The Matrix Sign Decomposition and Its Relation to the Polar Decomposition

Nicholas J. Higham*

Department of Mathematics
University of Manchester
Manchester, M13 9PL, England

Submitted by Dianne P. O'Leary

ABSTRACT

The sign function of a square matrix was introduced by Roberts in 1971. We show that it is useful to regard $S = \operatorname{sign}(A)$ as being part of a matrix sign decomposition $A = SN$, where $N = (A^2)^{1/2}$. This decomposition leads to the new representation $\operatorname{sign}(A) = A(A^2)^{-1/2}$. Most results for the matrix sign decomposition have a counterpart for the polar decomposition $A = UH$, and vice versa. To illustrate this, we derive best approximation properties of the factors U, H, and S, determine bounds for $\|A - S\|$ and $\|A - U\|$, and describe integral formulas for S and U. We also derive explicit expressions for the condition numbers of the factors S and N. An important equation expresses the sign of a block 2×2 matrix involving A in terms of the polar factor U of A. We apply this equation to a family of iterations for computing S by Pandey, Kenney, and Laub, to obtain a new family of iterations for computing U. The iterations have some attractive properties, including suitability for parallel computation.

1. INTRODUCTION

The matrix sign function was introduced by Roberts [41] in 1971 as a tool for solving the Lyapunov equation and the algebraic Riccati equation. Roberts defined it via a contour integral, but most authors favor a definition based on the Jordan canonical form $A = ZJZ^{-1}$ of $A \in \mathbb{C}^{n\times n}$. If we

*This work was supported by Science and Engineering Research Council grant GR/H52139, and, in part, by a Nuffield Science Research Fellowship. E-mail: na.nhigham@na-net.ornl.gov.

655 Avenue of the Americas, New York, NY 10010 0024-3795/94/$7.00

arrange that

$$J = \begin{bmatrix} J_1 & 0 \\ 0 & J_2 \end{bmatrix},$$

where the eigenvalues of J_1 lie in the open left half plane and those of J_2 lie in the open right half plane, then

$$\operatorname{sign}(A) = Z \begin{bmatrix} -I & 0 \\ 0 & I \end{bmatrix} Z^{-1}. \tag{1.1}$$

The sign function is undefined if A has an eigenvalue on the imaginary axis. While this definition is convenient to work with, it does not provide a reliable way to compute sign(A); nor does Roberts's integral definition. For computation, Roberts proposed the Newton iteration

$$X_{k+1} = \tfrac{1}{2}(X_k + X_k^{-1}), \qquad X_0 = A, \tag{1.2}$$

which he showed converges quadratically to sign(A) for any $A \in \mathbf{C}^{n\times n}$ having no pure imaginary eigenvalues. This iteration is Newton's method applied to the equation $X^2 = I$.

The utility of the sign function is easily seen from Roberts's observation that the Sylvester equation

$$AX + XB = C, \qquad A \in \mathbf{C}^{m\times m}, \quad B \in \mathbf{C}^{n\times n}, \quad C \in \mathbf{C}^{m\times n},$$

is equivalent to the equation

$$\begin{bmatrix} A & -C \\ 0 & -B \end{bmatrix} = \begin{bmatrix} I & X \\ 0 & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & -B \end{bmatrix} \begin{bmatrix} I & X \\ 0 & I \end{bmatrix}^{-1}.$$

If sign(A) $= I$ and sign(B) $= I$ then

$$\operatorname{sign}\begin{bmatrix} A & -C \\ 0 & -B \end{bmatrix} = \begin{bmatrix} I & X \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} \begin{bmatrix} I & X \\ 0 & I \end{bmatrix}^{-1} = \begin{bmatrix} I & -2X \\ 0 & -I \end{bmatrix}, \tag{1.3}$$

so the solution X can be read from the sign of the block upper triangular matrix

$$\begin{bmatrix} I & -C \\ 0 & B \end{bmatrix}.$$

The conditions that sign(A) and sign(B) are identity matrices are certainly satisfied for the Lyapunov equation ($B = A^*$) in the common case where A is positive stable, that is, $\operatorname{Re}\lambda_i(A) > 0$ for all i.

The matrix sign function was the subject of a steady stream of papers throughout the 1970s and 1980s. It has been widely used to solve the algebraic Riccati equation; see, for example, [8], and see [36] for a survey. It has also been applied to eigensystem computations (see, for example, [5, 11, 28]), though some of the proposed algorithms are of dubious computational merit. Recently there has been a resurgence of interest in the matrix sign function because of its suitability for constructing parallel algorithms [9, 14, 15], particularly in the context of the nonsymmetric eigenproblem [3, 37].

The polar decomposition is much older than the matrix sign function. It was introduced by Autonne in 1902 [2]. It is the decomposition $A = UH$ of $A \in \mathbf{C}^{m\times n} (m \geq n)$, where $U \in \mathbf{C}^{m\times n}$ has orthonormal columns and $H \in \mathbf{C}^{n\times n}$ is Hermitian and positive semidefinite. If A has full rank, then H is nonsingular and U is unique. For a thorough discussion of the history of the polar decomposition see [27, Section 3.0].

The purpose of this work is to show that there are several relationships and analogies between the matrix sign function and the polar decomposition and that by exploiting them we can derive useful insights and new results. In Section 2 we introduce the matrix sign decomposition, $A = SN$, and use it to derive the new formula $S = \operatorname{sign}(A) = A(A^2)^{-1/2}$. We summarize some best approximation properties of the polar and sign decompositions in Section 3. We investigate the conditioning of the matrix sign decomposition in Section 4, deriving explicit expressions for the condition numbers of S and N. We obtain estimates for the distance from a matrix to its sign function and its polar factor U in Section 5. Finally, in Section 6 we derive a new family of iterations for computing the polar decomposition by adapting a family obtained by Pandey, Kenney, and Laub for the sign function. Since the iterations are in partial fraction form, they are very amenable to parallel computation.

2. THE MATRIX SIGN DECOMPOSITION

If a is a real scalar, then $a = \operatorname{sign}(a)|a|$, where $\operatorname{sign}(a) = \pm 1$ is the familiar sign of a scalar. The polar decomposition is one generalization of this scalar decomposition to complex matrices, with $\operatorname{sign}(a)$ becoming the factor U with orthonormal columns, and $|a|$ the Hermitian positive semidefinite matrix H. Another generalization, which has apparently not been explored before (except incidentally in [10], as described below) translates $\operatorname{sign}(a)$ to $\operatorname{sign}(A)$. We define the *matrix sign decomposition*

$$A = SN, \qquad S = \operatorname{sign}(A), \quad A \in \mathbf{C}^{n\times n}.$$

Here, and throughout this section, we assume that A has no pure imaginary eigenvalues. This decomposition is uniquely defined because $S = \operatorname{sign}(A)$ is uniquely defined and nonsingular (its eigenvalues are ± 1), so that $N = S^{-1}A$. In fact, since S is involutary ($S^2 = I$), $N = SA$.

In the case of a Jordan block the decomposition can be written as

$$J(\lambda) = \operatorname{sign}(\lambda) I \cdot \operatorname{sign}(\lambda) J(\lambda) \equiv SN.$$

More generally, if $A = Z \operatorname{diag}(J(\lambda_k)) Z^{-1}$ is a Jordan decomposition, then

$$S = Z \operatorname{diag}(\operatorname{sign}(\lambda_k)) Z^{-1}, \qquad N = Z \operatorname{diag}(\operatorname{sign}(\lambda_k)) \operatorname{diag}(J(\lambda_k)) Z^{-1}.$$

It is clear from these expressions that S, N, and A always commute with each other.

Since the matrix sign decomposition and the polar decomposition are generalizations of the same scalar decomposition, we might expect there to be analogies between them. Immediately apparent are spectral analogies:

$$\begin{array}{llll} A = SN: & S^2 = I, & \lambda_i(S) = \pm 1, & \operatorname{Re} \lambda_i(N) > 0, \\ A = UH: & U^*U = I, & |\lambda_i(U)| = 1, & \lambda_i(H) > 0. \end{array}$$

As might be guessed from these properties, the matrix sign decomposition and the polar decomposition are the same decomposition when A is Hermitian. Unlike the matrix sign factors S and N, the polar factors of a square matrix do not always commute: $UH = HU$ if and only if A is normal (see, for example, [26, Theorem 7.3.4]).

It is well known (and easy to see) that $H = (A^*A)^{1/2}$, where the square root is the unique Hermitian positive semidefinite square root of a Hermitian positive semidefinite matrix; therefore $U = A(A^*A)^{-1/2}$. For the matrix sign decomposition, $A^2 = SNSN = S^2N^2 = N^2$, and since we are assuming A has no pure imaginary eigenvalues, A^2 is nonsingular and has no real, negative eigenvalues. Therefore, $N = (A^2)^{1/2}$, where for a nonsingular matrix B with no real, negative eigenvalues, $B^{1/2}$ denotes the unique square root all of whose eigenvalues lie in the open right half plane. The existence and uniqueness of this square root follow from Theorem 4 in [18] (or see Lemma 1 in [12]). This characterization of N provides a new and particularly concise definition of the sign function:

$$\operatorname{sign}(A) = A(A^2)^{-1/2}.$$

The properties $S^2 = I$ and $SA = AS$ are immediate from this definition, given the fact that $(A^2)^{-1/2}$ is a polynomial in A^2 [18, Theorem 4]. The theory of square roots guarantees that if B is real and has no real, negative eigenvalues, then $B^{1/2}$ is real [18; 27, Section 6.4]. Hence it is clear from this formula that $\operatorname{sign}(A)$ is real when A is real; this is not obvious from (1.1).

Another analogy between S and U concerns exponential representations. It is well known that any unitary matrix U can be written $U = \exp(iG)$, where $G = G^*$. Similarly, any involutary matrix S can be written $S = \exp(i\pi W)$, where W has eigenvalues 0 and 1. Using the Jordan canonical form, it is straightforward to show that $W = (I - S)/2$ and

$$S^p = \tfrac{1}{2}(I + S) + \frac{e^{ip\pi}}{2}(I - S), \qquad p \in \mathbb{R}.$$

Denman makes this observation in [10] and proposes a method for computing matrix pth roots based on the equations $A^p = (SN)^p = S^p N^p$.

The matrix sign decomposition and the polar decomposition can be found explicitly when $A \in \mathbb{R}^{2\times 2}$. Uhlig [42] shows that

$$U = \gamma(A + |\det A| A^{-T}), \quad H = \gamma\big(A^T A + |\det A| I\big) \qquad (A \in \mathbb{R}^{2\times 2}),$$

where

$$\gamma = \big|\det\big(A + |\det A| A^{-T}\big)\big|^{-1/2}.$$

Kenney and Laub [31, Lemma 3.4] give the following prescription for S: $S = \operatorname{sign}(\operatorname{trace} A) I$ if $\det A > 0$ and $\operatorname{trace} A \neq 0$; if $\det A < 0$ then

$$S = \mu\big[A - (\det A) A^{-1}\big], \quad N = \mu\big[A^2 - (\det A) I\big] \qquad (A \in \mathbb{R}^{2\times 2}), \quad (2.1)$$

where

$$\mu = \big(-\det\big[A - (\det A) A^{-1}\big]\big)^{-1/2};$$

and otherwise S is undefined. These explicit formulas for $n = 2$ again reveal similarities between the matrix sign and polar decompositions. It is interesting to note that S in (2.1) and X_1 from (1.2) are both linear combinations of A and A^{-1}.

3. APPROXIMATION PROPERTIES

The widespread use of the polar decomposition stems from the best approximation properties of its factors: If $A \in \mathbb{C}^{n\times n}$, then for any unitarily invariant norm,

$$\min\{\|A - Q\| : Q^* Q = I\} = \|A - U\|$$

and

$$\begin{aligned}&\min\{\|A - X\|_F : X \text{ Hermitian positive semidefinite}\}\\ &\qquad = \left\|A - \tfrac{1}{2}(B + H_B)\right\|_F,\end{aligned}$$

where $B = (A + A^*)/2$ and $B = U_B H_B$ is a polar decomposition. The first property is proved by Fan and Hoffman [13] and the second by Higham [19]. The first property also holds for rectangular A in the 2-norm and the Frobenius norm, and the second property holds for the 2-norm when A is Hermitian [17]. The matrix sign decomposition is less well endowed with approximation properties, largely because the sign function ignores valuable information present in the imaginary parts of the eigenvalues of A. Nevertheless, it is straightforward to derive the result, for $A \in \mathbf{C}^{n\times n}$,

$$\min\{\|A - X\|_F : X^2 = I,\ X = X^*\} = \|A - \operatorname{sign}(B)\|_F, \tag{3.1}$$

where B is defined as above, and where, for the purposes of (3.1) only, we extend the definition of sign to singular Hermitian matrices by defining $\operatorname{sign}(0) = 1$ or -1. The general problem

$$\min\{\|A - X\|_F : X^2 = I\} \tag{3.2}$$

appears to be an analytically intractable optimization problem. Three plausible conjectures can be ruled out by counter example. It is possible to find examples with $n = 2$ where $\operatorname{sign}(A)$ does not solve (3.2), where the solution is not triangular when A is triangular, and where the solution is not $X = I$ when A is positive stable.

4. PERTURBATION THEORY AND CONDITIONING

Perturbation theory for the polar decomposition is well understood. Let $A \in \mathbf{C}^{n\times n}$ be nonsingular, and let $\epsilon = \|\Delta A\|_F/\|A\|_F$. Higham [17, Theorem 2.5] shows that $A + \Delta A$ has the polar decomposition

$$A + \Delta A = (U + \Delta U)(H + \Delta H),$$

where

$$\frac{\|\Delta H\|_F}{\|H\|_F} \le \sqrt{2}\epsilon + O(\epsilon^2).$$

In fact , the $O(\epsilon^2)$ term can be dropped from this bound [1, 6, 39]. Kenney and Laub [31, Theorem 2.2] and Barrlund [4, Theorem 2.6] show that if

$\sigma_1(\Delta A) < \sigma_n(A)$, where σ_1 and σ_n denote the largest and smallest singular values, respectively, then

$$\frac{\|\Delta U\|_F}{\|U\|_F} \le \theta\epsilon\frac{\|A\|_F}{\|U\|_F} + O(\epsilon^2),$$

where $\theta = \sigma_n(A)^{-1}$. Moreover, if A is real and ΔA is restricted to be real, then we can take $\theta = 2/[\sigma_n(A) + \sigma_{n-1}(A)]$ [31, Theorem 2.3; 4, Theorem 2.4], which can be much smaller than the value $\sigma_n(A)^{-1}$ obtained for complex perturbations. These bounds are sharp, so $\kappa_H = \sqrt{2}$ and $\kappa_U = \theta\|A\|_F/\|U\|_F$ are condition numbers for the polar factors H and U, respectively. Mathias [39] shows that, more generally, for any unitarily invariant norm, $\kappa_U = \theta\|A\|/\|U\|$, with θ defined as for the Frobenius norm.

We now make use of the perturbation analysis technique from [23] to derive a sharp perturbation result for the matrix sign decomposition. We use the Kronecker product $\otimes$ [27].

THEOREM 4.1. *Suppose that $A \in \mathbf{C}^{n\times n}$ and $A + \Delta A$ both have no pure imaginary eigenvalues, and let their matrix sign decompositions be $A = SN$ and $A + \Delta A = (S + \Delta S)(N + \Delta N)$. Define $\epsilon = \|\Delta A\|_F/\|A\|_F$ and $P = I \otimes N + N^T \otimes I$. Then*

$$\frac{\|\Delta S\|_F}{\|S\|_F} \le \big\|P^{-1}(I_{n^2} - S^T \otimes S)\big\|_2 \frac{\|A\|_F}{\|S\|_F}\epsilon + O(\epsilon^2), \qquad (4.1)$$

$$\frac{\|\Delta N\|_F}{\|N\|_F} \le \big\|P^{-1}(I \otimes A + A^T \otimes I)\big\|_2 \frac{\|A\|_F}{\|N\|_F}\epsilon + O(\epsilon^2), \qquad (4.2)$$

and both bounds are attainable, to first order.

Proof. Expanding the equation $(N + \Delta N)^2 = (A + \Delta A)^2$, we have

$$N\Delta N + \Delta N N = A\,\Delta A + \Delta A\,A + O(\epsilon^2). \qquad (4.3)$$

By applying the vec operator, which stacks the columns of a matrix into a vector, and using the relation $\mathrm{vec}(AXB) = (B^T \otimes A)\mathrm{vec}(X)$ [27, Lemma 4.3.1], we obtain

$$P\,\mathrm{vec}(\Delta N) = (I \otimes A + A^T \otimes I)\,\mathrm{vec}(\Delta A) + O(\epsilon^2).$$

Hence we have the bound

$$\|\mathrm{vec}(\Delta N)\|_2 \le \big\|P^{-1}(I \otimes A + A^T \otimes I)\big\|_2 \|\mathrm{vec}(\Delta A)\|_2 + O(\epsilon^2),$$

which is sharp, to first order. The required bound (4.2) follows on noting that $\|\mathrm{vec}(\Delta A)\|_2 = \|\Delta A\|_F = \epsilon\|A\|_F$ and dividing by $\|N\|_F$.

The bound for $\|\Delta S\|_F$ can be deduced using the equations for ΔN, but it follows more simply from the equations

$$N\Delta S + \Delta S N = \Delta A - S\Delta A\, S + O(\epsilon^2), \tag{4.4}$$

which are obtained by writing $(A+\Delta A)(S+\Delta S) = (S+\Delta S)(A+\Delta A)$ as

$$A\,\Delta S - \Delta S A = S\Delta A - \Delta A\, S + O(\epsilon^2),$$

premultiplying by S, and using $(S+\Delta S)^2 = I$ to replace $S\Delta S$ by $-\Delta S S$ $+\ O(\epsilon^2)$. The rest of the analysis is similar to that for ΔN. ■

Since the bounds of the theorem are sharp, the condition numbers for S and N are, respectively,

$$\kappa_S = \left\|P^{-1}(I_{n^2} - S^T \otimes S)\right\|_2 \frac{\|A\|_F}{\|S\|_F}, \tag{4.5}$$

$$\kappa_N = \left\|P^{-1}(I \otimes A + A^T \otimes I)\right\|_2 \frac{\|A\|_F}{\|N\|_F}. \tag{4.6}$$

Note that if all the eigenvalues of A lie in the right half plane, then $S = I$ and $N = A$, so that $\kappa_S = 0$ and $\kappa_N = 1$. This is what we would expect, since S is unaffected by small perturbations to A in this case.

These explicit formulas for the condition numbers κ_S and κ_N are new. Kenney and Laub derive (4.4) in [31, Theorem 3.1]; they show how to estimate and bound κ_S, but do not give an explicit expression for it, except for normal A (see below).

An interesting aspect of the theorem is that N plays a key role in the conditioning of $S = \mathrm{sign}(A)$. To clarify the conditioning it is helpful to bound the condition numbers. We will assume that A is diagonalizable: $A = ZDZ^{-1}$, where $D = \mathrm{diag}(\lambda_i)$. Then $S = Z\mathrm{diag}(s_i)Z^{-1}$ and $N = Z\mathrm{diag}(s_i\lambda_i)Z^{-1}$, where $s_i = \mathrm{sign}(\lambda_i)$, and (4.4) can be rewritten

$$\mathrm{diag}(s_i\lambda_i)\,\widetilde{\Delta S} + \widetilde{\Delta S}\,\mathrm{diag}(s_i\lambda_i) = \widetilde{\Delta A} - \mathrm{diag}(s_i)\,\widetilde{\Delta A}\,\mathrm{diag}(s_i) + O(\epsilon^2),$$

where $\widetilde{\Delta S} = Z^{-1}\Delta S Z$ and $\widetilde{\Delta A} = Z^{-1}\Delta A\, Z$. Thus

$$\widetilde{\Delta s}_{ij} = \frac{1 - s_i s_j}{s_i\lambda_i + s_j\lambda_j}\widetilde{\Delta a}_{ij} + O(\epsilon^2).$$

The first-order term is zero if $s_is_j = 1$, so

$$\|\widetilde{\Delta S}\|_F \le \max\left\{\left|\frac{1-s_is_j}{s_i\lambda_i + s_j\lambda_j}\right| : s_is_j = -1\right\}\|\widetilde{\Delta A}\|_F + O(\epsilon^2),$$

which implies

$$\|\Delta S\|_F \le 2\kappa_2(Z)^2 \max\left\{\frac{1}{|\lambda_i - \lambda_j|} : \operatorname{Re}\lambda_i \operatorname{Re}\lambda_j < 0\right\}\|\Delta A\|_F + O(\epsilon^2).$$

Here we have used the result that for any unitarily invariant norm [27, Corollary 3.5.10]

$$\|ABC\| \le \|A\|_2\|B\|\|C\|_2, \qquad A \in \mathbf{C}^{r\times m}, \quad B \in \mathbf{C}^{m\times n}, \quad C \in \mathbf{C}^{n\times s}. \tag{4.7}$$

(in fact, any two of the norms on the right-hand side can be 2-norms). Thus

$$\kappa_S \le 2\kappa_2(Z)^2 \max\left\{\frac{1}{|\lambda_i - \lambda_j|} : \operatorname{Re}\lambda_i \operatorname{Re}\lambda_j < 0\right\}\frac{\|A\|_F}{\|S\|_F}. \tag{4.8}$$

Similar analysis using (4.3) yields the bound

$$\begin{aligned}\kappa_N &\le \kappa_2(Z)^2 \max_{i,j}\left|\frac{\lambda_i+\lambda_j}{s_i\lambda_i + s_j\lambda_j}\right|\frac{\|A\|_F}{\|N\|_F} \\ &= \kappa_2(Z)^2 \max\left\{\left|\frac{\lambda_i+\lambda_j}{\lambda_i-\lambda_j}\right| : \operatorname{Re}\lambda_i \operatorname{Re}\lambda_j < 0\right\}\frac{\|A\|_F}{\|N\|_F},\end{aligned} \tag{4.9}$$

where we set the maximum of a set to 1 if the set is empty or contains only zero elements. If A is normal, then we can take $\kappa_2(Z) = 1$ in (4.8) and (4.9), and both bounds are equalities in this case. For normal A, (4.8) is derived as an equality by Kenney and Laub [31, Theorem 3.1] (note that $|\lambda_i + \lambda_j|$ in (3.12) of [31] should be $|\lambda_i - \lambda_j|$).

The gist of (4.8) and (4.9) is that the sensitivity of both S and N is bounded in terms of the minimum distance between eigenvalues across the imaginary axis. If the eigenvalues are all real, then $(\lambda_i + \lambda_j)/(\lambda_i - \lambda_j)$ does not exceed 1 in modulus when $\operatorname{Re}\lambda_i\operatorname{Re}\lambda_j < 0$, so in this case κ_N is bounded solely in terms of the condition of the eigenvectors.

For general A, we can estimate κ_S and κ_N in (4.5) and (4.6) without explicitly solving an $n^2 \times n^2$ system of equations, as might be thought necessary from the formulas. We can use the matrix-norm estimator of Hager and Higham [16, 20, 22], which estimates $\|B\|_1$ by computing a few matrix-vector products Bx and B^*y (typically four or five in total). For

κ_S, the matrix $B = P^{-1}(I_{n^2} - S^T \otimes S)$. Assuming that we know S and N, we can compute $z = Bx$ as follows:

(1) Form $W = X - SXS$, where $x = \mathrm{vec}(X)$.
(2) Solve $NZ + ZN = W$, and set $z = \mathrm{vec}(Z)$.

Thus we have to solve a Sylvester equation involving N; this can be done via a matrix sign evaluation using (1.3), since N is a positive stable matrix. We can compute B^*y in a similar fashion, solving a Sylvester equation of the same form. The condition number κ_N can be estimated in a similar way, by solving Sylvester equations with the same coefficient matrices. As an alternative to this 1-norm estimation procedure, we could apply the power method to B^*B. Kenney and Laub [31] show how to estimate κ_S by using its characterization as the Fréchet derivative of $\mathrm{sign}(A)$ and applying the power method (see also [30] for a detailed description of the role of the Fréchet derivative in condition estimation for matrix functions). They show how to compute the Fréchet derivative explicitly: by solving a Sylvester equation [in fact, our (4.4), minus the second-order term] or by an iteration, and they also show how to approximate it by finite differences. Mathias [38] shows that the Fréchet derivative can be computed efficiently via the Schur decomposition. All these condition estimators require $O(n^3)$ operations.

To emphasize the dependence of κ_S and κ_N on the eigenvector conditioning (or equivalently, on the nonnormality of A), we give in Table 1 condition numbers for the upper triangular 6×6 matrix $T_6(\alpha)$ with diagonal elements equally spaced between -1 and 1 and off-diagonal elements all equal to α. The last two columns of the table show that the upper bounds (4.8) and (4.9) are not too far from being equalities in this example.

5. ERROR BOUNDS

It is useful to have easily computable estimates of $\|A - U\|$ and $\|A - S\|$, where $A = UH$ and $A = SN$. By taking A in these estimates to be an iterate from an iteration for computing U or S we can decide when to terminate the iteration, because for the standard iterations the iterates have the same factor U or S as the starting matrix. First, we present two lemmas that bound $\|A - U\|$. Lemma 5.1 generalizes Lemma 4.2 in [21] from the 2-norm and the Frobenius norm to an arbitrarily unitarily invariant norm.

TABLE 1.

Condition numbers for a 6×6 matrix $T_6(\alpha)$

α	κ_S	κ_N	(4.8)	(4.9)
1.00e − 01	4.02e + 00	1.20e + 00	9.88e + 00	2.97e + 00
1.67e − 01	5.29e + 00	1.52e + 00	1.99e + 01	6.24e + 00
2.78e − 01	9.96e + 00	2.37e + 00	6.15e + 01	2.13e + 01
4.64e − 01	3.09e + 01	5.24e + 00	3.33e + 02	1.39e + 02
7.74e − 01	1.57e + 02	2.75e + 01	2.83e + 03	1.75e + 03
1.29e + 00	1.35e + 03	4.60e + 02	3.11e + 04	3.79e + 04
2.15e + 00	1.96e + 04	9.56e + 03	3.79e + 05	7.33e + 05
3.59e + 00	3.75e + 05	1.84e + 05	5.64e + 06	1.57e + 07
5.99e + 00	7.79e + 06	3.77e + 06	1.02e + 08	4.49e + 08
1.00e + 01	1.66e + 08	7.98e + 07	2.06e + 09	1.48e + 10

Lemma 5.1. *Let $A \in \mathbf{C}^{m\times n}$ $(m \geq n)$ have the polar decomposition $A = UH$. Then*

$$\frac{\|A^*A - I\|}{\|A\|_2 + 1} \leq \|A - U\| \leq \|A^*A - I\|$$

for any unitarily invariant norm $\|\cdot\|$.

Proof. It is straightforward to show that $A^*A - I = (A - U)^*(A + U)$. Taking norms and using (4.7) gives the lower bound, since $\|A^*\| = \|A\|$ for any unitarily invariant norm. Since $A + U = U(H + I)$, we have, from the previous relation,

$$(A - U)^*U = (A^*A - I)(H + I)^{-1}.$$

Using (4.7) again,

$$\begin{aligned} \|A - U\| &= \|(A - U)^*U\| \\ &\leq \|A^*A - I\|\|(H + I)^{-1}\|_2 \\ &\leq \|A^*A - I\|. \end{aligned}$$

■

The next result bounds the distance $\|A-U\|_2$ in terms of the Newton correction $\frac{1}{2}(A-A^{-*})$ [see (6.4) below].

LEMMA 5.2. *Let the nonsingular matrix $A \in \mathbf{C}^{n\times n}$ have the polar decomposition $A = UH$. If $\|A-U\|_2 = \epsilon < 1$, then for any unitarily invariant norm*

$$\frac{1-\epsilon}{2+\epsilon}\|A-A^{-*}\| \le \|A-U\| \le \frac{1+\epsilon}{2-\epsilon}\|A-A^{-*}\|.$$

Proof. Let $E = A-U$. It is straightforward to show that

$$E = (A-A^{-*})(I+E^*U)(2I+E^*U)^{-1}.$$

Since $\|E^*U\|_2 = \|E\|_2 = \epsilon < 1$, (4.7) yields

$$\|E\| \le \|A-A^{-*}\|\frac{1+\epsilon}{2-\epsilon}.$$

The lower bound for E is obtained by taking norms in

$$A-A^{-*} = E(2I+E^*U)(I+E^*U)^{-1}.$$

■

We mention that if all the singular values of A are at least 1 [as is the case for the iterates X_k ($k \ge 1$) from the Newton iteration (6.4) below], then $\|A-U\| \le \|A-A^{-*}\|$ for any unitarily invariant norm, with no restriction on $\|A-U\|$; this inequality was noted for the 2-norm in [33, Lemma 2.1].

Analogous results are already known for the matrix sign function.

LEMMA 5.3 [40]. *Let $A \in \mathbf{C}^{n\times n}$, let $S = \mathrm{sign}(A)$, and let $\|\cdot\|$ be any subordinate matrix norm. If $\|S(A-S)\| < 1$, then*

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

The lower bound always holds.

Proof. The lower bound is obtained by taking norms in $A^2-I = (A-S)(A+S)$. The upper bound is obtained by manipulating the equation $A-S = (A^2-I)(A+S)^{-1}$; for details, see the proof of Lemma 3.1 in [40].

■

LEMMA 5.4 [34]. *Let $A \in \mathbf{C}^{n\times n}$, let $S = \operatorname{sign}(A)$, and let $\|\cdot\|$ be any subordinate matrix norm. If $\|S(A-S)\| = \epsilon < 1$ then*

$$\frac{1-\epsilon}{2+\epsilon}\|A - A^{-1}\| \le \|A - S\| \le \frac{1+\epsilon}{2-\epsilon}\|A - A^{-1}\|.$$

Proof. The matrix $E = A - S$ satisfies

$$E(2I + ES) = (A - A^{-1})(I + ES).$$

The proof is entirely analogous to that of Lemma 5.2. ■

6. ITERATIONS

A direct link between the matrix sign function and the polar decomposition is provided by the relation, for nonsingular $A \in \mathbf{C}^{n\times n}$,

$$\operatorname{sign}\left(\begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix}\right) = \begin{bmatrix} 0 & U \\ U^* & 0 \end{bmatrix}, \tag{6.1}$$

where $A = UH$ is a polar decomposition. This relation, which is easily verified using the formula $\operatorname{sign}(A) = A(A^2)^{-1/2}$, was pointed out to us by R. Byers (private communication, 1984) and was used implicitly in [17]. It allows us to take various formulas and iterations for the matrix sign function and convert them to iterations for the polar decomposition. For example, consider Roberts's integral formula [41],

$$\operatorname{sign}(A) = \frac{2}{\pi}A\int_0^\infty (t^2 I + A^2)^{-1}dt. \tag{6.2}$$

Replacing A in this formula by $\left[\begin{smallmatrix} 0 & A \\ A^* & 0 \end{smallmatrix}\right]$ and invoking (6.1), we obtain

$$\begin{bmatrix} 0 & U \\ U^* & 0 \end{bmatrix} = \frac{1}{2\pi}\begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix}\int_0^\infty \begin{bmatrix} t^2 I + AA^* & 0 \\ 0 & 0 \quad t^2 I + A^*A \end{bmatrix}^{-1} dt,$$

and the (1, 2) and (2, 1) blocks both yield the formula

$$U = \frac{2}{\pi}A\int_0^\infty (t^2 I + A^*A)^{-1}dt. \tag{6.3}$$

This appears to be a new representation for the unitary polar factor U. Although (6.1) is strictly valid only for square nonsingular A (as otherwise $\left[\begin{smallmatrix} 0 & A \\ A^* & 0 \end{smallmatrix}\right]$ is singular), (6.3) holds for any rectangular matrix. We note that (6.3) can also be obtained using an integral formula given by Kato for $C^{-1/2}$, where C has positive definite Hermitian part [29, Section V.3.11, (3.43)]. Although we will not explore it here, one use for (6.3) is to derive perturbation bounds for U.

On applying the Newton iteration (1.2) to $\left[\begin{smallmatrix} 0 & A \\ A^* & 0 \end{smallmatrix}\right]$, we find that the iterates have the form

$$\begin{bmatrix} 0 & X_k \\ X_k^* & 0 \end{bmatrix},$$

and we obtain the following Newton iteration for computing U:

$$X_{k+1} = \tfrac{1}{2}(X_k + X_k^{-*}), \qquad X_0 = A. \tag{6.4}$$

This iteration has been studied in [17] and [33].

In [32] Kenney and Laub derive a family of Padé iterations for computing $\operatorname{sign}(A)$. They have the form

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

where p_r and q_s are polynomials of degree r and s respectively. Pandey, Kenney, and Laub [40] derive an explicit partial fraction form for the iteration with $r = s - 1$ (more recently, Kenney and Laub [35] have found a shorter derivation). The iteration is

$$X_{k+1} = \frac{1}{p} X_k \sum_{i=1}^{p} \frac{1}{\xi_i} \left(X_k^2 + \alpha_i^2 I\right)^{-1}, \qquad X_0 = A, \tag{6.5}$$

where

$$\xi_i = \frac{1}{2}\left(1 + \cos\frac{(2i-1)\pi}{2p}\right), \qquad \alpha_i^2 = \frac{1}{\xi_i} - 1, \quad i = 1{:}p.$$

As shown in [40], this iteration is very suitable for parallel computation. By exploiting (6.1) in the same way as in the derivation of (6.4), we obtain from (6.5) the new iteration for computing U

$$X_{k+1} = \frac{1}{p} X_k \sum_{i=1}^{p} \frac{1}{\xi_i} \left(X_k^* X_k + \alpha_i^2 I\right)^{-1}, \qquad X_0 = A \in \mathbf{C}^{m\times n}. \tag{6.6}$$

It is interesting to note that (6.5) can be obtained from (6.2), and (6.6) from (6.3), by changing the variable of integration according to $t^2 = (1+$

$y)/(1-y)$, applying the Gauss-Chebyshev quadrature rule, and iterating on the rule. From the theory for (6.5) [32, 35, 40] we deduce the following properties of (6.6):

(1) The iteration (6.6) converges to U for any full-rank $A \in \mathbf{C}^{m\times n}$, with order of convergence $2p$.

(2) One step of iteration (6.6) with $p=1$,

$$X_{k+1} = 2X_k(X_k^*X_k + I)^{-1}, \qquad X_0 = A, \tag{6.7}$$

yields the conjugate transpose of the inverse of the matrix from one step of the Newton iteration (6.4), assuming X_k is a square, nonsingular matrix.

(3) If p is a power of 2, one step of iteration (6.6) yields the matrix from $\log_2 p + 1$ steps of the same iteration with $p=1$.

The second and third properties tell us that the iteration (6.6) is a convenient way of combining several Newton iterations into one.

For the $p=1$ iteration (6.7), it is easy to derive the relation

$$X_{k+1}^*X_{k+1} - I = -(X_k^*X_k + I)^{-2}(X_k^*X_k - I)^2.$$

As long as A has full rank, this implies that $\|X_k^*X_k - I\|_2 < 1$ for all iterates X_k $(k \geq 1)$. Using property (3), we can state a fourth property. (Strictly, we have shown this only for p a power of 2, but for general p the property can be established by an adaptation of the proof of Theorem 3.2 in [32].)

(4) If A has full rank, every iterate from (6.6) satisfies $\|X_k^*X_k - I\|_2 < 1$ (and hence $\|X_k - U\|_2 < 1$ by Lemma 5.1).

This last property is important because certain iterations for computing U can be guaranteed to converge only if $\|X_0^*X_0 - I\|_2 < 1$ [7]. An example is the quadratically convergent Schulz iteration

$$X_{k+1} = X_k\big[I + \tfrac{1}{2}(I - X_k^*X_k)\big], \tag{6.8}$$

which is attractive because it involves only matrix multiplication. Property (4) opens the possibility of carrying out a fixed number of iterations of (6.6) and then switching to (6.8); a similar idea of switching from (6.4) to (6.8) is explored in [25].

Implementation of the iteration (6.6) on a parallel computer, including scaling to improve the speed of convergence, is described in a separate paper [24].

7. CONCLUSIONS

We have shown that it is profitable to regard the matrix sign function $S = \operatorname{sign}(A)$ as part of a decomposition, $A = SN$. Analogies between S and N and the polar factors U and H suggest that most results for one decomposition will have a counterpart for the other. Motivated by this observation, we have derived some new results for both decompositions, including a new iteration for computing the polar decomposition. Of particular interest is the formula $S = A(A^2)^{-1/2}$, the analogue of $U = A(A^*A)^{-1/2}$, which is a concise way to define $\operatorname{sign}(A)$ that readily yields some of its key properties.

I thank Des Higham for suggesting improvements to the manuscript.

REFERENCES

1 Huzihiro Araki and Shigeru Yamagami, An inequality for Hilbert-Schmidt norm, *Comm. Math. Phys.* 81:89–96 (1981).

2 L. Autonne, Sur les groupes linéaires, réels et orthogonaux, *Bull. Soc. Math. France* 30:121–134 (1902).

3 Zhaojun Bai and James W. Demmel, Design of a parallel nonsymmetric eigenroutine toolbox, part I, in *Proceedings of the Sixth SIAM Conference on Parallel Processing for Scientific Computing,* Vol. I, SIAM, Philadelphia 1993, pp. 391–398.

4 Anders Barrlund, Perturbation bounds on the polar decomposition, *BIT,* 30:101–113 (1990).

5 A. N. Beavers, Jr., and E. D. Denman, A computational method for eigenvalues and eigenvectors of a matrix with real eigenvalues, *Numer. Math.* 21: 389–396 (1973).

6 Rajendra Bhatia, Matrix factorisations and their perturbations, *Linear Algebra Appl.* 197/198:245–276 (1994).

7 Å. Björck and C. Bowie, An iterative algorithm for computing the best estimate of an orthogonal matrix, *SIAM J. Numer. Anal.* 8(2):358–364 (1971).

8 Ralph Byers, Solving the algebraic Riccati equation with the matrix sign function, *Linear Algebra Appl.* 85:267–279 (1987).

9 J.-P. Charlier and P. Van Dooren, A systolic algorithm for Riccati and Lyapunov equations, *Math. Control Signals Systems* 2:109–136 (1989).

10 Eugene D. Denman, Roots of real matrices, *Linear Algebra Appl.* 36:133–139 (1981).

11 Eugene D. Denman and Alex N. Beavers, Jr., The matrix sign function and computations in systems, *Appl. Math. Comput.* 2:63–94 (1976).

12 C. R. DePrima and C. R. Johnson, The range of $A^{-1}A^*$ in $GL(n, C)$, *Linear Algebra Appl.* 9:209–222 (1974).

13 Ky Fan and A. J. Hoffman, Some metric inequalities in the space of matrices, *Proc. Amer. Math. Soc.* 6:111–116 (1955).

14 Judith D. Gardiner and Alan J. Laub, Solving the algebraic Riccati equation on a hypercube multiprocessor, in *Hypercube Concurrent Computers and Applications,* Vol. II, (G. Fox, Ed.), ACM Press, New York, 1988, pp. 1562–1568.

15 Judith D. Gardiner and Alan J. Laub, Parallel algorithms for algebraic Riccati equations, *Internat. J. Control* 54(6):1317–1333 (1991).

16 W. W. Hager, Condition estimates, *SIAM J. Sci. Statist. Comput.* 5:311–316 (1984).

17 Nicholas J. Higham, Computing the polar decomposition—with applications, *SIAM J. Sci. Statist. Comput.* 7(4):1160–1174 (Oct. 1986).

18 Nicholas J. Higham, Computing real square roots of a real matrix, *Linear Algebra Appl.* 88/89:405–430 (1987).

19 Nicholas J. Higham, Computing a nearest symmetric positive semidefinite matrix, *Linear Algebra Appl.* 103:103–118 (1988).

20 Nicholas J. Higham, FORTRAN codes for estimating the one-norm of a real or complex matrix, with applications to condition estimation (Algorithm 674), *ACM Trans. Math. Software* 14(4):381–396 (Dec. 1988).

21 Nicholas J. Higham, Matrix nearness problems and applications, in *Applications of Matrix Theory,* (M. J. C. Gover and S. Barnett, Eds.), Oxford U.P., 1989, pp. 1–27.

22 Nicholas J. Higham, Experience with a matrix norm estimator, *SIAM J. Sci. Statist. Comput.* 11(4):804–809 (July 1990).

23 Nicholas J. Higham, Perturbation Theory and Backward Error for $AX - XB = C$, *BIT* 33:124–136 (1993).

24 Nicholas J. Higham and Pythagoras Papadimitriou, A Parallel Algorithm for Computing the Polar Decomposition, *Parallel Computing* 20(8):1161–1173 (1994).

25 Nicholas J. Higham and Robert S. Schreiber, Fast polar decomposition of an arbitrary matrix, *SIAM J. Sci. Statist. Comput.* 11(4):648–655 (July 1990).

26 Roger A. Horn and Charles R. Johnson, *Matrix Analysis,* Cambridge U.P., 1985.

27 Roger A. Horn and Charles R. Johnson, *Topics in Matrix Analysis,* Cambridge U.P., 1991.

28 James Lucien Howland, The sign matrix and the separation of matrix eigenvalues, *Linear Algebra Appl.* 49:221–232 (1983).

29 Tosio Kato, *Perturbation Theory for Linear Operators,* 2nd ed., Springer-Verlag, Berlin, 1976.

30 Charles Kenney and Alan J. Laub, Condition estimates for matrix functions, *SIAM J. Matrix Anal. Appl.* 10(2):191–209 (1989).

31 Charles Kenney and Alan J. Laub, Polar decomposition and matrix sign

function condition estimates, *SIAM J. Sci. Statist. Comput.* 12:488–504 (1991).

32 Charles Kenney and Alan J. Laub, Rational iterative methods for the matrix sign function, *SIAM J. Matrix Anal. Appl.* 12(2):273–291 (1991).

33 Charles Kenney and Alan J. Laub, On scaling Newton's method for polar decomposition and the matrix sign function, *SIAM J. Matrix Anal. Appl.* 13(3):688–706 (1992).

34 Charles Kenney, Alan J. Laub, and P. M. Papadopoulos, A Newton-squaring algorithm for computing the negative invariant subspace of a matrix, *IEEE Trans. Automat. Control* 38(8):1284–1289 (1993).

35 Charles S. Kenney and Alan J. Laub, A hyperbolic tangent identity and the geometry of Padé sign function iterations, Preprint, Dept. of Electrical and Computer Engineering, Univ. of California, Santa Barbara, 1993.

36 Charles S. Kenney, Alan J. Laub, and Philip M. Papadopoulos, Matrix sign algorithms for Riccati equations, in *Proceedings of IMA Conference on Control: Modelling, Computation, Information,* (Southend-On-Sea, 1992), Inst. Mathematics and its Applications, pp. 1–10.

37 Chih-Chang Lin and Earl Zmijewski, A Parallel Algorithm for Computing the Eigenvalues of an Unsymmetric Matrix on an SIMD Mesh of Processors, Report TRCS 91-15, Dept. of Computer Science, Univ. of California, Santa Barbara, July 1991.

38 Roy Mathias, Condition estimation for matrix functions via the Schur decomposition, Manuscript, Inst. for Mathematics and its Applications, Univ. of Minnesota, 1993 to appear in *SIAM J. Matrix Anal. Appl.*

39 Roy Mathias, Perturbation bounds for the polar decomposition, *SIAM J. Matrix Anal. Appl.*, 14(2):588–597 (1993).

40 Pradeep Pandey, Charles Kenney, and Alan J. Laub, A Parallel algorithm for the matrix sign function, *Internat, J. High Speed Comput.* 2(2):181–191 (1990).

41 J. D. Roberts, Linear model reduction and solution of the algebraic Riccati equation by use of the sign function, *Internat. J. Control* 32(4):677–687 (1980); first issued as Report CUED/B-Control/TR13, Dept. of Engineering, Univ. of Cambridge, 1971.

42 Frank Uhlig, Explicit polar decomposition and a near-characteristic polynomial: The 2 × 2 case, *Linear Algebra Appl.* 38:239–249 (1981).

Received 22 April 1993; final manuscript accepted 11 July 1993