rev}p(Z)]^\star\big)^{-1} \cdot p(Z^\star) \cdot p(Z) \cdot [\mathrm{rev}p(Z)]^{-1} \\
&= \big([\mathrm{rev}p(Z^\star)]\big)^{-1} \cdot p(Z^{-1}) \cdot p(Z) \cdot [\mathrm{rev}p(Z)]^{-1} \\
&= \big(Z^{-m}p(Z)\big)^{-1} \cdot p(Z) \cdot p(Z^{-1}) \cdot [Z^m p(Z^{-1})]^{-1} \\
&= I. \qquad \square
\end{aligned}
$$

We mention that a converse of part (a) in Theorem 2.1 is proved in [12], from which it follows that any rational function f that maps $\mathbb{G}$ into itself for *all* $\mathbb{G}$ can be expressed in the form (2.1), with p a polynomial with real coefficients.

Theorem 2.1 says nothing about the convergence of the iteration $X_{k+1} = f(X_k)$, so further restrictions on f are needed to obtain a useful iteration. By using only elementary means, one can construct rational iteration functions of the form (2.1) with any specified odd order of convergence. The first two functions in this sequence are

$$x_{k+1} = f_{11}(x_k), \qquad f_{11}(x) = \frac{x(3+x^2)}{1+3x^2}, \tag{2.2}$$

$$x_{k+1} = f_{22}(x_k), \qquad f_{22}(x) = \frac{x(5+10x^2+x^4)}{1+10x^2+5x^4}, \tag{2.3}$$

which, for $x_0 \in \mathbb{C}$ not on the imaginary axis, converge to $\mathrm{sign}(x_0)$ at a cubic[1] and quintic rate, respectively. See [20] for details of this approach. It turns out that the functions f_{ii} thus constructed belong to the family of rational iterations

$$x_{k+1} = f_{\ell m}(x_k) = x_k \frac{P_{\ell m}(1-x_k^2)}{Q_{\ell m}(1-x_k^2)} \tag{2.4}$$

studied by Kenney and Laub [15], where $P_{\ell m}(t)/Q_{\ell m}(t)$ is the $[\ell/m]$ Padé approximant to $(1-t)^{-1/2}$, with the polynomials $P_{\ell m}$ and $Q_{\ell m}$ having degrees ℓ and m, respectively. These iterations are designed to compute $\mathrm{sign}(x_0)$, and the iterations with $\ell = m$ and $\ell = m-1$ are shown in [15] to be globally convergent, that is, they converge to $\mathrm{sign}(x_0)$ for any $x_0 \in \mathbb{C}$ not on the imaginary axis. For $\ell = m$ and $\ell = m-1$ it is also noted in [15] that $-xP_{\ell m}(1-x^2)$ and $Q_{\ell m}(1-x^2)$ are, respectively, the odd and even parts of $(1-x)^{\ell+m+1}$. For $\ell = m-1$, the iteration is easily verified not to be structure-preserving for $\mathbb{G}$. But from the odd-even property just mentioned, the iteration for $\ell = m$ can be seen to have the form (2.1) with $Y = Z = X$, and therefore by part (a) of Theorem 2.1 the iteration is structure-preserving for all automorphism groups $\mathbb{G}$.

THEOREM 2.2. *Let $A \in \mathbb{K}^{n\times n}$ and consider the iterations*

$$Y_{k+1} = Y_k\, P_{mm}(I - Y_k^2)\, Q_{mm}(I - Y_k^2)^{-1}, \qquad Y_0 = A, \tag{2.5}$$

and

$$Z_{k+1} = Z_k\, P_{mm}(I - Z_k^* Z_k)\, Q_{mm}(I - Z_k^* Z_k)^{-1}, \qquad Z_0 = A, \tag{2.6}$$

with $m \geq 1$. Assume that A has no eigenvalues on the imaginary axis for (2.5).

(a) *If $\mathbb{G}$ is any automorphism group and $A \in \mathbb{G}$, then $Y_k \in \mathbb{G}$ for all k, and Y_k converges to* $\mathrm{sign}(A)$.

[1] The iteration (2.2) is Halley's method for $x^2 - 1 = 0$ [7].

(b) *If $\mathbb{G}$ is any automorphism group in $\mathfrak{U}$ and $A \in \mathbb{G}$, then $Z_k \in \mathbb{G}$ for all k, and Z_k converges to the unitary polar factor of A.*

Moreover, both sequences have order of convergence $2m+1$.

Proof. The preservation of structure has already been shown. It remains to prove convergence. The existence of the inverse in (2.5) and the global convergence of (2.5) to $\operatorname{sign}(A)$ with order $2m+1$ are established in [15, Thm. 5.3]. That (2.6) converges globally to the unitary polar factor at the same rate can be shown by using the singular value decomposition of A to reduce (2.6) to n independent scalar iterations on the singular values, whose convergence to 1 follows from that of (2.5). ∎

A proof of Theorem 1.3 now follows immediately from part (a) of Theorem 2.2: since $\mathbb{G}$ is a closed set, $\lim Y_k = \operatorname{sign}(A)$ belongs to $\mathbb{G}$, and since $\mathbb{G}$ is a group under multiplication, the factor N in the matrix sign decomposition of A must also belong to $\mathbb{G}$. In an entirely analogous way, a proof of Theorem 1.2 follows from part (b) of Theorem 2.2.

In the next four sections we restrict our attention to the polar decomposition. In section 7 we explain to what extent our analysis for the polar decomposition can be adapted to the matrix sign decomposition.

3. Iterations for the polar decomposition. We begin by examining the first two iterations of the previous section and their computational cost.

The cubically convergent iteration (2.2) is, in matrix form for computing the polar decomposition,

$$X_{k+1} = X_k(3I + X_k^*X_k)(I + 3X_k^*X_k)^{-1}, \qquad X_0 = A. \tag{3.1}$$

We will measure the cost of iterations by counting the number of (general) matrix multiplications, *mult*, and the number of (general) matrix inversions, *inv*. When evaluating a term of the form AB^{-1} it is less expensive to factor B and then solve a multiple right-hand-side linear system than to explicitly invert B, so we will assume the former is done and record the cost as a corresponding multiple of *inv*. In our iterations, B is Hermitian positive definite and AB^{-1} is Hermitian; if we exploit this structure the cost of computing AB^{-1} is $(5/6)inv$.

One iteration of (3.1) costs $(3/2)\,mult + (5/6)\,inv$ per iteration. By rewriting the iteration the cost can be reduced: for

$$X_{k+1} = \frac{1}{3}X_k\big[I + 8\big(I + 3X_k^*X_k\big)^{-1}\big], \qquad X_0 = A, \tag{3.2}$$

the cost per iteration is $(3/2)\,mult + (1/2)inv$.

The quintically convergent iteration (2.3) becomes

$$X_{k+1} = X_k\big[5I + 10X_k^*X_k + (X_k^*X_k)^2\big]\big[I + 10X_k^*X_k + 5(X_k^*X_k)^2\big]^{-1}, \quad X_0 = A, \tag{3.3}$$

which costs $2\,mult + (5/6)\,inv$ per iteration. This iteration can be rewritten in various more efficient ways. We state the two of most interest in scalar form, for readability; for matrices, x^2 should be replaced by $X_k^*X_k$ and the divisions by matrix inversions. First, we have the continued fraction form

$$x_{k+1} = \frac{1}{5}x_k\left[1 + \cfrac{8}{x^2 + \cfrac{7}{5 + \cfrac{16}{7x^2+1}}}\right], \qquad x_0 = a, \tag{3.4}$$

which costs $(3/2)(mult + inv)$ per iteration. The alternative form

$$x_{k+1} = x_k \left[\frac{1}{5} + \frac{8}{5x_k^2 + 7 - \dfrac{16}{5x_k^2 + 3}} \right], \qquad x_0 = a, \tag{3.5}$$

requires just $(3/2)\,mult + inv$ per iteration, so is the least expensive of the three variants.

For any m, the iterations (2.6) can also be expressed in partial fraction form [17, (35)]. The cost of evaluation is $(3/2)\,mult + (m/2)\,inv$ per iteration, which in the cases $m = 1$ and $m = 2$ is the same as the cost of evaluating (3.2) and (3.5), respectively.

Also of interest is the well-known Newton iteration

$$X_{k+1} = \frac{1}{2}(X_k + X_k^{-*}), \qquad X_0 = A. \tag{3.6}$$

This iteration is not structure-preserving for automorphism groups $\mathbb{G}$, since $\mathbb{G}$ is not closed under addition, but, as we will see, the iteration is nevertheless of interest when $A \in \mathbb{G}$. In practice, the Newton iteration is usually implemented with scaling to accelerate the initial speed of convergence. The scaled iteration is

$$X_{k+1} = \frac{1}{2}\left[\gamma^{(k)} X_k + \frac{1}{\gamma^{(k)}} X_k^{-*} \right], \qquad X_0 = A, \tag{3.7}$$

where the scaling parameter $\gamma^{(k)} \in \mathbb{R}$ is intended to make X_{k+1} closer to U. Higham [8] identified the scaling

$$\gamma_{\mathrm{opt}}^{(k)} = \bigl(\sigma_{\min}(X_k)\sigma_{\max}(X_k)\bigr)^{-1/2},$$

where $\sigma_{\min}$ and $\sigma_{\max}$ denote the smallest and largest singular values, respectively, as optimal in the sense of minimizing a bound on $\|U - X_{k+1}\|_2 / \|U + X_{k+1}\|_2$, and this scaling leads to convergence in s iterations, where s is the number of distinct singular values of A [16, Lem. 2.2]. Among more economical choices analyzed in [16] is the Frobenius norm scaling

$$\gamma_F^{(k)} = \left(\frac{\|X_k^{-1}\|_F}{\|X_k\|_F} \right)^{1/2}, \tag{3.8}$$

which has the property that it minimizes $\|X_{k+1}\|_F$ over all $\gamma^{(k)}$ [5]. Both these scalings have the property that

$$X_0 \in \mathbb{G} \in \mathfrak{U} \quad \Rightarrow \quad \gamma^{(0)} = 1, \tag{3.9}$$

by virtue of the reciprocal pairing of the singular values when $\mathbb{G} \in \mathfrak{U}$ (see Theorem 1.1).

We do not investigate scaling for the structure-preserving iterations, because for $X_k \in \mathbb{G}$ and f in (2.1), $f(\gamma^{(k)} X_k) \notin \mathbb{G}$ in general, and so scaling destroys group structure.

We first ask which of the three iterations (3.2), (3.5), and (3.6) is the most computationally efficient, independent of structure considerations. In answering this question we need to take account of the fact that the iterations comprise two phases: the

TABLE 3.1
Cost estimates for (3.2), (3.5), *and* (3.6), *assuming* $\|X_0 - U\|_2 = 0.25$.

Iterations	Resulting error	Cost
2 quintic iterations	$0.25^{25} = 9 \times 10^{-16}$	$3\,mult + 2\,inv$
3 cubic iterations	$0.25^{27} = 6 \times 10^{-17}$	$(9/2)mult + (3/2)inv$
5 Newton iterations	$0.25^{32} = 5 \times 10^{-20}$	$5\,inv$

initial phase in which the error $\|X_k - U\|_2$ is reduced to safely less than 1, and then the phase in which asymptotic convergence sets in at a quadratic, cubic, or quintic rate. Working in IEEE double precision arithmetic, the furthest we iterate is until the error reaches the unit roundoff $u \approx 10^{-16}$, so the higher order iterations are not in their asymptotic phase for long. In the initial phase, our three iterations converge essentially linearly, with rate constants 1/2, 1/3, and 1/5, respectively (this can be seen by considering the scalar iterations with $0 < x_0 \ll 1$ and $x_0 \gg 1$). Hence for large $\|X_0 - U\|_2$ the quintic iteration requires the least work to reduce the error below 1, followed by the cubic and then the Newton iterations. Once the error is safely below 1, the three iterations cost roughly the same amount to reduce the error to the unit roundoff level; see Table 3.1. Our conclusion is that if $\|X_0 - U\|_2 \lesssim 1$ there is little to choose between the iterations in cost, but for $\|X_0 - U\|_2 \gg 1$ the quintic iteration has the advantage. In the scaled Newton iteration (3.7), with the Frobenius norm scaling (3.8), the first phase of convergence is shortened considerably. Practical experience, supported by theory [16], shows that about 9 or 10 iterations at most are required for any A in IEEE double precision arithmetic. Therefore scaled Newton is competitive in cost with the quintic iteration, albeit not structure-preserving.

4. Structure in the Newton iteration. We have seen that the cubic iteration (3.2) and the quintic iteration (3.5) are structure-preserving for automorphism groups $\mathbb{G} \in \mathfrak{U}$, while the Newton iteration (3.6) and its scaled form (3.7) are not. We now consider precisely how the Newton iteration affects structure, and to do so we first develop a measure of departure from $\mathbb{G}$-structure. Throughout the rest of this section we assume that $\mathbb{G} \in \mathfrak{U}$, that is, that M is unitary.

The characterization (1.2) says that A is in the automorphism group $\mathbb{G}$ if $A^\star = A^{-1}$. To obtain a measure of departure from $\mathbb{G}$-structure that is less dependent on the conditioning of A, we rewrite this relation as $A^\star A = I$. Consider $A + \Delta A$, where $A \in \mathbb{G}$ and $\|\Delta A\|_2 \le \epsilon \|A\|_2$. Using the fact that M is unitary, we have $\|A^\star\|_2 = \|A\|_2$ for all A, and hence

$$\begin{aligned}\|(A+\Delta A)^\star(A+\Delta A) - I\|_2 &= \|A^\star \Delta A + \Delta A^\star A + \Delta A^\star \Delta A\|_2 \\ &\le 2\|A\|_2\|\Delta A\|_2 + \|\Delta A\|_2^2 \\ &\le (2\epsilon + \epsilon^2)\|A\|_2^2.\end{aligned}$$

This inequality suggests that an appropriate relative measure of departure from $\mathbb{G}$-structure is

$$\mu_{\mathbb{G}}(A) = \frac{\|A^\star A - I\|_2}{\|A\|_2^2}. \tag{4.1}$$

In the particular case $\mathbb{G} = \mathbb{O}$, the unitary group, we have

$$\mu_{\mathbb{O}}(A) = \frac{\|A^* A - I\|_2}{\|A\|_2^2}, \tag{4.2}$$

which is a standard measure of departure from unitarity. Further justification for this measure is given by showing that if $\mu_{\mathbb{G}}(A)$ is small then A is close to a matrix in $\mathbb{G}$. For this, we use the *generalized* polar decomposition, which is closely related to the polar decompositions in indefinite scalar product spaces studied by Bolshakov et al. [2].

THEOREM 4.1 (generalized polar decomposition [22]). *Let $\mathbb{G}$ be an automorphism group corresponding to a bilinear or sesquilinear form for which $(A^\star)^\star = A$ for all $A \in \mathbb{K}^{n\times n}$. For any $A \in \mathbb{K}^{n\times n}$ such that $A^\star A$ has no eigenvalues on the nonpositive real axis, A has a unique decomposition $A = WS$, where $W \in \mathbb{G}$ (that is, $W^\star = W^{-1}$), $S^\star = S$, and $\operatorname{sign}(S) = I$.*

We note that the condition in this theorem on the adjoint being involutory holds precisely when $M^T = \pm M$ for bilinear forms and $M^* = \alpha M$ with $|\alpha| = 1$ for sesquilinear forms [22], and that these conditions hold for all the groups in Table 1.1.

LEMMA 4.2. *Let $A \in \mathbb{K}^{n\times n}$ have a generalized polar decomposition $A = WS$ with respect to an automorphism group $\mathbb{G} \in \mathfrak{U}$. If $\|W^{-1}(A - W)\|_2 < 1$, or equivalently $\|S - I\|_2 < 1$, then*

$$\frac{\|A^\star A - I\|_2}{\|A\|_2(\|A\|_2 + \|W\|_2)} \le \frac{\|A - W\|_2}{\|A\|_2} \le \frac{\|A^\star A - I\|_2}{\|A\|_2^2}\ \|A\|_2\|W\|_2.$$

The lower bound always holds.

Proof. 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 follows immediately. For the upper bound, we need to bound $\|(A + W)^{-\star}\|_2$. Note that $W^{-1}(A - W) = W^{-1}(WS - W) = S - I$ and $A + W = 2W(I + (S - I)/2)$. Hence, using the fact that $\mathbb{G} \in \mathfrak{U}$,

$$\begin{aligned}\|(A + W)^{-\star}\|_2 = \|(A + W)^{-1}\|_2 &= \left\| \frac{1}{2}\big(I + (S - I)/2\big)^{-1} W^{-1} \right\|_2 \\ &\le \frac{1}{2}\|W^{-1}\|_2 \frac{1}{1 - \frac{1}{2}\|S - I\|_2} \\ &\le \|W^{-1}\|_2,\end{aligned}$$

which yields the result. ∎

Lemma 4.2 shows that there is a matrix $W \in \mathbb{G}$ within relative distance $\mu_{\mathbb{G}}(A)$ of A, modulo a factor $\|A\|_2\|W\|_2$, as we wanted to show.

We now present a numerical experiment in which we compute the orthogonal polar factor of a random symplectic matrix $A \in \mathbb{R}^{12\times 12}$ with $\|A\|_2 = 3.1 \times 10^2 = \|A^{-1}\|_2$. All our experiments were performed in MATLAB, for which $u \approx 1.1\times 10^{-16}$. Table 4.1 reports the behavior of the Newton iteration, both without scaling and with Frobenius norm scaling, the cubic iteration (3.2), and the quintic (3.5). We report iterations up to the last one for which there was a significant decrease in $\|X_k^* X_k - I\|_2$. First, note that the convergence is entirely consistent with our description earlier, with the quintic and scaled Newton iterations spending the least time in the first phase. Next, we see from the first line of the table that the matrix A is indeed symplectic to machine precision, but far from orthogonal. The Newton iterations destroy the symplectic structure on the first iteration, but gradually restore it, as they must,

TABLE 4.1
Results for a symplectic matrix $A \in \mathbb{R}^{12\times 12}$ with $\kappa_2(A) = 9.6 \times 10^4$. Here, $\mu_{\mathbb{G}}$ and $\mu_{\mathbb{O}}$ are defined in (4.1) *and* (4.2)*, and $E = \min_k \|U - X_k\|_2$.*

k	Newton		Newton (scaled)		Cubic, (3.2)		Quintic, (3.5)	
	$\mu_{\mathbb{O}}(X_k)$	$\mu_{\mathbb{G}}(X_k)$	$\mu_{\mathbb{O}}(X_k)$	$\mu_{\mathbb{G}}(X_k)$	$\mu_{\mathbb{O}}(X_k)$	$\mu_{\mathbb{G}}(X_k)$	$\mu_{\mathbb{O}}(X_k)$	$\mu_{\mathbb{G}}(X_k)$
0	1.0e+0	7.0e-18	1.0e+0	7.0e-18	1.0e+0	7.0e-18	1.0e+0	7.0e-18
1	1.0e+0	1.0e+0	1.0e+0	1.0e+0	1.0e+0	8.9e-17	1.0e+0	1.1e-15
2	1.0e+0	1.0e+0	8.6e-01	8.6e-01	1.0e+0	8.1e-16	9.9e-01	1.7e-14
3	1.0e+0	1.0e+0	2.0e-01	2.0e-01	9.9e-01	6.3e-15	8.5e-01	3.0e-13
4	1.0e+0	1.0e+0	3.2e-03	3.2e-03	9.4e-01	5.0e-14	7.0e-02	1.7e-12
5	9.9e-01	9.9e-01	9.0e-07	9.0e-07	5.7e-01	2.8e-13	7.6e-09	1.8e-12
6	9.6e-01	9.6e-01	6.0e-14	1.3e-13	3.6e-02	5.2e-13	4.8e-16	1.8e-12
7	8.5e-01	8.5e-01	4.3e-16	1.1e-13	3.2e-06	5.3e-13		
8	5.4e-01	5.4e-01			3.8e-16	5.3e-13		
9	1.4e-01	1.4e-01						
10	5.5e-03	5.5e-03						
11	7.7e-06	7.7e-06						
12	1.5e-11	1.5e-11						
13	4.4e-16	1.1e-13						
E	4.4e-13		4.4e-13		7.3e-13		1.9e-12	

since the limit U is symplectic. However, we see that for these two iterations the relation $\mu_{\mathbb{O}}(X_k) = \mu_{\mathbb{G}}(X_k)$, that is,

$$\|X_k^* X_k - I\|_2 = \|X_k^\star X_k - I\|_2,$$

holds from iteration 1 until close to convergence, at which point rounding errors vitiate the relation—thus the approach to symplecticity is precisely tethered to the approach to orthogonality in this example. In fact, this is always true, as is an even stronger condition: the Newton iterates satisfy $X_k^\star = X_k^*$ for $k \geq 1$, for both the unscaled and Frobenius norm scaled iterations! Hence although the Newton iteration destroys the group structure, from this structure it creates and preserves a *different* kind of structure.

THEOREM 4.3. *Let $\mathbb{G} \in \mathfrak{U}$, $A \in \mathbb{G}$, and X_k be defined by the Newton iteration* (3.6) *or by a scaled Newton iteration* (3.7) *for which* (3.9) *holds. Then, for $k \geq 1$, $X_k^\star = X_k^*$.*

Proof. We will use two properties that we recall from the proof of Theorem 2.1 and that hold for all $B \in \mathbb{K}^{n\times n}$: $(B^{-1})^\star = (B^\star)^{-1}$, and $\mathbb{G} \in \mathfrak{U}$ implies $(B^*)^\star = (B^\star)^*$. For the scaled iteration, (3.9) implies $\gamma^{(0)} = 1$, and hence for both the scaled and unscaled iterations

$$\begin{aligned} X_1^\star &= \frac{1}{2}(A + A^{-*})^\star = \frac{1}{2}\big(A^\star + (A^{-*})^\star\big) \\ &= \frac{1}{2}\big(A^{-1} + (A^{-\star})^*\big) = \frac{1}{2}(A^{-1} + A^*) = X_1^*. \end{aligned}$$

Now assume that $X_{k-1}^\star = X_{k-1}^*$. Then, writing $\gamma = \gamma^{(k-1)}$,

$$\begin{aligned} X_k^\star &= \frac{1}{2}\big(\gamma X_{k-1} + \gamma^{-1} X_{k-1}^{-*}\big)^\star \\ &= \frac{1}{2}\big((\gamma X_{k-1})^\star + (\gamma^{-1} X_{k-1}^{-*})^\star\big) \\ &= \frac{1}{2}\big(\gamma X_{k-1}^* + \gamma^{-1} X_{k-1}^{-1}\big) = X_k^*. \end{aligned}$$

The result follows by induction. ∎

COROLLARY 4.4. *Under the conditions of Theorem* 4.3*, for* $k \geq 1$*,*

(a) $MX_k = X_k M$ *and* $MX_k^* = X_k^* M$ *for real bilinear and complex sesquilinear forms,*

(b) $MX_k = \overline{X_k} M$ *and* $MX_k^* = X_k^T M$ *for complex bilinear forms.*

Proof. Theorem 4.3 gives $X_k^\star = X_k^*$ for $k \geq 1$.

(a) We therefore have $M^{-1} X_k^* M = X_k^*$, or $X_k^* M = M X_k^*$. Taking the conjugate transpose and using $M^* = M^{-1}$ gives $MX_k = X_k M$.

(b) Similarly, $M^{-1} X_k^T M = X_k^*$, or $X_k^T M = M X_k^*$. Taking the conjugate transpose and using $M^* = M^{-1}$ gives $MX_k = \overline{X_k} M$. ∎

While Theorem 4.3 establishes the tethering between our measures of departure from $\mathbb{G}$-structure and unitarity, Corollary 4.4 has some further implications. In the case where A is pseudo-orthogonal, $M = \Sigma_{p,q}$, and to commute with $\Sigma_{p,q}$ is to be block diagonal! So all iterates X_k, $k \geq 1$, and the unitary polar factor itself, are block diagonal. For symplectic A, all the Newton iterates have the block structure $\left[\begin{smallmatrix} E & F \\ -F & E \end{smallmatrix}\right]$, and for perplectic A all the Newton iterates are centrosymmetric, that is, $a_{i,j} = a_{n-i+1,n-j+1}$ for $1 \leq i, j \leq n$. Computational savings can readily be made in all these cases. For example, in the symplectic case we need compute only the first n rows of the iterates, since the last n rows can be obtained from them.

5. Numerical stability. All the iterations under consideration involve matrix inversion, either explicitly or via the solution of linear systems with multiple right-hand sides, and when the corresponding matrices are ill conditioned numerical instability is a concern. Many years of experience have shown that the Newton iteration (3.7) is less prone to instability than might be expected. Indeed, it performs better than the best available bounds suggest; for a recent rounding error analysis of the iteration see [18]. Table 4.1 provides some insight and is representative of the typical behavior of the four iterations it illustrates: the computed iterates converge to a matrix that is orthogonal to working precision, and the error $\|U - X_k\|_2$ is of order at most $\kappa_2(A)u$, as is the measure $\mu_{\mathbb{G}}(X_k)$ in (4.1) of departure from $\mathbb{G}$-structure.

Of the other iterations we have found empirically that (3.1) and (3.4) are numerically stable, but (3.3) is not; the latter iteration produces an error $\|U - X_k\|_2$ and loss of structure $\mu_{\mathbb{G}}(X_k)$ observed to be of order $\kappa_2(A)^2 u$ and fails to converge when $\kappa_2(A) \gtrsim u^{-1/2}$.

We have found the partial fraction form of the quintic, mentioned in section 3, to have the same numerical stability as (3.5).

6. Convergence tests. An important question is how to terminate these matrix iterations. Since the Padé-based iterations compute $X_k^* X_k$, a convergence test of the form $\|X_k^* X_k - I\| \leq \mathrm{tol}$ can be used at no extra cost. For small tol, this test directly controls the error, since from Lemma 4.2 with $\mathbb{G} = \mathbb{O}$, using in the upper bound a refinement specific to this case from [10, Prob. 19.14],

$$\frac{\|X_k^* X_k - I\|_2}{\sigma_{\max}(X_k) + 1} \leq \|U - X_k\|_2 \leq \frac{\|X_k^* X_k - I\|_2}{\sigma_{\min}(X_k) + 1}.$$

The Padé-based iterations also have special properties that can be exploited. For the iteration function $f_{\ell m}$ in (2.4), it can be shown that f_{mm} has the properties

$$f_{mm}(\sigma^{-1}) = f_{mm}(\sigma)^{-1}, \tag{6.1a}$$

$$1 < \sigma \;\Rightarrow\; 1 < f_{mm}(\sigma) < \sigma, \tag{6.1b}$$

$$1 \le \mu < \sigma \;\Rightarrow\; f_{mm}(\mu) < f_{mm}(\sigma), \tag{6.1c}$$

$$f_{mm}(1) = 1. \tag{6.1d}$$

Let $A \in \mathbb{G}$ and $\mathbb{G} \in \mathfrak{U}$. Then, by Theorem 1.1, A has singular values that we may index

$$\sigma_1^{-1} \le \cdots \le \sigma_q^{-1} < \sigma_{q+1} = \cdots = \sigma_n = 1 < \sigma_q \le \cdots \le \sigma_1.$$

Using (6.1), we find that Z_k from (2.6) has singular values

$$\begin{aligned} f_{mm}^{(k)}(\sigma_1)^{-1} \le \cdots \le f_{mm}^{(k)}(\sigma_q)^{-1} < \sigma_{q+1} = \cdots \\ = \sigma_n = 1 < f_{mm}^{(k)}(\sigma_q) \le \cdots \le f_{mm}^{(k)}(\sigma_1). \end{aligned} \tag{6.2}$$

Applying this argument repeatedly, we deduce that Z_k from (2.6) satisfies

$$\|U - Z_k\|_2 = f_{mm}^{(k)}(\sigma_1) - 1. \tag{6.3}$$

The practical significance of this equality is that we can precisely predict the convergence of the matrix iteration simply by performing the iteration on σ_1, which is a scalar computation. If σ_1 is not known, or is too expensive to compute or estimate, then we can instead use

$$\|U - Z_k\|_2 \le f_{mm}^{(k)}(\|A\|_F) - 1.$$

The scalar computations can be done in advance of the matrix iteration, if required.

Another useful property of the iterates Z_k when $A \in \mathbb{G}$ and $\mathbb{G} \in \mathfrak{U}$ can be derived from (6.1) and (6.2): the sequence $\|Z_k\|_F$ decreases monotonically to $\sqrt{n}$. This means that the iteration can be terminated when the computed iterates $\widehat{Z}_k$ satisfy

$$\frac{\|\widehat{Z}_{k+1}\|_F}{\|\widehat{Z}_k\|_F} \ge 1 - \delta, \tag{6.4}$$

for some tolerance δ depending on u, that is, when rounding errors start to dominate.

Similar techniques apply to the Newton iteration. Convergence prediction can be done for the unscaled Newton iteration (3.6) for any A, as observed by Kenney and Laub [16], though with a simplification when $A \in \mathbb{G}$ and $\mathbb{G} \in \mathfrak{U}$. The iteration $h(x) = (x + 1/x)/2$ shares the properties (6.1b)–(6.1d) of f_{mm} and satisfies $h(\sigma^{-1}) = h(\sigma)$ in place of (6.1a). Therefore[2]

$$\|U - X_k\|_2 = h^{(k)}(\sigma_1) - 1; \tag{6.5}$$

again, we can use $\|A\|_F$ in place of σ_1 and obtain an upper bound. Convergence prediction based on a scalar iteration is not possible for the Newton iteration with Frobenius norm scaling.

For any A, the Newton sequence norms $\|X_k\|_F$ decrease monotonically for $k \ge 1$, both for the unscaled iteration and for the iteration with Frobenius norm scaling. This follows from the properties of h in the unscaled case and for Frobenius norm scaling can readily be proved from its definition, as shown by Dubrulle [5]. Therefore the stopping criterion (6.4) is applicable, and indeed it is advocated by Dubrulle [5] for the Frobenius norm scaling.

[2]For general A, (6.5) holds for $k \ge 1$ with $h^{(k)}$ replaced by $h^{(k-1)}$ and with σ_1 now the largest singular value of X_1.

7. Matrix sign decomposition. Much, but not all, of the analysis of the previous four sections applies with minor modification to the matrix sign decomposition.

The rewritten forms of the cubic and quintic iterations remain valid, with $X_k^* X_k$ replaced by X_k^2. Their costs are slightly higher than in the polar decomposition case, since X_k^2 is not Hermitian. The scaled Newton iteration for the matrix sign function is

$$X_{k+1} = \frac{1}{2}\left[\gamma^{(k)} X_k + \frac{1}{\gamma^{(k)}} X_k^{-1}\right], \qquad X_0 = A. \tag{7.1}$$

Among many proposed scalings is the determinantal scaling, $\gamma_k = |\det(X_k)^{-1/n}|$. This scaling satisfies (3.9), which continues to be an important property.

Theorem 4.3, which shows that the (suitably scaled) Newton iterates satisfy $X_k^\star = X_k^*$, has the following analogue.

THEOREM 7.1. *Let $A \in \mathbb{G}$, where $\mathbb{G}$ is any automorphism group. Let X_k be defined by the Newton iteration* (7.1)*, either unscaled ($\gamma^{(k)} = 1$) or with a scaling for which $\gamma^{(0)} = 1$. Then $X_k^\star = X_k$ for $k \geq 1$.*

COROLLARY 7.2. *Under the conditions of Theorem* 7.1*, for $k \geq 1$,*

(a) *$MX_k = X_k^T M$ for bilinear forms,*
(b) *$MX_k = X_k^* M$ for sesquilinear forms.*

Theorem 7.1 implies that the Newton iterates for the matrix sign function satisfy the condition $X_k^2 - I = X_k^\star X_k - I$, and so the approach to the group structure is tethered to the approach to involutory structure.

The convergence tests discussed in section 6 are not applicable to the sign iteration. In particular, since A is generally nonnormal the errors are not determined solely by the eigenvalues of the iterates.

8. Numerical experiments. Returning to the polar decomposition, we now compare experimentally the quintic iteration (3.5) with the Newton iteration (3.7) with Frobenius norm scaling. We generated random complex orthogonal $A_1, A_{16} \in \mathbb{R}^{16\times 16}$, where A_k denotes a product of k random complex orthogonal $\mathbb{G}$-reflectors [21]. Results are shown in Tables 8.1 and 8.2. In these tables the term $\mathrm{err}_k = \|U - X_k\|_2$ is computed from (6.3) using a scalar recurrence. Also shown is the value $1 - \|X_{k+1}\|_F / \|X_k\|_F$ arising in the convergence test (6.4).

The results, and others from similar experiments, reveal a number of interesting features.

1. The monotonicity test (6.4), and, for the quintic iteration, convergence prediction based on (6.3), both provide reliable termination criteria. For the former, $\delta \approx \sqrt{u}$ seems an appropriate choice, and for the latter, $\mathrm{err}_k \approx u$.
2. The Newton iterations (scaled and unscaled) can produce a computed unitary polar factor with smaller errors and better structure preservation than the quintic iteration (by a factor of up to 10^4 in Table 8.2), though all these quantities are empirically bounded by about $\kappa_2(A)u$.
3. The quintic iteration's faster initial linear convergence and faster asymptotic convergence enable it to require fewer iterations than scaled Newton when $\|A - U\|_2 \lesssim 1$, but nevertheless the scaled Newton iteration usually requires the fewest flops.

TABLE 8.1

Results for a complex orthogonal matrix $A_1 \in \mathbb{R}^{16\times 16}$ with $\kappa_2(A) = 6.6$. Here, $\mu_{\mathbb{G}}$ and $\mu_{\mathbb{O}}$ are defined in (4.1) *and* (4.2), $\mathrm{err}_k = \|U - X_k\|_2$ *is computed from a scalar recurrence, and* $E = \min_k \|U - X_k\|_2$.

k	Newton (scaled)			Quintic, (3.5)			
	$\mu_{\mathbb{O}}(X_k)$	$\mu_{\mathbb{G}}(X_k)$	$1 - \frac{\|X_{k+1}\|_F}{\|X_k\|_F}$	$\mu_{\mathbb{O}}(X_k)$	$\mu_{\mathbb{G}}(X_k)$	err_k	$1 - \frac{\|X_{k+1}\|_F}{\|X_k\|_F}$
0	9.8e-01	6.7e-17		9.8e-01	6.7e-17	7.1e+0	
1	9.4e-01	9.4e-01	2.3e-01	7.0e-01	3.3e-15	8.2e-01	5.3e-01
2	6.5e-01	6.5e-01	3.4e-01	8.3e-03	7.5e-15	4.2e-03	4.7e-02
3	1.5e-01	1.5e-01	1.2e-01	1.5e-13	7.6e-15	7.7e-14	2.2e-06
4	4.8e-03	4.8e-03	1.2e-02	5.6e-16	7.5e-15	0.0e+0	0.0e+0
5	4.3e-06	4.3e-06	3.5e-04				
6	3.4e-12	3.4e-12	3.1e-07				
7	4.7e-16	1.1e-15	2.4e-13				
8	5.5e-16	1.2e-15	0.0e+0				
E	1.4e-15			5.6e-15			

TABLE 8.2

Results for a complex orthogonal matrix $A_{16} \in \mathbb{R}^{16\times 16}$ with $\kappa_2(A) = 6.5 \times 10^9$. Here, $\mu_{\mathbb{G}}$ and $\mu_{\mathbb{O}}$ are defined in (4.1) *and* (4.2), $\mathrm{err}_k = \|U - X_k\|_2$ *is computed from a scalar recurrence, and* $E = \min_k \|U - X_k\|_2$.

k	Newton (scaled)			Quintic, (3.5)			
	$\mu_{\mathbb{O}}(X_k)$	$\mu_{\mathbb{G}}(X_k)$	$1 - \frac{\|X_{k+1}\|_F}{\|X_k\|_F}$	$\mu_{\mathbb{O}}(X_k)$	$\mu_{\mathbb{G}}(X_k)$	err_k	$1 - \frac{\|X_{k+1}\|_F}{\|X_k\|_F}$
0	1.0e+0	2.9e-16		1.0e+0	2.9e-16	8.1e+04	
1	1.0e+0	1.0e+0	2.9e-1	1.0e+0	7.9e-15	1.6e+04	8.0e-1
2	1.0e+0	1.0e+0	1.0e+0	1.0e+0	1.9e-13	3.2e+03	8.0e-1
3	9.7e-01	9.7e-1	9.5e-1	1.0e+0	4.8e-12	6.4e+02	8.0e-1
4	5.3e-01	5.3e-1	6.2e-1	1.0e+0	1.2e-10	1.3e+02	8.0e-1
5	5.9e-02	5.9e-2	1.3e-1	1.0e+0	3.0e-09	2.5e+01	8.0e-1
6	4.8e-04	4.8e-4	8.7e-3	9.6e-01	7.3e-08	4.2e+0	7.5e-1
7	3.8e-08	3.8e-8	4.3e-5	4.4e-01	1.1e-06	3.4e-01	3.7e-1
8	6.1e-16	1.1e-8	2.9e-9	2.5e-04	1.9e-06	1.2e-04	1.1e-2
9	6.3e-16	1.1e-8	0.0e+0	1.5e-15	1.9e-06	2.2e-16	2.0e-9
10				7.3e-16	1.9e-06	0.0e+0	0.0e+0
E	3.6e-10			1.8e-6			

9. Conclusions. When a problem has structure it is important to exploit it to advantage. This work was motivated by the discovery of Mackey, Mackey, and Tisseur [22] that the polar and matrix sign factors of matrices from automorphism groups $\mathbb{G}$ also lie in the group: unconditionally for the sign decomposition, and provided the matrix of the underlying form is unitary for the polar decomposition. We have identified a family of globally convergent rational iterations that preserve group structure and shown how structure preservation leads to particularly convenient convergence tests in the case of the polar decomposition.

The most surprising results in this work concern Newton's method. Although Newton's method for the polar decomposition immediately destroys the underlying group structure, when $\mathbb{G} \in \mathfrak{U}$ it forces equality between the adjoint and the conjugate transpose of each iterate. This implies that the Newton iterates approach the group at the same rate that they approach unitarity. It also yields "commutativity" relations that for certain groups imply a different, exploitable structure. Similar properties hold for Newton's method for the matrix sign function, here with no restrictions on $\mathbb{G}$.

We have identified various pros and cons in the "structured iteration versus scaled Newton" comparison, including the slightly better empirically observed numerical stability of Newton, the convergence prediction possible with the structured iterations,

and the fact that, in practice, scaled Newton usually requires the fewest flops.

Our conclusion is that the Newton iteration (3.7) with Frobenius norm scaling (3.8) and the cubic (3.2) and quintic (3.5) structure-preserving iterations are all well-suited to computing the polar decomposition of a matrix from one of the automorphism groups under consideration. Likewise, for the matrix sign decomposition the scaled Newton iteration (7.1) and the obvious analogues of the cubic and quintic iterations are all suitable. Which of the iterations is to be preferred depends on the matrix A, the group $\mathbb{G}$, and the user's accuracy requirements.

REFERENCES

[1] Z. Bai and J. W. Demmel, *Using the matrix sign function to compute invariant subspaces*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 205–225.

[2] Y. Bolshakov, C. V. M. van der Mee, A. C. M. Ran, B. Reichstein, and L. Rodman, *Polar decompositions in finite dimensional indefinite scalar product spaces: General theory*, Linear Algebra Appl., 261 (1997), pp. 91–141.

[3] J. R. Cardoso, C. 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.

[4] J. R. Cardoso and F. Silva Leite, *Extending Results from Orthogonal Matrices to the Class of P-Orthogonal Matrices*. manuscript, 2002.

[5] A. A. Dubrulle, *An optimum iteration for the matrix polar decomposition*, Electron. Trans. Numer. Anal., 8 (1999), pp. 21–25.

[6] K. Fan and A. J. Hoffman, *Some metric inequalities in the space of matrices*, Proc. Amer. Math. Soc., 6 (1955), pp. 111–116.

[7] W. Gander, *On Halley's iteration method*, Amer. Math. Monthly, 92 (1985), pp. 131–134.

[8] N. J. Higham, *Computing the polar decomposition—with applications*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1160–1174.

[9] N. J. Higham, *The matrix sign decomposition and its relation to the polar decomposition*, Linear Algebra Appl., 212/213 (1994), pp. 3–20.

[10] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., SIAM, Philadelphia, 2002.

[11] N. J. Higham, *J-orthogonal matrices: Properties and generation*, SIAM Rev., 45 (2003), pp. 504–519.

[12] N. J. Higham, D. S. Mackey, N. Mackey, and F. Tisseur, *Functions Preserving Matrix Groups and Iterations for the Matrix Square Root*, Numerical Analysis Report 446, Manchester Centre for Computational Mathematics, Manchester, England, 2004.

[13] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[14] N. Jacobson, *Basic Algebra* I, W. H. Freeman and Company, San Francisco, 1974.

[15] C. S. Kenney and A. J. Laub, *Rational iterative methods for the matrix sign function*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 273–291.

[16] C. S. Kenney and A. J. Laub, *On scaling Newton's method for polar decomposition and the matrix sign function*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 688–706.

[17] C. S. Kenney and A. J. Laub, *The matrix sign function*, IEEE Trans. Automat. Control, 40 (1995), pp. 1330–1348.

[18] A. Kiełbasiński and K. Ziętak, *Numerical behaviour of Higham's scaled method for polar decomposition*, Numer. Algorithms, 32 (2003), pp. 105–140.

[19] S. Lang, *Algebra*, 3rd ed., Addison-Wesley, Reading, MA, 1993.

[20] D. S. Mackey and N. Mackey, *On the Construction of Structure-Preserving Matrix Iterations*, in preparation.

[21] D. S. Mackey, N. Mackey, and F. Tisseur, *$\mathbb{G}$-Reflectors: Analogues of Householder Transformations in Scalar Product Spaces*, Numerical Analysis Report 420, Manchester Centre for Computational Mathematics, Manchester, England, Feb. 2003. Linear Algebra Appl., to appear.

[22] D. S. Mackey, N. Mackey, and F. Tisseur, *Structured Factorizations in Scalar Product Spaces*, Numerical Analysis Report 432, Manchester Centre for Computational Mathematics, Manchester, England, 2004.

[23] 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.

[24] R. Shaw, *Linear Algebra and Group Representations. Vol.* I, Academic Press, London, 1982.