Minimal distortion problems for classes of unitary matrices

Vladimir Bolotnikov, Chi-Kwong Li, Leiba Rodman *

Abstract

Given two chains of subspaces in $\mathbb{C}^n$, we study the set of unitary matrices that map the subspaces in the first chain onto the corresponding subspaces in the second chain, and minimize the value $\|\mathbf{U} - I_n\|$ for various unitarily invariant norms $\| \cdot \|$ on $\mathbb{C}^{n\times n}$. In particular, we give formula for the minimum value $\|\mathbf{U} - I_n\|$, and describe the set of all the unitary matrices in the set attaining the minimum, for the Frobenius norm. For other unitarily invariant norms, we obtain the results if the subspaces have special structure. Several related matrix minimization problems are also considered.

Key Words: Unitary matrix, matrix optimization.
AMS Subject Classification: 15A99.

1 Introduction

Let

$$\{0\} \subsetneqq \mathcal{M}_1 \subsetneqq \mathcal{M}_2 \subsetneqq \ldots \subsetneqq \mathcal{M}_\ell \subsetneqq \mathbb{C}^n \quad \text{and} \quad \{0\} \subsetneqq \mathcal{N}_1 \subsetneqq \mathcal{N}_2 \subsetneqq \ldots \subsetneqq \mathcal{N}_\ell \subsetneqq \mathbb{C}^n \tag{1.1}$$

be two chains of nonzero proper subspaces in $\mathbb{C}^n$, the complex vector space of n-component column vectors, with

$$\dim \mathcal{M}_j = \dim \mathcal{N}_j = r_j \qquad (j = 1, \ldots, \ell). \tag{1.2}$$

It is easily seen that there exists a unitary matrix $\mathbf{U}$ on $\mathbb{C}^n$ such that

$$\mathbf{U}\mathcal{M}_j = \mathcal{N}_j \qquad (j = 1, \ldots, \ell). \tag{1.3}$$

In this paper we consider the problem of minimizing the deviation of $\mathbf{U}$ from the identity transformation, i.e., minimizing of the value $\|\mathbf{U}-I_n\|$ on the set of unitary transformations satisfying (1.3), for unitarily invariant norms $\| \cdot \|$ on $\mathbb{C}^{n\times n}$, the complex vector space of $n \times n$ matrices. Recall that a norm $\| \cdot \|$ on $\mathbb{C}^{m\times n}$ is called *unitarily invariant* if the equality $\|UAV\| = \|A\|$ holds for every $A \in \mathbb{C}^{m\times n}$ and every choice of unitary matrices $U \in \mathbb{C}^{m\times m}$, $V \in \mathbb{C}^{n\times n}$.

*Department of Mathematics, College of William and Mary, Williamsburg VA 23187-8795, USA. E-mail: vladi@math.wm.edu, ckli@math.wm.edu, lxrodm@math.wm.edu
The research of the second and third author was supported by NSF grants.

Problem 1.1 *Given two chains* (1.1) *of subspaces in* $\mathbb{C}^n$ *and given a unitarily invariant norm* $\|\cdot\|$ *on* $\mathbb{C}^{n\times n}$*, compute the value*

$$\min\{\|\mathbf{U}-I\| \,:\, \mathbf{U} \quad is \quad unitary \quad and \quad \mathbf{U}\mathcal{M}_j=\mathcal{N}_j \quad for \quad j=1,\ldots,\ell\}, \tag{1.4}$$

find a unitary matrix $\mathbf{U}_{\min}$ *for which the minimum in* (1.4) *is attained, and describe the set of all such matrices* $\mathbf{U}_{\min}$.

Let $x_1,\ldots,x_{r_\ell}$ and $y_1,\ldots,y_{r_\ell}$ be two orthonormal sets of vectors in $\mathbb{C}^n$ such that

$$\mathrm{Span}\,\{x_1,\ldots,x_{r_j}\}=\mathcal{M}_j \quad \text{and} \quad \mathrm{Span}\,\{y_1,\ldots,y_{r_j}\}=\mathcal{N}_j, \quad j=1,\ldots,\ell. \tag{1.5}$$

(Here and elsewhere in the paper we denote by Span $\{z_1,\ldots,z_m\}$ the subspace spanned by the vectors $z_1,\ldots,z_m$.) Clearly, a unitary matrix $\mathbf{U}$ satisfies (1.3) if and only if $\mathbf{U}$ maps Span $\{x_{r_{j-1}+1},\ldots,x_{r_j}\}$ onto Span $\{y_{r_{j-1}+1},\ldots,y_{r_j}\}$, for $j=1,\ldots,\ell$ (we set $r_0=0$).

In fact, by the above observation, we can formulate Problem 1.1 entirely in matrix language as follows. Let X and Y be unitary matrices with columns $x_1,\ldots,x_n$ and $y_1,\ldots,y_n$, respectively, i.e., the last $n-r_\ell$ columns of X (respectively, Y) span the orthogonal complement of $\mathcal{M}_\ell$ (repectively, $\mathcal{N}_\ell$). Then clearly the unitary matrix $U=YX^*$ satisfies (1.3). Furthermore, it is easy to show that a unitary U satisfies (1.3) if and only if $U=YVX^*$, where

$$V=V_1\oplus\cdots\oplus V_\ell\oplus V_{\ell+1}, \tag{1.6}$$

where V_j are $(r_j-r_{j-1})\times(r_j-r_{j-1})$ unitary matrices for $j=1,\ldots,\ell,\ell+1$ with $r_0=0$ and $r_{\ell+1}=n$. Let $\mathcal{S}$ be the set of unitary matrices in block form as V above. Then Problem 1.1 can be restated as finding

$$\min_{V\in\mathcal{S}}\|YVX^*-I\|=\min_{V\in\mathcal{S}}\|Y^*(YVX^*)X-Y^*X\|=\min_{V\in\mathcal{S}}\|Y^*X-V\| \tag{1.7}$$

and characterizing the matrices $V\in\mathcal{S}$ for which the minimum is attained.

A particular case (corresponding to $\ell=1$) of Problem 1.1 appears in guidance control; see [3], where a complete solution of this particular case for real matrices and the Frobenius norm is given. More generally, several cases of Problem 1.1, and of closely related problems have been studied in the literature (see, for example, [5], [6, Section 4]). In turn, Problem 1.1 belongs to a large class of extremal problems in matrix analysis, many of which have been studied extensively in connection with numerical algorithms (see, e.g., [10], [7], and references there), statistics (see Chapter 10 in [13]), semidefinite programming, etc.

Besides the introduction, the paper consists of three sections. In Section 2, we present some preliminary results on unitarily invariant norms. The main result here, Theorem 2.3, characterizes the minimizers of the distance between a given positive semidefinite matrix and the unitary group, for strictly increasing Schur convex unitarily invariant norms. We address Problem 1.1 in its matrix formulation in Section 3. We show (cf. Theorem 3.1) that if there exist $U,V\in\mathcal{S}$ such that UAV is a block matrix with some nice properties, then one can easily determine $V\in\mathcal{S}$ satisfying (1.7), and the corresponding optimal value $\|YVX^*-I\|$. When $\ell=1$, the most convenient approach is to reduce YVX^* to its

CS-*decomposition*, i.e., finding $X, Y \in \mathcal{S}$ so that

$$YVX^* = \begin{bmatrix} C & 0 & -S & 0 \\ 0 & I_p & 0 & 0 \\ S & 0 & C & 0 \\ 0 & 0 & 0 & I_q \end{bmatrix}, \tag{1.8}$$

where C and S are $k \times k$ nonnegative diagonal matrices satisfying $C^2+S^2 = I_k$, and $p+q+2k = n$. If $\ell > 2$, we generally do not have the nice canonical form. Nevertheless, we can still study Problem 1.1 in the operator context with the help of the CS decomposition as shown in Section 4. Here, for the Frobenius norm, we describe completely the minimizers of the distortion problem 1.1 in terms of the CS decomposition of the matrices X and Y (Theorem 4.2); for unitarily invariant norms that are not scalar multiples of the Frobenius norm, we have a less complete result (Theorem 4.5).

Although we formulate and prove our results for complex vector spaces and matrices only, the results and their proofs remain valid also in the context of real vector spaces and matrices.

We shall denote by $\sigma_1(A) \geq \cdots \geq \sigma_n(A)$ the singular values of a matrix $A \in \mathbb{C}^{n\times n}$. Unitarily invariant norms $\|\cdot\|_\Phi$ on $\mathbb{C}^{n\times n}$ are associated with symmetric gauge functions ϕ in a standard fashion, so that

$$\|A\|_\Phi = \phi\left(\sigma_1(A),\, \sigma_2(A), \ldots, \sigma_n(A)\right)$$

for every $A \in \mathbb{C}^{n\times n}$; see, e.g., [13], [9] for background and basic results on this association. The size n will be fixed throughout the paper, and the unitarily invariant norms are considered on the algebra $\mathbb{C}^{n\times n}$ of $n \times n$ complex matrices. The Schatten p-norms are

$$\|A\|_p = \left(\sum_{j=1}^{\infty} (\sigma_j(A))^p\right)^{1/p}, \quad 1 \leq p < \infty, \qquad \|A\|_\infty = \sigma_1(A) \ \text{ (the operator norm)}.$$

The Schatten 2-norm is known as the *Frobenius norm*:

$$\|A\|_2 = \left(\sum_{i,j=1}^{n} |a_{i,j}|^2\right)^{1/2},$$

where $a_{i,j}$ are the entries of A. If $X \in \mathbb{C}^{n\times n}$ is Hermitian, $\lambda(X) = (\lambda_1(X), \ldots, \lambda_n(X))$ will denote the vector of eigenvalues of X, where $\lambda_1(X) \geq \cdots \geq \lambda_n(X)$. A^t stands for the transpose of a matrix A. The block diagonal matrix with diagonal blocks $A_1, \ldots, A_p$ (in that order) is denoted $\operatorname{diag}(A_1, \ldots, A_p)$, or $A_1 \oplus A_2 \oplus \ldots \oplus A_p$. We use the notation $\Sigma(A) = \operatorname{diag}(\sigma_1(A), \ldots, \sigma_n(A))$ for the $n \times n$ diagonal matrix with the singular values of A (in the non-increasing order) on the main diagonal. Denote by $\{E_{11}, E_{12}, \ldots, E_{nn}\}$ the standard basis for $\mathbb{C}^{n\times n}$.

2 Unitarily invariant norms

In this section we present some results on unitarily invariant norms that are needed for solution of Problem 1.1. A unitarily invariant norm $\|\cdot\|$ is a called a *Q-norm* if there is a unitarily invariant norm $\|\cdot\|_\Phi$, which will be called the *associated norm* of $\|\cdot\|$, such that $\|A\|^2 = \|A^*A\|_\Phi$ for every $A \in \mathbb{C}^{n\times n}$. For example, the Schatten p-norm $\|\cdot\|_p$ is a Q-norm if and only if $2 \le p \le \infty$ because $\|A\|_p^2 = \|A^*A\|_{p/2}$; we refer to [4, Chapter 4] for more information on Q-norms.

A unitarily invariant norm $\|\cdot\|_\Phi$ is called *strictly convex* if the unit ball with respect to that norm is strictly convex: If $A \ne B$ satisfy

$$\|A\|_\Phi = \|B\|_\Phi = 1,$$

then

$$\|tA + (1-t)B\|_\Phi < 1 \quad \text{for } 0 < t < 1.$$

For example, a Schatten p-norm is strictly convex if and only if $1 < p < \infty$. We need the following property of strictly convex norms:

Proposition 2.1 *Let $\|\cdot\|_\Phi$ be a unitarily invariant norm which is strictly convex. If $A, B \in \mathbb{C}^{n\times n}$ are such that*

$$\|A\| \le \|B\| \quad \text{for every unitarily invariant norm,} \tag{2.1}$$

and $\|A\|_\Phi = \|B\|_\Phi$, then $\sigma_j(A) = \sigma_j(B)$ for $j = 1, 2, \ldots$.

We will use the well-known fact that (2.1) holds if and only if $\|A\|_{K,k} \le \|B\|_{K,k}$, $k = 1, 2, \ldots$, where $\|\cdot\|_{K,k}$ denotes the k-th *Ky Fan norm*, i.e., the sum of k largest singular values of A.

Proof. Let $D_1 = \operatorname{diag}(\sigma_1(A), \ldots, \sigma_n(A))$ and $D_2 = \operatorname{diag}(\sigma_1(B), \ldots, \sigma_n(B))$. Since $\|A\|_{K,k} \le \|B\|_{K,k}$ for all $k = 1, \ldots, n$, [16, Corollary 5],

$$D_1 = \sum_{i=1}^m t_i U_i P_i D_2 P_i^t,$$

where for $i = 1, \ldots, m$, U_i is a diagonal unitary matrix, P_i is a permutation matrix, $t_i > 0$, such that $t_1 + \cdots + t_m = 1$. But then,

$$\|A\|_\Phi = \|D_1\|_\Phi \le \sum_{i=1}^m t_i \|U_i P_i D_2 P_i^t\|_\Phi = \|B\|_\Phi.$$

By the strict convexity, all $U_i P_i D_2 P_i^t$ are equal, and hence must be equal to D_1. Thus, $D_1 = D_2$ as asserted. ∎

In connection with Proposition 2.1 note that there exist non-strictly convex unitarily invariant norms that have the property described in Proposition 2.1, as the following example shows:

Example 2.2 Define

$$\|A\|_c = \sum_{j=1}^n (n-j+1)\sigma_j(A) = \sum_{j=1}^n \sum_{i=1}^j \sigma_i(A), \quad A \in \mathbb{C}^{n\times n}.$$

If $\|A\| \le \|B\|$ for every unitarily invariant norm $\|\cdot\|$ and if $\|A\|_c = \|B\|_c$, then we have

$$\sum_{j=1}^k \sigma_j(A) \le \sum_{j=1}^k \sigma_j(B), \qquad k = 1, 2, \ldots, n,$$

and

$$\sum_{j=1}^{n}\sum_{i=1}^{j}\sigma_i(A) = \sum_{j=1}^{n}\sum_{i=1}^{j}\sigma_i(B),$$

which imply

$$\sum_{j=1}^{k}\sigma_j(A) = \sum_{j=1}^{k}\sigma_j(B), \qquad k = 1, 2, \ldots, n,$$

i.e., $\sigma_j(A) = \sigma_j(B)$, for $j = 1, 2, \ldots, n$. On the other hand, $\|\cdot\|_c$ is not strictly convex; for example,

$$\|rI + sE_{11}\|_c = r\|I\|_c + s\|E_{11}\|_c \quad \text{for every} \quad r, s > 0.$$

In the sequel we will use the property of strictly convex norms that is described in Proposition 2.1. By the result in [13, Chapter 3, A6-A8], it follows that the symmetric gauge function ϕ that corresponds to the unitarily invariant norm $\|\cdot\|_\Phi$ on $\mathbb{C}^{n\times n}$ is *strictly Schur convex and strictly increasing* if and only if for every pair of $n \times n$ matrices A and B such that $\|A\| \leq \|B\|$ for every unitarily invariant norm $\|\cdot\|$ and $\|A\|_\Phi = \|B\|_\Phi$, we actually have $\sigma_j(A) = \sigma_j(B)$ for $j = 1, 2, \cdots$. For simplicity, we shall call such a unitarily invariant norm *strictly increasing Schur convex.*

Theorem 2.3 *Let $\|\cdot\|_\Phi$ be a strictly increasing Schur convex unitarily invariant norm on $\mathbb{C}^{n\times n}$. Then, for a given positive semidefinite P, we have $\|P - I\|_\Phi \leq \|P - U\|_\Phi$ for every unitary U; the equality $\|P - I\|_\Phi = \|P - U\|_\Phi$ holds if and only if the unitary U is such that $Ux = x$ for every $x \in$* Range *P.*

The result of Theorem 2.3 is known for positive definite P [8] (see [2] for generalizations to infinite dimensional operators, with Schatten p-norms).

The proof of Theorem 2.3 is based on the following lemma.

Lemma 2.4 *Let $P \in \mathbb{C}^{n\times n}$ be positive semidefinite. Then a unitary matrix U has the property that $P - U$ has singular values $|\sigma_1(P) - 1|, \ldots, |\sigma_n(P) - 1|$ if and only if $Ux = x$ for every $x \in$* Range *P.*

Proof: The result is obviousely true for $P = 0_{n\times n}$. Assume that $P \neq 0$ and prove the statement by induction. The statement is clear when $n = 1$. Suppose the result is true for matrices of sizes up to $n - 1$ with $n > 1$. For $X \in \mathbb{C}^{n\times n}$, let $\tilde{X} = \begin{bmatrix} 0 & X \\ X^* & 0 \end{bmatrix}$, and let $C = P - U$. Then $\tilde{P}$ and $\tilde{C}$ have eigenvalues $\pm\sigma_j(P)$ and $\pm|\sigma_j(P) - 1|$ $(j = 1, \ldots, n)$, respectively. Now, we have

$$\tilde{C} = \tilde{P} + (-\tilde{U}).$$

Note that (i) $\sigma_1(C) = \sigma_1(P) - 1$, or (ii) $\sigma_1(C) = 1 - \sigma_n(P)$. Thus, $\lambda_1(\tilde{C}) = \lambda_r(\tilde{P}) + \lambda_s(-\tilde{U})$ with

$$(r, s) = \begin{cases} (1, 2n) & \text{if (i) holds,} \\ (n + 1, n) & \text{if (ii) holds.} \end{cases}$$

By [12, Theorem 3.1], there is a unit vector $\tilde{x} \in \mathbb{C}^{2n}$ such that

$$\tilde{C}\tilde{x} = \lambda_1(\tilde{C})\tilde{x}, \quad \tilde{P}\tilde{x} = \lambda_r(\tilde{P})\tilde{x}, \quad \text{and} \quad -\tilde{U}\tilde{x} = \lambda_s(-\tilde{U})\tilde{x}.$$

If (i) holds, then the eigenvectors of $\tilde{P}$ corresponding to $\lambda_r(\tilde{P}) = \sigma_1(P)$ are of the form $\frac{1}{\sqrt{2}}\begin{bmatrix} x \\ x \end{bmatrix}$ for some unit eigenvector $x \in \mathbb{C}^n$ of $\lambda_1(P)$. Since $\tilde{C}\tilde{x} = \lambda_1(\tilde{C})\tilde{x}$ and $\tilde{U}\tilde{x} = \lambda_s(\tilde{U})\tilde{x}$, if X is a unitary matrix with x as the first column, then

$$X^*CX = [\sigma_1(C)] \oplus C_1, \quad X^*PX = [\sigma_1(P)] \oplus P_1, \quad X^*UX = [1] \oplus U_1.$$

If (ii) holds, then the eigenvectors of $\tilde{P}$ corresponding to $\lambda_r(\tilde{P}) = -\sigma_n(P)$ is of the form $\frac{1}{\sqrt{2}}\begin{bmatrix} x \\ -x \end{bmatrix}$ for some unit eigenvector $x \in \mathbb{C}^n$ of $\lambda_n(P)$. Since $\tilde{C}\tilde{x} = \lambda_1(\tilde{C})\tilde{x}$ and $\tilde{U}\tilde{x} = \lambda_s(\tilde{U})\tilde{x}$, if X is a unitary matrix with x as the first column, then

$$X^*CX = [\sigma_1(C)] \oplus C_1, \quad X^*PX = [-\sigma_n(P)] \oplus P_1, \quad X^*UX = [-1] \oplus U_1.$$

In both cases, one can then apply induction assumption on the matrices $C_1 = P_1 - U_1$ to get the conclusion. ∎

Proof of Theorem 2.3. Suppose P is positive semidefinite. Then for any unitary U, we have (see [8])

$$(|\sigma_1(P) - 1|, \ldots, |\sigma_n(P) - 1|) \prec_w \sigma(P - U),$$

where $\sigma(P-U)$ is the vector of singular values of $P-U$, and $\prec_w$ is the weak majorization relation (see, e.g., [13] for this relation and its properties). It follows that

$$\|P - I\|_\Phi = \phi(|\sigma_1(P) - 1|, \ldots, |\sigma_n(P) - 1|) \leq \phi(\sigma(P - U)) = \|P - U\|_\Phi.$$

Since ϕ is strictly increasing Schur convex, the equality holds if and only if $P - U$ has singular values $|\sigma_1(P) - 1|, \ldots, |\sigma_n(P) - 1|$. Now, the result follows from Lemma 2.4. ∎

If we omit the hypothesis that $\|\cdot\|_\Phi$ is strictly increasing Schur convex in Theorem 2.3, then the result of that theorem is no longer valid. However, omitting the hypothesis of strict increasing Schur convexity of $\|\cdot\|_\Phi$ and simultaneously omitting "and only if" will produce a correct statement. In other words, the set of best approximants may only become larger if $\|\cdot\|_\Phi$ is not assumed to be strictly increasing Schur convex. The proof of this statement is easily obtained using a continuity argument and the fact that the set of strictly increasing Schur convex symmetric gauge functions is dense in the set of all symmetric gauge functions.

Note that there can be a much larger set of minimizers if the unitarily invariant norm is not strictly Schur convex and strictly increasing. For example, if $P = \text{diag}\,(3, 1, \ldots, 1)$ then $2 = \|P - I\|_\infty = \|P - U\|_\infty$ for any unitary $U = [1] \oplus U_1$.

Next, we mention several simple optimization results which will be used later.

Lemma 2.5 *Let $A \in \mathbb{C}^{n\times n}$ have singular values $\sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq \sigma_n(A) \geq 0$ and let*

$$A = PU = V\Sigma(A)W \qquad (P \geq 0, \;\; UU^* = VV^* = I_n),$$

where $W = V^*U$, be its polar and singular values decompositions. Then, for every unitarily invariant norm,

$$\min_{X \text{ unitary}} \|I_n - AX\| = \|I_n - \Sigma(A)\| = \|I_n - P\| \tag{2.2}$$

and the minimum is attained for $X = U^* = W^*V^*$. Moreover, if $\| \cdot \|$ is strictly increasing Schur convex, then

$$\|I_n - AX\| = \|I_n - P\|$$

for a unitary matrix X if and only if $Xx = U^*x$ for every $x \in \operatorname{Range} A$.

Proof: For every unitary matrix X, it follows that $\Sigma(AX) = \Sigma(A)$ and therefore, by the inequality $\|A - B\| \geq \|\Sigma(A) - \Sigma(B)\|$ (see, e. g., [11, Theorem 7.4.51]), we have

$$\|I_n - AX\| \geq \|I_n - \Sigma(AX)\| = \|I_n - \Sigma(A)\|.$$

The equality is attained for $X = W^*V^*$, since the norm is unitarily invariant. The second equality in (2.2) follows because $P = \tilde{U}^*\Sigma(A)\tilde{U}$ for some unitary $\tilde{U}$. For the second part of the lemma, observe that

$$\|I - AX\| = \|I - PUX\| = \|X^*U^* - P\| \quad \text{and} \quad \operatorname{Range} A = \operatorname{Range} P,$$

and apply Theorem 2.3. ∎

Lemma 2.6 *For every matrix $A \in \mathbb{C}^{n\times n}$, for every pair of orthogonal projections $P, Q \in \mathbb{C}^{n\times n}$ and for every unitarily invariant norm $\| \cdot \|$ on $\mathbb{C}^{n\times n}$,*

$$\|A\| \geq \|PAQ\|. \tag{2.3}$$

Moreover, if $\| \cdot \|$ is the Frobenius norm, then $\|A\|_2 > \|PAQ\|_2$, unless $(I-P)A+PA(I-Q) = 0$.

Proof: Let $U, V \in \mathbb{C}^{n\times n}$ be unitary such that $U^*PU = I_r \oplus 0_{n-r}$ and $V^*QV = I_s \oplus 0_{n-s}$. Then one readily checks that

$$\|A\| = \|UAV^*\| \geq \|(I_r \oplus 0_{n-r})UAV^*(I_s \oplus 0_{n-s})\| = \|PAQ\|.$$

The second assertion is clear from the above calculation. ∎

Lemma 2.7 *Let A_t and B_s be two families of $n \times m$ and $n \times k$ matrices, respectively. If*

$$\min_t \|A_t\|_2 = \|A_{t_0}\|_2 \quad \text{and} \quad \min_s \|B_s\|_2 = \|B_{s_0}\|_2,$$

then

$$\min_{t,s} \| \, [A_t, \;\; B_s] \, \|_2 = \| \, [A_{t_0}, \;\; B_{s_0}] \, \|_2.$$

Proof: By the definition of the Frobenius norm $\| \, [A_t, \;\; B_s] \, \|_2^2 = \|A_t\|_2^2 + \|B_s\|_2^2$, which implies immediately the desired result. ∎

3 Results in the matrix formulation

In this section, we describe solutions of the problem (1.7) for Q-norms and the Frobenius norm. We shall let $A = YX^*$ in (1.7), and let $q = \ell + 1$ in the following discussion. Also, we shall continue to use $\mathcal{S} = \mathcal{S}(r_1, \ldots, r_q)$ to represent the set of unitary matrices of the form

$$U = U_1 \oplus \cdots \oplus U_q, \qquad U_1 \oplus \cdots \oplus U_r \in \mathbb{C}^{r_j \times r_j}.$$

Note that

$$\min_{U \in \mathcal{S}} \|A - U\| = \min_{U,V,W \in \mathcal{S}} \|VAW - U\|$$

for any unitarily invariant norm $\|\cdot\|$. Thus, we can always replace A by VAW with $V, W \in \mathcal{S}$ to determine the minimum norm, and $\tilde{U}$ is a minimizer for

$$\min_{U \in \mathcal{S}} \|A - U\|$$

if and only if $V\tilde{U}W$ is a mimimizer for

$$\min_{U \in \mathcal{S}} \|VAW - U\|.$$

We first present the following result on Q-norms.

Theorem 3.1 *Let $\|\cdot\|$ be a Q-norm, i.e., there is a unitarily invariant norm $\|\cdot\|_\Phi$ such that $\|A\|^2 = \|A^*A\|_\Phi$ for every $A \in \mathbb{C}^{n\times n}$, and let A be an $n \times n$ unitary matrix. Suppose there exist*

$$V = V_1 \oplus \cdots \oplus V_q \quad \text{and} \quad W = W_1 \oplus \cdots \oplus W_q \in \mathcal{S}$$

with $VAW = [A_{i,j}]_{i,j=1}^q$ such that the matrices $A_{i,i} \in \mathbb{C}^{(r_i - r_{i-1})\times(r_i - r_{i-1})}$ are positive semidefinite for $i = 1, \ldots, q$, and $A_{i,j} = -(A_{j,i})^*$ for $i \neq j$. Then

$$\|VAW - I\| = \|A - V^*W^*\| \leq \|A - U\| \qquad \text{for all } U \in \mathcal{S}.$$

Moreover, $\tilde{U} = \tilde{U}_1 \oplus \cdots \oplus \tilde{U}_q \in \mathcal{S}$ satisfies the equality

$$\|A - \tilde{U}\| = \|A - V^*W^*\| \tag{3.1}$$

for some Q-norm $\|\cdot\|$ whose associate norm is strictly increasing Schur convex if and only if

$$V_j\tilde{U}_jW_jx = x \quad \text{for every} \quad x \in \text{Range } A_{j,j}, \quad j = 1, \ldots, q. \tag{3.2}$$

As mentioned in the introduction (cf. formula (1.8)), the hypothesis on existence of $V \in \mathcal{S}$ and $W \in \mathcal{S}$ with the indicated properties is always satisfied when $q = 2$.

Proof: Without loss of generality, we may assume that $V = W = I$, otherwise, we may replace A by VAW. Also, we may assume that $A_{i,i}$ are diagonal, and thus $A = A_1 + iA_2$, where $A_1 = \text{diag}\,(t_1, \ldots, t_n)$ and A_2 are Hermitian. Let $U = U_1 \oplus \cdots \oplus U_q \in \mathcal{S}$. Then

$$(A - U)^*(A - U) = 2I - (A^*U + U^*A) \quad \text{and} \quad (A - I)^*(A - I) = 2I - (A + A^*).$$

We claim that

$$\sum_{j=1}^m \sigma_j((A - I)^*(A - I)) \leq \sum_{j=1}^m \sigma_j((A - U)^*(A - U)), \qquad m = 1, \ldots, n, \tag{3.3}$$

or equivalently, that

$$\sum_{j=m}^{n} \lambda_j(A+A^*) \geq \sum_{j=m}^{n} \lambda_j(A^*U+U^*A), \qquad m=1,\ldots,n.$$

Suppose $U = X+iY$ so that X and Y are Hermitian. Then the vector of diagonal entries of the matrix A^*U+U^*A equals to that of A_1X+XA_1, i.e., to $2(d_1t_1,\ldots,d_nt_n)$, where $d_1\ldots,d_n$ are the diagonal entries of X and satisfy $|d_j| \leq 1$ for all j. For a fixed r, let $1 \leq i_1 < \cdots < i_r \leq n$ be the indices so that $t_{i_1},\ldots,t_{i_r}$ are the r smallest diagonal entries of A_1. Set

$$P = \sum_{j=1}^{r} E_{i_j,i_j}. \tag{3.4}$$

Since $|d_{i_j}| \leq 1$, we see that (the first inequality below follows by the interlacing properties of eigenvalues of a Hermitian matrix and of its principal submatrix, whereas the last equality is valid since the matrix $A+A^*$ is diagonal)

$$\begin{aligned}\sum_{k=n-r+1}^{n} & \lambda_k(A^*U+U^*A) \leq \operatorname{trace}(P(A^*X+X^*A)P) \\ = \ & 2\sum_{i=1}^{r} d_{i_j}t_{i_j} \leq 2\sum_{i=1}^{r} t_{i_j} = \operatorname{trace}(P(A+A^*)P) = \sum_{k=n-r+1}^{n} \lambda_k(A+A^*),\end{aligned}$$

which proves (3.3). We obtain that $\|(A-I)^*(A-I)\| \leq \|(A-U)^*(A-U)\|$ for every unitarily invariant norm $\|\cdot\|$. Consequently, $|||A-I||| \leq |||A-U|||$ for every Q-norm $|||\cdot|||$.

Now, suppose $\tilde{U} \in \mathcal{S}$ and

$$|||A-I||| = |||A-\tilde{U}|||$$

for some Q-norm $|||\cdot|||$ whose associate norm $\|\cdot\|$ is strictly increasing Schur convex. Thus,

$$\|(A-I)^*(A-I)\| = \|\left(A-\tilde{U}\right)^*\left(A-\tilde{U}\right)\| \tag{3.5}$$

and thus the singular values of $(A-I)^*(A-I)$ coincide with those of $\left(A-\tilde{U}\right)^*\left(A-\tilde{U}\right)$. Therefore,

$$\lambda_j(A+A^*) = \lambda_j(A^*\tilde{U}+\tilde{U}^*A), \qquad j=1,2,\ldots n.$$

Using the notation introduced in the first part of the proof, we have

$$\sum_{j=1}^{n} t_j = \operatorname{trace} A_1 = \operatorname{trace}(A_1^*X+X^*A_1) = \sum_{j=1}^{n} d_jt_j.$$

Therefore, $d_j = 1$ for every j for which $t_j > 0$. This condition is easily seen to be equivalent to (3.2). ∎

We give explicit formulas for the Schatten norms for 2×2 real matrices A of the form described in Theorem 3.1.

Lemma 3.2 *Let $a \geq 0$, b be real numbers such that $a^2+b^2=1$. Then*

$$\min\left\|\begin{bmatrix} a-t & -b \\ b & a-r\end{bmatrix}\right\|_p = \left\|\begin{bmatrix} a-1 & -b \\ b & a-1\end{bmatrix}\right\|_p = 2^{1/p}\sqrt{2-2a}, \quad 2 \leq p \leq \infty, \tag{3.6}$$

where the minimum is taken over the set of (ordered) pairs $\{t, r\}$*,* $t, r \in \mathbb{C}$*, such that* $|t| = |r| = 1$*.*
Moreover, if $a > 0$*, the minimum in* (3.6) *is achieved only for* $t = r = 1$*, whereas if* $a = 0$ *and* $p > 2$*, the minimum in* (3.6) *is achieved precisely for those pairs* $\{t, r\}$ *(*$|t| = |r| = 1$*) for which* $t = \overline{r}$*. If* $a = 0$ *and* $p = 2$*, then* $\left\|\left[\begin{array}{cc} -t & -1 \\ 1 & -r \end{array}\right]\right\|_2$ *has constant value* 2*.*

Proof: A calculation shows that the singular values of the matrix

$$A = \left[\begin{array}{cc} a-t & -b \\ b & a-r \end{array}\right]$$

are $\sqrt{2 - a\mathrm{Re}\, t - a\mathrm{Re}\, r \pm \sqrt{q}}$, where $q = (a\mathrm{Re}\, t - a\mathrm{Re}\, r)^2 + b^2|t - \overline{r}|^2$. So we have to prove that

$$(2 - a\mathrm{Re}\, t - a\mathrm{Re}\, r + \sqrt{q})^{p/2} + (2 - a\mathrm{Re}\, t - a\mathrm{Re}\, r - \sqrt{q})^{p/2} \geq 2(2-2a)^{p/2} \tag{3.7}$$

if $p \geq 2$ and $|t| = |r| = 1$. Treating $u = \sqrt{q}$ as an independent variable, we need only to prove that

$$(2 - a\mathrm{Re}\, t - a\mathrm{Re}\, r + u)^{p/2} + (2 - a\mathrm{Re}\, t - a\mathrm{Re}\, r - u)^{p/2} \geq 2(2-2a)^{p/2} \tag{3.8}$$

for $p \geq 2$, $|t| = |r| = 1$, and $0 \leq u \leq 2 - a\mathrm{Re}\, r - a\mathrm{Re}\, t$. The inequality (3.8) is valid for $u = 0$, and the derivative with respect to u of the left hand side of (3.8) is positive for $u > 0$ (here the hypothesis $p \geq 2$ is used). Thus, (3.8) is proved. An examination of the proof of (3.6) shows that the equality in (3.8) is achieved only in the situations indicated in Lemma 3.2. ∎

Lemma 3.2 (applied with $a = 0$) shows in particular, that the condition (3.2) is generally not sufficient to guarantee the equality in (3.1).

The result of Lemma 3.2 fails for $1 \leq p < 2$. More precisely:

Lemma 3.3 *Let* $1 \leq p < 2$*. Then*

$$\min \left\|\left[\begin{array}{cc} -t & -1 \\ 1 & -r \end{array}\right]\right\|_p = 2^p, \tag{3.9}$$

where the minimum is taken over the set of pairs $\{t, r\}$*,* $t, r \in \mathbb{C}$*, such that* $|t| = |r| = 1$*. Moreover, the minimum in* (3.9) *is achieved precisely for those pairs* $\{t, r\}$ *for which* $t = -\overline{r}$*.*

The proof is elementary and relies on the explicit formulas for the singular values of the matrix $\left[\begin{array}{cc} -t & -1 \\ 1 & -r \end{array}\right]$ obtained in the proof of Lemma 3.2.

In particular, Lemma 3.3 shows that Theorem 3.1 is generally false for unitarily invariant norms that are not Q-norms.

For the Frobenius norm, we have the following general result.

Theorem 3.4 *Let A be an $n \times n$ unitary matrix partitioned as a $q \times q$ block matrix $A = [A_{i,j}]_{i,j=1}^{q}$, where $A_{i,i} \in \mathbb{C}^{(r_i - r_{i-1}) \times (r_i - r_{i-1})}$, and let $W = W_1 \oplus \cdots \oplus W_q \in \mathcal{S}$ (as defined in the paragraph following the statement of Problem 1.1), be such that $A_{i,i}W_i^*$ is positive semidefinite. Then*

$$2n - \sum_{j=1}^{q} \operatorname{trace}(A_{ii}W_i^*) = 2n - \sum_{j=1}^{q} \operatorname{trace}\sqrt{A_{ii}A_{ii}^*} = \|A - W\|_2^2 \tag{3.10}$$

and

$$\|A - W\|_2 \le \|A - U\|_2 \qquad \textit{for all } U \in \mathcal{S}. \tag{3.11}$$

Moreover, $\tilde{U} = \tilde{U}_1 \oplus \cdots \oplus \tilde{U}_q \in \mathcal{S}$ satisfies the equality

$$\|A - W\|_2 = \|A - \tilde{U}\|_2 \tag{3.12}$$

if and only if

$$\tilde{U}_j W_j^* x = x \quad \text{for} \quad \text{every} \quad x \in \text{Range } A_{j,j}W_j^*, \quad j = 1, \ldots, q. \tag{3.13}$$

Proof: The proof of the first part follows the same pattern as the proof of Theorem 3.1, except that (3.3) needs to be proven only for $m = n$, and therefore we take $P = I$ in (3.4).

For the second part, in view of Theorem 3.1, we need only to show that if (3.12) holds for some unitary matrices $\tilde{U} = \tilde{U}_1 \oplus \cdots \oplus \tilde{U}_q \in \mathcal{S}$, then (3.13) holds. To this end notice that

$$\begin{aligned} \|A - \tilde{U}\|_2^2 &= \|(A_{1,1} - \tilde{U}_1) \oplus \ldots \oplus (A_{q,q} - \tilde{U}_q)\|_2^2 + \sum_{j \neq k} \|A_{j,k}\|^2 \\ &= \sum_{j=1}^{q} \|A_{j,j} - \tilde{U}_j\|_2^2 + \sum_{j \neq k} \|A_{j,k}\|_2^2, \end{aligned}$$

and therefore, the proof reduces to showing that, for a fixed j, $\|A_{j,j} - \tilde{U}_j\|_2 = \|A_{j,j} - W_j\|_2$ as soon as the unitary matrix $\tilde{U}_j$ has the property that $\tilde{U}_j W_j^* x = x$ for every $x \in \text{Range}\, A_{j,j}W_j^*$. But every such unitary matrix $U_j W_j^*$ decomposes into the orthogonal sum $U_j W_j^* = X_j \oplus Y_j$ with respect to the orthogonal decomposition $\mathbb{C}^{r_j - r_{j-1}} = \text{Range}\, A_{jj}W_j^* \oplus \text{Ker}\, A_{jj}W_j^*$, and the equality $\|A_{j,j} - \tilde{U}_j\|_2 = \|A_{j,j} - W_j\|_2$ is obvious. ∎

The result of Theorem 3.4 does not hold for all Q-norms, as the following example (produced by Matlab) shows.

Example 3.5 Let

$$Q = \begin{bmatrix} 0.4104 & -0.5789 - 0.2985i & -0.5773 + 0.2722i \\ -0.1678 + 0.7165i & 0.2186 & -0.0369 + 0.6397i \\ -0.5337 + 0.0721i & -0.5740 - 0.4455i & 0.4266 \end{bmatrix}.$$

The matrix Q is unitary (up to Matlab precision), and $Q - I$ has singular values 1.9328, 0.3665, 0.1367. On the other hand, let

$$E = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -0.9999 + 0.0150i & 0 \\ 0 & 0 & 0.4749 - 0.8800i \end{bmatrix}.$$

Then E is unitary (up to Matlab precision), and the singular values of $Q - E$ are 1.6706, 1.5250, 0.3076. Thus, $\|Q - I\|_\infty > \|Q - E\|_\infty$.

4 Cosine-sine decomposition approach

In this section we treat Problem 1.1 in the operator form, using the cosine-sine decompositions as the main tools. Although in principle the operator formulation of Problem 1.1 is equivalent to the matrix formulation which was dealt with in Section 3, the cosine-sine decompositions allow us to obtain the main result in a more detailed geometric form (using canonical angles between subspaces).

We recall these decompositions of partitioned unitary matrices.

Theorem 4.1 *Let X, $Y \in \mathbb{C}^{n\times r}$ be two isometric matrices: $X^*X = Y^*Y = I_r$. Then:*

1. *If $2r \le n$, there exist unitary matrices $Q \in \mathbb{C}^{n\times n}$ and V, $W \in \mathbb{C}^{r\times r}$ such that*

$$QXW = \begin{array}{c} r \\ r \\ n-2r \end{array} \begin{bmatrix} I_r \\ 0 \\ 0 \end{bmatrix}, \qquad QYV = \begin{array}{c} r \\ r \\ n-2r \end{array} \begin{bmatrix} \Gamma \\ \Delta \\ 0 \end{bmatrix}, \tag{4.1}$$

where

$$\Gamma = \mathrm{diag}\,(\gamma_1, \ldots, \gamma_r) \quad \textit{and} \quad \Delta = \mathrm{diag}\,(\delta_1, \ldots, \delta_r) \tag{4.2}$$

satisfy

$$0 \le \gamma_1 \le \ldots \le \gamma_r, \qquad \delta_1 \ge \ldots \ge \delta_r \ge 0, \qquad \gamma_j^2 + \delta_j^2 = 1 \qquad (j = 1, \ldots, r). \tag{4.3}$$

2. *If $2r > n$, then Q, V and $W \in \mathbb{C}^{r\times r}$ can be chosen so that*

$$QXW = \begin{array}{c} n-r \\ 2r-n \\ n-r \end{array} \begin{bmatrix} I_{n-r} & 0 \\ 0 & I_{2r-n} \\ 0 & 0 \end{bmatrix}, \qquad QYV = \begin{array}{c} n-r \\ 2r-n \\ n-r \end{array} \begin{bmatrix} \Gamma & 0 \\ 0 & I_{2r-n} \\ \Delta & 0 \end{bmatrix}, \tag{4.4}$$

where

$$\Gamma = \mathrm{diag}\,(\gamma_1, \ldots, \gamma_{n-r}) \quad \textit{and} \quad \Delta = \mathrm{diag}\,(\delta_1, \ldots, \delta_{n-r}) \tag{4.5}$$

satisfy

$$0 \le \gamma_1 \le \ldots \le \gamma_{n-r}, \qquad \delta_1 \ge \ldots \ge \delta_{n-r} \ge 0, \qquad \gamma_j^2 + \delta_j^2 = 1 \qquad (j = 1, \ldots, n-r). \tag{4.6}$$

The proof (see e.g., [15]) relies on the CS (cosine-sine) decomposition of a partitioned unitary matrix which in turn, was introduced in [6] and [14]. See also [10] and [7], where the CS decomposition is used in the context of numerical algorithms and geometry of subspaces. Since γ_j and δ_j satisfy $\gamma_j^2 + \delta_j^2 = 1$, they can be regarded as cosines and sines of certain angles θ_j: $\gamma_j = \cos\theta_j$, $\delta_j = \sin\theta_j$, which are called the *canonical angles* between subspaces $\mathcal{M} = \mathrm{Range}\, X_\ell$ and $\mathcal{N} = \mathrm{Range}\, Y_\ell$ (see e.g., [1]). Note also the equalities

$$X^*Y = W\Gamma V^* \;\; (2r \le n) \quad \text{and} \quad X^*Y = W \begin{bmatrix} \Gamma & 0 \\ 0 & I_{2r-n} \end{bmatrix} V^* \;\; (2r > n), \tag{4.7}$$

which follow from (4.1) and (4.4), respectively, and present in fact the singular value decomposition of the matrix X^*Y.

Consider the chains of subspaces (1.1) and (1.2) It is clear from (1.1) and (1.2) that

$$0 = r_0 < r_1 < r_2 < \cdots < r_\ell = r < n.$$

Let $x_1, \dots, x_{r_\ell}$ and $y_1, \dots, y_{r_\ell}$ be two orthonormal sets of vectors in $\mathbb{C}^n$ such that

$$\{x_1, \dots, x_{r_j}\} \quad \text{and} \quad \{y_1, \dots, y_{r_j}\}$$

form bases of $\mathcal{M}_j$ and $\mathcal{N}_j$, respectively. Let

$$X_j = \begin{bmatrix} x_{r_j+1} & x_{r_j+2} & \dots & x_{r_{j+1}} \end{bmatrix} \quad \text{and} \quad Y_j = \begin{bmatrix} y_{r_j+1} & y_{r_j+2} & \dots & y_{r_{j+1}} \end{bmatrix} \qquad (j = 0, \dots, \ell)$$

be the $n \times (r_{j+1} - r_j)$ matrices with orthonormal columns which span the subspaces $\mathcal{M}_{j+1} \ominus \mathcal{M}_j$ and $\mathcal{N}_{j+1} \ominus \mathcal{N}_j$, respectively. Then a unitary matrix $\mathbf{U}$ satisfies (1.3) if and only if

$$\mathbf{U}X_j = Y_j D_j \quad \text{for some unitary matrix} \quad D_j \in \mathbb{C}^{(r_{j+1}-r_j)\times(r_{j+1}-r_j)} \quad (j = 1, \dots, \ell). \tag{4.8}$$

Introducing the matrices

$$X = [X_1 \;\; X_2 \;\; \dots \;\; X_\ell] \quad \text{and} \quad Y = [Y_1 \;\; Y_2 \;\; \dots \;\; Y_\ell], \tag{4.9}$$

we rewrite (4.8) as

$$\mathbf{U}X = YD \quad \text{for} \quad D = \operatorname{diag}(D_1, \dots, D_\ell). \tag{4.10}$$

The set of all unitary matrices $\mathbf{U}$ satisfying (4.10) we denote by $\mathcal{U}$. We formulate first our result for the Frobenius norm.

Theorem 4.2 *Let* $X,\ Y \in \mathbb{C}^{n\times r}$ *be two isometric matrices of the form* (4.9) *with CS decompositions* (4.1) – (4.3) *(for* $2r \le n$*) or* (4.4) – (4.6) *(for* $2r > n$*) with unitary matrices* Q, V *and* W*. Let*

$$X_j^* Y_j = P_j Z_j \qquad (P_j \ge 0, \quad Z_j Z_j^* = Z_j^* Z_j = I_{r_{j+1}-r_j}) \tag{4.11}$$

be the polar decompositions of matrices $X_j^* Y_j$*. Then*

$$\min_{\mathbf{U}\in\mathcal{U}} \|\mathbf{U} - I_n\|_2^2 = 4r - 2\text{trace}\left(\sqrt{X^*YY^*X}\right) - 2\sum_{i=1}^{\ell} \text{trace}\left(\sqrt{X_j^*Y_jY_j^*X_j}\right). \tag{4.12}$$

Moreover, a matrix $\mathbf{U}_{\min} \in \mathcal{U}$ *is a minimizer for the unitary distortion problem, i.e., it satisfies*

$$\min_{\mathbf{U}\in\mathcal{U}} \|\mathbf{U} - I_n\|_2 = \|\mathbf{U}_{\min} - I_n\|_2 \tag{4.13}$$

if and only if it is of the form

$$\mathbf{U}_{\min} = Q^* \begin{bmatrix} \Gamma T & -\Delta R & 0 \\ \Delta T & \Gamma R & 0 \\ 0 & 0 & I_{n-2r} \end{bmatrix} Q, \tag{4.14}$$

if $2r \le n$ *and*

$$\mathbf{U}_{\min} = Q^* \begin{bmatrix} \Gamma & 0 & -\Delta \\ 0 & I_{2r-n} & 0 \\ \Delta & 0 & \Gamma \end{bmatrix} \begin{bmatrix} T & 0 \\ 0 & R \end{bmatrix} Q, \tag{4.15}$$

if $2r > n$*, where* R *is an arbitrary* $r \times r$ *(if* $2r \leq n$*) or* $(n-r) \times (n-r)$ *(if* $2r > n$*) unitary matrix such that*

$$Rx = x \quad \textit{for every} \quad x \in \mathrm{Range}\,\Gamma \tag{4.16}$$

and T *is an* $r \times r$ *matrix of the form*

$$T = V^* \mathrm{diag}\,(D_1, \ldots, D_\ell) W, \tag{4.17}$$

where $D_j \in \mathbb{C}^{(r_{j+1}-r_j)\times(r_{j+1}-r_j)}$ *are arbitrary unitary matrices such that*

$$D_j x = Z_j^* x \quad \textit{for every} \quad x \in \mathrm{Range}\, P_j \qquad (j = 1, \ldots, \ell). \tag{4.18}$$

It is interesting to compare this result and Theorem 3.4, in which $X_i^* Y_j = A_{ij}$ for $1 \leq i, j \leq \ell = q - 1$. The formula (4.12) is the same as (3.10) because

$$n - \mathrm{trace}\sqrt{X_q^* Y_q Y_q^* X_q} = 2r - \mathrm{trace}\left(\sqrt{X^* Y Y^* X}\right)$$

is just the dimension of the space spanned by the columns of X and Y.
We establish two lemmas to prove Theorem 4.2.

Lemma 4.3 *Let* Γ *and* Δ *be positive semidefinite matrices such that* $\Gamma^2 + \Delta^2 = I$*, and let* $T \in \mathbb{C}^{r\times r}$ *be a unitary matrix. Then the matrix*

$$\mathbf{V} = \begin{bmatrix} \Gamma T & U_1 \\ \Delta T & U_2 \end{bmatrix}. \tag{4.19}$$

is unitary if and only if $U_1 = -\Delta R$ *and* $U_2 = \Gamma R$ *for some unitary matrix* $R \in \mathbb{C}^{r\times r}$*.*

Proof: The "if" part is clear. For the "only if" part, observe that

$$W = \begin{bmatrix} \Gamma & \Delta \\ -\Delta & \Gamma \end{bmatrix} \qquad \text{and} \qquad W\mathbf{V} = \begin{bmatrix} T & S \\ 0 & R \end{bmatrix}$$

are unitary, and hence $S = 0$. It follows that

$$\mathbf{V} = W^*(T \oplus R) = \begin{bmatrix} \Gamma T & -\Delta R \\ \Delta T & \Gamma R \end{bmatrix}$$

as asserted. ∎

Lemma 4.4 *Under the hypotheses and notation of Theorem 4.2, let* $\mathcal{D}$ *denote the set of all block diagonal unitary matrices* $D = \mathrm{diag}\,(D_1, \ldots, D_\ell)$ *with the blocks* $D_j \in \mathbb{C}^{(r_{j+1}-r_j)\times(r_{j+1}-r_j)}$ *for* $j = 1, \ldots, \ell$*. Then*

$$\|I_r - X^* Y D\|_2 = \min_{D \in \mathcal{D}} \|I_r - X^* Y D\|_2 \tag{4.20}$$

(i.e., D *minimizes the value of* $\|I_r - X^* Y D\|_2$ *over the set* $\mathcal{D}$*) if and only if the blocks* D_j *satisfy conditions* (4.18)*.*

Proof: In view of the block decompositions (4.9) and (4.10) of X, Y and D,

$$I_r - X^*YD = \left[\delta_{ij} I_{r_{j+1}-r_j} - X_i^* Y_j D_j\right]_{i,j=1}^{\ell}, \tag{4.21}$$

where δ_{ij} is the Kronecker symbol. Decomposing the last matrix as

$$I_r - X^*YD = [A_1(D_1), \ \ldots \ , \ A_\ell(D_\ell)], \qquad A_j(D_\ell) \in \mathbb{C}^{n\times(r_{j+1}-r_j)}, \tag{4.22}$$

we conclude by Lemma 2.7 that the minimum on the right hand side of (4.20) is attained for a unitary matrix $D = \text{diag}\,(D_1, \ldots, D_\ell)$ if and only if its blocks D_j's are minimizers for

$$\|A_j(D_j)\|_2 = \min_{D_j \text{ unitary}} \|A_j(D_j)\|_2 \qquad (j = 1, \ldots, \ell). \tag{4.23}$$

Comparing (4.22) and (4.21) and taking into account that the Frobenius norm is unitarily invariant, we get

$$\|A_j(D_j)\|_2 = \left\| \begin{bmatrix} -X_1^*Y_jD_j \\ \vdots \\ -X_{j-1}^*Y_jD_j \\ I - X_j^*Y_jD_j \\ -X_{j+1}^*Y_jD_j \\ \vdots \\ -X_\ell^*Y_jD_j \end{bmatrix} \right\|_2 = \left\| \begin{bmatrix} -X_1^*Y_j \\ \vdots \\ -X_{j-1}^*Y_j \\ D_j^* - X_j^*Y_j \\ -X_{j+1}^*Y_j \\ \vdots \\ -X_\ell^*Y_j \end{bmatrix} \right\|_2 .$$

Only the j-th block in the latter expression depends on D_j and therefore, again by Lemma 2.7, the extremal matrix D_j in (4.23) has to satisfy

$$\|D_j^* - X_j^*Y_j\|_2 = \min_{D_j \text{ unitary}} \|D_j^* - X_j^*Y_j\|_2 = \min_{D_j \text{ unitary}} \|I - X_j^*Y_jD_j\|_2.$$

Making use of the polar decompositions (4.11), we obtain

$$\|I - X_j^*Y_jD_j\|_2 = \|I - P_jZ_jD_j\|_2 = \|D_j^*Z_j^* - P_j\|_2 = \|P_j - Z_jD_j\|_2$$

and by Theorem 2.3, we conclude that D_j minimizes the value of $\|I - X_j^*Y_jD_j\|_2$ if and only if $Z_jD_jx = x$ for every vector $x \in \text{Range}\,P_j$. It follows now by Lemma 2.7, that $D = \text{diag}\,(D_1, \ldots, D_\ell)$ satisfies (4.20) if and only if

$$Z_jD_jx = x \qquad \text{for every } x \in \text{Range}\,P_j \quad (j = 1, \ldots, \ell).$$

Since Z_j's are unitary, the latter conditions are equivalent to (4.18). ∎

Proof of Theorem 4.2:
The case $2r \le n$. Making use of representations (4.1) we rewrite (4.10) as

$$Q\mathbf{U}Q^*QXW = QYVV^*DW \tag{4.24}$$

and get, in view of (4.1),

$$Q\mathbf{U}Q^* \begin{bmatrix} I_r \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \Gamma \\ \Delta \\ 0 \end{bmatrix} T, \quad \text{where} \quad T = V^*\text{diag}\,(D_1, \ldots, D_\ell)W. \tag{4.25}$$

We regard T as a function of unitary matrices $D_1, \ldots, D_\ell$, with V and W fixed. Thus, $Q\mathbf{U}Q^*$ necessarily has the form

$$Q\mathbf{U}Q^* = \begin{bmatrix} \Gamma T & U_{12} & U_{13} \\ \Delta T & U_{22} & U_{23} \\ 0 & U_{32} & U_{33} \end{bmatrix}.$$

Since we are minimizing the value of $\|\mathbf{U} - I_n\|_2 = \|Q\mathbf{U}Q^* - I\|_2$, upon applying Lemma 2.6 to the matrices $A = Q\mathbf{U}Q^* - I_n$ and $P = \begin{bmatrix} I_{2r} & 0 \\ 0 & 0 \end{bmatrix}$, we conclude that the minimal value of $\|\mathbf{U} - I\|_2$ is attained *only* for unitary matrices $\mathbf{U}$ of the form

$$\mathbf{U} = Q^* \begin{bmatrix} \Gamma T & U_{12} & 0 \\ \Delta T & U_{22} & 0 \\ 0 & 0 & I_{n-2r} \end{bmatrix} Q.$$

By Lemma 4.3, $U_{12} = -\Delta R$ and $U_{22} = \Gamma R$ for some unitary matrix $R \in \mathbb{C}^{r\times r}$ and thus,

$$\mathbf{U} = Q^* \begin{bmatrix} \Gamma T & -\Delta R & 0 \\ \Delta T & \Gamma R & 0 \\ 0 & 0 & I_{n-2r} \end{bmatrix} Q, \tag{4.26}$$

which provides the representation formula (4.14) for minimizers. It remains to show that $\mathbf{U}$ of the form (4.26) is a minimizer if and only if the matrices T and R are subjects to (4.16)–(4.18). Since the norm is unitarily invariant and since Q, T and R are unitary, it follows by (4.26), that

$$\min_{D_1,\ldots,D_\ell,R} \|\mathbf{U}-I_n\|_2 = \min_{D_1,\ldots,D_\ell,R} \left\| \begin{bmatrix} \Gamma T - I_r & -\Delta R \\ \Delta T & \Gamma R - I_r \end{bmatrix} \right\|_2 = \min_{D_1,\ldots,D_\ell,R} \left\| \begin{bmatrix} \Gamma - T^* & -\Delta \\ \Delta & \Gamma - R^* \end{bmatrix} \right\|_2. \tag{4.27}$$

By Lemma 2.7, it remains to find separately unitary matrices $D_1, \ldots, D_\ell$ and R which minimize the values of $\|\Gamma - T^*\|_2$ and $\|\Gamma - R^*\|_2$. By Theorem 2.3,

$$\|\Gamma - R^*\|_2 = \min_{R \text{ unitary}} \|\Gamma - R^*\|_2$$

if and only if $R^*x = x$ for every vector $x \in \operatorname{Range}\Gamma$. Since R is unitary, this condition is equivalent to (4.16). On the other hand, in view of (4.25) and the first relation in (4.7) and since the norm is unitarily invariant,

$$\|\Gamma - T^*\|_2 = \|I_r - \Gamma T\|_2 = \|I_r - \Gamma V^* D W\|_2 = \|I_r - W\Gamma V^* D\|_2 = \|I_r - X^* Y D\|_2. \tag{4.28}$$

and by Lemma 4.4, a matrix T of the form (4.17) minimizes $\|\Gamma - T^*\|_2$ if and only if the matrices D_j are subject to (4.18).

The case $2r > n$. Making use of representation (4.4) and of equality (4.24), we get, in view of (4.10),

$$Q\mathbf{U}Q^* \begin{bmatrix} I_{n-r} & 0 \\ 0 & I_{2r-n} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \Gamma & 0 \\ 0 & I_{2r-n} \\ \Delta & 0 \end{bmatrix} T, \quad \text{where} \quad T = V^*\operatorname{diag}(D_1, \ldots, D_\ell)W. \tag{4.29}$$

Thus, $Q\mathbf{U}Q^*$ necessarily has the form

$$Q\mathbf{U}Q^* = \begin{bmatrix} \Gamma & 0 & U_{13} \\ 0 & I_{2r-n} & 0 \\ \Delta & 0 & U_{33} \end{bmatrix} \begin{bmatrix} T & 0 \\ 0 & I_{n-r} \end{bmatrix}.$$

By Lemma 4.3, $U_{13} = -\Delta R$ and $U_{33} = \Gamma R$ for some unitary matrix $R \in \mathbb{C}^{(n-r)\times(n-r)}$ and thus,

$$\mathbf{U} = Q^* \begin{bmatrix} \Gamma & 0 & -\Delta \\ 0 & I_{2r-n} & 0 \\ \Delta & 0 & \Gamma \end{bmatrix} \begin{bmatrix} T & 0 \\ 0 & R \end{bmatrix} Q, \tag{4.30}$$

which leads to the representation formula (4.15). As in the case $2r \le n$, we have to minimize

$$\|\mathbf{U} - I_n\| = \left\| \begin{bmatrix} \begin{bmatrix} \Gamma & 0 \\ 0 & I_{2r-n} \end{bmatrix} - T^* & \begin{matrix} -\Delta \\ 0 \end{matrix} \\ \begin{matrix} \Delta & \quad 0 \end{matrix} & \Gamma - R^* \end{bmatrix} \right\|_2 \tag{4.31}$$

or equivalently (by Lemma 2.7), to minimize

$$\|\Gamma - R^*\|_2 \quad \text{and} \quad \left\| \begin{bmatrix} \Gamma & 0 \\ 0 & I_{2r-n} \end{bmatrix} - T^* \right\|_2$$

by the appropriate choice of T (of the form (4.29), where $D_1, \ldots, D_\ell$ are unitary) and unitary R. As in the case $2r \le n$, for $\mathbf{U}$ of the form (4.30) to be a minimizer, R has to satisfy (4.16). Furthermore, in view of (4.29) and the second relation in (4.7),

$$\begin{aligned} \left\| \begin{bmatrix} \Gamma & 0 \\ 0 & I_{2r-n} \end{bmatrix} - T^* \right\|_2 &= \left\| I_r - \begin{bmatrix} \Gamma & 0 \\ 0 & I_{2r-n} \end{bmatrix} T \right\|_2 \\ &= \left\| I_r - \begin{bmatrix} \Gamma & 0 \\ 0 & I_{2r-n} \end{bmatrix} V^*DW \right\|_2 \\ &= \left\| I_r - W \begin{bmatrix} \Gamma & 0 \\ 0 & I_{2r-n} \end{bmatrix} V^*D \right\|_2 = \|I_r - X^*YD\|_2, \end{aligned} \tag{4.32}$$

and by Lemma 4.4, a matrix T of the form (4.17) minimizes $\|\Gamma - T^*\|_2$ if and only if the matrices D_j are satisfy (4.18).

Finally, to compute explicitly the minimal value of $\|\mathbf{U} - I\|_2$, where $\mathbf{U} \in \mathcal{U}$, we chose a special minimizer corresponding to

$$R = \begin{cases} I_r, & \text{if } 2r \le n \\ I_{n-r}, & \text{if } 2r > n \end{cases}, \quad D^\circ = \operatorname{diag}(Z_1^*, \ldots, Z_\ell^*) \quad \text{and} \quad T = V^*D^\circ W.$$

Let $0 \le \gamma_1 \le \gamma_2 \le \ldots \le \gamma_r$ be the singular values of the matrix X^*Y (i.e., γ_i are the cosines of canonical angles between subspaces $\mathcal{M}$ and $\mathcal{N}$), let matrices Γ and Δ be defined by (4.2), (4.3) (for $2r \le n$) or by (4.5), (4.6) (for $2r > n$) and let (4.7) be the polar decomposition of the matrix X^*Y. Then for $2r \le n$, we get from (4.27) and (4.28)

$$\begin{aligned} &\min_{\mathbf{U}\in\mathcal{U}} \|\mathbf{U} - I_n\|_2^2 \\ = \;& \|\Gamma - T^*\|_2^2 + 2\|\Delta\|_2^2 + \|\Gamma - R^*\|_2^2 \\ = \;& \|I_r - X^*YD^\circ\|_2^2 + 2\|\Delta\|_2^2 + \|\Gamma - I_r\|_2^2 \\ = \;& \operatorname{trace}\left\{(I_r - (D^\circ)^*Y^*X)(I_r - X^*YD^\circ) + 2\Delta^2 + (\Gamma - I_r)^2\right\} \\ = \;& \operatorname{trace}\left\{2I_r - X^*YD^\circ - (D^\circ)^*Y^*X + (D^\circ)^*Y^*XX^*YD^\circ + 2\Delta^2 + \Gamma^2 - 2\Gamma\right\}. \end{aligned} \tag{4.33}$$

Since D°, V and W are unitary, it follows from (4.7) that

$$\text{trace}\ \{(D^\circ)^* Y^* X X^* Y D^\circ\} = \text{trace}\ \{(D^\circ)^* V \Gamma W^* W \Gamma V D^\circ\} = \text{trace}\, \Gamma^2,$$

and since Γ is positive semidefinite, we have also

$$\text{trace}\, \Gamma = \text{trace}\, \sqrt{X^* Y Y^* X}.$$

Furthermore, in view of (4.11),

$$\text{trace}\ (X^* Y D^\circ) = \sum_{j=1}^{\ell} \text{trace}\ \left(X_j^* Y_j Z_j^* \right) = \sum_{j=1}^{\ell} \text{trace}\, P_j = \sum_{i=1}^{\ell} \text{trace}\ \left(\sqrt{X_j^* Y_j Y_j^* X_j} \right).$$

Substituting the three last equalities into (4.33) and taking into account that $\Gamma^2 + \Delta^2 = I_r$, we get (4.12). If $2r > n$, then we get from (4.31) and (4.32)

$$\begin{aligned} \|\mathbf{U} - I_n\|_2^2 &= \| \begin{bmatrix} \Gamma & 0 \\ 0 & I_{2r-n} \end{bmatrix} - T^* \|_2^2 + 2\|\Delta\|_2^2 + \|\Gamma - R^*\|_2^2 \\ &= \|I_r - X^* Y D^\circ\|_2^2 + 2\|\Delta\|_2^2 + \|\Gamma - I_{n-r}\|_2^2 \end{aligned}$$

and using the preceding arguments we come again to (4.12). ∎

Recall that the positive semidefinite square root $\sqrt{A}$ of a positive semidefinite matrix A is a real analytic function of the real and imaginary parts of the entries of A, provided that the rank of A remains constant; this follows from the functional calculus formula

$$\sqrt{A} = \frac{1}{2\pi i} \int_{|\lambda| = \varepsilon} (\lambda I - A)^{-1} d\lambda + \int_{\Gamma} \sqrt{\lambda} (\lambda I - A)^{-1} d\lambda.$$

Here $\varepsilon > 0$ is such that that the positive semidefinite matrix A has no nonzero eigenvalues inside the circle $|\lambda| = \varepsilon$, and Γ is a suitable contour that surrounds the nonzero part of the spectrum of A. Using the analyticity of the square root, one obtains from (4.12) that $\min_{\mathbf{U} \in \mathcal{U}} \|\mathbf{U} - I_n\|_2^2$ is a real analytic function of the real and imaginary parts of the components of the vectors x_j and y_j, as long as the numbers $\ell, r_1, \ldots, r_\ell$ and the ranks of the matrices $X_j^* Y_j$ $(j = 1, \ldots, \ell)$ and of $X^* Y$ are kept constant.

For norms other than scalar multiples of the Frobenius norm, the proof of Theorem 4.2 breaks down. The reason is that Lemma 2.7 is valid only for unitarily invariant norms that are scalar multiples of the Frobenius norm. Indeed, let $\| \cdot \|_\Phi$ be such a unitarily invariant norm, and Let ϕ be the corresponding symmetric gauge function on $\mathbb{R}_+^n$, the convex cone of real vectors with n nonnegative components. We may assume that $\phi(1, 0, \ldots, 0) = 1$ by a suitable normalization. By assumption, $\phi \neq l_2$, where $l_2(x_1, \ldots, x_n) = \sqrt{x_1^2 + \ldots + x_n^2}$, $x_j \geq 0$ is the symmetric gauge function corresponding ot the Frobenius norm. Then there exists $1 \leq k < n$ such that $\phi(x) = l_2(x)$ for all $x \in \mathbb{R}_+^n$ with at most k nonzero entries, but $\phi(y) \neq l_2(y)$, for a certain $y = (y_1, \ldots, y_{k+1}, 0, \ldots, 0) \in \mathbb{R}_+^n$ with $y_1 \geq \cdots \geq y_{k+1} > 0$. Consider the family F_1 of $n \times n$ matrices with exactly k nonzero singular values $y_1, \ldots, y_k$, and the family F_2 of $n \times n$ matrices with the only nonzero singular value y_{k+1}. We see that $A_0 = y_1 E_{11} + \cdots + y_k E_{kk}$ is a minimizer of $\| \cdot \|_\Phi$ in F_1, and $B_0 = y_{k+1} E_{k+1,k+1}$ and $\tilde{B}_0 = y_{k+1} E_{11}$ are are both minimizers of $\| \cdot \|_\Phi$ in F_2. But then

$$\begin{aligned} \|[A_0 \ \ B_0]\|_\Phi &= \phi(y_1, \ldots, y_{k+1}, 0, \ldots, 0) \ \neq \ l_2(y_1, \ldots, y_{k+1}, 0, \ldots, 0) \\ &= l_2(\sqrt{y_1^2 + y_{k+1}^2}, y_2, \ldots, y_k, 0, \ldots, 0) \ = \ \phi(\sqrt{y_1^2 + y_{k+1}^2}, y_2, \ldots, y_k, 0, \ldots, 0) \\ &= \|[A_0 \ \ \tilde{B}_0]\|_\Phi. \end{aligned}$$

So, $[A_0 \;\; B_0]$ and $[A_0 \;\; \tilde{B}_0]$ cannot be both minimizers of $\|\cdot\|_\Phi$ in the set of $n \times 2n$ matrices $\{[A \;\; B] : A \in F_1, B \in F_2\}$.

In view of the observation made in the preceding paragraph, we need an additional hypothesis to deal with norms other than scalar multiples of the Frobenius norm. For Q-norms we have the following result (recall that $\mathcal{U}$ is the set of all unitary matrices $\mathbf{U}$ satisfying (4.10)).

Theorem 4.5 *Let X, $Y \in \mathbb{C}^{n\times r}$ be two isometric matrices of the form* (4.9) *with CS decompositions* (4.1) − (4.3) *(for $2r \le n$) or* (4.4) − (4.6) *(for $2r > n$) with unitary matrices Q, V and W. Assume further that the matrix VW^* is block diagonal: $VW^* = \operatorname{diag}(W_1, \ldots, W_\ell)$, where W_j is $r_j \times r_j$ ($j = 1, \ldots, \ell$). Then for any Q-norm,*

$$\min_{\mathbf{U}\in\mathcal{U}} \|\mathbf{U} - I_n\| = \left\| \begin{bmatrix} \Gamma - I & -\Delta \\ \Delta & \Gamma - I \end{bmatrix} \right\|. \tag{4.34}$$

The proof proceeds similarly to that of Theorem 4.2. We omit the details. It is worth noting that the above theorem can be viewed as a special case of Theorem 3.1 when there exist $V, W \in \mathcal{S}$ so that up to a permutation VAW is of the form

$$\begin{bmatrix} \Gamma & -\Delta \\ \Delta & \Gamma \end{bmatrix} \oplus I_{2m},$$

(using the notation of Theorem 3.1). One can translate the conclusion on the minimizers in Theorem 3.1 as well.

References

[1] S. N. Afriat, *Orthogonal and oblique projectors and the characteristics of pairs of vector spaces,* Proc. Cambridge Philos. Soc. **53** (1957), 800–816.

[2] J. G. Aiken, J. A. Erdos, and J. A. Goldstein, *Unitary approximation of positive operators*, Illinois J. Math. **24** (1981), 61–72.

[3] I. Y. Bar–Itzhack, D. Hershkowitz, and L. Rodman, *The pointing in real Euclidean space*, Journal of Guidance, Control, and Dynamics, **20** (1997), 916–922.

[4] R. Bhatia, *Matrix Analysis*, Springer-Verlag, New York, 1997.

[5] C. Davis, *Separation of two linear subspaces*, Acta Math. Szeged **19** (1958), 172–187.

[6] C. Davis and W. M. Kahan, *The rotation of eigenvectors by a perturbation. III,* SIAM J. Numer. Anal. **7** (1970), 1–46.

[7] A. Edelman, T. A. Arias, and S. T. Smith, *Geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. and Appl. **20** (1998), 303–353.

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

[9] I. C. Gohberg and M. G. Krein, *Introduction to the Theory of Linear Nonselfadjoint Operators*, Translations of Mathematical Monographs, Vol. 18, Amer. Math. Soc., Providence, R.I., 1969 (translation from Russian)

[10] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1989.

[11] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge-New York, 1985.

[12] R. A. Horn, N. H. Rhee, and W. So, *Eigenvalue inequalities and equalities*, Linear Algebra Appl. **270** (1998), 29–44.

[13] A.W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and its Applications*, Academic Press, London, 1979.

[14] G. W. Stewart, *On the perturbation of pseudo-inverses, projections and linear least squares problems,* SIAM Rev. **19** (1977), no. 4, 634–662.

[15] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press Inc., Boston, 1990.

[16] R. C. Thompson, *Singular values, diagonal elements, and convexity,* SIAM J. Appl. Math. **32** (1977), 39-63.