VARIATION OF THE UNITARY PART OF A MATRIX*

RAJENDRA BHATIA[†] AND KALYAN MUKHERJEA[‡]

Abstract. The derivative of the map that takes an invertible matrix A to the unitary factor U in the polar decomposition $A = UP$ is evaluated. The same is done for the map that takes A to the unitary factor Q in the QR decomposition $A = QR$. These results lead to perturbation bounds for these maps. Other applications of the method developed are discussed.

Key words. polar decomposition, QR decomposition, Cholesky factorisation, perturbation, manifold, tangent space, unitarily invariant norm, singular value, Fréchet derivative

AMS subject classifications. 15A45, 65F99

1. Introduction. Let $\mathbf{M}(n)$ be the space of all $n \times n$ (complex) matrices; let $\mathbf{GL}(n)$ be the group consisting of all invertible matrices and let $\mathbf{U}(n)$ be the subgroup of unitary matrices. Every matrix A has a *polar decomposition* $A = UP$, where $U \in \mathbf{U}(n)$ and P is positive semidefinite. The positive part P, written as $|A|$, is unique and is equal to $(A^*A)^{1/2}$. If $A \in \mathbf{GL}(n)$ then the polar part U is also unique, since $U = AP^{-1}$.

Let $F : \mathbf{GL}(n) \longrightarrow \mathbf{U}(n)$ be the map $F(UP) = U$, which takes an invertible matrix to its polar part. Our first result, Theorem 2.1 below, gives an explicit expression for the Fréchet derivative of this map. As corollaries we obtain the value of the norm of this derivative with respect to any unitarily invariant norm on $\mathbf{M}(n)$, and then a perturbation bound for the polar part.

Another expression for the derivative of F has been obtained by Barrlund [1]. Using this and some results on Hadamard products, Mathias [12] has obtained the perturbation bound (13) derived below. Our coordinate-free approach to these questions is in line with some of our earlier work [3], [6], and [2, Chaps. 4, 5]. This approach has two merits. First, it is adaptable to more general contexts such as the KAK decomposition in semisimple Lie groups. We do not pursue that direction in this paper. Second, it works well for other matrix decompositions like the QR factorisation and the Cholesky factorisation. We illustrate this in later sections of this paper. Results similar to these have been obtained by Stewart [13] and, more recently, by Sun [15]. Here our approach clarifies some of the issues, unifies the work on these different questions, and clearly brings out the similarities and the differences between them.

We will denote by $|||\cdot|||$ any norm on $\mathbf{M}(n)$ that is *unitarily invariant*, i.e., a norm that satisfies the condition $|||UAV||| = |||A|||$ for all $A \in \mathbf{M}(n)$ and $U, V \in \mathbf{U}(n)$. Basic properties of such norms may be found in [2]. The singular values of A will be denoted as $s_1(A) \geq s_2(A) \geq \cdots \geq s_n(A)$. The *operator bound norm*, also called the *spectral norm* in the numerical analysis literature, will be denoted by $||\cdot||$ and the *Frobenius norm* by $||\cdot||_F$. We have

$$||A|| = s_1(A),$$

* Received by the editors September 30, 1992; accepted for publication (in revised form) January 31, 1993.

† Indian Statistical Institute, New Delhi 110016, India (rbh@isid.ernet.in). The work of this author was supported by a Department of Atomic Energy research grant.

‡ Indian Statistical Institute, Calcutta 700035, India.

$$\|A\|_F = \left[\sum_j s_j^2(A)\right]^{1/2}.$$

If $\mathcal{T}$ is a transformer, i.e., a linear map on the space $\mathbf{M}(n)$, then for any norm $|||\cdot|||$ on $\mathbf{M}(n)$, we define

$$|||\mathcal{T}||| = \sup\{|||\mathcal{T}(X)||| : |||X||| = 1\}.$$

We will use some elementary facts of calculus on manifolds that the reader may find in texts such as [7].

2. Variation of the unitary part. Let $\mathbf{T}_A\mathbf{GL}(n)$ be the tangent space to the manifold $\mathbf{GL}(n)$ at a point A in it. Since $\mathbf{GL}(n)$ is an open subset of $\mathbf{M}(n)$ we have $\mathbf{T}_A\mathbf{GL}(n) = \mathbf{M}(n)$. This is a special instance of the correspondence between a Lie group and its Lie algebra. Here the Lie algebra corresponding to the group $\mathbf{GL}(n)$ is $gl(n) = \mathbf{M}(n)$. The Lie algebra corresponding to the group $\mathbf{U}(n)$ is $u(n)$, the set of all skew-Hermitian matrices. This is the tangent space to $\mathbf{U}(n)$ at the point I. The tangent space to $\mathbf{U}(n)$ at a point U is $\mathbf{T}_U\mathbf{U}(n) = U\cdot \mathrm{u}(n) = \{US : S \in \mathrm{u}(n)\}$. The derivative of F at a point $A = UP$ of $\mathbf{GL}(n)$, denoted by $DF(UP)$, is a linear map from $\mathbf{M}(n)$ to $U\cdot \mathrm{u}(n)$.

Let $h(n)$ denote the space of all Hermitian matrices. We have $h(n) = \imath\cdot \mathrm{u}(n)$. We have a vector space decomposition

$$\mathbf{M}(n) = u(n) + h(n), \tag{1}$$

in which every matrix splits uniquely as

$$X = S + H, \tag{2}$$

where

$$S = \frac{X - X^*}{2}, \qquad H = \frac{X + X^*}{2}. \tag{3}$$

We can now state our first main result.

THEOREM 2.1. *Let $F : \mathbf{GL}(n) \longrightarrow \mathbf{U}(n)$ be the map defined above as $F(UP) = U$. Let X be any element of $\mathbf{M}(n)$ and let $X = S + H$ be its splitting into skew-Hermitian and Hermitian parts. Then the value of the derivative $DF(UP)$ on the tangent vector UX is given by*

$$DF(UP)(UX) = 2U\int_0^\infty e^{-tP}Se^{-tP}dt. \tag{4}$$

Proof. Let $\mathbf{P}(n)$ be the set of all $n\times n$ positive definite matrices. This is an open subset of the real vector space $h(n)$. Hence for every $P \in \mathbf{P}(n)$ the tangent space $\mathbf{T}_P\mathbf{P}(n) = h(n)$.

Let $\Psi : \mathbf{U}(n)\times \mathbf{P}(n) \longrightarrow \mathbf{GL}(n)$ be the map $\Psi(U,P) = UP$ and let Φ be the inverse map $\Phi(UP) = (U,P)$. Then writing $\Phi = (\Phi_1, \Phi_2)$, we have $F = \Phi_1$.

The derivative $D\Psi(U,P)$ is a linear map with domain $\mathbf{T}_U\mathbf{U}(n) + \mathbf{T}_P\mathbf{P}(n) = U\cdot u(n) + h(n)$ and range $\mathbf{M}(n) = U\cdot u(n) + U\cdot h(n)$. By definition, this derivative is evaluated as

$$D\Psi(U,P)(US,H) = \frac{d}{dt}[\Psi(Ue^{tS}, P + tH)]_{t=0} = USP + UH \tag{5}$$

for all $S \in u(n), H \in h(n)$.

Now note that for small values of t, $P+tH$ is positive for any $H \in h(n)$, and hence we have $\Phi_1(UP+tUH) = \Phi_1(UP) = U$. So the kernel of $D\Phi_1(UP)$ contains $U \cdot h(n)$. In fact, ker $D\Phi_1(UP) = U \cdot h(n)$, since Φ is a diffeomorphism from $\mathbf{GL}(n)$ onto $\mathbf{U}(n) \times \mathbf{P}(n)$ and each of $u(n)$ and $h(n)$ has half the dimension of $\mathbf{M}(n)$. So we need to compute the value of $D\Phi_1(UP)$ only on tangent vectors of the form $US, S \in u(n)$. Let

$$D\Phi(UP)(US) = (UM, N), \qquad M \in u(n), N \in h(n). \tag{6}$$

Since $\Phi = \Psi^{-1}$, we have using (5)

$$US = D\Psi(U,P)(UM,N) = UMP + UN. \tag{7}$$

We want to determine M from this equation. So, we must solve the equation

$$MP + N = S.$$

Taking adjoints we have

$$-PM + N = -S.$$

From these two equations we obtain

$$MP + PM = 2S. \tag{8}$$

This is the familiar Lyapunov equation and its solution (see [9], [10]) is

$$M = 2\int_0^\infty e^{-tP} S e^{-tP} dt. \tag{9}$$

Equation (4) now follows from (6) and (9). □

Corollary 2.1. *For every unitarily invariant norm* $|||\cdot|||$ *on* $\mathbf{M}(n)$ *we have*

$$|||DF(UP)||| = \|P^{-1}\| = s_n^{-1}(A). \tag{10}$$

Proof. This follows from (4) by a familiar argument that we repeat for the reader's convenience.

Since the norm is unitarily invariant, we have

$$|||DF(UP)(UX)||| \le 2\int_0^\infty |||e^{-tP} S e^{-tP}|||dt. \tag{11}$$

Then, since $|||BCD||| \le \|B\| \cdot |||C||| \cdot \|D\|$ for all B, C, D, we have

$$\begin{aligned} |||e^{-tP} S e^{-tP}||| &\le \|e^{-tP}\| \cdot |||S||| \cdot \|e^{-tP}\| \\ &\le e^{-2ts_n(A)}|||S||| \\ &\le e^{-2ts_n(A)}|||X||| \end{aligned} \tag{12}$$

using the fact $|||S||| \le |||X|||$.

From (11) and (12) we obtain

$$|||DF(UP)||| = \sup_{|||X|||=1} |||DF(UP)(UX)||| \le s_n^{-1}(A).$$

Choosing $X = iI/|||I|||$ one sees that this is actually an equality. □

Using the mean value theorem, we obtain from Corollary 2.2.

COROLLARY 2.2. *Let A_0 and A_1 be two elements of $\mathbf{GL}(n)$ with polar parts U_0 and U_1, respectively. Assume that the line segment $A(t) = (1-t)A_0 + tA_1, 0 \le t \le 1$, joining A_0 and A_1 lies inside $\mathbf{GL}(n)$. Then for every unitarily invariant norm*

$$|||U_0 - U_1||| \le \max_{0\le t\le 1} \|A(t)^{-1}\|.|||A_0 - A_1|||. \tag{13}$$

These statements can be expressed in another language by saying that in any unitarily invariant norm, the *condition* of the function F at any point A of $\mathbf{GL}(n)$ is given by $s_n^{-1}(A)$.

We should remark that the solution of (8) can also be expressed as a Hadamard product [10], [11]; from this we can obtain estimates like ours either directly or by converting this formula to the integral expression (9). We have chosen the integral form of the solution because it might be useful in analysing infinite dimensional problems as well. An effective use of such integrals was made earlier in [5].

3. The QR decomposition. Every square complex matrix A can be written as a product $A = QR$ where Q is unitary and R is upper triangular. If A is invertible then so is R. Furthermore, we can choose the diagonal entries of R to be positive and with this added restriction this product decomposition is unique for every $A \in GL(n)$. This decomposition called the *QR decomposition* is extremely important in numerical analysis. See [14] for details.

We will now analyse the variation of the unitary part in this decomposition in the same way as for the polar decomposition.

Let $\mathbf{B}(n)$ denote the set of all upper triangular matrices with positive diagonal entries and let $b(n)$ be the set of all upper triangular matrices with real diagonal entries. Then $b(n)$ is a real vector space and $\mathbf{B}(n)$ is an open subset of it. So, the tangent space $\mathbf{T}_R\mathbf{B}(n)$ to $\mathbf{B}(n)$ at any point R of it is the space $b(n)$. (One may note here that $\mathbf{B}(n)$ is a Lie group and $b(n)$ is its Lie algebra.)

The QR decomposition associates with every element A of $\mathbf{GL}(n)$ a unique element Q of $\mathbf{U}(n)$ and a unique element R of $\mathbf{B}(n)$. Let $F : \mathbf{GL}(n) \longrightarrow \mathbf{U}(n)$ now be the map $F(QR) = Q$. The derivative of F at $A = QR$ is a linear map from $\mathbf{M}(n)$ to $Q{\cdot}\mathbf{U}(n)$.

The subspaces $u(n)$ and $b(n)$ are complementary to each other in $\mathbf{M}(n)$ and we have a vector space decomposition

$$\mathbf{M}(n) = u(n) + b(n). \tag{14}$$

This decomposition is not as familiar as the one in (1) and it has some different features. If a matrix X splits as

$$X = K + T \tag{15}$$

in the above decomposition then we must have the following relations between the entries of these matrices

$$k_{jj} = i\operatorname{Im} x_{jj} \quad \text{for all } j, \quad k_{ij} = -\overline{x}_{ji} \quad \text{for } j > i, \quad k_{ij} = x_{ij} \quad \text{for } i > j,$$

$$t_{jj} = \operatorname{Re} x_{jj} \quad \text{for all } j, \quad t_{ij} = x_{ij} + \overline{x}_{ji} \quad \text{for } j > i, \quad t_{ij} = 0 \quad \text{for } i > j. \tag{16}$$

Whereas, in the case of the decomposition (1) the projections onto both the components are norm-reducing for every unitarily invariant norm (just use the triangle inequality), this is not the case for the decomposition (14). Instead, we have for the Frobenius norm the following lemma.

LEMMA 3.1. *Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be the complementary projection operators in* $\mathbf{M}(n)$ *corresponding to the decomposition* (14). *Then*

$$||\mathcal{P}_1||_F = ||\mathcal{P}_2||_F = \sqrt{2}. \tag{17}$$

Proof. From (15) and (16) one can easily see that $||K||_F^2 \leq 2||X||_F^2$ and $||T||_F^2 \leq 2||X||_F^2$. The first inequality becomes an equality when $X = \left(\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right)$, the second when $X = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. □

Remark 3.1. If instead of the Frobenius norm the operator norm is used then the norms of the projections $\mathcal{P}_1$ and $\mathcal{P}_2$ grow with the dimension n. To see this, note that if X is Hermitian then

$$T = 2\Delta(X) - \text{diag } X, \tag{18}$$

where Δ is the *triangular truncation* operator, i.e., for any matrix $A, \Delta(A)$ is the matrix obtained from A by replacing the entries below the main diagonal by zeros. It is well known that the norm $||\Delta||$ grows as $\log n$. For example, if X is the $n \times n$ Hermitian matrix whose diagonal entries are zero and whose off-diagonal entries are $x_{ij} = \sqrt{-1}/(i-j)$, then $||X|| \leq \pi$ and $||\Delta(X)|| \geq \frac{4}{5}\log n$. (See [8, p. 39].) On the other hand, $||\text{diag}X|| \leq ||X||$. So $||\mathcal{P}_2||$ must grow at least as $\log n$. Hence so must $||\mathcal{P}_1||$.

Returning to the map $F(QR) = Q$, let us see how far an analysis similar to the one in §2 takes us. Now define $\Psi : \mathbf{U}(n) \times \mathbf{B}(n) \longrightarrow \mathbf{GL}(n)$ to be the map $\Psi(Q,R) = QR$ and let Φ be its inverse map $\Phi(QR) = (Q,R)$. If Φ is written as $\Phi = (\Phi_1, \Phi_2)$ then $F = \Phi_1$. The derivative $D\Psi(Q,R)$ is a linear map whose domain is $\mathbf{T}_Q\mathbf{U}(n) + \mathbf{T}_R\mathbf{B}(n) = Q\cdot \mathrm{u}(n) + b(n)$, and whose range is $\mathbf{M}(n) = Q\cdot\mathrm{u}(n) + Q\cdot b(n)$. The derivative is evaluated as

$$D\Psi(Q,R)(QK,T) = \frac{d}{dt}[\Psi(Qe^{tK}, R+tT)]_{t=0} = QKR + QT \tag{19}$$

for all $K \in u(n), T \in b(n)$.

If $R \in \mathbf{B}(n)$ and $T \in b(n)$ then for small values of t, $R + tT$ is in $\mathbf{B}(n)$. By the uniqueness of the QR factorisation, $\Phi_1(QR + tQT) = \Phi_1(QR) = Q$. Hence the space $Q\cdot b(n)$ is contained in $\ker D\Phi_1(QR)$. But, then counting their dimensions we can conclude that $Q\cdot b(n) = \ker D\Phi_1(QR)$. So we need to compute the values of $D\Phi_1(QR)$ only on tangent vectors of the form $QK, K \in \mathrm{u}(n)$. Let

$$D\Phi(QR)(QK) = (QM, Y), \quad \text{where } M \in u(n), \qquad Y \in b(n). \tag{20}$$

Since $\Phi = \Psi^{-1}$, we have from (19) and (20)

$$QK = QMR + QY. \tag{21}$$

To determine M from this we need to solve the equation

$$MR + Y = K. \tag{22}$$

Here the similarity with the analysis in §2 ends. In (3) P and N were selfadjoint, so taking adjoints we could achieve a major simplification by eliminating the redundant variable N. We cannot do that here. However, we can still obtain an expression for M from (22). Rewrite this equation as

$$M + YR^{-1} = KR^{-1}.$$

Note that $M \in u(n)$ and $YR^{-1} \in b(n)$. So $M = \mathcal{P}_1(KR^{-1})$ in the notation used earlier. We have thus proved the following theorem.

THEOREM 3.1. *Let $F : \mathbf{GL}(n) \longrightarrow \mathbf{U}(n)$ be the map defined as $F(QR) = Q$. Let X be any element of $\mathbf{M}(n)$ and let $X = K + T$ be its splitting in the decomposition $\mathbf{M}(n) = u(n) + b(n)$. Then the value of the derivative $DF(QR)$ on the tangent vector QX is given by*

$$DF(QR)(QX) = Q\mathcal{P}_1(KR^{-1}), \tag{23}$$

where $\mathcal{P}_1$ is the projection operator in $\mathbf{M}(n)$ projecting onto $u(n)$ along the complementary space $b(n)$.

Note that the quantities occurring in the above formula can be explicitly computed from the relations (16).

COROLLARY 3.1. *For every matrix $A = QR$ in $\mathbf{GL}(n)$, we have*

$$||DF(QR)||_F \leq \sqrt{2}||R^{-1}|| = \sqrt{2}||A^{-1}||. \tag{24}$$

Proof. Use Theorem 3.1, Lemma 3.1, the unitary invariance of the Frobenius norm, and the inequality $||ST||_F \leq ||S||_F||T||$ that is valid for any two matrices S and T. □

Using the mean value theorem we obtain the following corollary.

COROLLARY 3.2. *Let $A_0 = Q_0R_0$ and $A_1 = Q_1R_1$ be any two elements of $\mathbf{GL}(n)$. Suppose that the line segment $A(t) = (1-t)A_0 + tA_1$, $0 \leq t \leq 1$, joining A_0 and A_1 lies entirely inside $\mathbf{GL}(n)$. Then*

$$||Q_0 - Q_1||_F \leq \sqrt{2} \max_{0 \leq t \leq 1} ||A(t)^{-1}|| \; ||A_0 - A_1||_F. \tag{25}$$

We should remark that from (23) we could surely derive some estimates for $|||DF(QR)|||$ for any unitarily invariant norm. These would, however, involve $|||\mathcal{P}_1|||$ and for this we have good estimates only in the case of the Frobenius norm.

4. The Cholesky factorisation. A common feature of our analysis of the polar decomposition and the QR decomposition is that we replaced the study of the map Φ, which takes a matrix to its factors, by that of its inverse map Ψ. This, being a multiplication map, is easier to handle. A similar idea is useful in the perturbation analysis of the Cholesky factorisation.

Every positive definite matrix A has a unique factorisation $A = R^*R$, where R is an upper triangular matrix with positive diagonal entries. This is called the *Cholesky factorisation.*

In our notation, we now have a map $\Phi : \mathbf{P}(n) \longrightarrow \mathbf{B}(n)$ defined as $\Phi(A) = R$, where R is the Cholesky factor of A. The inverse map is $\Psi(R) = R^*R$. The derivative $D\Psi(R)$ is a linear map from the tangent space $\mathbf{T}_R\mathbf{B}(n) = b(n)$ to the tangent space $\mathbf{T}_A\mathbf{P}(n) = h(n)$. This derivative is evaluated as

$$D\Psi(R)(T) = \frac{d}{dt} \; [\Psi(R + tT)]_{t=0} = R^*T + T^*R, \tag{26}$$

for every $T \in b(n)$.

Now, for $H \in h(n)$ let

$$D\Phi(A)(H) = T, \quad \text{where } T \in b(n). \tag{27}$$

Then since $\Phi = \Psi^{-1}$, we must have

$$R^*T + T^*R = H. \tag{28}$$

To estimate $|||D\Phi(A)|||$, we need to estimate T in terms of H and R. Rewrite (28) as

$$TR^{-1} + (TR^{-1})^* = (R^*)^{-1}HR^{-1}. \tag{29}$$

Since $TR^{-1} \in b(n)$, we have from (29)

$$||TR^{-1}||_F \leq \frac{1}{\sqrt{2}}||(R^*)^{-1}HR^{-1}||_F \leq \frac{1}{\sqrt{2}}||R^{-1}||^2||H||_F.$$

Since $||T||_F \leq ||TR^{-1}||_F||R||$, this gives

$$||T||_F \leq \frac{1}{\sqrt{2}}||R||\ ||R^{-1}||^2\ ||H||_F. \tag{30}$$

From (27) and (30), we get

$$||D\Phi(A)||_F \leq \frac{1}{\sqrt{2}}||R||\ ||R^{-1}||^2 \ = \ \frac{1}{\sqrt{2}}||A||^{1/2}\ ||A^{-1}||. \tag{31}$$

For the map Ψ, we could write from (26)

$$|||D\Psi(R)||| \ = \ \sup_{|||T|||=1} |||R^*T + T^*R||| \leq 2||R||, \tag{32}$$

for every unitarily invariant norm.

Inequalities (31) and (32) can be used to write perturbation bounds for Φ and Ψ as before.

Finally, we remark that from results of §§2 and 3, we can obtain some information about the variation of the positive part P in the polar decomposition and the upper triangular part R in the QR decomposition.

Note. In a sequel to this paper [4], the above analysis has been carried further to obtain perturbation bounds for several other matrix decompositions.

REFERENCES

[1] A. Barrlund, *Perturbation bounds on the polar decomposition*, BIT, 30 (1989), pp. 101–113.

[2] R. Bhatia, *Perturbation Bounds for Matrix Eigenvalues*, Longman Scientific and Technical, Essex, UK, 1987.

[3] ———, *Analysis of spectral variation and some inequalities*, Trans. Amer. Math. Soc., 272 (1982), pp. 323–331.

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

[5] R. Bhatia, C. Davis, and A. McIntosh, *Perturbation of spectral subspaces and solution of linear operator equations*, Linear Algebra Appl., 52 (1983), pp. 45–67.

[6] R. Bhatia and K. K. Mukherjea, *On the rate of change of spectra of operators*, Linear Algebra Appl., 27 (1979), pp. 147–157.

[7] R. L. Bishop and R. J. Crittenden, *Geometry of Manifolds*, Academic Press, New York, 1964.

[8] K. R. DAVIDSON, *Nest Algebras*, Longman Scientific and Technical, Essex, UK, 1988.

[9] E. HEINZ, *Beiträge zur Störungstheorie der Spectralzerlegung*, Math. Ann., 123 (1951), pp. 415–438.

[10] R. A. HORN, *The Hadamard product*, Proc. Symp. Appl. Math., Vol. 40, Amer. Math. Soc., 1990.

[11] R. A. HORN AND C. R. JOHNSON, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991.

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

[13] G. W. STEWART, *Perturbation bounds for the QR factorization of a matrix*, SIAM J. Numer. Anal., 14 (1977), pp. 509–518.

[14] G. W. STEWART AND J. -G. SUN, *Matrix Perturbation Theory*, Academic Press, New York, 1990.

[15] J. -G. SUN, *Perturbation bounds for the Cholesky and QR factorizations*, BIT, 31 (1991), pp. 341–352.