SIAM J. MATRIX ANAL. APPL.
Vol. 38, No. 4, pp. 1354–1379

# ITERATIVE COMPUTATION OF THE FRÉCHET DERIVATIVE OF THE POLAR DECOMPOSITION[*]

EVAN S. GAWLIK[†] AND MELVIN LEOK[†]

**Abstract.** We derive iterative methods for computing the Fréchet derivative of the map which sends a full-rank matrix $A$ to the factor $U$ in its polar decomposition $A = UH$, where $U$ has orthonormal columns and $H$ is Hermitian positive definite. The methods apply to square matrices as well as rectangular matrices having more rows than columns. Our derivation relies on a novel identity that relates the Fréchet derivative of the polar decomposition to the matrix sign function $\operatorname{sign}(X) = X(X^2)^{-1/2}$ applied to a certain block matrix $X$.

**Key words.** polar decomposition, Fréchet derivative, matrix function, matrix iteration, Newton iteration, Newton–Schulz iteration, matrix sign function

**AMS subject classifications.** 65F30, 15A23, 15A24

**DOI.** 10.1137/16M108971X

## 1. Introduction.

The polar decomposition theorem asserts that every matrix $A \in \mathbb{C}^{m\times n}$ $(m \geq n)$ can be written as the product $A = UH$ of a matrix $U \in \mathbb{C}^{m\times n}$ having orthonormal columns times a Hermitian positive semidefinite matrix $H \in \mathbb{C}^{n\times n}$ [18, Theorem 8.1]. If $A$ is full-rank, then this decomposition is unique and $H$ is positive definite, allowing one to define a map $\mathcal{P}$ which sends a full-rank matrix $A \in \mathbb{C}^{m\times n}$ to the factor $\mathcal{P}(A) = U \in \mathbb{C}^{m\times n}$ in its polar decomposition $A = UH$. We refer to $U$ as the unitary factor in the polar decomposition of $A$, bearing in mind that this is a slight abuse of terminology when $A$ (and hence $U$) is rectangular. The aim of this paper is to derive iterative algorithms for computing the Fréchet derivative of $\mathcal{P}$.

Our interest in differentiating the polar decomposition stems from several sources. First, differentiating the polar decomposition gives precise information about the sensitivity of the polar decomposition to perturbations. This is a topic of longstanding interest in numerical analysis [25, 26, 6, 21, 28, 7, 4, 24], where much of the literature has focused on bounding the deviations in the perturbed factors in the polar decomposition of $A$ after a small-normed perturbation of $A$. These analyses often rely on a formula for the Fréchet derivative of $\mathcal{P}$ that involves the singular value decomposition (SVD) of $A$ [21, equation 2.18]. While theoretically useful, such a formula loses some of its appeal in the numerical setting, where computing the SVD tends to be costly. As a second source of motivation, differentiating the polar decomposition has proven necessary in the design of certain schemes for interpolating functions which take values in the special orthogonal group [12], the group of real square matrices with orthonormal columns and positive determinant. These interpolation schemes have applications in computer animation, mechanics, and other areas in which continuously varying rotation matrices play a role.

[*] Received by the editors August 16, 2016; accepted for publication (in revised form) by J. L. Barlow May 19, 2017; published electronically November 14, 2017.
http://www.siam.org/journals/simax/38-4/M108971.html
**Funding:** The work of the first author was supported in part by the NSF under grants DMS-1411792 and DMS-1345013. The work of the second author was supported in part by the NSF under grants DMS-1010687, CMMI-1029445, DMS-1065972, CMMI-1334759, DMS-1411792, and DMS-1345013.

[†] Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0112 (egawlik@ucsd.edu, mleok@math.ucsd.edu).

A number of authors have addressed the computation of the Fréchet derivatives of other functions of matrices, such as the matrix exponential [1, 27, 30], the matrix logarithm [3, 23], the matrix square root [1, section 2], the matrix $p$th root [19, 9, 8], and the matrix sign function $\mathrm{sign}(X) = X(X^2)^{-1/2}$ [21]. The aforementioned functions, unlike the map $\mathcal{P}$, are examples of *primary matrix functions*. Roughly speaking, a primary matrix function is a scalar function that has been extended to square matrices in a canonical way; for a precise definition, see [18, section 1.2] and [20]. The polar decomposition is not a primary matrix function, which is perhaps the main reason that the computation of its Fréchet derivative (a quantity whose existence is justified in section 3.1) has largely evaded scrutiny until now.

Formally, iterative schemes for computing the Fréchet derivatives of matrix functions (be they primary or nonprimary) can be derived as follows. Let $f : \mathbb{C}^{m\times n} \to \mathbb{C}^{m\times n}$ be a function with Fréchet derivative $L_f$. That is, given $X \in \mathbb{C}^{m\times n}$, the map $L_f(X,\cdot) : \mathbb{C}^{m\times n} \to \mathbb{C}^{m\times n}$ is a linear map satisfying

$$f(X+E) - f(X) - L_f(X,E) = o(\|E\|) \tag{1}$$

for every $E \in \mathbb{C}^{m\times n}$, where $\|\cdot\|$ denotes any matrix norm. Let $A \in \mathbb{C}^{m\times n}$, and suppose that

$$X_{k+1} = g(X_k), \quad X_0 = A, \tag{2}$$

is an iterative scheme for computing $f(A)$; that is, $X_k \to f(A)$ as $k \to \infty$. Then, as noted in [1, section 2], differentiation of (2) with respect to $A$ in the direction $E \in \mathbb{C}^{m\times n}$ yields the coupled iteration

$$X_{k+1} = g(X_k), \qquad X_0 = A, \tag{3}$$

$$E_{k+1} = L_g(X_k, E_k), \quad E_0 = E, \tag{4}$$

for computing $f(A)$ and $L_f(A,E)$. The validity of this formal derivation, of course, depends on the commutativity of $\lim_{k\to\infty}$ with differentiation, which is generally nontrivial to establish.

For a primary matrix function $f$, proving the validity of this formal derivation is greatly simplified by the following identity. For any primary matrix function $f$ and any square matrices $A$ and $E$,

$$f\begin{pmatrix} A & E \\ 0 & A \end{pmatrix} = \begin{pmatrix} f(A) & L_f(A,E) \\ 0 & f(A) \end{pmatrix}, \tag{5}$$

provided that $f$ is $2p-1$ times continuously differentiable on an open subset of $\mathbb{C}$ containing the spectrum of $A$, where $p$ is the size of the largest Jordan block of $A$ [29]. From this it follows that if (2) is an iterative scheme for computing $f(A)$, and if $g$ maps block upper triangular matrices to block upper triangular matrices, then

$$\begin{pmatrix} X_{k+1} & E_{k+1} \\ 0 & X_{k+1} \end{pmatrix} = g\begin{pmatrix} X_k & E_k \\ 0 & X_k \end{pmatrix}, \quad \begin{pmatrix} X_0 & E_0 \\ 0 & X_0 \end{pmatrix} = \begin{pmatrix} A & E \\ 0 & A \end{pmatrix} \tag{6}$$

defines an iterative scheme for computing $\left(\begin{smallmatrix} f(A) & L_f(A,E) \\ 0 & f(A) \end{smallmatrix}\right)$, provided that it converges and provided that $f$ has the requisite regularity to apply (5). Using (5) again to isolate each block of the iteration (6), one obtains the coupled iteration (3)–(4). Details behind this argument, as well as an example of its application, can be found in [1, section 2].

Our main result in this paper, Theorem 2.1, establishes the validity of schemes like (3)–(4) when the function $f$ under consideration is the function $\mathcal{P}$ which sends $A$ to the unitary factor $U$ in its polar decomposition, *even though $\mathcal{P}$ is not a primary matrix function.* In particular,

$$\mathcal{P}\begin{pmatrix} A & E \\ 0 & A \end{pmatrix} \neq \begin{pmatrix} \mathcal{P}(A) & L_{\mathcal{P}}(A,E) \\ 0 & \mathcal{P}(A) \end{pmatrix},$$

so the argument in the preceding paragraph does not apply. For example, if $A = 2$ and $E = 3$, then it is not hard to check that

$$\mathcal{P}\begin{pmatrix} A & E \\ 0 & A \end{pmatrix} = \begin{pmatrix} 4/5 & 3/5 \\ -3/5 & 4/5 \end{pmatrix},$$

but

$$\begin{pmatrix} \mathcal{P}(A) & L_{\mathcal{P}}(A,E) \\ 0 & \mathcal{P}(A) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Instead of using (5), our derivation relies on a novel identity that relates the Fréchet derivative of $\mathcal{P}$ to the matrix sign function $\operatorname{sign}(X) = X(X^2)^{-1/2}$ applied to a certain block matrix $X$; see Theorem 2.3.

One notable corollary of Theorem 2.1 is that the popular Newton iteration [16]

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

for computing the unitary factor $\mathcal{P}(A) = U$ in the polar decomposition $A = UH$ of a square matrix $A$ extends to a coupled iteration for computing $\mathcal{P}(A)$ and its Fréchet derivative. In particular, Corollary 2.4 shows that for any nonsingular $A \in \mathbb{C}^{n\times n}$ and any $E \in \mathbb{C}^{n\times n}$, the scheme

$$\begin{aligned} X_{k+1} &= \frac{1}{2}(X_k + X_k^{-*}), & X_0 &= A, \\ E_{k+1} &= \frac{1}{2}(E_k - X_k^{-*}E_k^*X_k^{-*}), & E_0 &= E, \end{aligned}$$

produces iterates $X_k$ and $E_k$ that converge to $\mathcal{P}(A) = U$ and $L_{\mathcal{P}}(A,E)$, respectively, as $k \to \infty$.

The fact that the matrix sign function will play a role in our study of Fréchet derivatives of the polar decomposition should come as no surprise, given the sign function's intimate connection with the polar decomposition. The sign function and polar decomposition are linked via the identity

$$\operatorname{sign}\begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix} = \begin{pmatrix} 0 & \mathcal{P}(A) \\ \mathcal{P}(A)^* & 0 \end{pmatrix}, \tag{7}$$

which holds for any square nonsingular matrix $A$ [18]. This identity has been used, among other things, to derive iterative schemes for computing the polar decomposition. The essence of this approach is to write an iterative scheme for computing $\operatorname{sign}\left(\begin{smallmatrix} 0 & A \\ A^* & 0 \end{smallmatrix}\right)$, check that its iterates retain the relevant block structure, and read off the $(1,2)$-block of the resulting algorithm. In principle, one can adopt a similar strategy to derive iterative schemes for computing the Fréchet derivatives of the polar decomposition. Indeed, any iterative scheme that computes

$$\operatorname{sign}\begin{pmatrix} 0 & A & 0 & E \\ A^* & 0 & E^* & 0 \\ 0 & 0 & 0 & A \\ 0 & 0 & A^* & 0 \end{pmatrix}$$

while retaining its block structure will suffice, owing to the following observation. By appealing to the definition (1) of the Fréchet derivative, the identity (7) can be used to verify that

$$L_{\mathrm{sign}}\left(\begin{pmatrix}0 & A\\ A^* & 0\end{pmatrix}, \begin{pmatrix}0 & E\\ E^* & 0\end{pmatrix}\right) = \begin{pmatrix}0 & L_{\mathcal{P}}(A,E)\\ L_{\mathcal{P}}(A,E)^* & 0\end{pmatrix}. \tag{8}$$

Now since the sign function is a primary matrix function, (5), (7), and (8) imply that

$$\begin{aligned}\mathrm{sign}\begin{pmatrix}0 & A & 0 & E\\ A^* & 0 & E^* & 0\\ 0 & 0 & 0 & A\\ 0 & 0 & A^* & 0\end{pmatrix} &= \begin{pmatrix}\mathrm{sign}\begin{pmatrix}0 & A\\ A^* & 0\end{pmatrix} & L_{\mathrm{sign}}\left(\begin{pmatrix}0 & A\\ A^* & 0\end{pmatrix}, \begin{pmatrix}0 & E\\ E^* & 0\end{pmatrix}\right)\\ 0 & \mathrm{sign}\begin{pmatrix}0 & A\\ A^* & 0\end{pmatrix}\end{pmatrix}\\ &= \begin{pmatrix}0 & \mathcal{P}(A) & 0 & L_{\mathcal{P}}(A,E)\\ \mathcal{P}(A)^* & 0 & L_{\mathcal{P}}(A,E)^* & 0\\ 0 & 0 & 0 & \mathcal{P}(A)\\ 0 & 0 & \mathcal{P}(A)^* & 0\end{pmatrix}.\end{aligned}$$

A drawback of this approach is that it is valid only for square matrices $A$. The strategy we adopt in the present paper will be quite different and will be valid not just for square matrices $A$ but also for rectangular matrices $A$ having more rows than columns.

*Organization.* This paper is organized as follows. We begin in section 2 by giving statements of our main results, deferring their proof to section 3. In section 4, we discuss several practical aspects of the iterative schemes, including stability, scaling, and termination criteria. We compare the iterative schemes to other methods for computing the Fréchet derivative of the polar decomposition in section 5. We finish with some numerical experiments in section 6.

## 2. Statement of results.

In this section, we give a presentation of this paper's main result, which is a theorem that details a class of iterative schemes for computing the Fréchet derivative $L_{\mathcal{P}}$ of the map $\mathcal{P}$ which sends a matrix $A$ to the unitary factor $U$ in its polar decomposition $A = UH$. A proof of the theorem is given in section 3.

The class of iterative schemes to be considered comprises schemes of the form (3)–(4), with a mild constraint on the form of the function $g$. To understand this constraint, it is helpful to develop some intuition concerning iterative schemes for computing the polar decomposition and their relationship to iterative schemes for computing the matrix sign function. Fundamental to that intuition are the identities

$$\mathrm{sign}(A) = A(A^2)^{-1/2}, \quad \mathcal{P}(A) = A(A^*A)^{-1/2}, \tag{9}$$

and the integral representation formulas [17, equations 6.2 and 6.3]

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

which hint at two rules of thumb. First, iterative schemes for computing the matrix sign function tend to have the form $X_{k+1} = X_k h(X_k^2)$, where $h$ is a primary matrix function. Second, to each iterative scheme $X_{k+1} = X_k h(X_k^2)$ for computing the matrix sign function, there corresponds an iterative scheme $X_{k+1} = X_k h(X_k^* X_k)$ for computing the polar decomposition. The first of these rules of thumb appears to hold empirically to our knowledge. The second is made precise in [18, Theorem

8.13]. The theorem below extends [18, Theorem 8.13] by showing, in essence, that to each iterative scheme $X_{k+1} = X_k h(X_k^2)$ for computing the matrix sign function, there corresponds an iterative scheme for computing the polar decomposition *and its Fréchet derivative*. This iterative scheme is given by (3)–(4) with $g(X) = Xh(X^*X)$.

In what follows, we denote by $\operatorname{skew}(B) = \frac{1}{2}(B - B^*)$ and $\operatorname{sym}(B) = \frac{1}{2}(B + B^*)$ the skew-Hermitian and Hermitian parts, respectively, of a square matrix $B$. We denote the spectrum of $B$ by $\Lambda(B)$.

THEOREM 2.1. *Let $A \in \mathbb{C}^{m\times n}$ $(m \geq n)$ be a full-rank matrix having polar decomposition $A = UH$, where $U \in \mathbb{C}^{m\times n}$ has orthonormal columns and $H \in \mathbb{C}^{n\times n}$ is Hermitian positive definite. Let $E \in \mathbb{C}^{m\times n}$, and define $\Omega = \operatorname{skew}(U^*E)$ and $S = \operatorname{sym}(U^*E)$. Let $h$ be a primary matrix function satisfying $h(Z^*) = h(Z)^*$ for every $Z$, and suppose that the iteration $Z_{k+1} = Z_k h(Z_k^2)$ produces iterates $Z_k$ that converge to* $\operatorname{sign}(Z_0)$ *as $k \to \infty$ when the initial condition is*

$$Z_0 = \begin{pmatrix} H & \Omega \\ 0 & -H \end{pmatrix}, \tag{10}$$

*as well as when the initial condition is*

$$Z_0 = \begin{pmatrix} H & S \\ 0 & H \end{pmatrix}. \tag{11}$$

*Assume that in both cases, $h$ is smooth on an open subset of $\mathbb{C}$ containing $\cup_{k=0}^{\infty}\Lambda(Z_k)$. Let $g(X) = Xh(X^*X)$. Then the iteration*

$$X_{k+1} = g(X_k), \qquad X_0 = A, \tag{12}$$

$$E_{k+1} = L_g(X_k, E_k), \quad E_0 = E, \tag{13}$$

*produces iterates $X_k$ and $E_k$ that converge to $\mathcal{P}(A) = U$ and $L_{\mathcal{P}}(A, E)$, respectively, as $k \to \infty$.*

*Remark* 2.2. Taking $E = 0$ in the preceding theorem, one recovers [18, Theorem 8.13], up to the following modification: Instead of requesting that $h$ is a primary matrix function satisfying $h(Z^*) = h(Z)^*$, [18, Theorem 8.13] makes the weaker assumption that the function $\widetilde{g}(Z) = Zh(Z^2)$ satisfies $\widetilde{g}(Z^*) = \widetilde{g}(Z)^*$ for every $Z$. It is easily checked using elementary properties of primary matrix functions [18, Theorem 1.13] that the latter is implied by the former.

Note that it is sometimes the case that the convergence of the matrix sign function iteration $Z_{k+1} = Z_k h(Z_k^2)$ referenced in Theorem 2.1 is dictated by the spectrum of $Z_0$. If this is the case, then the hypothesis that the iteration converges when $Z_0$ is given by (10) or (11) is equivalent to the simpler hypothesis that the iteration converges when $Z_0 = \pm H$. This follows from the fact that the spectra of (10) and (11) are given by $\Lambda(H) \cup \Lambda(-H)$ and $\Lambda(H)$, respectively.

Central to the proof of Theorem 2.1 is an identity that relates the Fréchet derivative of the polar decomposition to the sign of the block matrix $Z_0$ appearing in (10). We state the identity below to emphasize its importance. A proof is given in section 3.1.

THEOREM 2.3. *Let $A$, $U$, $H$, $E$, and $\Omega$ be as in Theorem* 2.1. *Then*

$$\operatorname{sign}\begin{pmatrix} H & \Omega \\ 0 & -H \end{pmatrix} = \begin{pmatrix} I & U^*L_{\mathcal{P}}(A, E) \\ 0 & -I \end{pmatrix}. \tag{14}$$

*In particular, if $U^*E$ is skew-Hermitian, then*

$$\operatorname{sign}\left(\begin{pmatrix} U^* & 0 \\ 0 & -U^* \end{pmatrix}\begin{pmatrix} A & E \\ 0 & A \end{pmatrix}\right) = \begin{pmatrix} U^* & 0 \\ 0 & -U^* \end{pmatrix}\begin{pmatrix} \mathcal{P}(A) & L_{\mathcal{P}}(A,E) \\ 0 & \mathcal{P}(A) \end{pmatrix}. \tag{15}$$

In addition to being useful in the proof of Theorem 2.1, the identity (15) bears an interesting resemblance to (5).

Theorem 2.1 has several corollaries, each corresponding to a different choice of iterative scheme $Z_{k+1} = Z_k h(Z_k^2)$ for computing the matrix sign function. The simplest is the well-known Newton iteration

$$Z_{k+1} = \frac{1}{2}(Z_k + Z_k^{-1}), \tag{16}$$

which corresponds to the choice $h(Z) = \frac{1}{2}(I + Z^{-1})$. It is known that this iteration converges quadratically to $\operatorname{sign}(Z_0)$ for any $Z_0$ having no pure imaginary eigenvalues [18, Theorem 5.6]. Since (10) and (11) have eigenvalues equal to plus or minus the eigenvalues of $H$, all of which are nonzero real numbers, we obtain the following corollary. In it, we restrict the discussion to square matrices, since this leads to a particularly simple iterative scheme.

Corollary 2.4. *Let $A \in \mathbb{C}^{n\times n}$ be a nonsingular matrix having polar decomposition $A = UH$, where $U \in \mathbb{C}^{n\times n}$ is unitary and $H \in \mathbb{C}^{n\times n}$ is Hermitian positive definite. Let $E \in \mathbb{C}^{n\times n}$. Then the iteration*

$$X_{k+1} = \frac{1}{2}(X_k + X_k^{-*}), \qquad X_0 = A, \tag{17}$$

$$E_{k+1} = \frac{1}{2}(E_k - X_k^{-*}E_k^*X_k^{-*}), \quad E_0 = E, \tag{18}$$

*produces iterates $X_k$ and $E_k$ that converge to $\mathcal{P}(A) = U$ and $L_{\mathcal{P}}(A,E)$, respectively, as $k \to \infty$.*

*Remark* 2.5. If $A$ is rectangular, then the iteration obtained from (16) reads

$$X_{k+1} = \frac{1}{2}X_k(I + (X_k^*X_k)^{-1}), \qquad X_0 = A, \tag{19}$$

$$\begin{aligned} E_{k+1} = \frac{1}{2}\Big[&E_k(I + (X_k^*X_k)^{-1}) \\ &- X_k(X_k^*X_k)^{-1}(E_k^*X_k + X_k^*E_k)(X_k^*X_k)^{-1}\Big], \quad E_0 = E. \end{aligned} \tag{20}$$

This scheme simplifies to (17)–(18) when $A$ is square.

A second corollary of Theorem 2.1 is obtained by considering the Newton–Schulz iteration

$$Z_{k+1} = \frac{1}{2}Z_k(3I - Z_k^2), \tag{21}$$

which corresponds to the choice $h(Z) = \frac{1}{2}(3I - Z)$. It is known that this iteration converges to $\operatorname{sign}(Z_0)$ provided that (i) $Z_0$ has no pure imaginary eigenvalues and (ii) the eigenvalues of $I - Z_0^2$ all have magnitude strictly less than one [22, Theorem 5.2]. Note that [22, Theorem 5.2] replaces the latter condition with $\|I - Z_0^2\| < 1$, but it is evident from their proof that this condition can be relaxed to what we have written here. Since the eigenvalues of

$$\begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} - \begin{pmatrix} H & \Omega \\ 0 & -H \end{pmatrix}^2 = \begin{pmatrix} I - H^2 & \Omega H - H\Omega \\ 0 & I - H^2 \end{pmatrix}$$

and

$$\begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} - \begin{pmatrix} H & S \\ 0 & H \end{pmatrix}^2 = \begin{pmatrix} I - H^2 & -HS - SH \\ 0 & I - H^2 \end{pmatrix}$$

coincide with those of $I - H^2 = I - A^*A$, we obtain the following corollary.

Corollary 2.6. *Let $A \in \mathbb{C}^{m\times n}$ $(m \geq n)$ be a full-rank matrix having polar decomposition $A = UH$, where $U \in \mathbb{C}^{m\times n}$ has orthonormal columns and $H \in \mathbb{C}^{n\times n}$ is Hermitian positive definite. Let $E \in \mathbb{C}^{m\times n}$. If all of the singular values of $A$ lie in the interval $(0, \sqrt{2})$, then the iteration*

$$X_{k+1} = \frac{1}{2}X_k(3I - X_k^*X_k), \qquad X_0 = A, \tag{22}$$

$$E_{k+1} = \frac{1}{2}E_k(3I - X_k^*X_k) - \frac{1}{2}X_k(E_k^*X_k + X_k^*E_k), \quad E_0 = E, \tag{23}$$

*produces iterates $X_k$ and $E_k$ that converge to $\mathcal{P}(A) = U$ and $L_{\mathcal{P}}(A,E)$, respectively, as $k \to \infty$.*

*Remark* 2.7. A more direct analysis of (22), without appealing to its relationship to a matrix sign function iteration, shows that $X_k \to U$ under the less stringent requirement that all of the singular values of $A$ lie in the interval $(0, \sqrt{3})$ [18, Problem 8.20]. Our numerical experiments suggest that the coupled iteration (22)–(23) enjoys convergence under the same condition, but Theorem 2.1 alone appears inadequate to conclude such a claim.

Other corollaries to Theorem 2.1 can be derived in a similar fashion. For instance, iterative schemes based on Padé approximations of $\operatorname{sign}(Z) = Z(I - (I - Z^2))^{-1/2}$ (of which (21) is a special case) can be used; see [18, Chapter 5.4] for further details.

## 3. Proofs.

In this section, we present proofs of Theorems 2.1 and 2.3. Our presentation is divided into two parts. First, in section 3.1, we derive a few identities involving the Fréchet derivative of the polar decomposition, proving Theorem 2.3 in the process. Then, in section 3.2, we use the aforementioned identites to prove convergence of the iteration (12)–(13), thereby proving Theorem 2.1.

### 3.1. Identities involving the Fréchet derivative of the polar decomposition.

This section studies the Fréchet derivative of the polar decomposition and its relationship to the matrix sign function, culminating in a proof of Theorem 2.3. A couple of main observations will be made. First, as will be seen in Lemma 3.3, the task of evaluating $L_{\mathcal{P}}(A,E)$ can essentially be reduced to the case in which $A$ is Hermitian positive definite and $E$ is skew-Hermitian. This is relatively simple to show when $A$ is square, but the rectangular case turns out to be more subtle, requiring that some attention be paid to the relationship between the column space of $A$ and that of $E$. This observation will be followed with a proof of Theorem 2.3, which reveals that the value of $U^*L_{\mathcal{P}}(A,E)$ can be read off of the $(1,2)$-block of the matrix sign function applied to a certain block matrix.

Before studying the derivatives of $\mathcal{P}$ in detail, it is worth pointing out that $\mathcal{P}$ is a smooth map from the set of full-rank $m \times n$ $(m \geq n)$ matrices to the set of $m \times n$ matrices with orthonormal columns. This follows from two facts: (1) the latter set of matrices constitutes a smooth, compact manifold, the Stiefel manifold $V_n(\mathbb{C}^m) = \{U \in \mathbb{C}^{m\times n} \mid U^*U = I\}$, and (2) the map $\mathcal{P}$ coincides with the closest point projection onto $V_n(\mathbb{C}^n)$. That is, in the Frobenius norm $\|\cdot\|_F$,

$$\mathcal{P}(A) = \underset{U\in V_n(\mathbb{C}^m)}{\arg\min}\ \|A - U\|_F$$

for any full-rank $A \in \mathbb{C}^{m\times n}$ [18, Theorem 8.4]. It is a classical result from differential geometry that the closest point projection onto a smooth, compact manifold embedded in Euclidean space is a smooth map [11]. In particular, $\mathcal{P}$ is Fréchet differentiable at any full-rank $A \in \mathbb{C}^{m\times n}$. (For a different justification of this fact, see [10, section 2.3(c)].)

We now turn our attention to the differentiation of $\mathcal{P}$. We begin by recording a useful formula for the Fréchet derivative of a function of the form $g(X) = Xh(X^*X)$. Along the way, we make some observations concerning the column space $\mathcal{R}(A)$ of a matrix $A \in \mathbb{C}^{m\times n}$ and the column space $\mathcal{R}(L_g(A,E))$ of the Fréchet derivative $L_g(A,E)$ of $g$ at $A$ in a direction $E \in \mathbb{C}^{m\times n}$. We denote by $\mathcal{N}(A^*)$ the null space of $A^*$; equivalently, $\mathcal{N}(A^*)$ is the orthogonal complement to $\mathcal{R}(A)$ in $\mathbb{C}^m$.

Lemma 3.1. *Let $A \in \mathbb{C}^{m\times n}$ $(m \geq n)$, let $h : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ be Fréchet differentiable at $A^*A$, and define $g(X) = Xh(X^*X)$. Then for any $E \in \mathbb{C}^{m\times n}$,*

$$L_g(A,E) = Eh(A^*A) + AL_h(A^*A, A^*E + E^*A). \tag{24}$$

*In particular, if $\mathcal{R}(E) \subseteq \mathcal{R}(A)$, then $\mathcal{R}(L_g(A,E)) \subseteq \mathcal{R}(A)$. On the other hand, if $\mathcal{R}(E) \subseteq \mathcal{N}(A^*)$, then*

$$L_g(A,E) = Eh(A^*A), \tag{25}$$

*and hence $\mathcal{R}(L_g(A,E)) \subseteq \mathcal{N}(A^*)$.*

*Proof.* The formula (24) is a consequence of the product rule and the chain rule [18, Theorems 3.3 and 3.4]. The implication $\mathcal{R}(E) \subseteq \mathcal{R}(A) \implies \mathcal{R}(L_g(A,E)) \subseteq \mathcal{R}(A)$ is immediate since the columns of $L_g(A,E)$ are linear combinations of the columns of $A$ and $E$. Equation (25) follows from the fact that $A^*E + E^*A = 0$ whenever $\mathcal{R}(E) \subseteq \mathcal{N}(A^*)$. □

The preceding lemma has several important consequences. The first of these is an application of Lemma 3.1 to the function $g(X) = \mathcal{P}(X)$, which has the requisite functional form in view of (9).

Lemma 3.2. *Let $A \in \mathbb{C}^{m\times n}$ $(m \geq n)$ be a full-rank matrix having polar decomposition $A = UH$, where $U = \mathcal{P}(A) \in \mathbb{C}^{m\times n}$ has orthonormal columns and $H \in \mathbb{C}^{n\times n}$ is Hermitian positive definite. Let $E \in \mathbb{C}^{m\times n}$, and write*

$$E = E^{\parallel} + E^{\perp}, \quad E^{\parallel} = UU^*E, \quad E^{\perp} = (I - UU^*)E. \tag{26}$$

*Then*

$$UU^*L_{\mathcal{P}}(A, E^{\parallel}) = L_{\mathcal{P}}(A, E^{\parallel}) \tag{27}$$

*and*

$$L_{\mathcal{P}}(A, E^{\perp}) = E^{\perp}H^{-1}. \tag{28}$$

*Proof.* Apply Lemma 3.1 with the choice $h(X) = X^{-1/2}$, so that $g(X) = X(X^*X)^{-1/2} = \mathcal{P}(X)$. Equation (27) is a restatement of the fact that $\mathcal{R}(L_{\mathcal{P}}(A,E^{\parallel})) \subseteq \mathcal{R}(A) = \mathcal{R}(U)$, while (28) follows from (25) together with the identity $H = (A^*A)^{1/2}$. □

We will now show, with the help of Lemma 3.2, that the task of evaluating $L_{\mathcal{P}}(A,E)$ can essentially be reduced to the case in which $A$ is Hermitian positive definite and $E$ is skew-Hermitian.

Lemma 3.3. *Let $A \in \mathbb{C}^{m\times n}$ $(m \geq n)$ be a full-rank matrix having polar decomposition $A = UH$, where $U \in \mathbb{C}^{m\times n}$ has orthonormal columns and $H \in \mathbb{C}^{n\times n}$ is Hermitian positive definite. Then for any $E \in \mathbb{C}^{m\times n}$,*

$$\operatorname{skew}(U^* L_{\mathcal{P}}(A,E)) = L_{\mathcal{P}}(H,\Omega), \tag{29}$$

$$\operatorname{sym}(U^* L_{\mathcal{P}}(A,E)) = L_{\mathcal{P}}(H,S) = 0, \tag{30}$$

*where* $\Omega = \operatorname{skew}(U^*E)$ *and* $S = \operatorname{sym}(U^*E)$. *Hence,*

$$U^* L_{\mathcal{P}}(A,E) = L_{\mathcal{P}}(H,\Omega).$$

*Proof.* Let $E^{\parallel}$ and $E^{\perp}$ be as in (26). The linearity of the Fréchet derivative implies that

$$L_{\mathcal{P}}(A,E) = L_{\mathcal{P}}(A,E^{\parallel}) + L_{\mathcal{P}}(A,E^{\perp}).$$

The formula (28) and the identities $A = UH$ and $UU^*E^{\parallel} = E^{\parallel}$ then give

$$L_{\mathcal{P}}(A,E) = L_{\mathcal{P}}(UH, UU^*E^{\parallel}) + E^{\perp}H^{-1}.$$

Now note that the map $\mathcal{P}$ clearly satisfies $\mathcal{P}(VB) = V\mathcal{P}(B)$ for any nonsingular $B \in \mathbb{C}^{n\times n}$ and any $V \in \mathbb{C}^{m\times n}$ $(m \geq n)$ with orthonormal columns. From this it follows that for any such $V$ and $B$, and any $F \in \mathbb{C}^{n\times n}$,

$$L_{\mathcal{P}}(VB, VF) = VL_{\mathcal{P}}(B,F). \tag{31}$$

Applying this identity to the case in which $B = H$, $V = U$, and $F = U^*E^{\parallel}$, we obtain

$$\begin{aligned} L_{\mathcal{P}}(UH, UU^*E^{\parallel}) &= UL_{\mathcal{P}}(H, U^*E^{\parallel}) \\ &= UL_{\mathcal{P}}(H, U^*E), \end{aligned}$$

where the second line follows from the fact that $U^*E^{\perp} = 0$. Thus,

$$L_{\mathcal{P}}(A,E) = UL_{\mathcal{P}}(H, U^*E) + E^{\perp}H^{-1}.$$

Multiplying from the left by $U^*$ gives

$$U^* L_{\mathcal{P}}(A,E) = L_{\mathcal{P}}(H, U^*E)$$

since $U^*U = I$ and $U^*E^{\perp} = 0$. Equivalently, in terms of $\Omega = \operatorname{skew}(U^*E)$ and $S = \operatorname{sym}(U^*E)$,

$$U^* L_{\mathcal{P}}(A,E) = L_{\mathcal{P}}(H,\Omega) + L_{\mathcal{P}}(H,S).$$

The proof will be complete if we can show that $L_{\mathcal{P}}(H,\Omega)$ is skew-Hermitian and

$$L_{\mathcal{P}}(H,S) = 0. \tag{32}$$

In fact, (32) holds for any Hermitian matrix $S$ since, for all sufficiently small $\varepsilon$, $H + \varepsilon S$ is Hermitian positive definite, showing that $\mathcal{P}(H + \varepsilon S) = I$. The skew-Hermiticity of $L_{\mathcal{P}}(H,\Omega)$ follows from differentiating the identity

$$\mathcal{P}(H + \varepsilon\Omega)^* \mathcal{P}(H + \varepsilon\Omega) = I$$

with respect to $\varepsilon$ and using the fact that $\mathcal{P}(H) = I$. □

Another consequence of Lemma 3.1 is the following identity that relates the Fréchet derivative of the polar decomposition of a Hermitian positive definite matrix to the matrix sign function applied to a certain block matrix.

LEMMA 3.4. *Let $H \in \mathbb{R}^{n\times n}$ be Hermitian positive definite, and let $\Omega \in \mathbb{R}^{n\times n}$ be skew-Hermitian. Then*

$$\operatorname{sign}\begin{pmatrix} H & \Omega \\ 0 & -H \end{pmatrix} = \begin{pmatrix} I & L_{\mathcal{P}}(H,\Omega) \\ 0 & -I \end{pmatrix}, \tag{33}$$

$$\operatorname{sign}\begin{pmatrix} H & S \\ 0 & H \end{pmatrix} = \begin{pmatrix} I & L_{\mathcal{P}}(H,S) \\ 0 & I \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}. \tag{34}$$

*Proof.* By definition,

$$\begin{aligned}\operatorname{sign}\begin{pmatrix} H & \Omega \\ 0 & -H \end{pmatrix} &= \begin{pmatrix} H & \Omega \\ 0 & -H \end{pmatrix}\begin{pmatrix} H^2 & H\Omega - \Omega H \\ 0 & H^2 \end{pmatrix}^{-1/2} \\ &= \begin{pmatrix} H & \Omega \\ 0 & -H \end{pmatrix}\begin{pmatrix} H^2 & H\Omega + \Omega^* H \\ 0 & H^2 \end{pmatrix}^{-1/2}.\end{aligned}$$

Now apply (5) to the primary matrix function $f(X) = X^{-1/2}$ to obtain

$$\begin{pmatrix} H^2 & H\Omega + \Omega^* H \\ 0 & H^2 \end{pmatrix}^{-1/2} = \begin{pmatrix} H^{-1} & L_{x^{-1/2}}(H^2, H\Omega + \Omega^* H) \\ 0 & H^{-1} \end{pmatrix},$$

where the identity $(H^2)^{-1/2} = H^{-1}$ follows from the positive-definiteness of $H$. Thus,

$$\begin{aligned}\operatorname{sign}\begin{pmatrix} H & \Omega \\ 0 & -H \end{pmatrix} &= \begin{pmatrix} H & \Omega \\ 0 & -H \end{pmatrix}\begin{pmatrix} H^{-1} & L_{x^{-1/2}}(H^2, H\Omega + \Omega^* H) \\ 0 & H^{-1} \end{pmatrix} \\ &= \begin{pmatrix} I & HL_{x^{-1/2}}(H^2, H\Omega + \Omega^* H) + \Omega H^{-1} \\ 0 & -I \end{pmatrix}.\end{aligned}$$

The identity (33) follows upon observing that, by (24),

$$L_{\mathcal{P}}(H,\Omega) = \Omega H^{-1} + HL_{x^{-1/2}}(H^2, H\Omega + \Omega^* H).$$

The proof of (34) is simpler, since, by (5) and the identity $L_{\text{sign}}(H,S) = 0$,

$$\operatorname{sign}\begin{pmatrix} H & S \\ 0 & H \end{pmatrix} = \begin{pmatrix} I & L_{\text{sign}}(H,S) \\ 0 & I \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}.$$

□

We remark that an alternative proof of (33) exists. It is based on the observation that $L_{\mathcal{P}}(H,\Omega)$ is the solution of a Lyapunov equation which can be solved by reading off the $(1,2)$-block of $\operatorname{sign}\left(\begin{smallmatrix} H & \Omega \\ 0 & -H \end{smallmatrix}\right)$. For details, see section 5.

Combining Lemma 3.4 with Lemma 3.3 proves Theorem 2.3.

**3.2. Convergence of the iteration.** We now focus our efforts on proving convergence of the iteration (12)–(13), thereby proving Theorem 2.1. The cornerstone of the proof is Lemma 3.7, where a relationship is established between certain blocks of the matrices $Z_k$ defined by the matrix sign function iteration $Z_{k+1} = Z_k h(Z_k)^2$ and the matrices $X_k$ and $E_k$ defined by the iteration (12)–(13). Once this has been shown, convergence of the iteration (12)–(13) will follow from the convergence of $Z_k$ to $\operatorname{sign}(Z_0)$, together with the knowledge (from Theorem 2.3) that the Fréchet derivative of the polar decomposition is related to the $(1,2)$-block of $\operatorname{sign}(Z_0)$ for certain values of $Z_0$.

We begin by examining the block structure of the iterates $Z_k$.

Lemma 3.5. *The iterates $Z_k$ produced by the iteration $Z_{k+1} = Z_k h(Z_k^2)$ with initial condition* (10) *have the form*

$$Z_k = \begin{pmatrix} H_k & \Omega_k \\ 0 & -H_k \end{pmatrix},$$

*where $H_k$ is Hermitian and $\Omega_k$ is skew-Hermitian.*

*Proof.* Assume the statement is true at iteration $k$. Then by (5),

$$\begin{aligned} Z_{k+1} &= \begin{pmatrix} H_k & \Omega_k \\ 0 & -H_k \end{pmatrix} h \begin{pmatrix} H_k^2 & H_k\Omega_k - \Omega_k H_k \\ 0 & H_k^2 \end{pmatrix} \\ &= \begin{pmatrix} H_k & \Omega_k \\ 0 & -H_k \end{pmatrix} \begin{pmatrix} h(H_k^2) & L_h(H_k^2, H_k\Omega_k - \Omega_k H_k) \\ 0 & h(H_k^2) \end{pmatrix} \\ &= \begin{pmatrix} H_k h(H_k^2) & H_k L_h(H_k^2, H_k\Omega_k - \Omega_k H_k) + \Omega_k h(H_k^2) \\ 0 & -H_k h(H_k^2) \end{pmatrix}. \end{aligned} \tag{35}$$

By the remark following Theorem 2.1, $H_k h(H_k^2) = \left[H_k h(H_k^2)\right]^*$, showing that $H_{k+1} = H_k h(H_k^2)$ is Hermitian. On the other hand, the fact that $h$ is a primary matrix function implies that $Z_k$ commutes with $h(Z_k^2)$, so, by a calculation similar to that above, we also have

$$Z_{k+1} = \begin{pmatrix} h(H_k^2)H_k & h(H_k^2)\Omega_k - L_h(H_k^2, H_k\Omega_k - \Omega_k H_k)H_k \\ 0 & -h(H_k^2)H_k \end{pmatrix}. \tag{36}$$

Denote $C_k = H_k\Omega_k - \Omega_k H_k$. Since $H_k$ is Hermitian and $\Omega_k$ is skew-Hermitian, $C_k$ is Hermitian. Hence, since $h(Z^*) = h(Z)^*$ for every $Z$,

$$\begin{aligned} L_h(H_k^2, C_k)^* &= L_h((H_k^2)^*, C_k^*) \\ &= L_h(H_k^2, C_k). \end{aligned}$$

Comparing the $(1, 2)$ blocks of (35) and (36) then shows that

$$\begin{aligned} 0 &= H_k L_h(H_k^2, C_k) + \Omega_k h(H_k^2) - h(H_k^2)\Omega_k + L_h(H_k^2, C_k)H_k \\ &= H_k L_h(H_k^2, C_k) + \Omega_k h(H_k^2) + h(H_k^2)^*\Omega_k^* + L_h(H_k^2, C_k)^* H_k^* \\ &= \Omega_{k+1} + \Omega_{k+1}^*. \end{aligned}$$

It follows that $\Omega_k = -\Omega_k^*$ for every $k$. □

The proof above also reveals a recursion satisfied by $H_k$ and $\Omega_k$, namely,

$$H_{k+1} = H_k h(H_k^2), \tag{37}$$

$$\Omega_{k+1} = \Omega_k h(H_k^2) + H_k L_h(H_k^2, H_k\Omega_k - \Omega_k H_k). \tag{38}$$

Next, we examine the block structure of the iterates $Z_k$ with initial condition (11).

Lemma 3.6. *The iterates $Z_k$ produced by the iteration $Z_{k+1} = Z_k h(Z_k^2)$ with initial condition* (11) *have the form*

$$Z_k = \begin{pmatrix} H_k & S_k \\ 0 & H_k \end{pmatrix},$$

*where $H_k$ is the same Hermitian matrix as in Lemma* 3.5 *and $S_k$ is Hermitian.*

*Proof.* We omit the proof, which is very similar to the proof of Lemma 3.5. ∎

In analogy with (38), the iterates $S_k$ satisfy the recursion

$$S_{k+1} = S_k h(H_k^2) + H_k L_h(H_k^2, S_k H_k + H_k S_k). \tag{39}$$

We now relate the matrices $H_k$, $\Omega_k$, and $S_k$ defined in the preceding pair of lemmas to the matrices $X_k$ and $E_k$ defined by the coupled iteration (12)–(13).

LEMMA 3.7. *The iterates $H_k$, $\Omega_k$, and $S_k$ are related to $X_k$ and $E_k$ via*

$$UH_k = X_k, \tag{40}$$

$$\Omega_k = \operatorname{skew}(U^* E_k), \tag{41}$$

$$S_k = \operatorname{sym}(U^* E_k). \tag{42}$$

*Proof.* The first of these equalities follows easily by induction, for if it holds at iteration $k$, then

$$\begin{aligned} X_{k+1} &= g(X_k) \\ &= X_k h(X_k^* X_k) \\ &= UH_k h(H_k^* U^* U H_k) \\ &= UH_k h(H_k^2) \\ &= UH_{k+1}. \end{aligned}$$

Furthermore, $X_0 = A = UH = UH_0$, which proves (40). To prove (41) and (42), we will show that if $\Omega_k = \operatorname{skew}(U^*E_k)$ and $S_k = \operatorname{sym}(U^*E_k)$ for a given $k$, and if $E_{k+1}$, $\Omega_{k+1}$, and $S_{k+1}$ are given by (13), (38), and (39), respectively, then $\Omega_{k+1} = \operatorname{skew}(U^*E_{k+1})$ and $S_{k+1} = \operatorname{sym}(U^*E_{k+1})$. Recalling (24), we have

$$\begin{aligned} U^*E_{k+1} &= U^* L_g(X_k, E_k) \\ &= U^* E_k h(X_k^* X_k) + U^* X_k L_h(X_k^* X_k, E_k^* X_k + X_k^* E_k) \\ &= U^* E_k h(H_k^2) + H_k L_h(H_k^2, E_k^* U H_k + H_k U^* E_k) \\ &= \Omega_k h(H_k^2) + H_k L_h(H_k^2, \Omega_k^* H_k + H_k \Omega_k) + S_k h(H_k^2) + H_k L_h(H_k^2, S_k H_k + H_k S_k) \\ &= \Omega_{k+1} + S_{k+1}, \end{aligned}$$

where we have used (38), (39), and the decomposition $U^*E_k = \Omega_k + S_k$. By Lemmas 3.5 and 3.6, $\Omega_{k+1}$ is skew-Hermitian and $S_{k+1}$ is Hermitian, proving (41) and (42). ∎

The proof of Theorem 2.1 is now almost complete, since by Lemma 3.5 and Theorem 2.3,

$$\begin{pmatrix} H_k & \Omega_k \\ 0 & -H_k \end{pmatrix} \to \operatorname{sign}\begin{pmatrix} H & \Omega \\ 0 & -H \end{pmatrix} = \begin{pmatrix} I & U^* L_{\mathcal{P}}(A, E) \\ 0 & -I \end{pmatrix}$$

as $k \to \infty$. Likewise, by (34) and Lemma 3.6,

$$\begin{pmatrix} H_k & S_k \\ 0 & H_k \end{pmatrix} \to \operatorname{sign}\begin{pmatrix} H & S \\ 0 & H \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix},$$

as $k \to \infty$. These observations, together with (40)–(42), show that

$$\begin{aligned} X_k &\to U, \\ \operatorname{skew}(U^*E_k) &\to U^* L_{\mathcal{P}}(A, E), \\ \operatorname{sym}(U^*E_k) &\to 0 \end{aligned}$$

as $k \to \infty$. In other words,

$$X_k \to \mathcal{P}(A), \tag{43}$$

$$U^* E_k \to U^* L_{\mathcal{P}}(A, E) \tag{44}$$

as $k \to \infty$. The latter limit implies that $E_k \to L_{\mathcal{P}}(A, E)$ when $U$ is square, but not when $U$ is rectangular. To handle the rectangular case, consider the decompositions (26) and

$$E_k = E_k^{\|} + E_k^{\perp}, \quad E_k^{\|} = UU^* E_k, \quad E_k^{\perp} = (I - UU^*)E_k.$$

By Lemma 3.2 and the linearity of the Fréchet derivative, the statement (44) is equivalent to the statement that

$$\begin{aligned} U^* E_k^{\|} &\to U^* L_{\mathcal{P}}(A, E^{\|}) + U^* L_{\mathcal{P}}(A, E^{\perp}) \\ &= U^* L_{\mathcal{P}}(A, E^{\|}). \end{aligned}$$

Multiplying from the left by $U$ and recalling that $UU^* E_k^{\|} = E_k^{\|}$ and $UU^* L_{\mathcal{P}}(A, E^{\|}) = L_{\mathcal{P}}(A, E^{\|})$ (by (27)), we conclude that

$$E_k^{\|} \to L_{\mathcal{P}}(A, E^{\|}).$$

The proof of Theorem 2.1 will be complete if we can show that

$$E_k^{\perp} \to L_{\mathcal{P}}(A, E^{\perp}).$$

This is carried out in the following lemma.

Lemma 3.8. *As $k \to \infty$, $E_k^{\perp} \to L_{\mathcal{P}}(A, E^{\perp})$.*

*Proof.* By (28), it suffices to show that

$$E_k^{\perp} \to E^{\perp} H^{-1}.$$

Using Lemma 3.1, it is straightforward to see that $E_k^{\|}$ and $E_k^{\perp}$ satisfy independent recursions of the form

$$\begin{aligned} E_{k+1}^{\|} &= L_g(X_k, E_k^{\|}), \\ E_{k+1}^{\perp} &= L_g(X_k, E_k^{\perp}). \end{aligned}$$

Now since $\mathcal{R}(E_k^{\perp})$ is orthogonal to $\mathcal{R}(U) \supseteq \mathcal{R}(UH_k) = \mathcal{R}(X_k)$, it follows from (25) that

$$L_g(X_k, E_k^{\perp}) = E_k^{\perp} h(X_k^* X_k),$$

so

$$E_{k+1}^{\perp} = E_k^{\perp} h(X_k^* X_k).$$

If we introduce the matrix $B_k \in \mathbb{C}^{n \times n}$ defined by the recursion

$$B_{k+1} = B_k h(X_k^* X_k), \quad B_0 = I,$$

then an inductive argument shows that

$$E_k^{\perp} = E^{\perp} B_k.$$

We claim that $B_k \to H^{-1}$ as $k \to \infty$. To see this, observe that (12) implies that

$$X_k = X_0 B_k = A B_k.$$

Since $X_k \to U$ as $k \to \infty$, we conclude that

$$I = U^* U = U^* \lim_{k\to\infty} X_k = U^* A \lim_{k\to\infty} B_k = H \lim_{k\to\infty} B_k.$$

It follows that $E_k^\perp = E^\perp B_k \to E^\perp H^{-1}$ as $k \to \infty$. ∎

## 4. Practical considerations.

This section discusses several practical considerations concerning the iterative schemes detailed in Theorem 2.1.

### 4.1. Scaling.

Scaling the iterates $X_k$ in the Newton iteration (17) often reduces the number of iterations required to achieve convergence [18, Chapter 8.6]. If this strategy is generalized to the coupled iteration (17)–(18), then the resulting iteration reads

$$X_{k+1} = \frac{1}{2}(\mu_k X_k + \mu_k^{-1} X_k^{-*}), \qquad X_0 = A, \tag{45}$$

$$E_{k+1} = \frac{1}{2}(\mu_k E_k - \mu_k^{-1} X_k^{-*} E_k^* X_k^{-*}), \quad E_0 = E, \tag{46}$$

where $\mu_k > 0$ is a scaling factor chosen heuristically. Practical choices for $\mu_k$ include [18]

$$\mu_k = \left( \frac{\|X_k^{-1}\|_1 \|X_k^{-1}\|_\infty}{\|X_k\|_1 \|X_k\|_\infty} \right)^{1/4} \tag{47}$$

and

$$\mu_k = \left( \frac{\|X_k^{-1}\|_F}{\|X_k\|_F} \right)^{1/2}, \tag{48}$$

where $\|\cdot\|_1$, $\|\cdot\|_\infty$, and $\|\cdot\|_F$ denote the matrix 1-, ∞- and Frobenius norms, respectively.

More generally, scaling can be applied to other iterative schemes of the form (12)–(13), leading to iterative schemes of the form

$$\begin{aligned} X_{k+1} &= g(\mu_k X_k), & X_0 &= A, \\ E_{k+1} &= L_g(\mu_k X_k, \mu_k E_k), & E_0 &= E. \end{aligned}$$

Note that if $A$ is rectangular, then (47) and (48) are inapplicable. We have found

$$\mu_k = \left( \frac{\|(X_k^* X_k)^{-1}\|_1 \|(X_k^* X_k)^{-1}\|_\infty}{\|X_k^* X_k\|_1 \|X_k^* X_k\|_\infty} \right)^{1/8} \tag{49}$$

and

$$\mu_k = \left( \frac{\|(X_k^* X_k)^{-1}\|_F}{\|X_k^* X_k\|_F} \right)^{1/4} \tag{50}$$

to be effective alternatives to (47) and (48) in our numerical experiments with rectangular $A$.

**4.2. Termination criteria.** Determining when to terminate the iteration (12)–(13) is a delicate task. Termination criteria for (12) by itself are, of course, well-studied, but the accuracy of $E_k$ should be taken into account when choosing termination criteria for the coupled iteration (12)–(13).

One possibility is to appeal to the relationship between $X_k$ and $E_k$ and the sign function iterates $Z_k$ referenced in the statement of Theorem 2.1. Convergence of the sign function iterates to $\operatorname{sign}(Z_0) = \lim_{k\to\infty} Z_k$ can be readily verified with the aid of the inequality

$$\frac{\|Z_k^2 - I\|}{\|\operatorname{sign}(Z_0)\|(\|Z_k\| + \|\operatorname{sign}(Z_0)\|)} \le \frac{\|Z_k - \operatorname{sign}(Z_0)\|}{\|\operatorname{sign}(Z_0)\|} \le \|Z_k^2 - I\|,$$

which holds in any submultiplicative matrix norm, so long as $\|\operatorname{sign}(Z_0)(Z_k - \operatorname{sign}(Z_0))\| < 1$ and $Z_0$ has no pure imaginary eigenvalues [18, Lemma 5.12]. In other words, $\|Z_k^2 - I\|$ provides an estimate for the accuracy of $Z_k$.

For the iterates $Z_k$ with initial condition (10), we have, in the notation of Lemma 3.5,

$$Z_k^2 - I = \begin{pmatrix} H_k^2 - I & H_k\Omega_k - \Omega_k H_k \\ 0 & H_k^2 - I \end{pmatrix}.$$

Likewise, for the iterates $Z_k$ with initial condition (11), we have, in the notation of Lemma 3.6,

$$Z_k^2 - I = \begin{pmatrix} H_k^2 - I & H_k S_k + S_k H_k \\ 0 & H_k^2 - I \end{pmatrix}.$$

Thus, accuracy is ensured when the quantities $\|H_k^2 - I\|$, $\|H_k\Omega_k - \Omega_k H_k\|$, and $\|H_k S_k + S_k H_k\|$ are small. Of course, $H_k$, $\Omega_k$, and $S_k$ are never computed explicitly in the iteration (12)–(13), so we must relate these quantities to $X_k$ and $E_k$ using Lemma 3.7. By (40), we have

$$H_k^2 - I = X_k^* X_k - I.$$

The quantities $H_k\Omega_k - \Omega_k H_k$ and $H_k S_k + S_k H_k$ are more difficult to relate to $X_k$ and $E_k$ in a computable way (i.e., a way that does not involve knowing $U$ in advance). However, second-order accurate approximations to $H_k\Omega_k - \Omega_k H_k$ and $H_k S_k + S_k H_k$ are available. As shown in Appendix A, we have

$$H_k\Omega_k - \Omega_k H_k = \frac{1}{2}\left(X_k^* X_k X_k^* E_k - X_k^* E_k X_k^* X_k\right) + F_k, \tag{51}$$

$$H_k S_k - S_k H_k = X_k^* E_k + E_k^* X_k - \frac{1}{2}\left(X_k^* X_k X_k^* E_k - X_k^* E_k X_k^* X_k\right) - F_k, \tag{52}$$

where

$$\|F_k\| = O\left(\|H_k^2 - I\|^2 + \|H_k^2 - I\|\|H_k S_k + S_k H_k\|\right).$$

Roughly speaking, (51) arises from the approximations $H_k \approx \frac{1}{2}(I + X_k^* X_k)$ and $\Omega_k \approx X_k^* E_k$. It turns out that only the first of these approximations is second-order accurate (see Lemma A.3), but delicate cancellations detailed in Appendix A lead to the validity of (51). One then deduces (52) by noting that $X_k^* E_k + E_k^* X_k = (H_k\Omega_k - \Omega_k H_k) + (H_k S_k - S_k H_k)$ (see Lemma A.1).

In summary, the quantities

$$\mathcal{A}_k = X_k^* X_k - I, \tag{53}$$

$$\mathcal{B}_k = \frac{1}{2}\left(X_k^* X_k X_k^* E_k - X_k^* E_k X_k^* X_k\right), \tag{54}$$

$$\mathcal{C}_k = X_k^* E_k + E_k^* X_k - \mathcal{B}_k \tag{55}$$

are computable approximations to $H_k^2 - I$, $H_k\Omega_k - \Omega_k H_k$, and $H_k S_k + S_k H_k$, respectively. These are small in norm if and only if $\|Z_k - \mathrm{sign}(Z_0)\|$ is small (for each of the initial conditions (10) and (11)), which is true if and only if $\|X_k - U\|$ and $\|E_k - L_{\mathcal{P}}(A,E)\|$ are small. As a practical note, these arguments appear to break down if $A$ is very ill-conditioned, as illustrated in section 6.

Based on these considerations, we propose that the iterations be terminated when

$$\|\mathcal{A}_k\| \le \delta\|X_k\| \quad \text{and} \quad \|\mathcal{B}_k\| + \|\mathcal{C}_k\| \le \varepsilon\|E_k\|, \tag{56}$$

where $\delta$ and $\varepsilon$ are relative error tolerances for $\|X_k - U\|$ and $\|E_k - L_{\mathcal{P}}(A,E)\|$, respectively.

As an alternative approach to terminating the iterations, one could consider basing the decision to terminate on the smallness of the step lengths $\|X_{k+1} - X_k\|$ and $\|E_{k+1} - E_k\|$. Details of this approach, for the case in which $E_k$ is absent, can be found in [18, Chapter 8.7].

**4.3. Stability.** Stability of the iterative schemes detailed in Theorem 2.1 is relatively easy to establish. Indeed, the map

$$\mathcal{F}\begin{pmatrix} A \\ E \end{pmatrix} = \begin{pmatrix} \mathcal{P}(A) \\ L_{\mathcal{P}}(A,E) \end{pmatrix} \tag{57}$$

is idempotent, since $\mathcal{P}(\mathcal{P}(A)) = \mathcal{P}(A)$ and $L_{\mathcal{P}}(\mathcal{P}(A), L_{\mathcal{P}}(A,E)) = L_{\mathcal{P}}(A,E)$ by the chain rule. It follows that any superlinearly convergent iteration for computing $\left(\begin{smallmatrix} \mathcal{P}(A) \\ L_{\mathcal{P}}(A,E) \end{smallmatrix}\right)$ is automatically stable [18, Therorem 4.19]. More precisely, if

$$\begin{pmatrix} X_{k+1} \\ E_{k+1} \end{pmatrix} = \begin{pmatrix} g(X_k) \\ L_g(X_k, E_k) \end{pmatrix} \tag{58}$$

converges superlinearly to $\left(\begin{smallmatrix} \mathcal{P}(X_0) \\ L_{\mathcal{P}}(X_0,E_0) \end{smallmatrix}\right)$ for all $X_0$ and $E_0$ sufficiently close to $A$ and $E$, respectively, then the iteration is stable in the sense of [18, Definition 4.17]. Moreover, the Fréchet derivative of the map (57) coincides with the Fréchet derivative of the map (58) at the fixed point $\left(\begin{smallmatrix} \mathcal{P}(A) \\ L_{\mathcal{P}}(A,E) \end{smallmatrix}\right)$ [18, Therorem 4.19].

As an example, the Newton iteration (17)–(18) is superlinearly convergent by virtue of the superlinear (indeed, quadratic) convergence of the corresponding matrix sign function iteration (16). The Newton–Schulz iteration (22)–(23) is likewise superlinearly (indeed, quadratically) convergent, provided that the singular values of $A$ lie in the interval $(0, \sqrt{2})$. Thus, both iterations are stable. Using, for instance, (22)–(23), we find that the Fréchet derivative of the map (58) (and hence of the map (57)) at $\left(\begin{smallmatrix} U \\ K \end{smallmatrix}\right) = \left(\begin{smallmatrix} \mathcal{P}(A) \\ L_{\mathcal{P}}(A,E) \end{smallmatrix}\right)$ is given by

$$L_{\mathcal{F}}\Big(\begin{pmatrix} U \\ K \end{pmatrix}, \begin{pmatrix} F \\ G \end{pmatrix}\Big) = \begin{pmatrix} F - \frac{1}{2}U(U^*F + F^*U) \\ G - \frac{1}{2}\left[U(U^*G + G^*U + K^*F + F^*K) + K(U^*F + F^*U)\right] \end{pmatrix}.$$

Note that when $U$ is square, the identities $UU^* = I$ and $U^*K = -K^*U$ (by (30)) imply that this formula reduces to

$$L_{\mathcal{F}}\left(\begin{pmatrix} U \\ K \end{pmatrix}, \begin{pmatrix} F \\ G \end{pmatrix}\right) = \begin{pmatrix} \frac{1}{2}(F - UF^*U) \\ \frac{1}{2}(G - UG^*U - UF^*K - KF^*U) \end{pmatrix},$$

in agreement with [18, Theorem 8.19].

**4.4. Condition number estimation.** A seemingly natural application of Theorem 2.1 is to leverage the iterative scheme (12)–(13) to estimate the (absolute) condition number

$$\kappa(\mathcal{P}, A) = \|L_{\mathcal{P}}(A, \cdot)\| = \sup_{\substack{E \in \mathbb{C}^{m\times n}, \\ E \neq 0}} \frac{\|L_{\mathcal{P}}(A, E)\|}{\|E\|}$$

of the map $\mathcal{P}$ at $A$ (or its counterpart $\kappa_{rel}(\mathcal{P}, A) = \kappa(\mathcal{P}, A)\|A\|/\|\mathcal{P}(A)\|$, the relative condition number of $\mathcal{P}$ at $A$). As tempting as it may seem, a much simpler (and undoubtedly more efficient) algorithm is available for estimating $\kappa(\mathcal{P}, A)$. As explained in [18, Theorem 8.9], the value of $\kappa(\mathcal{P}, A)$ at $A \in \mathbb{C}^{m\times n}$ $(m \geq n)$ is $\sigma_n^{-1}$, where $\sigma_n$ denotes the smallest singular value of $A$. This quantity can be estimated efficiently by applying the power method [18, Algorithm 3.19] to $(A^*A)^{-1}$. In most iterative algorithms for computing the polar decomposition, this matrix (or $A^{-1}$) is computed in the first iteration, so the additional cost of computing $\kappa(\mathcal{P}, A)$ is negligible.

Before finishing our discussion of condition number estimation, it is worth pointing out a subtlety that arises when considering the polar decomposition of a real square matrix. If $A$ is real and square $(m = n)$, then it can be shown that the (absolute) condition number of $\mathcal{P}$ at $A$ with respect to *real* perturbations is $2(\sigma_n + \sigma_{n-1})^{-1}$ [18, Theorem 8.9]. This fact will play a role in our interpretation of certain numerical experiments in section 6.

## 5. Comparison with other methods.

There are several other methods that can be used to compute the Fréchet derivative of the polar decomposition. Below, we describe a few and compare them with iterative schemes of the form (12)–(13).

One alternative is to recognize that $L_{\mathcal{P}}(A, E)$ is the solution to a Lyapunov equation. Indeed, upon noting that $\mathcal{P}(A)^*A = U^*A = H$ is Hermitian, one can differentiate the relation

$$\operatorname{skew}(\mathcal{P}(A)^*A) = 0$$

with the aid of the product rule to obtain

$$\operatorname{skew}(L_{\mathcal{P}}(A, E)^*A + \mathcal{P}(A)^*E) = 0$$

for any $E \in \mathbb{C}^{m\times n}$. Substituting $A = UH$ and $\mathcal{P}(A) = U$, and denoting $Y := U^*L_{\mathcal{P}}(A, E) = -L_{\mathcal{P}}(A, E)^*U$, we obtain

$$HY + YH = U^*E - E^*U. \tag{59}$$

Given $H$, $U$, and $E$, this is a Lyapunov equation in the unknown $Y$, which, by the positive-definiteness of $H$, has a unique solution. It can be solved using standard algorithms for the solution of Lyapunov and Sylvester equations [5, 13]. It also has theoretical utility, offering an alternative proof of part of Theorem 2.3, owing to a well-known connection between the solution of Lyapunov and Sylvester equations and the matrix sign function [31], [18, Chapter 2.4]. Indeed, (59) is equivalent to the equation

$$\begin{pmatrix} H & E^*U - U^*E \\ 0 & -H \end{pmatrix} = \begin{pmatrix} I & Y \\ 0 & I \end{pmatrix} \begin{pmatrix} H & 0 \\ 0 & -H \end{pmatrix} \begin{pmatrix} I & Y \\ 0 & I \end{pmatrix}^{-1}.$$

Taking the sign of both sides, noting that $\operatorname{sign}(H) = I$, and using the fact that the matrix sign function commutes with similarity transformations, we conclude that

$$\begin{aligned}\operatorname{sign}\begin{pmatrix} H & E^*U - U^*E \\ 0 & -H \end{pmatrix} &= \begin{pmatrix} I & Y \\ 0 & I \end{pmatrix} \left[\operatorname{sign}\begin{pmatrix} H & 0 \\ 0 & -H \end{pmatrix}\right] \begin{pmatrix} I & Y \\ 0 & I \end{pmatrix}^{-1} \\ &= \begin{pmatrix} I & -2Y \\ 0 & -I \end{pmatrix}.\end{aligned}$$

This is precisely the identity (14), up to a rescaling of $E$. Its connection with the Lyapunov equation (59) reveals that the coupled iteration (12)–(13) is effectively solving (59) and computing the polar decomposition simultaneously. In comparison to a naive approach in which (59) is solved after first computing the polar decomposition, the coupled iteration (12)–(13) is attractive, as it computes $L_{\mathcal{P}}(A, E)$ at the expense of a few extra matrix-matrix multiplications and additions on top of the computation of $\mathcal{P}(A)$.

When $A$ and $E$ are real, another method for computing the Fréchet derivative of a matrix function $f$ is to use the complex step approximation [2]

$$L_f(A, E) \approx \operatorname{Im}\left(\frac{f(A + ihE) - f(A)}{h}\right), \tag{60}$$

where $h$ is a small positive scalar and $\operatorname{Im}(B)$ denotes the imaginary part of a matrix $B$. By using a pure imaginary step $ih$, this approximation does not suffer from cancellation errors that plague standard finite differencing, allowing $h$ to be taken arbitrarily small [2]. This approximation can be applied to the polar decomposition, but care must be exercised in order to do so correctly. In particular, a meaningful approximation is obtained only if the conjugate transposes $X_k^*$ appearing in the algorithm are interpreted as transposes $X_k^T$ when evaluating the "polar decomposition" of $A + ihE$. We have put "polar decomposition" in quotes since the result of such a computation is the matrix $(A + ihE)\left[(A + ihE)^T (A + ihE)\right]^{-1/2}$, not $\mathcal{P}(A + ihE) = (A + ihE)\left[(A + ihE)^*(A + ihE)\right]^{-1/2}$. The cost of this approximation is close to the cost of computing two polar decompositions.

Another approach is to appeal to the relation $\mathcal{P}(A) = A(A^*A)^{-1/2}$. By (24), the Fréchet derivative of $\mathcal{P}$ at $A$ in the direction $E$ is given by

$$\begin{aligned} L_{\mathcal{P}}(A, E) &= E(A^*A)^{-1/2} + AL_{x^{-1/2}}(A^*A, E^*A + A^*E) \\ &= EH^{-1} + AL_{x^{-1/2}}(A^*A, E^*A + A^*E). \end{aligned}$$

Evaluating the second term, the Fréchet derivative of the inverse square root, can be reduced to the task of solving a Lyapunov equation, so this approach is essentially of the same complexity as the one based on (59).

Any of the aforementioned methods, including our own, can be applied in two different ways when $A$ is rectangular ($m \times n$ with $m > n$). One way is to apply the methods verbatim, working at all times with rectangular matrices. The alternative is to first compute a reduced $QR$ decomposition $A = QR$, where $Q \in \mathbb{C}^{m\times n}$ has orthonormal columns and $R \in \mathbb{C}^{n\times n}$ is upper triangular. Then, one can compute $\mathcal{P}(R)$ and $L_{\mathcal{P}}(R, Q^*E)$ (which are square matrices) and invoke the identities

$$U = \mathcal{P}(A) = Q\mathcal{P}(R), \quad H = \mathcal{P}(R)^*R$$

and

$$\begin{aligned} L_{\mathcal{P}}(A, E) &= L_{\mathcal{P}}(A, QQ^*E) + L_{\mathcal{P}}(A, (I - QQ^*)E) \\ &= QL_{\mathcal{P}}(R, Q^*E) + (I - QQ^*)EH^{-1} \end{aligned}$$

to recover $\mathcal{P}(A)$ and $L_{\mathcal{P}}(A, E)$. The validity of the latter identity is a consequence of (31), (28), and the fact that $QQ^* = UU^*$. In summary, computations for rectangular $A$ can be reduced to the square case by performing a reduced $QR$ decomposition of $A$ at the outset.

Finally, when $A$ is square, one more method for computing $L_{\mathcal{P}}(A, E)$ is available, as noted in, for instance, [21]. The idea is to make use of the SVD $A = P\Sigma Q^*$, where $P, Q \in \mathbb{C}^{n\times n}$ are unitary and $\Sigma \in \mathbb{C}^{n\times n}$ is diagonal. The SVD is related to the polar decomposition $A = UH$ via the relations $U = PQ^*$ and $H = Q\Sigma Q^*$. Moreover, the Lyapunov equation (59) is equivalent to

$$\Sigma G + G\Sigma = F - F^*,$$

where $F = P^*EQ$ and $G = P^*L_{\mathcal{P}}(A, E)Q$ [21, equation 2.18]. Given $\Sigma$ and $F$, this equation admits an explicit solution for the components of $G$. Namely,

$$G_{ij} = \frac{1}{\sigma_i + \sigma_j}(F_{ij} - \overline{F_{ji}}),$$

where $\sigma_i$ denotes the $i$th diagonal entry of $\Sigma$, and $\overline{F_{ji}}$ denotes the complex conjugate of $F_{ji}$. One then obtains $L_{\mathcal{P}}(A, E)$ from $L_{\mathcal{P}}(A, E) = PGQ^*$. This method is attractive if the SVD of $A$ has already been computed, but otherwise it is an expensive approach in general.

**5.1. Floating point operations.** Relative to the methods listed above, the iterative schemes derived in this paper are distinguished by their efficiency, at least when $n$ is large and the columns of $A$ are close to being orthonormal. To see this, consider the number of floating point operations needed to compute $\mathcal{P}(A)$ and $L_{\mathcal{P}}(A, E)$. For simplicity, assume that $A$ and $E$ are real and of size $n \times n$. Then, to leading order in $n$, and excluding the costs associated with termination criteria in the iterative schemes, the methods have the following computational costs:

- The iteration (17)–(18) requires $n_{iter}$ matrix inversions (each requiring $2n^3$ flops [18, Appendix C]) and $2n_{iter}$ matrix multiplications (each requiring $2n^3$ flops), where $n_{iter}$ denotes the number of iterations used. Its computational cost is thus $n_{iter}(2n^3) + 2n_{iter}(2n^3) = 6n_{iter}n^3$ flops.
- Solving the Lyapunov equation (59) with a direct method involves diagonalizing $H$ ($9n^3$ flops [18, Appendix C]) and performing four matrix multiplications, for a total of $9n^3 + 4(2n^3) = 17n^3$ flops. The additional cost of computing $U$, $H = U^*A$, $L_{\mathcal{P}}(A, E) = UY$, and $U^*E$ (assuming that (17) is used to compute $U$) is dominated by the cost of performing $n_{iter}$ matrix inversions and three matrix multiplications, bringing the total to $17n^3 + n_{iter}(2n^3) + 3(2n^3) = (23 + 2n_{iter})n^3$ flops.
- The complex step approximation (assuming that (17) is used to compute the polar decomposition of $A$ and $A + ihE$) requires $2n_{iter}$ matrix inversions, of which $n_{iter}$ involve complex arithmetic. Since each inversion of a complex matrix requires $n^3$ additions of complex scalars (2 real flops) and $n^3$ multiplications of complex scalars (6 real flops), the computational cost of the complex step approximation is $n_{iter}(2n^3) + n_{iter}(8n^3) = 10n_{iter}n^3$ flops.
- The method based on the SVD requires five matrix multiplications plus the computation of the SVD. Assuming, for instance, that the Golub–Reinsch algoirthm ($22n^3$ flops [14]) is used to compute the SVD, this method's total cost is $5(2n^3) + 22n^3 = 32n^3$ flops.

We conclude from this analysis that, for sufficienty large $n$, the iteration (17)–(18) requires fewer floating point operations than its competitors whenever $n_{iter} \leq 5$. Note that this is no longer the case if the costs of computing the residual estimates (53)–(55) are taken into account. However, if efficiency is the primary objective, then cheaper termination criteria (based, for instance, on $\|X_k - X_k^{-*}\|$, $\|X_{k+1} - X_k\|$, and/or $\|E_{k+1} - E_k\|$) may be appropriate.

It should be noted that these floating point operation counts are, of course, crude measures of efficiency that do not account for other factors—parallelizability and numerical stability, for instance—which may very well render the iterative methods in this paper even more attractive (for large matrices with nearly orthonormal columns). Parallelizability is particularly noteworthy since these iterative methods require only multiplication, inversion, and addition of matrices. In contrast, methods based on the SVD or the solution of the Lyapunov equation (59) involve matrix decompositions (unless the Lyapunov equation (59) is solved iteratively, a strategy which we have already argued to be inferior to the simultaneous computation of $\mathcal{P}(A)$ and $L_{\mathcal{P}}(A, E)$ via (17)–(18)). These considerations suggest that for large matrices with nearly orthonormal columns, the iterative methods in this paper are likely better suited for parallel computing environments than their competitors.

We emphasize that these comparisons are relevant only if the goal is to calculate $L_{\mathcal{P}}(A, E)$ for specific $A$ and $E$. If condition number estimation is the ultimate goal, then calculation of the Fréchet derivative of $\mathcal{P}$ is unnecessary, as explained in section 4.4. We refer the reader to [12] for an example of an application in which is desirable to calculate $L_{\mathcal{P}}(A, E)$ and not merely $\kappa(\mathcal{P}, A)$.

## 6. Numerical experiments.

To illustrate the performance of the iterative schemes derived in this paper, we have computed the Fréchet derivative of the polar decomposition for a collection of 69 matrices: 44 real matrices of size $10 \times 10$ from the Matrix Computation Toolbox [15] (we used all of the test matrices in the toolbox except those that are singular to working precision), as well as 25 complex matrices of size $10 \times n$ generated by the MATLAB command `crandn(1)*gallery('randsvd',[10 n], kappa,mode)`, where $n \in \{2, 4, 6, 8, 10\}$, `kappa=10^(8*rand(1))`, and `mode` $\in \{1, 2, 3, 4, 5\}$.

We computed $\mathcal{P}(A)$ and $L_{\mathcal{P}}(A, E)$ for each matrix $A$ described above, with $E$ a matrix (of the same dimensions as $A$) consisting of random entries sampled from a normal distribution with mean 0 and variance 1. We used the Newton iteration (45)–(46) with scaling parameter (47) for the square matrices and its generalization (19)–(20) with scaling parameter (49) for the rectangular matrices. To terminate the iterations, we used (56) with $\delta = \varepsilon = 10^{-14}$ and $\|\cdot\|$ equal to the Frobenius norm. Note that for simplicity, we used scaling throughout the entire iteration, even though the scaling parameter $\mu_k$ approaches 1 near convergence. A more efficient approach is to switch to an unscaled iteration after a certain point. A heuristic for deciding when to do so is detailed in [18, Chapter 8.9].

We compared the iterative methods with three alternatives described in section 5: solving the Lyapunov equation (59), using the complex step approximation (60) (when applicable), and using the SVD. We also computed the "exact" values of $\mathcal{P}(A)$ and $L_{\mathcal{P}}(A, E)$ using the SVD with 100-digit precision using the Symbolic Math Toolbox of MATLAB. Figure 1(a) shows the relative errors (in the Frobenius norm) in the computed values of $L_{\mathcal{P}}(A, E)$ for each of the 69 tests, arranged in order of decreasing condition number $\kappa_{rel}(L_{\mathcal{P}}, (A, E))$—the relative condition number of the map $(A, E) \mapsto L_{\mathcal{P}}(A, E)$. We estimated the latter quantity using finite differencing to

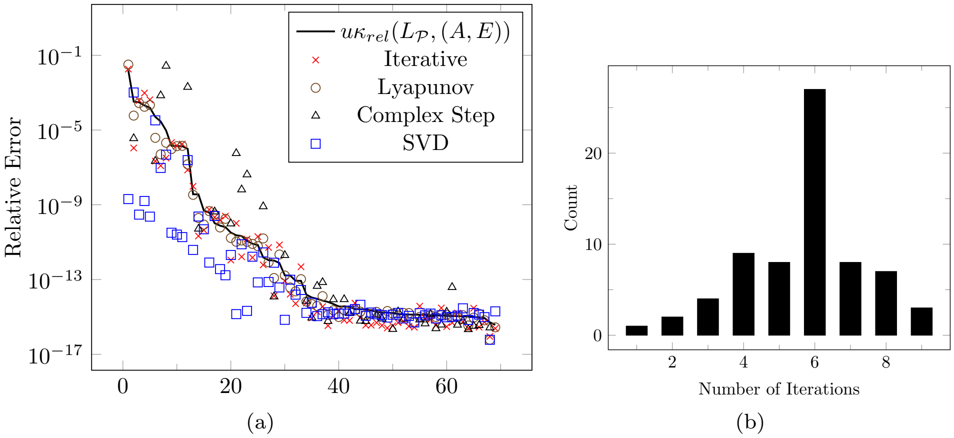

FIG. 1. (a) *Relative errors in the computed values of* $L_{\mathcal{P}}(A,E)$ *for various matrices* $A$ *and* $E$ *using four different methods.* (b) *Histogram showing the number of iterations used by the iterative method in these tests.*

approximate the derivative of this map in a randomly chosen direction, as advocated in [1]. The solid line in Figure 1(a) shows the estimated value of $u\kappa_{rel}(L_{\mathcal{P}},(A,E))$, where $u = 2^{-53}$ is the unit roundoff. The plot indicates that all of the methods under comparison behave in a forward stable way. Figure 1(b) shows a histogram of the number of iterations used by the iterative methods in these tests.

To study the convergence of the iterative methods in more detail, we focus now on a few representative matrices obtained from the MATLAB matrix gallery (note that the first three matrices are identical to those considered in [18, Chapter 8.9]):

1. a nearly orthogonal matrix, `orth(gallery('moler',16))+ones(16)*1e-3`,
2. a binomial matrix, `gallery('binomial',16)`,
3. the Frank matrix, `gallery('frank',16)`,
4. a modification of the Frank matrix obtained by setting its second smallest singular value equal to its smallest singular value. That is, $A = P\widetilde{\Sigma}Q^*$, where $P\Sigma Q^*$ is the SVD of the Frank matrix, $\widetilde{\Sigma}_{ii} = \Sigma_{ii}$ for $i \neq 15$, and $\widetilde{\Sigma}_{15,15} = \Sigma_{16,16}$.

We computed $\mathcal{P}(A)$ and $L_{\mathcal{P}}(A,E)$ for each $A$ listed above, again with $E$ a matrix consisting of random entries sampled from a normal distribution with mean 0 and variance 1. Tables 1–4 show the values of several quantities monitored during the iterations. The first two columns show the relative errors $\frac{\|X_k-U\|_F}{\|U\|_F}$ and $\frac{\|E_k-K\|_F}{\|K\|_F}$, where $U = \mathcal{P}(A)$ and $K = L_{\mathcal{P}}(A,E)$. The next three columns show the norms of (53)–(55), which are the quantities we used to determine when to terminate the iterations. Recall that (54) and (55) are computable approximations to $H_k\Omega_k - \Omega_k H_k$

TABLE 1
*Nearly orthogonal matrix,* $m = n = 16$, $\sigma_n(A) = 9.9\text{e-}1$, $\sigma_{n-1}(A) = 1.0\text{e}0$, $\kappa(A) = 1.0\text{e}0$.

| $k$ | $\frac{\|X_k-U\|_F}{\|U\|_F}$ | $\frac{\|E_k-K\|_F}{\|K\|_F}$ | $\|\mathcal{A}_k\|_F$ | $\|\mathcal{B}_k\|_F$ | $\|\mathcal{C}_k\|_F$ | $\|\widetilde{\mathcal{B}}_k\|_F$ | $\|\widetilde{\mathcal{C}}_k\|_F$ | $\mu_k$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 1.3e-5 | 2.3e-3 | 1.1e-4 | 1.6e-4 | 4.7e-2 | 1.6e-4 | 4.7e-2 | 1.0e0 |
| 2 | 3.0e-10 | 4.8e-8 | 2.4e-9 | 3.5e-9 | 9.9e-7 | 3.5e-9 | 9.9e-7 | 1.0e0 |
| 3 | 3.4e-16 | 5.0e-16 | 1.4e-15 | 2.3e-15 | 4.3e-15 | 5.0e-15 | 6.2e-15 | 1.0e0 |

and $H_kS_k + S_kH_k$, respectively. We have denoted $\widetilde{\mathcal{B}}_k = H_k\Omega_k - \Omega_kH_k$ and $\widetilde{\mathcal{C}}_k = H_kS_k+S_kH_k$ in the tables and recorded their norms in the seventh and eighth columns. Finally, the last column of the tables shows the value of the scaling parameter $\mu_k$. In the caption of each table, we have made note of the dimensions of the matrix $A$, the smallest and second smallest singular values $\sigma_n(A)$ and $\sigma_{n-1}(A)$ of $A$, respectively, and the condition number $\kappa(A)$ of $A$.

Tables 1 and 2 illustrate the effectiveness of the iteration on relatively well-conditioned matrices. In both cases, small relative errors in both $X_k$ and $E_k$ are achieved simultaneously, and convergence is detected appropriately by the termination criteria (56). Comparison of the columns labeled $\|\mathcal{B}_k\|$ and $\|\mathcal{C}_k\|$ with the columns labeled $\|\widetilde{\mathcal{B}}_k\|$ and $\|\widetilde{\mathcal{C}}_k\|$, respectively, lends credence to the asymptotic accuracy of the approximations $\mathcal{B}_k \approx \widetilde{\mathcal{B}}_k$ and $\mathcal{C}_k \approx \widetilde{\mathcal{C}}_k$, at least until roundoff errors begin to intervene.

Tables 3 and 4 illustrate what can go wrong when $A$ is ill-conditioned. In the case of Table 4, the matrix $A$ (the modified Frank matrix) has condition number $\kappa(A) = 2.3\text{e}14$, and its two smallest singular values are both close to zero: $\sigma_n(A) = \sigma_{n-1}(A) = 3.5\text{e-}13$. As a consequence, the (absolute) condition number of $\mathcal{P}$ with respect to real perturbations (as explained in section 4.4) is $2(\sigma_n + \sigma_{n-1})^{-1} = 2.9\text{e}12$, and we cannot expect much more than three or four digits of relative accuracy in double precision arithmetic when approximating $\mathcal{P}(A)$, much less $L_{\mathcal{P}}(A, E)$. This expectation is born out in Table 4. A more subtle phenomenon occurs in Table 3. There, the matrix $A$ (the Frank matrix) has condition number $\kappa(A) = 2.3\text{e}14$ as well, but only one of its singular values is close to zero. Namely, $\sigma_n(A) = 3.5\text{e-}13$, but $\sigma_{n-1}(A) = 8.7\text{e-}1$. As a consequence, $\mathcal{P}$ is very well-conditioned with respect

TABLE 2

*Binomial matrix, $m = n = 16$, $\sigma_n(A) = 2.6\text{e}0$, $\sigma_{n-1}(A) = 2.6\text{e}0$, $\kappa(A) = 4.7\text{e}3$.*

| $k$ | $\frac{\Vert X_k-U\Vert_F}{\Vert U\Vert_F}$ | $\frac{\Vert E_k-K\Vert_F}{\Vert K\Vert_F}$ | $\Vert\mathcal{A}_k\Vert_F$ | $\Vert\mathcal{B}_k\Vert_F$ | $\Vert\mathcal{C}_k\Vert_F$ | $\Vert\widetilde{\mathcal{B}}_k\Vert_F$ | $\Vert\widetilde{\mathcal{C}}_k\Vert_F$ | $\mu_k$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 1.7e1 | 7.4e1 | 2.4e3 | 8.5e4 | 8.5e4 | 1.2e2 | 2.7e3 | 1.8e-1 |
| 2 | 1.4e0 | 6.1e0 | 2.2e1 | 5.2e0 | 2.2e1 | 8.7e-1 | 2.1e1 | 5.9e-1 |
| 3 | 1.3e-1 | 1.2e0 | 1.1e0 | 2.4e-2 | 1.7e0 | 1.9e-2 | 1.7e0 | 9.2e-1 |
| 4 | 2.6e-3 | 8.7e-2 | 2.1e-2 | 5.3e-4 | 1.0e-1 | 5.3e-4 | 1.0e-1 | 1.0e0 |
| 5 | 1.4e-6 | 1.8e-4 | 1.1e-5 | 4.0e-7 | 2.1e-4 | 4.0e-7 | 2.1e-4 | 1.0e0 |
| 6 | 3.9e-13 | 2.1e-10 | 3.1e-12 | 6.9e-14 | 2.4e-10 | 6.9e-14 | 2.4e-10 | 1.0e0 |
| 7 | 2.1e-15 | 2.4e-15 | 1.4e-15 | 9.4e-17 | 1.3e-16 | 1.6e-15 | 1.6e-15 | 1.0e0 |

TABLE 3

*Frank matrix, $m = n = 16$, $\sigma_n(A) = 3.5\text{e-}13$, $\sigma_{n-1}(A) = 8.7\text{e-}1$, $\kappa(A) = 2.3\text{e}14$.*

| $k$ | $\frac{\Vert X_k-U\Vert_F}{\Vert U\Vert_F}$ | $\frac{\Vert E_k-K\Vert_F}{\Vert K\Vert_F}$ | $\Vert\mathcal{A}_k\Vert_F$ | $\Vert\mathcal{B}_k\Vert_F$ | $\Vert\mathcal{C}_k\Vert_F$ | $\Vert\widetilde{\mathcal{B}}_k\Vert_F$ | $\Vert\widetilde{\mathcal{C}}_k\Vert_F$ | $\mu_k$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 2.9e6 | 8.9e17 | 8.4e13 | 2.5e27 | 2.5e27 | 1.4e14 | 6.8e25 | 1.1e-6 |
| 2 | 1.9e0 | 4.6e11 | 4.5e1 | 1.2e3 | 1.9e13 | 4.8e1 | 1.9e13 | 4.2e-1 |
| 3 | 2.7e-1 | 6.7e10 | 2.5e0 | 3.2e0 | 7.5e11 | 1.8e0 | 7.5e11 | 8.3e-1 |
| 4 | 9.5e-3 | 6.0e8 | 7.7e-2 | 9.6e-2 | 5.7e9 | 9.4e-2 | 5.7e9 | 9.9e-1 |
| 5 | 3.9e-5 | 3.6e6 | 3.1e-4 | 5.3e-4 | 3.4e7 | 5.3e-4 | 3.4e7 | 1.0e0 |
| 6 | 1.6e-9 | 4.1e1 | 1.3e-8 | 2.6e-8 | 3.8e2 | 2.6e-8 | 3.8e2 | 1.0e0 |
| 7 | 7.2e-16 | 4.1e-5 | 1.1e-15 | 8.4e-16 | 4.0e-7 | 8.4e-15 | 4.0e-7 | 1.0e0 |
| 8 | 7.3e-16 | 4.1e-5 | 9.3e-16 | 7.1e-16 | 1.4e-15 | 8.5e-15 | 9.2e-15 | 1.0e0 |

TABLE 4

*Modified Frank matrix, $m = n = 16$, $\sigma_n(A) = 3.5\text{e-}13$, $\sigma_{n-1}(A) = 3.5\text{e-}13$, $\kappa(A) = 2.3\text{e}14$.*

| $k$ | $\frac{\lVert X_k-U\rVert_F}{\lVert U\rVert_F}$ | $\frac{\lVert E_k-K\rVert_F}{\lVert K\rVert_F}$ | $\lVert\mathcal{A}_k\rVert_F$ | $\lVert\mathcal{B}_k\rVert_F$ | $\lVert\mathcal{C}_k\rVert_F$ | $\lVert\widetilde{\mathcal{B}}_k\rVert_F$ | $\lVert\widetilde{\mathcal{C}}_k\rVert_F$ | $\mu_k$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 3.4e6 | 6.9e6 | 1.0e14 | 7.9e36 | 7.9e36 | 1.9e23 | 1.5e26 | 6.5e-7 |
| 2 | 1.2e0 | 1.4e0 | 2.2e1 | 9.8e10 | 1.6e13 | 2.0e10 | 1.6e13 | 5.1e-1 |
| 3 | 1.4e-1 | 4.4e-2 | 1.2e0 | 5.3e8 | 4.2e11 | 5.2e8 | 4.1e11 | 9.4e-1 |
| 4 | 7.2e-3 | 2.7e-3 | 5.8e-2 | 2.7e7 | 2.2e10 | 2.7e7 | 2.3e10 | 1.0e0 |
| 5 | 9.3e-5 | 3.0e-4 | 4.9e-4 | 1.3e4 | 1.0e7 | 1.3e4 | 1.9e9 | 1.0e0 |
| 6 | 7.0e-5 | 3.0e-4 | 5.5e-8 | 6.2e-2 | 4.9e1 | 6.2e-2 | 1.9e9 | 1.0e0 |
| 7 | 7.0e-5 | 3.0e-4 | 1.2e-15 | 1.0e-3 | 1.5e-3 | 3.4e-3 | 1.9e9 | 1.0e0 |

to real perturbations, having (absolute) condition number $2(\sigma_n + \sigma_{n-1})^{-1} = 1.2\text{e}0$. Curiously, the result is that $\mathcal{P}(A)$ is approximated very accurately, but $L_{\mathcal{P}}(A, E)$ is not. The fact that the performance of the Newton iteration (45) is largely unaffected by poorly conditioned $A$ (unless $A$ has two singular values close to zero) has been noted in [18, Chapter 8.9]. The observation that, in contrast, it takes only one near-zero singular value to corrupt the computation of $L_{\mathcal{P}}(A, E)$ via the iteration (45)–(46) deserves further study.

**7. Conclusion.** In this paper, we have derived iterative schemes for computing the Fréchet derivative of the polar decomposition. The structure of these iterative schemes lends credence to the mantra that differentiating an iteration for computing $f(A)$ leads to an iteration for computing $L_f(A, E)$. It would be interesting to determine what conditions on a matrix function $f$ ensure that this mantra bears out in practice. Certainly being a primary matrix function suffices, but the results of the present paper suggest that such a construction might work in a more general setting.

On a more specific level, several aspects of this paper warrant further consideration. While the termination criteria devised in section 4.2 appear to work well in practice, a more careful analysis of their effectiveness is lacking. In addition, it would be of interest to better understand the behavior of the iterative scheme (45)–(46) on ill-conditioned matrices.

**Appendix A. Approximate residuals.** In this section, we prove the validity of (51)–(52). Suppressing the subscript $k$ for the remainder of this section, our goal is to show that if

$$\mathcal{B} = \frac{1}{2}\left(X^*XX^*E - X^*EX^*X\right), \tag{61}$$

$$\mathcal{C} = (X^*E + E^*X) - \mathcal{B}, \tag{62}$$

then

$$\mathcal{B} = H\Omega - \Omega H + O(\|H^2 - I\|^2 + \|H^2 - I\|\|HS + SH\|),$$
$$\mathcal{C} = HS + SH + O(\|H^2 - I\|^2 + \|H^2 - I\|\|HS + SH\|).$$

Now since

$$H^2 - I = 2(H - I) + (H - I)^2,$$

the norms of $H^2 - I$ and $H - I$ are asymptotically equal, up to a factor of 2. Thus, it is enough to show that

$$\mathcal{B} = H\Omega - \Omega H + O(\|H-I\|^2 + \|H-I\|\|HS+SH\|), \tag{63}$$

$$\mathcal{C} = HS + SH + O(\|H-I\|^2 + \|H-I\|\|HS+SH\|). \tag{64}$$

The following lemma reduces this task to the verification of (63).

Lemma A.1. *We have*

$$(H\Omega - \Omega H) + (HS + SH) = X^*E + E^*X.$$

*Proof.* By (40) and the equalities $H = H^*$, $U^*E = \Omega + S$, $\Omega^* = -\Omega$, and $S^* = S$, we have

$$\begin{aligned} X^*E + E^*X &= HU^*E + E^*UH \\ &= H(\Omega + S) + (\Omega + S)^*H \\ &= H(\Omega + S) + (-\Omega + S)H \\ &= (H\Omega - \Omega H) + (HS + SH). \end{aligned}$$ □

It follows from the preceding lemma that if $\mathcal{B}$ satisfies (63), then $\mathcal{C} = (X^*E + E^*X) - \mathcal{B}$ automatically satisfies (64).

To prove (63), we begin by noting a few useful relations.

Lemma A.2. *For any $B \in \mathbb{C}^{n\times n}$,*

$$\begin{aligned} H(HB - BH) &= HB - BH + O(\|H-I\|^2), \\ (HB - BH)H &= HB - BH + O(\|H-I\|^2). \end{aligned}$$

*Proof.* These relations follow from the identities

$$\begin{aligned} H(HB - BH) &= HB - BH - (H-I)B(H-I) + (H-I)^2B, \\ (HB - BH)H &= HB - BH + (H-I)B(H-I) - B(H-I)^2. \end{aligned}$$ □

Lemma A.3. *We have*

$$X^*X = 2H - I + O(\|H-I\|^2).$$

*Proof.* Use the identity

$$H^2 = 2H - I + (H-I)^2$$

together with the fact that $X^*X = HU^*UH = H^2$. □

Now consider (61). Substituting $X^*X = 2H - I + O(\|H-I\|^2)$ and $X^*E = HU^*E = H(\Omega + S)$ gives, after simplification,

$$\mathcal{B} = H\left(H(\Omega + S) - (\Omega + S)H\right) + O(\|H-I\|^2).$$

Applying Lemma A.2 with $B = \Omega + S$ gives

$$\begin{aligned} \mathcal{B} &= H(\Omega + S) - (\Omega + S)H + O(\|H-I\|^2) \\ &= (H\Omega - \Omega H) + (HS - SH) + O(\|H-I\|^2). \end{aligned}$$

We will finish the proof of (63) by showing that

$$HS - SH = O\left(\|H-I\|^2 + \|H-I\|\|HS+SH\|\right).$$

Averaging the two equalities in Lemma A.2 with $B = S$ gives

$$HS - SH = \frac{1}{2}\left[H(HS - SH) + (HS - SH)H\right] + O(\|H - I\|^2),$$

Finally, an algebraic manipulation shows that the term in brackets above is equal to

$$H(HS - SH) + (HS - SH)H = (H - I)(HS + SH) - (HS + SH)(H - I),$$

and so it is of order $\|H - I\|\|HS + SH\|$.

## REFERENCES

[1] A. H. Al-Mohy and N. J. Higham, *Computing the Fréchet derivative of the matrix exponential, with an application to condition number estimation*, SIAM J. Matrix Anal. Appl., 30 (2009), pp. 1639–1657.

[2] A. H. Al-Mohy and N. J. Higham, *The complex step approximation to the Fréchet derivative of a matrix function*, Numer. Algorithms, 53 (2010), pp. 133–148.

[3] A. H. Al-Mohy, N. J. Higham, and S. D. Relton, *Computing the Fréchet derivative of the matrix logarithm and estimating the condition number*, SIAM J. Sci. Comput., 35 (2013), pp. C394–C410.

[4] A. Barrlund, *Perturbation bounds on the polar decomposition*, BIT, 30 (1990), pp. 101–113.

[5] R. H. Bartels and G. Stewart, *Solution of the matrix equation $AX + XB = C$*, Commun. ACM, 15 (1972), pp. 820–826.

[6] R. Bhatia, *Matrix factorizations and their perturbations*, Linear Algebra Appl., 197 (1994), pp. 245–276.

[7] R. Bhatia, *Matrix Analysis*, Grad. Texts in Math. 169, Springer Science & Business Media, New York, 2013.

[8] J. R. Cardoso, *Evaluating the Fréchet derivative of the matrix pth root*, Electron. Trans. Numer. Anal., 38 (2011), pp. 202–217.

[9] J. R. Cardoso, *Computation of the matrix pth root and its Fréchet derivative by integrals*, Electron. Trans. Numer. Anal., 39 (2012), pp. 414–436.

[10] L. Dieci and T. Eirola, *On smooth decompositions of matrices*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 800–819.

[11] R. L. Foote, *Regularity of the distance function*, Proc. Amer. Math. Soc., 92 (1984), pp. 153–155.

[12] E. S. Gawlik and M. Leok, *Embedding-Based Interpolation on the Special Orthogonal Group*, preprint, 2016.

[13] G. Golub, S. Nash, and C. Van Loan, *A Hessenberg-Schur method for the problem $AX + XB = C$*, IEEE Trans. Automat. Control, 24 (1979), pp. 909–913.

[14] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Vol. 3, Johns Hopkins University Press, Baltimore, 2012.

[15] N. J. Higham, *The Matrix Computation Toolbox*, http://www.ma.man.ac.uk/~higham/mctoolbox.

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

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

[18] N. J. Higham, *Functions of Matrices: Theory and Computation*, SIAM, Philadelphia, 2008.

[19] N. J. Higham and L. Lin, *An improved Schur–Padé algorithm for fractional powers of a matrix and their Fréchet derivatives*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 1341–1360.

[20] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 2012.

[21] C. Kenney and A. J. Laub, *Polar decomposition and matrix sign function condition estimates*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 488–504.

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

[23] C. S. Kenney and A. J. Laub, *A Schur–Fréchet algorithm for computing the logarithm and exponential of a matrix*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 640–663.

[24] P. Kunkel and V. Mehrmann, *Smooth factorizations of matrix valued functions and their derivatives*, Numer. Math., 60 (1991), pp. 115–131.

[25] R.-C. Li, *Relative perturbation bounds for the unitary polar factor*, BIT, 37 (1997), pp. 67–75.

[26] W. Li and W. Sun, *New perturbation bounds for unitary polar factors*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 362–372.

[27] R. Mathias, *Evaluating the Fréchet derivative of the matrix exponential*, Numer. Math., 63 (1992), pp. 213–226.

[28] R. Mathias, *Perturbation bounds for the polar decomposition*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 588–597.

[29] R. Mathias, *A chain rule for matrix functions and applications*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 610–620.

[30] I. Najfeld and T. F. Havel, *Derivatives of the matrix exponential and their computation*, Adv. Appl. Math., 16 (1995), pp. 321–375.

[31] 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.