# Matrix Factorizations and Their Perturbations

Rajendra Bhatia
*Indian Statistical Institute*
*7, S.J.S. Sansanwal Marg*
*New Delhi-110 016, India*

Submitted by Chi-Kwong Li

Dedicated to C. S. Seshadri

ABSTRACT

There are several ways in which a matrix can be factorized as a product of two special matrices. These factorizations are often used in numerical analysis, and many perturbation bounds useful in such analysis have been proved by various authors. In this paper a simple method which leads to these results and several new ones is discussed. In the last section related results on the Lipschitz continuity of the matrix absolute value are surveyed.

## 1. INTRODUCTION

A *matrix factorization theorem* is an assertion that a matrix $A$ can be factorized into a product $A = A_1 A_2$ of two special matrices $A_1$, $A_2$. Some conditions may be necessary for such a decomposition to exist, and some further conditions may ensure the uniqueness of the factorization.

Among the better-known examples of such factorizations are the polar decomposition into unitary and positive factors, the $QR$ decomposition into unitary and upper triangular factors, and the $LR$ decomposition into lower triangular and upper triangular factors.

Such factorizations are pleasing to the theorist and useful in a variety of applications. Thus, for instance, the $LR$ decomposition is related to Gaussian elimination and the $QR$ decomposition to the $QR$ algorithm in numerical analysis [see, for example, Golub and Van Loan (1989) or Stewart and Sun (1990)]. The polar decomposition is related to the Löwdin orthogonalization




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

in quantum chemistry [see Goldstein and Levy (1991) or Bhatia and Mukherjea (1986)].

A problem that has been studied in considerable detail over the last fifteen years in several papers, mainly by numerical analysts, is that of finding perturbation bounds for these factorizations. Such results tell us how much a given change in $A$ might affect its factors in these decompositions.

Many of these diverse results, and several new ones, can be derived in a unified way following ideas introduced in Bhatia and Mukherjea (1992). The aim of the present paper is to explain this method and to use it to derive many new results for the decompositions mentioned above and also for some others for which such an analysis has not been made before. This approach has three virtues. It provides a unification of earlier results, sometimes obtained by methods which seem ad hoc; the results obtained are comparable to or stronger than the ones obtained by earlier authors; new problems can be tackled together with the old ones.

The paper is organized as follows. In Section 2 we outline the general features of the method. In Section 3 we list the matrix decompositions to which the method will be applied. Each of these multiplicative decompositions gives rise to an additive decomposition of the space of matrices. In this section we will also write down the projection operators corresponding to these additive decompositions and evaluate their norms. In Section 4 we obtain perturbation bounds and compare them with those of earlier authors. Section 5 is purely expository. Mathematical physicists have obtained several interesting results about the perturbation of the positive factor in the polar decomposition. These seem to have escaped the attention of numerical analysts working on the same question. We give a brief summary of these results in this section.

## 2. THE GENERAL METHOD

We will use some elementary facts from calculus on manifolds and matrix Lie groups. A good reference for the former is Dieudonné (1960, Chapter 8), and for the latter Chevalley (1946, Chapter 1).

Let $\mathbf{M}$ denote the space of all $n \times n$ matrices. These matrices will have complex entries, except that they will be real when we talk of the $SR$ decomposition. The class of matrices $\mathbf{A}$ to be factorized will be an open subset of $\mathbf{M}$; sometimes it will be a Lie group. The factors will come from two special subsets $\mathbf{A}_1$ and $\mathbf{A}_2$ of $\mathbf{M}$. These will either be Lie groups or open subsets of a linear space in $\mathbf{M}$. A matrix factorization theorem will tell us that each element $A$ of $\mathbf{A}$ has a unique factorization

$$A = A_1 A_2 \qquad \text{where} \quad A_1 \in \mathbf{A}_1, \quad A_2 \in \mathbf{A}_2. \tag{2.1}$$

Thus, for example, in the polar decomposition theorem $\mathbf{A} = \mathbf{GL}$ is the general linear group consisting of all nonsingular matrices. This is a Lie group and an open dense subset of $\mathbf{M}$. The factors $U$ and $P$ in the decomposition $A = UP$ come from the set $\mathbf{U}$ consisting of all unitary matrices and the set $\mathbf{P}$ consisting of all positive definite matrices. The first is a Lie group, the second is an open subset of the space **Herm** consisting of all Hermitian matrices. This is a real linear subspace of $\mathbf{M}$.

The decomposition (2.1) thus gives a map

$$\mathbf{A} \xrightarrow{\Phi} \mathbf{A}_1 \times \mathbf{A}_2 \tag{2.2}$$

between manifolds. We will write

$$\Phi(A) = (\Phi_1(A), \Phi_2(A)) = (A_1, A_2). \tag{2.3}$$

To study the variation of $A_1$ and $A_2$ with $A$ it is most natural to study the derivatives of the maps $\Phi_1$ and $\Phi_2$. However, this is not easy, because these maps are complicated to describe. This difficulty can be circumvented by studying instead the inverse map

$$\mathbf{A}_1 \times \mathbf{A}_2 \xrightarrow{\Psi} \mathbf{A}$$

defined as

$$\Psi(A_1, A_2) = A_1 A_2 = A. \tag{2.4}$$

The map $\Psi$ is simpler, being just a product. Its derivative can be computed easily. By the implicit function theorem we can then obtain the derivative of $\Phi$. This simple idea is crucial for the success of this method.

The derivative of $\Phi$ at $A$, denoted as $D\Phi(A)$, is a linear map

$$\mathbf{T}_A\mathbf{A} \xrightarrow{D\Phi(A)} \mathbf{T}_{A_1}\mathbf{A}_1 + \mathbf{T}_{A_2}\mathbf{A}_2, \tag{2.5}$$

where $\mathbf{T}_A\mathbf{A}$ is the tangent space to the manifold $\mathbf{A}$ at the point $A$. When $\mathbf{A}$ is an open set in a linear space $\mathbf{X}$, then $\mathbf{T}_A\mathbf{A} = \mathbf{X}$ for every $A \in \mathbf{A}$. When $\mathbf{A}$ is a Lie group, then the tangent space at the identity is the Lie algebra corresponding to this Lie group. The tangent space at any other point is then $A \cdot \mathbf{T}_I\mathbf{A}$. Thus, for example, the tangent space to $\mathbf{GL}$ at any point is the space

**M**; that to **P** is **Herm**. The tangent space to **U** at $I$ is the Lie algebra **Sherm** consisting of all skew-Hermitian matrices; the tangent space at any other point $U$ is $U \cdot \mathbf{Sherm} = \{US : S \in \mathbf{Sherm}\}$. The derivative of $\Psi$ at $(A_1, A_2)$ is a linear map $D\Psi(A_1, A_2)$ which is inverse to (2.5).

The two tangent spaces to $\mathbf{A}_1$ and $\mathbf{A}_2$ provide an additive decomposition of the tangent space to **A**. Thus, in the above special example, $\mathbf{M} = \mathbf{Sherm} + \mathbf{Herm}$ is the familiar decomposition into skew-Hermitian and Hermitian parts.

Our method has four steps:

*Step 1.* Evaluate $D\Psi(A_1, A_2)$. This is easy. In case $\mathbf{A}_1$ is an open set in a linear space **X**, all tangential vectors at $A_1$ are of the form $A_1 + tB$, $B \in \mathbf{X}$. In case $\mathbf{A}_1$ is not an open set but a Lie group, then using the correspondence between Lie algebras and Lie groups via the exponential map, tangent vectors at $A_1$ are written as $A_1 e^{tB}$ where $B$ is from the Lie algebra. The derivative $D\Psi(A_1, A_2)$ is just the directional derivative in these tangential directions. Thus, for example, in the case of the polar decomposition

$$\begin{aligned} D\Psi(U, P)(US, H) &= \left[\frac{d}{dt}\Psi(Ue^{tS}, P + tH)\right]_{t=0} \\ &= \left[\frac{d}{dt}Ue^{tS}(P + tH)\right]_{t=0} \\ &= USP + UH \end{aligned} \tag{2.6}$$

for each $S \in \mathbf{Sherm}$, $H \in \mathbf{Herm}$.

*Step 2.* Use this to find $D\Psi(A)$ in a convenient form. This leads to a matrix equation. Thus, in the case of the polar decomposition, for any $X \in \mathbf{M}$ we want to find the value of $D\Phi(A)$ at the tangent vector $X$. We have $D\Phi(A)(UX) = (UM, N)$ for some $M \in \mathbf{Sherm}$, $N \in \mathbf{Herm}$. Since $\Phi = \Psi^{-1}$, we get from (2.6)

$$UX = UMP + UN.$$

Hence, $D\Phi(A)(UX)$ will be known if we can determine $M$ and $N$ from the equation

$$X = MP + N. \tag{2.7}$$

We will see that in this special case $M$, $N$ can be found explicitly. In other

cases we may not be able to find equally explicit expressions.

*Step 3.* Whether or not we are able to solve the equation arising in step 2, we will find the above information adequate to get bounds for the derivatives $D\Phi_1(A)$ and $D\Phi_2(A)$. The additional ingredient required is an estimate of the norms of the projection operators in $\mathbf{M}$ corresponding to the additive decomposition into the two tangent spaces. We are able to get this in all cases for the Frobenius norm and in some cases for all unitarily invariant norms. This is done in Section 3.

*Step 4.* These bounds on the norms of the derivatives then lead to perturbation bounds when we use standard theorems of calculus like Taylor's theorem or the mean value theorem.

We should point out that in several papers [Barrlund (1989), Higham (1986), Sun (1991), and especially Kenny and Laub (1991) and Mathias (1991)] the idea of estimating the derivative is very much to be found. It is the simplification achieved by going to the map $\Psi$ and the generality and the flexibility this allows which are new.

Let us fix some notation. The symbol $\|A\|$ will denote the *operator bound norm* (the *spectral norm*) of $A$, the symbol $\|A\|_F$ the *Frobenius norm* of $A$, and $|||A|||$ any *unitarily invariant norm* of $A$. If $A$ has matrix entries $a_{ij}$ and singular values $s_1(A) \geqslant s_2(A) \geqslant \cdots \geqslant s_n(A)$, then

$$\|A\| = s_1(A) = \sup\{\|Ax\| : x \in \mathbb{C}^n, \|x\| = 1\},$$

$$\|A\|_F = \left(\sum_{i,j} |a_{ij}|^2\right)^{1/2} = \left(\sum_j s_j^2(A)\right)^{1/2}.$$

Unitary invariance means the property

$$|||A||| = |||UAV||| \qquad \text{for all} \quad A \in \mathbf{M} \text{ and } U, V \in \mathbf{U}.$$

Properties of such norms may be found in Bhatia (1987), Horn and Johnson (1985), or Stewart and Sun (1990). Both $\|\cdot\|$ and $\|\cdot\|_F$ are unitarily invariant norms. Each unitarily invariant norm is a function only of the singular values of $A$. Hence if $A^*$ denotes the Hermitian conjugate of $A$, $A^T$ its transpose, and $\bar{A}$ its complex conjugate, then

$$|||A||| = |||A^*||| = |||A^T||| = |||\bar{A}|||. \tag{2.8}$$

Another property we will use is

$$|||ABC||| \leqslant \|A\| \, |||B||| \, \|C\| \tag{2.9}$$

for all $A, B, C$. Finally if $\mathscr{J}$ is a linear map on $\mathbf{M}$, we will put

$$|||\mathscr{J}||| = \sup\{|||\mathscr{J}X|||: X \in \mathbf{M}, |||X||| = 1\}. \tag{2.10}$$

## 3. SOME MATRIX FACTORIZATIONS

In most of the decompositions of interest in numerical analysis one of the factors is a triangular matrix. Let us fix some notation for classes of triangular matrices which will occur often.

We will denote by $\Delta$ the set of all upper triangular matrices and by $\Delta^*$ the set of all lower triangular matrices. A subscript will usually indicate the nature of the diagonal entries. Thus, $\Delta_{re}$, $\Delta_+$, $\Delta_1$, and $\Delta_0$ will stand for the set of upper triangular matrices whose diagonal entries are all real, positive, one, and zero, respectively. The symbol $\Delta_{ns}$ will be used for upper triangular matrices which are nonsingular, i.e., those which have no zero on the diagonal.

It is worth noting here that $\Delta$ is a linear space; $\Delta_{re}$ and $\Delta_0$ are linear subspaces of it. The set $\Delta_{ns}$ is a subgroup of $\mathbf{GL}$; so are its subsets $\Delta_+$ and $\Delta_1$. The Lie algebras corresponding to the Lie groups $\Delta_{ns}$, $\Delta_+$, and $\Delta_1$ are $\Delta$, $\Delta_{re}$, and $\Delta_0$, respectively. Also note that $\Delta_{ns}$ and $\Delta_+$ are open subsets of $\Delta$ and $\Delta_{re}$, respectively.

### 3.1. *The Polar Decomposition*

Most of the facts about this decomposition have already been recalled. In particular, we have noted that the tangent spaces to $\mathbf{U}$ and $\mathbf{P}$ give a decomposition

$$\mathbf{M} = \mathbf{Sherm} + \mathbf{Herm}. \tag{3.1}$$

If $\mathscr{P}_1$ and $\mathscr{P}_2$ are the complementary projection operators in $\mathbf{M}$ corresponding to (3.1), then

$$\mathscr{P}_1(A) = \frac{A - A^*}{2}, \qquad \mathscr{P}_2(A) = \frac{A + A^*}{2}.$$

Using (2.8) we can see that

$$|||\mathscr{P}_1||| = |||\mathscr{P}_2||| = 1. \tag{3.2}$$

*3.2. The QR Decomposition*

Every matrix $A$ can be factorized as $A = QR$, where $Q \in \mathbf{U}$ and $R \in \mathbf{\Delta}$. If $A$ is nonsingular, so must be $R$. Now if we restrict $R$ to lie in $\mathbf{\Delta}_+$, then the decomposition $A = QR$ is unique for every $A \in \mathbf{GL}$. We have thus an invertible map $\Phi : \mathbf{GL} \to \mathbf{U} \times \mathbf{\Delta}_+$. The tangent spaces to $\mathbf{U}$ and $\mathbf{\Delta}_+$ give a direct sum decomposition

$$\mathbf{M} = \mathbf{Sherm} + \mathbf{\Delta}_{\text{re}}. \tag{3.3}$$

The projection operators $\mathscr{P}_1, \mathscr{P}_2$ with respect to this decomposition can be written down explicitly, and we have for the Frobenius norm

$$\|\mathscr{P}_1\|_F = \|\mathscr{P}_2\|_F = \sqrt{2}. \tag{3.4}$$

Details of this may be found in Bhatia and Mukherjea (1992). It is also explained there why for the operator bound norm $\|\mathscr{P}_1\|$ and $\|\mathscr{P}_2\|$ grow as $\log n$. The cause of the different behavior is that for the triangular truncation operator $\mathscr{T}$ on $\mathbf{M}$ ($\mathscr{T}$ takes a matrix to its upper triangular part), $|||\mathscr{T}|||$ grows with $n$ for some unitarily invariant norms. Unitarily invariant norms for which $|||\mathscr{T}|||$ does not depend on $n$ have been studied by operator theorists. See, e.g., Gohberg and Krein (1970), Kwapien and Pelczynski (1970), and Arazy (1978). For such norms $|||\mathscr{P}_1|||$ and $|||\mathscr{P}_2|||$ can also be bounded independently of $n$. However, these norms may not be of interest in numerical analysis.

*3.3. The LR Decomposition*

Let **SNS** denote the set of all *strongly nonsingular matrices* (Fiedler, 1986). These are matrices all whose leading principal minors are nonzero. Then **SNS** is a dense subset of **M**; it is not a group, however. Every element $A$ of **SNS** has a unique factorization

$$A = LR, \qquad \text{where} \quad L \in \mathbf{\Delta}_1^*, \quad R \in \mathbf{\Delta}_{\text{ns}}.$$

We thus have an invertible map $\Phi$ from **SNS** to $\mathbf{\Delta}_1^* \times \mathbf{\Delta}_{\text{ns}}$. The tangent spaces to these two manifolds give a direct sum decomposition

$$\mathbf{M} = \mathbf{\Delta}_0^* + \mathbf{\Delta}. \tag{3.5}$$

Obviously, for the projection operators $\mathscr{P}_1$ and $\mathscr{P}_2$ for this decomposition we have

$$\|\mathscr{P}_1\|_F = \|\mathscr{P}_2\|_F = 1. \tag{3.6}$$

As explained earlier, in this case too, $\|\mathscr{P}_1\|$ and $\|\mathscr{P}_2\|$ will grow as $\log n$. However, it is worth noting here that in this case the equality (3.6) will hold for several other norms which are not unitarily invariant but which are of much use in numerical analysis. For example, it will hold for the maximum row sum norm and the maximum column sum norm.

*3.4. The SR Decomposition*

In this subsection $n$ is an even integer $n = 2r$ and all matrices are real. We will be considering matrices of order $n$ and of order $r$. To distinguish between them we will use symbols such as $\mathbf{M}(n)$ to denote all $n \times n$ matrices and $\mathbf{\Delta}(r)$ to denote all $r \times r$ upper triangular matrices. Let $I_r$ be the identity matrix of size $r$, and let

$$J = \begin{bmatrix} 0 & I_r \\ -I_r & 0 \end{bmatrix}.$$

A matrix $S \in \mathbf{GL}(n)$ is called *symplectic* if

$$S^T J S = J. \tag{3.7}$$

The set of all such matrices forms a group called the (*real*) *symplectic group*. We will denote this group as $\mathbf{Symp}(n)$. Consider also all matrices $G$ of the following special form:

$$G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix},$$

in which

$$G_{11}, G_{22} \in \mathbf{\Delta}_{ns}(r); \qquad G_{12}, G_{21} \in \mathbf{\Delta}_0(r); \qquad \operatorname{diag}(G_{11}) = \operatorname{diag}(G_{22}). \tag{3.8}$$

Such matrices form another subgroup of $\mathrm{GL}(n)$, which we will denote as $\mathbf{Cosymp}(n)$. Let $P \in \mathbf{M}(n)$ be the matrix of the *perfect shuffle*, i.e., the permutation which permutes the rows by sending the $k$th row to the $(2k-1)$th one if $k \leqslant r$ and to the $(2k-2r)$th one if $k > r$. Let $\mathbf{A}$ be the subset of $\mathbf{GL}(n)$ consisting of all matrices $A$ for which all even principal

minors of $PA^TJPA^T$ are nonzero. Then $\mathbf{A}$ is a dense open set in $\mathbf{M}(n)$. It can be shown that every element $A$ of $\mathbf{A}$ can be factorized uniquely as

$$A = SR, \qquad \text{where} \quad S \in \mathbf{Symp}(n), \quad R \in \mathbf{Cosymp}(n). \tag{3.9}$$

This is called the *SR decomposition*. For more details see Bunse-Gerstner (1986) and Watkins and Elsner (1988).

From (3.7) we can see, using standard arguments of Lie theory, that the Lie algebra corresponding to the symplectic group consists of all those matrices $X$ for which

$$(JX)^T = JX. \tag{3.10}$$

This condition gives the following special block decomposition for $X$:

$$X = \begin{bmatrix} B & C \\ D & -B^T \end{bmatrix}, \qquad B \in \mathbf{M}(r), \quad C, D \in \mathbf{Symm}(r), \tag{3.11}$$

where **Symm** stands for the set of all symmetric matrices. Such matrices $X$ are called *Hamiltonian*, and we will denote their collection by $\mathbf{Ham}(n)$.

From (3.8) one can see that the Lie algebra for the group $\mathbf{Cosymp}(n)$ consists of all matrices of the form

$$Y = \begin{bmatrix} K & L \\ M & N \end{bmatrix}$$

in which

$$K, N \in \mathbf{\Delta}(r); \qquad L, M \in \mathbf{\Delta}_0(r); \qquad \operatorname{diag}(K) = \operatorname{diag}(N). \tag{3.12}$$

This Lie algebra will be denoted as $\mathbf{Coham}(n)$. Note that this set is closed under matrix multiplication.

Thus the tangent spaces to the two spaces involved in the $SR$ factorization give a vector space decomposition

$$\mathbf{M}(n) = \mathbf{Ham}(n) + \mathbf{Coham}(n). \tag{3.13}$$

We will now determine the projection operators corresponding to this decomposition. Let $A \in \mathbf{M}(n)$ have the block decomposition

$$A = \begin{bmatrix} E & F \\ P & Q \end{bmatrix},$$

and let $A = X + Y$, where $X, Y$ are as in (3.11) and (3.12). Let lower$(E)$ and upper$(E)$ denote the parts of $E$ below the main diagonal and above the main diagonal, respectively. Then we must have

$$B = \tfrac{1}{2}\,\mathrm{diag}(E - Q) + \mathrm{lower}(E) - \mathrm{upper}(Q^T),$$

$$K = \tfrac{1}{2}\,\mathrm{diag}(E + Q) + \mathrm{upper}(E) + \mathrm{upper}(Q^T),$$

$$N = \tfrac{1}{2}\,\mathrm{diag}(E + Q) + \mathrm{upper}(E^T) + \mathrm{upper}(Q).$$

Then entries of the matrices $C$ and $L$ are obtained from those of $F$ as follows:

$$c_{ii} = f_{ii}, \quad l_{ii} = 0 \qquad \text{for all } i,$$

$$c_{ij} = f_{ji}, \quad l_{ij} = f_{ij} - f_{ji} \qquad \text{for } \; j > i,$$

$$c_{ij} = f_{ij}, \quad l_{ij} = 0 \qquad \text{for } \; j < i.$$

In the same way, the entries of $D$ and $M$ are obtained from those of $P$ as

$$d_{ii} = p_{ii}, \quad m_{ii} = 0 \qquad \text{for all } i,$$

$$d_{ij} = p_{ji}, \quad m_{ij} = p_{ij} - p_{ji} \qquad \text{for } \; j > i,$$

$$d_{ij} = p_{ij}, \quad m_{ij} = 0 \qquad \text{for } \; j < i.$$

From these relations we obtain

$$\|B\|_F^2 + \|B^T\|_F^2 \leqslant \|E\|_F^2 + \|Q\|_F^2,$$

$$\|K\|_F^2 + \|N\|_F^2 \leqslant \|E\|_F^2 + \|Q\|_F^2,$$

$$\|C\|_F^2 \leqslant 2\|F\|_F^2, \qquad \|L\|_F^2 \leqslant 2\|F\|_F^2,$$

$$\|D\|_F^2 \leqslant 2\|P\|_F^2, \qquad \|M\|_F^2 \leqslant 2\|P\|_F^2.$$

Hence,

$$\|X\|_F^2 \leqslant 2\|A\|_F^2, \qquad \|Y\|_F^2 \leqslant 2\|A\|_F^2. \tag{3.14}$$

Both these inequalities are sharp. For example, for $r = 2$, the first one is seen to be an equality for the choice

$$E = P = Q = 0, \qquad F = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix};$$

the second one is an equality for the choice

$$E = P = Q = 0, \qquad F = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

Thus, we conclude that for the complementary projections $\mathscr{P}_1, \mathscr{P}_2$ for the decomposition (3.13) we have

$$\|\mathscr{P}_1\|_F = \|\mathscr{P}_2\|_F = \sqrt{2}\,. \tag{3.15}$$

The *SR* decomposition is related to the *SR* algorithm for computing the eigenvalues of Hamiltonian matrices. Such eigenvalue problems arise in solving the algebraic Riccati equation in control theory (see Bunse-Gerstner and Mehrmann, 1986) and in some economics problems (see Medio, 1987). Finally we remark that though these are complex analogues of the groups **Symp**($n$) and **Cosymp**($n$), there is no complex version of the *SR* decomposition.

### 3.5. *The HR Decomposition*

Let $J$ be a given diagonal matrix whose diagonal entries are $\pm 1$. A matrix $H$ is called *J-unitary* if

$$H^*JH = J. \tag{3.16}$$

The set of all such matrices is a subgroup of **GL**; we will denote it by **JU**. Let **A** be the set of all those matrices $A$ for which the leading principal minors of $A^*JA$ have the same signs as the corresponding minors of $J$. Such matrices form a neighborhood of $I$ in **M**. Every element $A$ of **A** has a decomposition

$$A = HR, \qquad \text{where} \quad H \in \mathbf{JU}, \quad R \in \mathbf{\Delta}_+. \tag{3.17}$$

See Bunse-Gerstner (1981) and Watkins and Elsner (1988). This is called the *HR decomposition.*

From (3.16) one can see that the Lie algebra corresponding to the group **JU** consists of all matrices $X$ such that

$$(JX)^* = -JX. \tag{3.18}$$

Such matrices are called *J-skew-Hermitian*, and we denote their collection by **JS**.

These matrices can also be characterized as follows. Partition the set $\{1, 2, \ldots, n\}$ into $\mathscr{S}$ and $\mathscr{S}'$ in such a way that $j_{ii} = 1$ if $i \in \mathscr{S}$ and $j_{ii} = -1$ if $i \in \mathscr{S}'$. Then $X \in \mathbf{JS}$ iff all its diagonal entries are purely imaginary, $\bar{x}_{ij} = x_{ji}$ if only one of $i$ and $j$ is in $\mathscr{S}$, and $\bar{x}_{ij} = -x_{ji}$ otherwise. Corresponding to the $HR$ decomposition we thus have a vector space decomposition

$$\mathbf{M} = \mathbf{JS} + \mathbf{\Delta}_{\mathrm{re}}. \tag{3.19}$$

If a matrix $A$ splits as $A = X + Y$ in this decomposition, then we have

$$x_{ii} = \operatorname{Im} a_{ii}, \quad y_{ii} = \operatorname{Re} a_{ii} \qquad \text{for all } i,$$

$$x_{ij} = a_{ij}, \quad y_{ij} = 0 \qquad \text{for} \quad i > j,$$

$$x_{ij} = \bar{a}_{ji}, \quad y_{ij} = a_{ij} - \bar{a}_{ji} \qquad \text{if} \quad i < j \text{ and only one of } i \text{ and } j \text{ is in } \mathscr{S},$$

$$x_{ij} = -\bar{a}_{ji}, \quad y_{ij} = a_{ij} + \bar{a}_{ji} \qquad \text{otherwise.}$$

It is clear then that $\|X\|_F^2 \leqslant 2\|A\|_F^2$ and $\|Y\|_F^2 \leqslant 2\|A\|_F^2$ and that these inequalities are sharp. Hence, we have

$$\|\mathscr{P}_1\|_F = \|\mathscr{P}_2\|_F = \sqrt{2} \tag{3.20}$$

for the complementary projection operators corresponding to the decomposition (3.19).

### 3.6. *The Second Polar Decomposition*

A complex matrix $A$ is called *orthogonal* if $A^TA = I$, and *skew-symmetric* if $A^T = -A$. We will denote the collections of such matrices as **Orth** and **Ssym**, respectively. The set **Orth** is a Lie group, and its Lie algebra is **Ssym**. If $A \in \mathbf{GL}$, then $A$ can be factorized as $A = A_1 A_2$ where $A_1$ is orthogonal and $A_2$ is symmetric; see Gantmacher (1959, Chapter XI). The factor $A_2$ is a

square root of the matrix $A^TA$. We can choose $A_2$ so that the argument of each of its eigenvalues lies in the interval $(-\pi/2, \pi/2]$. With this restriction the above factorization is unique; see de Bruijn and Szekeres (1955). For our analysis we will restrict $A$ a little more. Let $\mathbf{A}$ be the class of all matrices $A$ such that no eigenvalue of $A^TA$ is on the negative real axis $(-\infty, 0]$. Then $A^TA$ has a unique square root all whose eigenvalues are in the open right half plane. Let $\mathbf{Symm}^+$ be the set of all symmetric matrices whose eigenvalues are in the open right half plane. Then for each $A \in \mathbf{A}$ we have a unique decomposition

$$A = A_1 A_2 \qquad \text{where} \quad A_1 \in \mathbf{Orth}, \quad A_2 \in \mathbf{Symm}^+. \tag{3.21}$$

We will call this the *second polar decomposition*. The set $\mathbf{Symm}^+$ is an open subset of the linear space **Symm**, and so the tangent space at any point of $\mathbf{Symm}^+$ is the space **Symm**. Thus corresponding to the decomposition (3.21) we have the vector space decomposition

$$\mathbf{M} = \mathbf{Ssym} + \mathbf{Symm}. \tag{3.22}$$

If $\mathscr{P}_1$ and $\mathscr{P}_2$ are the complementary projection operators for the above decomposition then it follows from (2.8) that

$$|||\mathscr{P}_1||| = |||\mathscr{P}_2||| = 1 \tag{3.23}$$

for every unitarily invariant norm.

### 3.7. *The Third Polar Decomposition*

Let $\bar{A}$ be the matrix obtained from $A$ by replacing each entry with its complex conjugate. A matrix $A$ is called *circular* if $\bar{A}A = I$. Let **Circ** be the collection of all such matrices. Let $\mathbf{M}_{\text{re}}$, $\mathbf{M}_{\text{im}}$, and $\mathbf{GL}_{\text{re}}$ denote the sets of all real matrices, all imaginary matrices, and all nonsingular real matrices, respectively. Each matrix $A \in \mathbf{GL}$ can be factorized as $A = RC$ where $R$ is a real matrix and $C$ is a circular matrix. See de Bruijn and Szekeres (1955), where this decomposition seems to have been discovered, Mehta (1989, p. 90), or Horn and Johnson (1991, p. 481). The matrix $C$ is a square root of the circular matrix $(\bar{A})^{-1}A$. If we restrict $C$ so that all its eigenvalues have their arguments in the interval $(-\pi/2, \pi/2]$, then the above factorization is unique. Let $\mathbf{A}$ be the set of all matrices $A$ such that $(\bar{A})^{-1}A$ has no eigenvalue on $(-\infty, 0]$, and let $\mathbf{Circ}^+$ be the set of all circular matrices whose eigenvalues are in the open right half plane. Then for each $A \in \mathbf{A}$ we have a

unique decomposition

$$A = RC \qquad \text{where} \quad R \in \mathbf{GL}_{\text{re}}, \quad C \in \mathbf{Circ}^{+}. \tag{3.24}$$

We will call this the *third polar decomposition.*

Note that the set **Circ** is not a subgroup of **GL.** However, it is a differentiable manifold, and $\mathbf{Circ}^{+}$ is an open subset in it. To determine the tangent space to **Circ** at the point $I$, consider any smooth curve $X(t)$ in **Circ** with $X(0) = I$. Then differentiate the product $\overline{X(t)}X(t)$ at $t = 0$ to see that the matrix $X'(0)$ must be in $\mathbf{M}_{\text{im}}$. This space, therefore, is the tangent space to **Circ** at the point $I$. This is related to the fact that if $J$ is an imaginary matrix then $\exp J$ is circular. To calculate the tangent space at any other point, consider a curve $X(t)$ in **Circ** with $X(0) = C$. The same argument shows that if $Y = X'(0)$ then $Y$ must satisfy the equation

$$\overline{Y}C + \overline{C}Y = 0. \tag{3.25}$$

Let $C^{1/2}$ be the circular matrix which is a square root of $C$ and whose eigenvalues have arguments in the interval $(-\pi/2, \pi/2]$. Then for every $J$ in $\mathbf{M}_{\text{im}}$ the matrix $Y = C^{1/2}JC^{1/2}$ satisfies the equation (3.25). A count of the dimensions then shows that the tangent space to **Circ** at $C$ is the real linear space $C^{1/2}\mathbf{M}_{\text{im}}C^{1/2}$. The same is true for the open subset $\mathbf{Circ}^{+}$.

Corresponding to this decomposition we have the vector space decomposition

$$\mathbf{M} = \mathbf{M}_{\text{re}} + \mathbf{M}_{\text{im}}. \tag{3.26}$$

In this decomposition a matrix $A$ splits as $A = A_{\text{re}} + A_{\text{im}}$, where $A_{\text{re}} = \frac{1}{2}(A + \overline{A})$ and $A_{\text{im}} = \frac{1}{2}(A - \overline{A})$. If $\mathscr{P}_1$ and $\mathscr{P}_2$ are the complementary projection operators with respect to this decomposition, then it follows from (2.8) that

$$|||\mathscr{P}_1||| = |||\mathscr{P}_2||| = 1 \tag{3.27}$$

for every unitarily invariant norm.

### 3.6. *The Cholesky Decomposition*

Every positive definite matrix $A$ can be factorized uniquely as $A = R^*R$ where $R$ is upper triangular with positive entries. We thus have an invertible map $\Phi : \mathbf{P} \to \mathbf{\Delta}_+$, where $\Phi(A) = R$. This situation is different from the earlier ones. First, $\mathbf{P}$ is not an open subset of $\mathbf{M}$. However, it is an open

subset of the linear space **Herm**. Second, now the two factors are not independent; it is enough to look at one of them. However, the same general method works here. We have the inverse map $\Psi : \mathbf{\Delta}_+ \to \mathbf{P}$. The derivative $D\Phi(A)$ is a linear map from **Herm** to $\mathbf{\Delta}_{re}$. To study this it is easier to study the map $D\Psi(R)$.

## 4. PERTURBATION BOUNDS

Following Stewart and Sun (1990), we will adopt the following notation. The matrix $\tilde{A}$ will represent a perturbation of $A$. If $A = A_1 A_2$ is a factorization of $A$, then $\tilde{A} = \tilde{A}_1 \tilde{A}_2$, will be the corresponding factorization of $\tilde{A}$. In the numerical analysis literature several *first order perturbation bounds* are expressed in the form

$$\frac{\|\tilde{A}_1 - A_1\|}{\|A_1\|} \lesssim f(\|A\|) \frac{\|\tilde{A} - A\|}{\|A\|}, \tag{4.1}$$

which is a symbolic notation for the inequality

$$\frac{\|\tilde{A}_1 - A_1\|}{\|A_1\|} \leqslant f(\|A\|) \frac{\|\tilde{A} - A\|}{\|A\|} + O(\|\tilde{A} - A\|^2), \tag{4.2}$$

valid for $\tilde{A}$ in a neighborhood of $A$.

The function $f$ is to be determined explicitly. We will do this for the decompositions listed in Section 3.

We will reserve the notation $\text{cond}(A)$ for the condition number

$$\text{cond}(A) = \|A\| \, \|A^{-1}\|. \tag{4.3}$$

### 4.1. *The Polar Decomposition*

THEOREM 4.1. *Let* $\Phi : \mathbf{GL} \to \mathbf{U} \times \mathbf{P}$ *be the map* $\Phi(A) = (\Phi_1(A), \Phi_2(A)) = (U, P)$ *defined by the polar decomposition. Then for every unitarily invariant norm the derivatives* $D\Phi_1(A)$ *and* $D\Phi_2(A)$ *are bounded as*

$$|||D\Phi_1(A)||| = \|A^{-1}\|, \tag{4.4}$$

$$|||D\Phi_2(A)||| \leqslant 1 + \text{cond}(A). \tag{4.5}$$

*Proof.* The first part was proved in Bhatia and Mukherjea (1992). For the reader's convenience we indicate the proof. We have noted in Section 2 that if for $X \in \mathbf{M}$ one has $D\Phi(A)(UX) = (UM, N)$ where $M \in \mathbf{Sherm}$ and $N \in \mathbf{Herm}$, then the equation (2.7) must be satisfied. Subtracting from this equation its adjoint, we get

$$MP + PM = X - X^* =: 2i \operatorname{Im} X. \tag{4.6}$$

This equation has an explicit solution

$$M = 2i\int_0^\infty e^{-tP} \operatorname{Im} X e^{-tP}\, dt. \tag{4.7}$$

See, e.g., Bhatia, Davis, and McIntosh (1983) or Lancaster and Tismenetsky (1985, p. 144). From this one easily obtains, using (2.8) and (2.9),

$$|||M||| \leq \|A^{-1}\|\, |||\operatorname{Im} X||| \leq \|A^{-1}\|\, |||X|||. \tag{4.8}$$

By definition,

$$|||D\Phi_1(A)||| = \sup\{|||D\Phi_1(A)(X)||| : |||X||| = 1\} \leq \|A^{-1}\|.$$

It is shown in the paper mentioned above that equality holds here.

From (2.7) we also obtain

$$\begin{aligned} |||N||| &\leq |||X||| + |||MP||| \leq |||X||| + |||M|||\, \|P\| \\ &\leq |||X||| + \|A^{-1}\|\, |||X|||\, \|P\| \\ &= [1 + \operatorname{cond}(A)]\, |||X|||. \end{aligned}$$

This proves (4.5). ∎

COROLLARY 4.1. *For each $A = UP$ in* **GL** *we have the first order perturbation bounds*

$$|||\tilde{U} - U||| \lesssim \|A^{-1}\|\, |||\tilde{A} - A||| = \operatorname{cond}(A)\, \frac{|||\tilde{A} - A|||}{\|A\|}, \tag{4.9}$$

$$\frac{|||\tilde{P} - P|||}{|||P|||} \lesssim [1 + \operatorname{cond}(A)]\, \frac{|||\tilde{A} - A|||}{|||A|||}. \tag{4.10}$$

*Proof.* By Taylor's formula

$$\tilde{U} = U + D\Phi_1(A)(\tilde{A} - A) + \text{higher order terms.}$$

So (4.9) is a consequence of (4.4); and by the same argument (4.10) follows from (4.5). ■

From (2.7) we have $N = \operatorname{Re} X + \frac{1}{2}(MP - PM)$. We have an explicit expression for $M$ in (4.7). Hence we can also write an explicit expression for $N$.

For the Frobenius norm (4.4) was proved also in Kenny and Laub (1991, Theorem 2.2), and for all unitarily invariant norms in Mathias (1991b, Corollary 3.4). A weaker result was proved in Barrlund (1989). For real matrices a little stronger results have been obtained by these authors.

Also, for the Frobenius norm the factor $1 + \operatorname{cond}(A)$ occurring in (4.10) can be replaced by the smaller quantity $\sqrt{2}$. This has been observed in Higham (1986, Theorem 2.5), Barrlund (1989, Theorem 2.6), and Mathias (1991b, Theorem 5.1). We will discuss this again in Section 5. For other unitarily invariant norms the bound (4.10) seems to be new.

### 4.2. *The QR Decomposition*

THEOREM 4.2. *Let* $\Phi : \mathbf{GL} \to \mathbf{U} \times \mathbf{\Delta}_+$ *be the map* $\Phi(A) = (\Phi_1(A), \Phi_2(A)) = (Q, R)$ *defined by the QR decomposition. Then*

$$\|D\Phi_1(A)\|_F \leqslant \sqrt{2}\,\|A^{-1}\|, \tag{4.11}$$

$$\|D\Phi_2(A)\|_F \leqslant \sqrt{2}\,\operatorname{cond}(A). \tag{4.12}$$

*Proof.* As in the proof of Theorem 4.1, one finds that if for $X \in \mathbf{M}$ one has $D\Phi(A)(QX) = (QM, N)$, where $M \in \mathbf{Sherm}$ and $N \in \mathbf{\Delta}_{\mathrm{re}}$, then we must have

$$X = MR + N.$$

This equation is similar to (2.7). The fact that $\mathbf{\Delta}_{\mathrm{re}}$ is closed under multiplication allows us to estimate $M$ and $N$ from this equation. Write this as

$$XR^{-1} = M + NR^{-1}$$

Then note that $M \in \mathbf{Sherm}$ and $NR^{-1} \in \mathbf{\Delta}_{re}$. So we have from (3.4)

$$\|M\|_F \leqslant \sqrt{2}\,\|XR^{-1}\|_F \leqslant \|X\|_F\|R^{-1}\| = \sqrt{2}\,\|X\|_F\|A^{-1}\|,$$

$$\|NR^{-1}\|_F \leqslant \sqrt{2}\,\|XR^{-1}\|_F \leqslant \sqrt{2}\,\|X\|_F\|R^{-1}\|,$$

and hence,

$$\|N\|_F \leqslant \|NR^{-1}\|_F\|R\| \leqslant \sqrt{2}\,\|X\|_F\|R^{-1}\|\,\|R\| = \sqrt{2}\ \mathrm{cond}(A).$$

These inequalities lead to (4.11) and (4.12). ■

COROLLARY 4.2. *For each $A = QR$ in* **GL** *we have*

$$\|\tilde{Q} - Q\|_F \lesssim \sqrt{2}\,\|A^{-1}\|\,\|\tilde{A} - A\|_F, \tag{4.13}$$

$$\|\tilde{R} - R\|_F \lesssim \sqrt{2}\ \mathrm{cond}(A)\,\|\tilde{A} - A\|_F. \tag{4.14}$$

The inequality (4.11) was proved in Bhatia and Mukherjea (1992). The first work on the perturbation analysis of the $QR$ decomposition is Stewart (1977). Two recent papers are Sun (1991) and Stewart (1992). For a comparison with results of these authors write (4.13) as

$$\|\tilde{Q} - Q\|_F \lesssim \sqrt{2}\ \mathrm{cond}(A)\,\frac{\|\tilde{A} - A\|_F}{\|A\|}. \tag{4.15}$$

This is stronger than the bounds (3.2) of Stewart (1992) and (1.14) of Sun (1991). The bound (4.14) is exactly the same as obtained in Sun (1991, Theorem 1.5) and in Stewart (1992). However, both Stewart and Sun consider rectangular matrices, and thus their results are more generally applicable.

### *4.3. The LR Decomposition*

THEOREM 4.3. *Let $\Phi : \mathbf{SNS} \to \mathbf{\Delta}_1^* \times \mathbf{\Delta}_{ns}$ be the map $\Phi(A) = (\Phi_1(A), \Phi_2(A)) = (L, R)$ given by the LR decomposition. Then*

$$\|D\Phi_1(A)\|_F \leqslant \mathrm{cond}(L)\,\|R^{-1}\|, \tag{4.16}$$

$$\|D\Phi_2(A)\|_F \leqslant \mathrm{cond}(R)\,\|L^{-1}\|. \tag{4.17}$$

*Proof.* There are two little differences between this and the earlier situations. First, none of the factors is unitary, and so there are fewer cancellations while calculating norms. Second, the tangent space to $\mathbf{\Delta}_1^*$ at any point $L$ is $L \cdot \mathbf{\Delta}_0^* = \mathbf{\Delta}_0^*$. Thus the derivative of the inverse map $\Psi$ is calculated as

$$D\Psi(L, R)(K, T) = \left[\frac{d}{dt}\Psi(L + tK, R + tT)\right]_{t=0} = LT + KR$$

for all $K \in \mathbf{\Delta}_0^*$, $T \in \mathbf{\Delta}$. So if for $X \in \mathbf{M}$ we have

$$D\Phi(A)(X) = (M, N), \qquad \text{where} \quad M \in \mathbf{\Delta}_0^*, \quad N \in \mathbf{\Delta},$$

then we must have

$$X = LN + MR. \tag{4.18}$$

Write this as

$$L^{-1}XR^{-1} = NR^{-1} + L^{-1}M;$$

then note that $NR^{-1} \in \mathbf{\Delta}$ and $L^{-1}M \in \mathbf{\Delta}_0^*$. Hence, by (3.6),

$$\|L^{-1}M\|_F \leqslant \|L^{-1}XR^{-1}\|_F \leqslant \|L^{-1}\|\,\|X\|_F\|R^{-1}\|$$

and hence

$$\|M\|_F \leqslant \|L\|\,\|L^{-1}\|\,\|X\|_F\|R^{-1}\|.$$

This gives (4.16). The inequality (4.17) has the same proof. ∎

COROLLARY 4.3. *For each $A = LR$ in* **SNS** *we have*

$$\|\tilde{L} - L\|_F \lesssim \text{cond}(L)\,\|R^{-1}\|\,\|\tilde{A} - A\|_F, \tag{4.19}$$

$$\|\tilde{R} - R\|_F \lesssim \text{cond}(R)\,\|L^{-1}\|\,\|\tilde{A} - A\|_F. \tag{4.20}$$

We could also write (4.19) as

$$\frac{\|\tilde{L} - L\|_F}{\|L\|} \lesssim \|L^{-1}\|\,\|R^{-1}\|\,\|A\|\frac{\|\tilde{A} - A\|_F}{\|A\|}.$$

If instead of (2.9) we had used the inequality $\|XY\|_F \leq \|X\|_F\|Y\|_F$, valid for any $X, Y$, we would have obtained instead

$$\frac{\|\tilde{L} - L\|_F}{\|L\|_F} \lesssim \|L^{-1}\|_F\|R^{-1}\|_F\|A\|_F\frac{\|\tilde{A} - A\|_F}{\|A\|_F}.$$

The same would hold for any norm which is submultiplicative and for which (3.6) holds. This is exactly the result in Stewart [1992, (3.1)].

*4.4. The SR Decomposition*

THEOREM 4.4. *Let* $\Phi : \mathbf{A} \to \mathbf{Symp} \times \mathbf{Cosymp}$ *be the map* $\Phi(A) = (\Phi_1(A), \Phi_2(A))$ *defined by the SR decomposition. Then*

$$\|D\Phi_1(A)\|_F \leq \sqrt{2}\ \mathrm{cond}(S)\,\|R^{-1}\|, \tag{4.21}$$

$$\|D\Phi_2(A)\|_F \leq \sqrt{2}\|S^{-1}\|\mathrm{cond}(R). \tag{4.22}$$

*Proof.* The derivative $D\Phi(A)$ maps $\mathbf{M}$ into $S \cdot \mathbf{Ham} + \mathbf{Coham}$. If $D\Phi(A)(X) = (SM, N)$ where $M \in \mathbf{Ham}$, $N \in \mathbf{Coham}$, then one can see, as in the proof of Theorem 4.1, that

$$X = SMR + SN.$$

So we have

$$S^{-1}XR^{-1} = M + NR^{-1}, \qquad M \in \mathbf{Ham}, \quad NR^{-1} \in \mathbf{Coham}.$$

Hence, by (3.15)

$$\|M\|_F \leq \sqrt{2}\|S^{-1}XR^{-1}\|_F \leq \sqrt{2}\|S^{-1}\|\,\|X\|_F\|R^{-1}\|.$$

Hence

$$\|SM\|_F \leq \sqrt{2}\ \mathrm{cond}(S)\,\|R^{-1}\|\,\|X\|_F,$$

which gives (4.21). In the same way, we obtain (4.22). ∎

COROLLARY 4.4. *For each $A = SR$ in $\mathbf{A}$ we have*

$$\|\tilde{S} - S\|_F \lesssim \sqrt{2}\ \mathrm{cond}(S)\,\|R^{-1}\|\,\|\tilde{A} - A\|_F, \tag{4.23}$$

$$\|\tilde{R} - R\|_F \lesssim \sqrt{2}\|S^{-1}\|\mathrm{cond}(R)\,\|\tilde{A} - A\|_F. \tag{4.24}$$

*4.5. The HR Decomposition*

THEOREM 4.5. *Let $\Phi : \mathbf{A} \to \mathbf{JU} \times \mathbf{\Delta}_+$ be the map $\Phi(A) = (\Phi_1(A), \Phi_2(A)) = (H, R)$ defined by the HR decomposition. Then*

$$\|D\Phi_1(A)\|_F \leqslant \sqrt{2}\ \mathrm{cond}(H)\,\|R^{-1}\|, \tag{4.25}$$

$$\|D\Phi_2(A)\|_F \leqslant \sqrt{2}\|H^{-1}\|\mathrm{cond}(R). \tag{4.26}$$

COROLLARY 4.5. *Let $A = HR$ be any element of $\mathbf{A}$. Then*

$$\|\tilde{H} - H\|_F \lesssim \sqrt{2}\ \mathrm{cond}(H)\,\|R^{-1}\|\,\|\tilde{A} - A\|_F, \tag{4.27}$$

$$\|\tilde{R} - R\|_F \lesssim \sqrt{2}\|H^{-1}\|\mathrm{cond}(R)\,\|\tilde{A} - A\|_F. \tag{4.28}$$

The proofs are exactly the same as those for the $SR$ decomposition.

*4.6. The Second Polar Decomposition*

THEOREM 4.6. *Let $\Phi : \mathbf{A} \to \mathbf{Orth} \times \mathbf{Symm}^+$ be the map $\Phi(A) = (\Phi_1(A), \Phi_2(A)) = (A_1, A_2)$ defined by the second polar decomposition. Let*

$$\gamma(A_2) = \int_0^\infty \|e^{-tA_2}\|^2\,dt. \tag{4.29}$$

*Then for every unitarily invariant norm,*

$$|||D\Phi_1(A)||| \leqslant 2\,\mathrm{cond}(A_1)\,\gamma(A_2), \tag{4.30}$$

$$|||D\Phi_2(A)||| \leqslant \|A_1^{-1}\|\big[1 + 2\gamma(A_2)\|A_2\|\big]. \tag{4.31}$$

*Proof.* The derivative $D\Phi(A)$ is a linear map from **M** into $A_1 \cdot$ **Ssym** + **Symm**. Let $D\Phi(A)(X) = (A_1M, N)$, where $M \in$ **Ssym**, $N \in$ **Symm**. Then one can see, as in earlier proofs, that

$$X = A_1MA_2 + A_1N. \tag{4.32}$$

From this equation one obtains

$$A_1^{-1}X - \left(A_1^{-1}X\right)^T = MA_2 + A_2M.$$

Since the eigenvalues of $A_2$ are all in the open right half plane, this equation can be solved for $M$. The solution is unique and is given by

$$M = \int_0^\infty e^{-tA_2}\left[A_1^{-1}X - \left(A_1^{-1}X\right)^T\right]e^{-tA_2}\,dt.$$

See Lancaster and Tismenetsky (1985, p. 414). From this, using (2.8) and (2.9), we get

$$|||M||| \leq 2\|A_1^{-1}\|\,|||X|||\gamma(A_2), \tag{4.33}$$

and hence,

$$|||AM||| \leq 2\,\mathrm{cond}(A_1)\,\gamma(A_2)|||X|||.$$

Taking the supremum over all $X$ with $|||X||| = 1$, one obtains (4.30). From (4.32) and (4.33) we also get

$$|||N||| \leq \|A_1^{-1}\|\left[1 + 2\gamma(A_2)\|A_2\|\right]|||X|||.$$

This leads to (4.31). ∎

COROLLARY 4.6. *Let $A = A_1A_2$ be the second polar decomposition of any matrix $A \in$ **A**. Then*

$$|||\tilde{A}_1 - A_1||| \leq 2\,\mathrm{cond}(A_1)\,\gamma(A_2)|||\tilde{A} - A|||, \tag{4.34}$$

$$|||\tilde{A}_2 - A_2||| \leq \|A_1^{-1}\|\left[1 + 2\gamma(A_2)\|A_2\|\right]|||\tilde{A} - A|||. \tag{4.35}$$

Convenient bounds for $\gamma(A_2)$ are known under additional conditions. For example, if the quantity $\delta(A_2)$ defined as the minimum eigenvalue of $\frac{1}{2}(A_2 + A_2^*)$ is positive, then

$$\|e^{-tA_2}\| \leqslant e^{-\delta(A_2)t}.$$

See, e.g., Bhatia, Davis, and McIntosh (1983, p. 53). Hence in this case

$$\gamma(A_2) \leqslant \frac{1}{2\delta(A_2)}.$$

The condition $\delta(A_2) > 0$ says that not only the spectrum but also the numerical range of $A_2$ lies in the open right half plane. The matrix $A_2$ is then called *accretive*.

### 4.7. *The Third Polar Decomposition*

THEOREM 4.7. *Let* $\Phi : \mathbf{A} \to \mathbf{GL}_{\mathrm{re}} \times \mathbf{Circ}^+$ *be the map* $\Phi(A) = (\Phi_1(A), \Phi_2(A)) = RC$ *given by the third polar decomposition. Let*

$$\gamma(C) = \int_0^\infty \|e^{-tC}\|^2 \, dt.$$

*Then for every unitarily invariant norm*

$$|||D\Phi_1(A)||| \leqslant \|C^{-1}\|[1 + 2\,\mathrm{cond}(R)\,\gamma(C)\|C\|], \tag{4.36}$$

$$|||D\Phi_2(A)||| \leqslant 2\|R^{-1}\|\,\|C\|\gamma(C). \tag{4.37}$$

*Proof.* The derivative $D\Phi(A)$ is a linear map from $\mathbf{M}$ into $\mathbf{M}_{\mathrm{re}} + C^{1/2}\mathbf{M}_{\mathrm{im}}C^{1/2}$. Let $\Psi = \Phi^{-1}$, $\Psi(R, C) = RC = A$. Then for any $E \in \mathbf{M}_{\mathrm{re}}$ and $J \in \mathbf{M}_{\mathrm{im}}$ we have

$$D\Psi(RC)(E, C^{1/2}JC^{1/2}) = \left[\frac{d}{dt}\Psi(R + tE, C^{1/2}e^{tJ}C^{1/2})\right]_{t=0}$$

$$= EC + RC^{1/2}JC^{1/2}.$$

Hence, if

$$D\Phi(A)(X) = (N, C^{1/2}MC^{1/2}) \qquad \text{where} \quad N \in \mathbf{M}_{\text{re}},\ M \in \mathbf{M}_{\text{im}},$$

then

$$X = NC + RC^{1/2}MC^{1/2}. \tag{4.38}$$

Rewrite this as

$$R^{-1}XC^{-1} = R^{-1}N + C^{1/2}MC^{-1/2}.$$

Subtract from this equation its complex conjugate, remembering that $R$ and $N$ are real, $M$ is imaginary, and $C$ is circular. This gives

$$R^{-1}XC^{-1} - R^{-1}\bar{X}C = C^{1/2}MC^{1/2}C^{-1} + C^{-1}C^{1/2}MC^{1/2}.$$

This is the same kind of equation as in Section 4.6, and we have from this

$$C^{1/2}MC^{1/2} = \int_0^\infty e^{-t\bar{C}}\left[R^{-1}XC^{-1} - R^{-1}\bar{X}C\right]e^{-t\bar{C}}\,dt.$$

From this we get

$$|||C^{1/2}MC^{1/2}||| \leqslant 2\|R^{-1}\|\,\|C\|\gamma(C)|||X|||, \tag{4.39}$$

using (2.8), (2.9), and the fact that $C$ is circular. This gives the inequality (4.37). From (4.38) and (4.39) we get

$$|||NC||| \leqslant \left[1 + 2\,\mathrm{cond}(R)\,\gamma(C)\|C\|\right]|||X|||.$$

This leads to (4.36). ∎

COROLLARY 4.7. *Let $A = RC$ be the third polar decomposition of any matrix $A \in \mathbf{A}$. Then*

$$|||\tilde{R} - R||| \lesssim \|C^{-1}\|\left[1 + 2\,\mathrm{cond}(R)\,\gamma(C)\|C\|\right]|||\tilde{A} - A|||, \tag{4.40}$$

$$|||\tilde{C} - C||| \lesssim 2\|R^{-1}\|\,\|C\|\gamma(C)|||\tilde{A} - A|||. \tag{4.41}$$

As observed in Section 4.6, $\gamma(C)$ can be estimated conveniently when $C$ is accretive.

*4.8. The Cholesky Decomposition*

In Bhatia and Mukherjea (1992) we proved the following result using the same ideas as above:

THEOREM 4.8. *Let $\Phi : \mathbf{P} \to \mathbf{\Delta}_+$ be the map $\Phi(A) = R$ where $A = R^*R$ is the Cholesky factorization of $A$, and let $\Psi(R) = A$ be the inverse map. Then*

$$\|D\Phi(A)\|_F \leq \frac{1}{\sqrt{2}}\|A\|^{1/2}\|A^{-1}\|, \tag{4.42}$$

$$|||D\Psi(R)||| \leq 2\|R\|. \tag{4.43}$$

So we have the following

COROLLARY 4.8. *For any $A = R^*R$ in $\mathbf{P}$ we have*

$$\|\tilde{R} - R\|_F \leq \frac{1}{\sqrt{2}}\|A\|^{1/2}\|A^{-1}\|\,\|\tilde{A} - A\|_F, \tag{4.44}$$

$$|||\tilde{A} - A||| \leq 2\|R\|\,|||\tilde{R} - R|||. \tag{4.45}$$

Let us compare these bounds with those of other authors. Since $\|A\| = \|R\|^2$, we obtain from (4.44)

$$\frac{\|\tilde{R} - R\|_F}{\|R\|} \leq \frac{1}{\sqrt{2}}\,\mathrm{cond}(A)\,\frac{\|\tilde{A} - A\|_F}{\|A\|}.$$

This is the same as the bound in Sun [1991, (1.9)] and in Stewart (1992). Since $\|A\|_F \leq \|R\|\,\|R\|_F$, we also get from (4.44)

$$\frac{\|\tilde{R} - R\|_F}{\|R\|_F} \leq \frac{1}{\sqrt{2}}\,\mathrm{cond}(A)\,\frac{\|\tilde{A} - A\|_F}{\|A\|_F},$$

which is the same as the bound in Sun, [1991, (1.90)]. From (4.45) we obtain

$$\frac{|||\tilde{A} - A|||}{\|A\|} \leq 2\frac{|||\tilde{R} - R|||}{\|R\|}.$$

The Frobenius norm case of this is proved in Sun [1991, (1.9)]. Since $R^* = AR^{-1}$, we have $|||R||| \leq |||A|||\,\|R^{-1}\|$. Using this, we get from (4.45)

$$\frac{|||\tilde{A} - A|||}{|||A|||} \leq 2[\mathrm{cond}(A)]^{1/2} \frac{|||\tilde{R} - R|||}{|||R|||}.$$

The Frobenius norm case of this is proved in Sun [1991, (1.10)].

## 5. MORE ON THE POLAR DECOMPOSITION

In this section we will use the symbol $|A|$ to denote the $P$ part in the polar decomposition of any matrix $A$. Thus

$$|A| = (A^*A)^{1/2} = P.$$

This *matrix absolute value* has several interesting properties much different from the absolute value of a complex number. See Thompson (1992, Section 7) for one such instance. Our interest here is in the Lipschitz continuity of the map $A \to |A|$, i.e. in inequalities of the type

$$|||\,|A| - |B|\,||| \leq c|||A - B|||, \tag{5.1}$$

where $c$ is a constant which may depend on the norm $|||\cdot|||$ but should not depend on the dimension $n$.

No such inequality can exist for the operator bound norm. There exist $n \times n$ Hermitian matrices $A$, $B$ such that

$$\|\,|A| - |B|\,\| \geq \tfrac{1}{2}(\log_2 n)^{1/2}\|A - B\|. \tag{5.2}$$

Examples of such matrices were constructed in McIntosh (1971). This question was studied by Kato (1975), who obtained a different kind of inequality for the operator bound norm:

$$\|\,|A| - |B|\,\| \leq \frac{2}{\pi}\|A - B\|\left(2 + \log\frac{\|A\| + \|B\|}{\|A - B\|}\right).$$

In this paper Kato attributes to W. Kahan a theorem saying that the map $A \to |A|$ is Lipschitz continuous in the Frobenius norm. No reference is

provided by Kato, nor is the constant $c$ occurring in (5.1) specified. Araki and Yamagami proved that

$$\| |A| - |B| \|_F \leqslant \sqrt{2} \| A - B \|_F, \tag{5.3}$$

that the constant $\sqrt{2}$ occurring in (5.3) is best possible, and that when $A$, $B$ are self-adjoint this could be improved to 1. Another proof of this result was given by Kittaneh, who obtained stronger statements as well. See Araki and Yamagami (1981), Kittaneh (1985), and Kittaneh (1986). These results are valid in infinite dimensions as well.

Numerical analysts have also proved (5.3), but sometimes as an asymptotic bound and sometimes under restrictions. See Higham (1986, Theorem 2.5), Barrlund (1989, Theorem 2.6). Following the ideas of Kittaneh, let us outline a simple proof of (5.3).

LEMMA 5.1. *Let $f$ be a Lipschitz continuous function on the complex plane satisfying the inequality*

$$|f(z) - f(w)| \leqslant k|z - w| \qquad \text{for all } z, w. \tag{5.4}$$

*Then for all $X \in \mathbf{M}$ and for all normal matrices $A$ we have*

$$\| f(A) X - X f(A) \|_F \leqslant k \| AX - XA \|_F. \tag{5.5}$$

*Proof.* Without loss of generality we can assume $A = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Then if $X = (x_{ij})$, we have

$$\begin{aligned} \| f(A) X - X f(A) \|_F^2 &= \sum_{i,j} \left| \left[ f(\lambda_i) - f(\lambda_j) \right] x_{ij} \right|^2 \\ &\leqslant k^2 \sum_{i,j} |\lambda_i - \lambda_j|^2 |x_{ij}|^2 \\ &= k^2 \| AX - XA \|_F^2. \end{aligned}$$

■

Now, using what is called "Berberian's trick", we can extend this lemma to the case of two normal matrices:

LEMMA 5.2. *Let $f$ be a function satisfying* (5.4). *Let $A$, $B$ be any two normal matrices. The for every $X \in \mathbf{M}$*

$$\| f(A) X - X f(B) \|_F \leqslant k \| AX - XB \|_F. \tag{5.6}$$

*Proof.* Put

$$T = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}, \qquad Y = \begin{bmatrix} 0 & X \\ 0 & 0 \end{bmatrix},$$

and apply Lemma 5.1 to these two operators instead of $A$ and $X$. ∎

If we choose $X = I$ and $f(z) = |z|$ in the above lemma, we obtain

COROLLARY 5.3. *If $A$ and $B$ are normal matrices, then*

$$\| |A| - |B| \|_F \leqslant \|A - B\|_F. \tag{5.7}$$

This result is due to Kittaneh (1985). Araki and Yamagami had proved it when $A$, $B$ are Hermitian. Using another trick involving $2 \times 2$ block matrices, we can now prove a result stronger than (5.3).

THEOREM 5.4 (Kittaneh's generalization of the Araki-Yamagami inequality). *Let $A$, $B$ be any two matrices. Then*

$$\| |A| - |B| \|_F^2 + \| |A^*| - |B^*| \|_F^2 \leqslant 2\|A - B\|_F^2. \tag{5.8}$$

*Proof.* Let

$$T = \begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix}, \qquad S = \begin{bmatrix} 0 & B \\ B^* & 0 \end{bmatrix}.$$

Then note that $T$ and $S$ are Hermitian and

$$|T| = \begin{bmatrix} |A^*| & 0 \\ 0 & |A| \end{bmatrix}.$$

So the inequality (5.8) follows from (5.7). ∎

Some remarks are in order here. First, the above inequalities are all valid in infinite dimensions. For example, it is clear that Lemma 5.1 can be extended to operators with pure point spectrum. Then to extend it to arbitrary operators one can appeal to a theorem of Voiculescu which says every normal operator can be expressed as a diagonal operator plus a Hilbert-Schmidt operator with an arbitrarily small Hilbert-Schmidt (Frobenius) norm. See Kittaneh (1985). Second, these results have other very

interesting consequences. Thus for example, one famous theorem of operator theory is the Putnam-Fuglede theorem. According to this, if $A, B$ are normal, then for any $X$ we have $AX - XB = 0$ iff $A^*X - XB^* = 0$. Lemma 5.2 with $f(z) = \bar{z}$ gives the much stronger result $\|AX - XB\|_F = \|A^*X - XB^*\|_F$. For infinite dimensional spaces this result is proved in Weiss (1981). Third, inequalities like (5.8) would be useful in obtaining perturbation results for singular vectors. See Bhatia and Kittaneh (1988).

Now let us return to the question whether an inequality like (5.1) is true for some other unitarily invariant norms. For $1 \leqslant p < \infty$ consider the Schatten $p$-norms defined as

$$\|A\|_p = \left( \sum_j s_j^p(A) \right)^{1/p}.$$

In this notation $\|A\|_F = \|A\|_2$. It is also customary to put

$$\|A\|_\infty = \|A\| = s_1(A).$$

It has been proved in Davies (1988) that for each $p$, $1 < p < \infty$, there exists a constant $\gamma_p$ independent of $n$ such that

$$\| |A| - |B| \|_p \leqslant \gamma_p \|A - B\|_p. \tag{5.9}$$

We have already noted that no such inequality can hold for $p = \infty$. By a duality argument none can hold for $p = 1$ either. However, in the same paper Davies has shown that there exists a constant $c_n$ which is $O(\log n)$ such that for all $n \times n$ matrices $A, B$,

$$\| |A| - |B| \|_p \leqslant c_n \|A - B\|_p, \qquad p = 1, \infty. \tag{5.10}$$

Further, the best constant occurring in this inequality must grow as $\log n$. Good estimates for the value of this constant, accurate to a factor 2, have been obtained in Mathias (1991a).

In a recent paper (Kosaki, 1992) those norms for which estimates like (5.1) are valid have been characterized. It turns out that these norms are precisely the ones for which the triangular truncation operator $\mathcal{T}$ on $\mathbf{M}$ has norm $\|\|\mathcal{T}\|\|$, which can be bounded independently of $n$. (See Section 3.2 above.)

Finally, let us mention another kind of results. It was shown by Kosaki that

$$\| |A| - |B| \|_1 \leqslant \sqrt{2} \left( \|A + B\|_1 \|A - B\|_1 \right)^{1/2} \tag{5.11}$$

and by Kittaneh and Kosaki that

$$\| |A| - |B| \|_p \leqslant \left(\|A + B\|_p \|A - B\|_p\right)^{1/2}, \qquad p \geqslant 2. \tag{5.12}$$

See Kosaki (1984), Kittaneh and Kosaki (1986). The picture was completed in Bhatia (1988), where it was shown that

$$\| |A| - |B| \|_p \leqslant 2^{1/p - 1/2}\left(\|A + B\|_p \|A - B\|_p\right)^{1/2}, \qquad 1 \leqslant p \leqslant 2, \tag{5.13}$$

and that

$$|||\, |A| - |B| \,||| \leqslant \sqrt{2}\left(|||A + B||| \, |||A - B|||\right)^{1/2} \tag{5.14}$$

for every unitarily invariant norm. All the inequalities (5.11)–(5.14) are sharp.

*One of the key ideas of this paper was developed in collaboration with Kalyan Mukherjea in 1985. Stimulus to carry out this detailed analysis came from preprints I received from Roy Mathias and G. W. Stewart, which made me aware of the work of several numerical analysts on these problems. This work was facilitated by the support I received from the DAE India and the SFB 343 at the University of Bielefeld and was accelerated by an invitation to talk at the ILAS Conference in Lisbon in August 1992. It is a pleasure to record my thanks to all of them.*

REFERENCES

Araki, H. and Yamagami, S. 1981. An inequality for the Hilbert-Schmidt norm, *Comm. Math. Phys.* 81:89–98.

Arazy, J. 1978. Some remarks on interpolation theorems and the boundedness of the triangular projection in unitary matrix spaces, *Integral Equations Operator Theory* 1:453–495.

Barrlund, A. 1989. Perturbation bounds on the polar decomposition, *BIT* 30:101–113.

Bhatia, R. 1987. *Perturbation Bounds for Matrix Eigenvalues*, Longman, Essex.

Bhatia, R. 1988. Perturbation inequalities for the absolute value map in norm ideals of operators, *J. Operator Theory* 19:129–136.

Bhatia, R., Davis, C., and McIntosh, A. 1983. Perturbation of spectral subspaces and solution of linear operator equations, *Linear Algebra Appl.* 52/53:45–67.

Bhatia, R. and Kittaneh F. 1988. On some perturbation inequalities for operators, *Linear Algebra Appl.* 106:271–279.

Bhatia, R. and Mukherjea, K. 1986. On weighted Löwdin orthogonalization, *Internat. J. Qtm. Chem.* 29:1775–1778.

Bhatia, R. and Mukherjea, K. 1992. Variation of the unitary part of a matrix, *SIAM J. Matrix Appl.*, to appear.

Bunse-Gerstner, A. 1981. An analysis of the *HR* algorithm for computing the eigenvalues of a matrix, *Linear Algebra Appl.* 35:155–178.

Bunse-Gerstner, A. 1986. Matrix factorizations for symplectic *QR*-like methods, *Linear Algebra Appl.* 83:49–77.

Bunse-Gerstner, A. and Mehrmann, V. 1986. A symplectic *QR* like algorithm for the solution of the real algebraic Riccati equation, *IEEE Trans. Automat. Control* 31:1004–113.

Chevalley, C. 1946. *Theory of Lie Groups*, Princeton U.P., Princeton.

Davies, E. B. 1988. Lipschitz continuity of operators in the Schatten classes, *J. London Math. Soc.* 37:148–157.

de Bruijn, N. G. and Szekeres, G. 1955. On some exponential and polar representations of matrices, *Nieuw Arch. Wisk.* 3:20–32.

Dieudonné, J. 1960. *Foundations of Modern Analysis*, Academic, New York.

Fiedler, M. 1986. *Special Matrices and Their Applications in Numerical Mathematics*, Martinus Nijhoff, Dordrecht.

Gantmacher, F. R. 1959. *The Theory of Matrices*, Chelsea, New York.

Gohberg, I. C. and Krein, M. G. 1970. *Theory and Applications of Volterra Operators in Hilbert Space*, Amer. Math. Society, Providence.

Goldstein, J. A. and Levy, M. 1991. Linear algebra and quantum chemistry, *Amer. Math. Monthly* 78:710–718.

Golub, G. and Van Loan, C. 1989. *Matrix Computations*, 2nd ed., Johns Hopkins U.P., Baltimore.

Higham, N. J. 1986. Computing the polar decomposition—with applications, *SIAM J. Sci. Statist. Comput.* 7:1160–1174.

Horn, R. A. and Johnson, C. R. 1985. *Matrix Analysis*, Cambridge U.P., Cambridge.

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

Kato, T. 1973. Continuity of the map $S \to |S|$ for linear operators, *Proc. Japan Acad. Sci.* 49:157–160.

Kenny, C. and Laub, A. J. 1991. Polar decomposition and matrix sign function condition estimates, *SIAM J. Sci. Statist. Comput.* 12:488–504.

Kittaneh, F. 1985. On Lipschitz functions of normal operators, *Proc. Amer. Math. Soc.* 94:416–418.

Kittaneh, F. 1986. Inequalities for the Schatten $p$-norm IV, *Comm. Math. Phys.* 106:581–585.

Kittaneh, F. and Kosaki, H. 1986. Inequalities for the Schatten $p$-norm V, *Publ. Res. Inst. Math. Sci.* 23:433–443.

Kosaki, H. 1984. On the continuity of the map $\phi \to |\phi|$ from the predual of a $W^*$-algebra, *J. Funct. Anal.* 59:123–131.

Kosaki, H. 1992. Unitarily invariant norms under which the map $A \to |A|$ is Lipschitz continuous, *Publ. Res. Inst. Math. Sci.* 28:299–313.

Kwapien, S. and Pelczynski, A. 1970. The main triangle projection in matrix spaces and its applications, *Studia Math.* 34:43–68.

Lancaster, P. and Tismenetsky, M. 1985. *The Theory of Matrices with Applications*, Academic, New York.

Mathias, R. 1991a. The Hadamard operator norm of a circulant and applications, preprint.

Mathias, R. 1991b. Perturbation bounds for the polar decomposition, preprint.

McIntosh, A. R. 1971. Counterexample to a question on commutators, *Proc. Amer. Math. Soc.* 29:337–340.

Medio, A. 1987. Oscillations in optimal growth models, *J. Econom. Theory* 11:201–206.

Mehta, M. L. 1989. *Matrix Theory*, Hindustan, Delhi.

Stewart, G. W. 1977. Perturbation bounds for the *QR* factorization of a matrix, *SIAM J. Numer. Anal.* 14:509–518.

Stewart, G. W. 1992. On the Perturbation of *LU*, Cholesky and *QR* Factorizations, TR-92-24, Univ. of Maryland, College Park.

Stewart, G. W. and Sun, J.-G. 1990. *Matrix Perturbation Theory*, Academic, New York.

Sun, J.-G. 1991. Perturbation bounds for the Cholesky and *QR* factorizations, *BIT* 31:341–353.

Thompson, R. C. 1992. High, low, and quantitative roads in linear algebra, *Linear Algebra Appl.* 162–164:23–64.

Watkins, D. S. and Elsner, L. 1988. Self-similar flows, *Linear Algebra Appl.* 110:213–242.

Weiss, G. 1981. The Fuglede commutativity theorem modulo the Hilbert-Schmidt class and generating functions for matrix operators II, *J. Operator Theory* 5:3–16.

*Received 4 September 1982; final manuscript accepted 15 April 1993*