Norm Bounds for Summation of Two Normal Matrices

Man-Duen Choi[1] and Chi-Kwong Li[2]
Department of Mathematics, University of Toronto,
Toronto, Ontario, Canada M5S 3G3
Department of Mathematics, College of William and Mary,
Williamsburg, VA 23187-8795, USA

Abstract

A sharp upper bound is obtained for $\|A+iB\|$, where A and B are $n\times n$ Hermitian matrices satisfying $a_1I \le A \le a_2I$ and $b_1I \le B \le b_2I$. Similarly, an optimal bound is obtained for $\|U+V\|$, where U and V are $n\times n$ unitary matrices with any specified spectra; the study leads to some surprising phenomena of discontinuity concerning the spectral variation of unitary matrices. Moreover, it is proven that for two (non-commuting) normal matrices A and B with spectra $\sigma(A)$ and $\sigma(B)$, the optimal norm bound for $A+B$ equals

$$\min_{\lambda\in\mathbb{C}}\{\max_{\alpha\in\sigma(A)}|\alpha+\lambda| + \max_{\beta\in\sigma(B)}|\beta-\lambda|\}.$$

Extensions of the results to infinite dimensional cases are also considered.

Keywords: Norm bound, spectrum, spectral variation, spectral inequality, non-commuting normal matrices.
AMS Classifications: 47A30, 15A60.

[1] Research supported by an NSERC grant.

[2] Corresponding author. Research supported by an NSF grant.

1 Introduction

Let M_n be the algebra of $n \times n$ square matrices equipped with the spectral norm

$$\|T\| = \max\{\|Tx\| : x \in \mathbb{C}^n,\ \|x\| = 1\}$$

satisfying the C^*-norm features

$$\|T^*T\| = \|T\|^2, \qquad \text{and } \|TS\| \leq \|T\|\|S\|.$$

Suppose $A, B \in M_n$ are Hermitian matrices subject to the conditions

$$a_1 I \leq A \leq a_2 I \qquad \text{and} \qquad b_1 I \leq B \leq b_2 I.$$

There has been considerable interest in getting an upper bound for $\|A + iB\|$. For instance, if $O \leq A \leq aI$ and $O \leq B \leq bI$, then (see [1, Problem I.6.18])

$$\|A + iB\| \leq \{a^2 + b^2\}^{1/2},$$

and the equality holds if $A = aI$ and $B = bI$. If $-I \leq A \leq I$ and $-I \leq B \leq I$, then

$$\|A + iB\| \leq \|A\| + \|B\| = 2,$$

and the equality holds if $A + iB = \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix}$. In this paper, we obtain the optimal upper bound for $\|A + iB\|$ in terms of the given four real numbers $a_1 \leq a_2$, and $b_1 \leq b_2$ (see Theorem 2.1).

The norm bound problem can be transformed to another basic question in operator (matrix) inequalities. Namely, let $T \in M_n$ subject to four affine inequalities

$$a_1 I \leq \operatorname{Re} T \leq a_2 I \qquad \text{and} \qquad b_1 I \leq \operatorname{Im} T \leq b_2 I;$$

we wish to find the optimal bound c for the norm inequality $\|T\| \leq c$, which is equivalent to a quadratic inequality $T^*T \leq c^2 I$.

Using similar techniques, we obtain optimal bound for $\|U + V\|$, where U and V are $n \times n$ unitary matrices with any specified spectra (see Theorem 3.2); the study leads to some surprising phenomena of discontinuity concerning the spectral variation of unitary matrices.

We then extend our analysis to the summation of two (non-commuting) normal matrices. In fact, for any two normal matrices A and B with spectra $\sigma(A)$ and $\sigma(B)$, the optimal norm bound for $\|A + B\|$ equals

$$\min_{\lambda \in \mathbb{C}} \{ \max_{\alpha \in \sigma(A)} |\alpha + \lambda| + \max_{\beta \in \sigma(B)} |\beta - \lambda| \}, \tag{1.1}$$

(see Theorem 4.3). Moreover, extensions of the results to infinite dimensional cases are considered.

This paper is organized as follows. In Section 2, we obtain the optimal norm bound for $\|A + iB\|$ for two Hermitian matrices A and B in terms of their spectra. In Section

3. we study norm bounds for the sum of two unitary matrices. In particular, we get the best estimate of $\|U-V\|$ for unitary U and V, and see some jump-discontinuity phenomena about the set of unitary matrices. In Section 4, we prove that the quantity (1.1) gives the optimal norm bound for $\|A+B\|$ if A and B are normal matrices. In Section 5, we discuss the extension of the results to infinite dimensional cases.

We thank Rajendra Bhatia and the referee for drawing our attention to some additional references and related work. In particular, when A and B are both unitary, or when A is Hermitian and B is skew-Hermitian, our results improve the known bound (see [2] and [1, Theorem VI. 3.14])

$$\|A+B\| \leq \sqrt{2}\max\{|\alpha+\beta| : \alpha \in \sigma(A),\ \beta \in \sigma(B)\}.$$

Our study is about the uppper bound of $\|A+B\|$ for two normal matrices A and B; lower bounds for $\|A+B\|$ have been studied; e.g., see [1, Chapter VI] and the references therein. Related studies on norm bounds of the sum of two matrices with respect to other norms can be found in [2, 3, 4, 5]. The paper [2] is very close in spririt to our Section 4.

2 The sum of a Hermitian matrix and a skew-Hermitian matrix

In this section, we obtain the ultimate bound for any matrix $T = A + iB$, where A and B are Hermitian matrices subject to $a_1 I \leq A \leq a_2 I$ and $b_1 I \leq B \leq b_2 I$. Since

$$\|A+iB\| = \|A-iB\| = \|-A+iB\| = \|-A-iB\|,$$

we may assume without loss of generality that $a_2 \geq |a_1|$ and $b_2 \geq |b_1|$.

Theorem 2.1 *Suppose A, B are $n \times n$ Hermitian matrices subject to $a_1 I \leq A \leq a_2 I$ and $b_1 I \leq B \leq b_2 I$. Assume further that $a_2 \geq |a_1|$ and $b_2 \geq |b_1|$.*

(i) *If $a_1 b_2 + a_2 b_1 \geq 0$, then*

$$\|A+iB\| \leq |a_2 + ib_2| = \sqrt{a_2^2 + b_2^2}.$$

(ii) *If $a_1 b_2 + a_2 b_1 \leq 0$, then*

$$\|A+iB\| \leq \tau + \tau',$$

where

$$\tau = |a_1 - z_0| = |a_2 - z_0| = \frac{1}{2}\sqrt{(a_1-a_2)^2 + (b_1+b_2)^2}$$

and

$$\tau' = |ib_1 - z_0| = |ib_2 - z_0| = \frac{1}{2}\sqrt{(a_1+a_2)^2 + (b_1-b_2)^2}$$

with $z_0 = \{(a_1+a_2) + i(b_1+b_2)\}/2$.

(iii) *The bounds in* (i) *and* (ii) *are sharp in the following sense: If* $\{a_1, a_2\} \subseteq \sigma(A)$ *and* $\{b_1, b_2\} \subseteq \sigma(B)$*, then there exists a unitary* W *such that* $\|A + iWBW^*\|$ *attains the bound in each case.*

Note that $\tau + \tau' = |a_2 - z_0| + |ib_2 - z_0| \geq |a_2 - ib_2| = \sqrt{a_2^2 + b_2^2}$ is always valid. If $a_1b_2 + a_2b_1 = 0$, then $\tau = (1/2 + c)\sqrt{a_2^2 + b_2^2}$ and $\tau' = (1/2 - c)\sqrt{a_2^2 + b_2^2}$ with $2c = -a_1/a_2 = b_1/b_2$; thus $\tau + \tau' = \sqrt{a_2^2 + b_2^2}$ as in case (i).

Proof. Since $\sigma(A) \subseteq [a_1, a_2]$ and $\sigma(B) \subseteq [b_1, b_2]$, it follows that $\|A - zI\| \leq \max_{j=1,2} |a_j - z|$ and $\|iB - zI\| \leq \max_{j=1,2} |ib_j - z|$ for each complex number z. Write $\lambda = -\bar{z}$. Then

$$\begin{aligned}
\|A + iB\| &= \|(A + \lambda I) + (iB - \lambda I)\| \\
&\leq \|A + \lambda I\| + \|iB - \lambda I\| \\
&= \|(A + \lambda I)^*\| + \| - (iB - \lambda I)^*\| \\
&= \|A - zI\| + \|iB - zI\| \\
&\leq \max_{j=1,2} |a_j - z| + \max_{j=1,2} |ib_j - z|
\end{aligned}$$

for all $z \in \mathbb{C}$. Specifically, letting $z_0 = [(a_1 + a_2) + i(b_1 + b_2)]/2$, we get $|a_1 - z_0| = |a_2 - z_0| = \tau$ and $|ib_1 - z_0| = |ib_2 - z_0| = \tau'$; thus the inequality $\|A + iB\| \leq \tau + \tau'$ is always valid.

In case of $a_1b_2 + a_2b_1 \geq 0$, we select a different point $z = (a_1 + a_2)/2 + i(a_2 - a_1)b_2/(2a_2)$ in order to get the better bound $\sqrt{a_2^2 + b_2^2}$. (Here we ignore the degenerate case $a_2 = 0$ when $A = O$.) In fact, $|a_1 - z| = |a_2 - z| = \frac{a_2 - a_1}{2a_2}\sqrt{a_2^2 + b_2^2}$, $|ib_2 - z| = \frac{a_1 + a_2}{2a_2}\sqrt{a_2^2 + b_2^2}$, and $|ib_2 - z|^2 - |ib_1 - z|^2 = (b_2 - b_1)(a_2b_1 + a_1b_2)/a_2 \geq 0$. Therefore,

$$\|A + iB\| \leq |a_2 - z| + |ib_2 - z| = \sqrt{a_2^2 + b_2^2}.$$

These upper bounds for $\|A + iB\|$ are sharp as they are attained by $n \times n$ matrices with 1×1 and/or 2×2 matrices as direct summands. In case of $a_1b_2 + a_2b_1 \geq 0$, the 1×1 matrix $(a_2 + ib_2)$ is of norm $\sqrt{a_2^2 + b_2^2}$. In case of $a_1b_2 + a_2b_1 < 0$, the 2×2 matrix $T = A_0 + iB_0$ with

$$A_0 = \begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix} \text{ and } B_0 = \frac{1}{a_2 - a_1}\begin{pmatrix} -a_1(b_1 + b_2) & \sqrt{d} \\ \sqrt{d} & a_2(b_1 + b_2) \end{pmatrix},$$

where $d = -(a_1b_1 + a_2b_2)(a_1b_2 + a_2b_1)$, serves the purpose. In fact, B_0 is a Hermitian matrix with $\operatorname{tr} B_0 = b_1 + b_2$, $\det(B_0) = b_1b_2$ and so $\sigma(B_0) = \{b_1, b_2\}$; furthermore, $\operatorname{tr}(T^*T) = \operatorname{tr}(A_0^2 + B_0^2) = a_1^2 + a_2^2 + b_1^2 + b_2^2$, $\det(T^*T) = |\det T|^2 = (a_1a_2 - b_1b_2)^2$, and thus

$$\|T\| = \frac{1}{2}\left\{\sqrt{\operatorname{tr}(T^*T) - 2|\det(T)|} + \sqrt{\operatorname{tr}(T^*T) + 2|\det(T)|}\right\} = \tau + \tau'.$$

□

Remark 2.2 Note that the setting of Theorem 2.1 admits a geometrical interpretation. Namely, the given four real numbers (a_1, a_2, b_1, b_2) subject to $a_2 \geq |a_1|$ and $b_2 \geq |b_1|$ determine a rectangle

$$R = \{a + ib : a \in [a_1, a_2], b \in [b_1, b_2]\}$$

whose center

$$z_0 = \frac{a_1 + a_2}{2} + i\frac{b_1 + b_2}{2}$$

is a point in the first quadrant. We pay special attention to the location of $\omega = a_2 + ib_2$ the center of R and the vertex farthest away from the origin, and the location of the center z_0 with respect to the line segment L joining a_2 with ib_2 in the first quadrant. Hence, the inequality $a_1b_2 + a_2b_1 > 0$ means that z_0 (the center of the rectangle R) lies above the line segment L; thus the asserted norm bound $\sqrt{a_2^2 + b_2^2}$ is just the length of the line segment L, which is the same as the distance from the origin to the farthest vertex of the rectangle R. On the other hand, the inequality $a_1b_2 + a_2b_1 < 0$ means that the center of R lies below the line segment L; thus the asserted norm bound is just $|z_0 - a_2| + |z_0 - ib_2|$, the sum of the distances from the center of R to two ends of the line segment L, which is certainly larger than the length of L. In the particular case of the equality $a_1b_2 + a_2b_1 = 0$, which means that z_0 lies on the line segment L, the two norm bounds $\sqrt{a_2^2 + b_2^2}$ and $|z_0 - a_2| + |z_0 - ib_2|$ coincide.

Remark 2.3 Recall that the numerical range of a matrix $T \in M_n$ is the set

$$W(T) = \{x^*Tx : x \in \mathbb{C}^n,\ x^*x = 1\}.$$

Let R be the rectangle with vertices $a_j + ib_k$, where $j, k \in \{1, 2\}$. Then a matrix T has numerical range $W(T)$ lying inside R if and only if $T = A + iB$ such that A and B are Hermitian matrices subject to $a_1I \leq A \leq a_2I$ and $b_1I \leq B \leq b_2I$. Let $\omega = a_2 + ib_2$ and let Ω be the right-angled triangle formed by the three vertices $\omega, \bar{\omega}$, and $-\bar{\omega}$ that are equi-distant from the origin. It turns out that the conditions $a_2 \geq |a_1|$ and $b_2 \geq |b_1|$ together with $a_1b_2 + a_2b_1 \geq 0$ give rise to the situation that the rectangle R is a subset of the triangle Ω. We can therefore apply a result of Mirman [7] (see also [8]) to conclude that $\|A + iB\| \leq |\omega|$. Otherwise, the result of Mirman is not applicable whereas Theorem 2.1 provides a better way to obtain the optimal norm bound for $\|A + iB\|$. Furtermore, our bound improves the result in [5] asserting

$$\|A + iB\| \leq \sqrt{\|A\|^2 + 2\|B\|^2}$$

if R is on the right half plane.

As a supplement to Theorem 2.1, we give below a detailed description of the situations when the norm bounds in Theorem 2.1 are actually attained.

Lemma 2.4 *Suppose A and B are 2×2 Hermitian matrices with spectra $\sigma(A) = \{a_1, a_2\}$ and $\sigma(B) = \{b_1, b_2\}$. Assume further that $|a_1| \leq a_2$ and $|b_1| \leq b_2$. Then*

$$\|A + iB\| = \max\{\|A + iWBW^*\| : W \in M_2,\ W^*W = I_2\}$$

if and only if $A+iB$ is unitarily similar to

(i) *the diagonal matrix* $\begin{pmatrix} a_1+ib_1 & 0 \\ 0 & a_2+ib_2 \end{pmatrix}$ *in case of* $a_1b_2+a_2b_1 \geq 0$,

(ii) *the non-normal matrix*

$$\begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix} + \frac{i}{a_2-a_1}\begin{pmatrix} -a_1(b_1+b_2) & \sqrt{d} \\ \sqrt{d} & a_2(b_1+b_2) \end{pmatrix}$$

with $d = -(a_1b_1+a_2b_2)(a_1b_2+a_2b_1)$ *in case of* $a_1b_2+a_2b_1 > 0$.

Proof. By unitary similarity, we may assume that $A = \begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix}$ and B is a real Hermitian matrix. As every 2×2 real Hermitian matrix with spectrum $\{b_1, b_2\}$ has the form

$$B_s = \frac{b_1+b_2}{2}I + \frac{b_2-b_1}{2}\begin{pmatrix} -s & \sqrt{1-s^2} \\ \sqrt{1-s^2} & s \end{pmatrix} \quad \text{with} \quad s \in [-1,1],$$

we proceed to find s that maximizes the norm of $T_s = A + iB_s$ which can be computed through the equality

$$\|T_s\|^2 = \|T_s^*T_s\| = \frac{1}{2}\left\{\sqrt{(\mathrm{tr}\,(T_s^*T_s))^2 - 4\det(T_s^*T_s)} + \mathrm{tr}\,(T_s^*T_s)\right\}.$$

Since $\mathrm{tr}\,(T_s^*T_s) = \mathrm{tr}\,(A_0^2+B_0^2) = a_1^2+a_2^2+b_1^2+b_2^2$ is independent of s and $\det(T_s^*T_s) = |\det(T_s)|^2 = (a_1a_2-b_1b_2)^2 + \frac{1}{4}[(a_1+a_2)(b_1+b_2) - s(a_2-a_1)(b_2-b_1)]^2$, we see that $\max_{s\in[-1,1]}\|T_s\|$ occurs exactly when $\min_{s\in[-1,1]}\det(T_s^*T_s)$ occurs. There are two cases:

(i) Suppose $a_1b_2+a_2b_1 \geq 0$; equivalently, $(a_1+a_2)(b_1+b_2) \geq (a_2-a_1)(b_2-b_1) \geq 0$. Then $\min_{s\in[-1,1]}\det(T_s^*T_s)$ occurs at $s=1$ and $T_s = \begin{pmatrix} a_1+ib_1 & 0 \\ 0 & a_2+ib_2 \end{pmatrix}$ has the maximal norm as desired. (For the degenerate case $a_1=a_2$ or $b_1=b_2$, s can be arbitrary as all T_s are unitarily similar, and the conclusion still holds.)

(ii) Suppose $a_1b_2+a_2b_1 < 0$; equivalently $(a_2-a_1)(b_2-b_1) > (a_1+a_2)(b_1+b_2) \geq 0$. Then $\min_{s\in[-1,1]}\det(T_s^*T_s) = (a_1a_2-b_1b_2)^2$ occurs at $s = (a_1+a_2)(b_1+b_2)/[(a_2-a_1)(b_2-b_1)]$ and the corresponding T_s is the 2×2 matrix as specified. □

Proposition 2.5 *Suppose A, B are $n\times n$ Hermitian matrices subject to $a_1I \leq a \leq a_2I$ and $b_1I \leq B \leq b_2I$. Assume further that $a_2 \geq |a_1|$ and $b_2 \geq |b_1|$.*

(i) *Suppose* $a_1b_2+a_2b_1 \geq 0$. *Then*

$$\|A+iB\| = \sqrt{a_2^2+b_2^2}$$

if and only if $A+iB$ is unitarily similar to $C_1 \oplus C_2$ with $C_1 \in M_k$, $C_2 \in M_{n-k}$, where k is a positive integer $\leq n$, $\|C_2\| < \sqrt{a_2^2+b_2^2}$ (here C_2 is absent if $k=n$), and

(a) *C_1 is a normal matrix subject to $\sigma(C_1) \subseteq \{a_2 - ib_2, a_2 + ib_2\}$ if $(a_1, b_1) = (a_2, -b_2)$,*

(b) *C_1 is a normal matrix subject to $\sigma(C_1) \subseteq \{-a_2 + ib_2, a_2 + ib_2\}$ if $(a_1, b_1) = (-a_2, b_2)$,*

(c) *$C_1 = (a_2 + ib_2)I_k$ for all other cases.*

(ii) *Suppose $a_1b_2 + a_2b_1 < 0$. Then*

$$\|A+iB\| = \frac{1}{2}\sqrt{(a_1-a_2)^2+(b_1+b_2)^2} + \frac{1}{2}\sqrt{(a_1+a_2)^2+(b_1-b_2)^2}$$

if and only if $A+iB$ is unitarily similar to $C_1 \oplus C_2$, where C_1 equals a direct sum of k copies of the 2×2 matrix

$$T_0 = \begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix} + \frac{i}{a_2 - a_1}\begin{pmatrix} -a_1(b_1+b_2) & \sqrt{d} \\ \sqrt{d} & a_2(b_1+b_2) \end{pmatrix}$$

with $1 \leq k \leq n/2$, $d = -(a_1b_1 + a_2b_2)(a_1b_2 + a_2b_1)$, and $C_2 \in M_{n-2k}$ satisfying $\|C_2\| < \|A+iB\|$ (here C_2 is absent if $k = n/2$).

Proof. By direct computation (or as calculated in the proof of Theorem 2.1 and Lemma 2.4), the 2×2 matrix T_0 attains the norm bound. Thus, all of the "if" cases can be verified readily.

Conversely, suppose $\|A+iB\|$ has attained the upper bound as in Theorem 2.1. From the proof of Theorem 2.1, there exists $\lambda_0 \in \mathbb{C}$ be such that

$$\|A+iB\| = \|A+\lambda_0 I\| + \|iB - \lambda_0 I\|;$$

so, there is a unit vector $x \in \mathbb{C}^n$ such that

$$\|A+iB\| = \|(A+iB)x\| = \|(A+\lambda_0 I)x + (iB-\lambda_0 I)x\| \leq \|A+\lambda_0 I\| + \|iB-\lambda_0 I\| = \|A+iB\|.$$

Assume further $\|A+\lambda_0 I\| \neq 0$, $\|iB - \lambda_0 I\| \neq 0$ in order to omit the trivial cases. Then $\|(A+\lambda_0 I)x\| = \|A+\lambda_0 I\|$, $\|(iB-\lambda_0 I)x\| = \|iB-\lambda_0 I\|$, and $(A+\lambda_0 I)x$ is a positive multiple of $(iB-\lambda_0 I)x$; thus, span $\{x, Ax\}$ = span $\{x, Bx\} = \mathcal{S}$, say. Since $(A+\lambda_0 I)^*(A+\lambda_0 I)x = \|A+\lambda_0 I\|^2 x$, we have $A^2x = (\|A+\lambda_0 I\|^2 - |\lambda_0|^2)x - (\bar{\lambda}_0 + \lambda_0)Ax$; thus, $\mathcal{S}$ is an invariant subspace, and hence a reducing subspace of the Hermitian matrix A. Similarly, since $(iB-\lambda_0 I)^*(iB-\lambda_0 I)x = \|iB-\lambda_0 I\|^2 x$, we see that $\mathcal{S}$ is a reducing subspace for B. Therefore, $\mathcal{S}$ is a reducing subspace for $A+iB$. There are two possible situation.

(a) $\mathcal{S}$ is of dimension 1. Then $A+iB$ can only attain the value a_2+ib_2 (or a_1+ib_2 or a_2+ib_1 in the degenerate cases) as the possible reducing eigenvalue of maximal modulus.

(b) $\mathcal{S}$ is of dimension 2. By Lemma 2.4, we get a diagonal matrix or a non-normal matrix as a direct summand of $A+iB$.

In both cases, we can extract as many copies of the norm attaining summand until the remaining part has norm strictly less than $\|A+iB\|$. □

3 The Sum of Two Unitary Matrices

Suppose U and V are $n \times n$ unitary matrices with spectra $\mathcal{U}$ and $\mathcal{V}$. Obviously,

$$\max\{|u+v| : u \in \mathcal{U}, v \in \mathcal{V}\} \le \max\{\|U + W^*VW\| : W \in M_n, W^*W = I_n\} \le 2. \qquad (3.1)$$

It turns out that at least one of these two inequalities must be an equality. We need the following observation to prove our main result.

Lemma 3.1 *Suppose u_1, u_2, v_1, v_2 are complex numbers on the unit circle such that the line segment joining u_1, u_2, and the line segment joining v_1, v_2, intersect at w with $|w| \le 1$. Then*

$$U = \begin{pmatrix} w & -u_1u_2\sqrt{1-|w|^2} \\ \sqrt{1-|w|^2} & u_1u_2\bar{w} \end{pmatrix} \quad \textit{and} \quad V = \begin{pmatrix} w & -v_1v_2\sqrt{1-|w|^2} \\ \sqrt{1-|w|^2} & v_1v_2\bar{w} \end{pmatrix}$$

are unitary matrices with spectra $\sigma(U) = \{u_1, u_2\}$, $\sigma(V) = \{v_1, v_2\}$ and $\|U+V\| = 2$.

Proof. It is readily seen that U is a unitary matrix with $\det(U) = u_1u_2$. Write $w = tu_1 + (1-t)u_2$, where $t \in [0,1]$. Then $u_1u_2\bar{w} = (1-t)u_1 + tu_2$, and hence $\operatorname{tr} U = u_1 + u_2$, and $\sigma(U) = \{u_1, u_2\}$. Similarly, V is unitary with eigenvalues v_1 and v_2. The assertion $\|U+V\| = \|(U+V)e_1\| = 2$ is clear. □

Theorem 3.2 *Let U and V be $n \times n$ unitary matrices with spectra $\sigma(U)$ and $\sigma(V)$.*

(i) *If there is an arc Γ of the unit circle $\mathbf{T}$ such that*

$$\sigma(U) \subseteq \Gamma \quad \textit{and} \quad \sigma(V) \subseteq \mathbf{T} \setminus \Gamma, \qquad (3.2)$$

then

$$\|U+V\| \le \max\{|u+v| : u \in \sigma(U), v \in \sigma(V)\}.$$

(ii) *If there does not exist an arc Γ of the unit circle $\mathbf{T}$ satisfying (3.2), then*

$$\|U+V\| \le 2.$$

(iii) *The bounds in* (i) *and* (ii) *are sharp as there exists a unitary matrix W such that $\|U + WVW^*\|$ attains the bound in each case.*

Proof. Case (i) Suppose $u_0 \in \sigma(U)$ and $v_0 \in \sigma(V)$ satisfy

$$|u_0 + v_0| = \max\{|u+v| : u \in \sigma(U) \text{ and } v \in \sigma(V)\}.$$

Since $|u-v|^2 = 2 - |u+v|^2$, it follows that

$$|u_0 - v_0| = \min\{|u-v| : u \in \sigma(U) \text{ and } v \in \sigma(V)\} > 0.$$

After a rotation, we may assume that there exists $\theta_2 \in [0, \pi)$ so that

$$\{e^{i\theta_2}, e^{-i\theta_2}\} \subseteq \sigma(V) \subseteq \{e^{i\theta} : \theta \in [-\theta_2, \theta_2]\}.$$

Thus, $v_0 = e^{i\theta_2}$ or $e^{-i\theta_2}$. Without loss of generality, we may assume further that $v_0 = e^{i\theta_2}$. Then $u_0 = e^{i\theta_1}$ with $\theta_1 \in (\theta_2, \pi]$, Thus,

$$\operatorname{Re} u \leq \operatorname{Re} u_0 < \operatorname{Re} v_0 \leq \operatorname{Re} v$$

for all $u \in \sigma(U)$ and $v \in \sigma(V)$. For any positive real numbers λ, we have

$$|u_0 + \lambda|^2 - |u + \lambda|^2 = 2\operatorname{Re}\,(u_0 - u)\lambda \geq 0 \quad \text{and} \quad |v_0 - \lambda|^2 - |v - \lambda|^2 = -2\operatorname{Re}\,(v_0 - v)\lambda \geq 0$$

for all $u \in \sigma(U)$ and $v \in \sigma(V)$; thus

$$\|U + \lambda I\| = |u_0 + \lambda| \qquad \text{and} \qquad \|V - \lambda I\| = |v_0 - \lambda|.$$

As the right half circle joining v_0 and $-v_0$ through the point 1 includes the points $-u_0$, we see that the line joining v_0 and $-u_0$ meets the real line at a positive real number λ_0. In fact, $\lambda_0 = \sin(\theta_1 - \theta_2)/(\sin\theta_1 + \sin\theta_2)$, or 1 in the degenerate case if $\theta_1 = \pi$ and $\theta_2 = 0$. Hence, $u_0 + \lambda_0$ and $v_0 - \lambda_0$ are two complex numbers of same argument. Therefore

$$\|U + V\| \leq \|U + \lambda_0 I\| + \|V - \lambda_0 I\| = |u_0 + \lambda_0| + |v_0 - \lambda_0| = |u_0 + v_0|$$

as desired.

Case (ii) is obvious.

(iii) The bound in (i) is sharp as we can find diagonal matrices U and V with matching eigenvalues to attain the norm. The bound in (ii) is sharp as the pair of 2×2 unitary matrices in Lemma 3.1 attain the norm $\|U + V\| = 2$. □

Corollary 3.3 *Let U and V be $n \times n$ unitary matrices with spectra $\sigma(U)$ and $\sigma(V)$, and let a and b be positive numbers.*

(i) *If there is an arc Γ of the unit circle $\mathbf{T}$ such that 3.2 holds, then*

$$\|aU + bV\| \leq \max\{|au + bv| : u \in \sigma(U), v \in \sigma(V)\}.$$

(ii) *If there does not exist an arc Γ of the unit circle $\mathbf{T}$ satisfying (3.2), then*

$$\|aU + bV\| \leq a + b.$$

(iii) *The bounds in (i) and (ii) are sharp as there exists a unitary matrix W such that $\|aU + bWVW^*\|$ attains the bound in each case.*

Proof. Since $(aU+bV)^*(aU+bV) = (a-b)^2+ab(U+V)^*(U+V)$, we see that $\|aU+bV\|^2 = (a-b)^2 + ab\|U+V\|^2$. So, all of the results in Theorem 3.2 apply. □

Evidently, Theorem 3.2 is useful to estimate $\|U - V\|$ for a pair of unitary matrices U and V.

Corollary 3.4 *Let U and V be $n \times n$ unitary matrices with spectra $\sigma(U)$ and $\sigma(V)$. If there exists an arc Γ of the unit circle $\mathbf{T}$ such that*

$$\sigma(U) \subseteq \Gamma \quad \textit{and} \quad \sigma(-V) \subseteq \mathbf{T} \setminus \Gamma, \tag{3.3}$$

then

$$\max_{W^*W=I} \|U - W^*VW\| = \max\{|u-v| : u \in \sigma(U),\ v \in \sigma(V)\}; \tag{3.4}$$

otherwise, (i.e., (3.3) is not valid), we have

$$\max_{W^*W=I} \|U - W^*VW\| = 2.$$

Putting $U = V$ in Corollary 3.4, we get the following formula for

$$\max_{W^*W=I} \|U - W^*UW\|,$$

which is the diameter of the unitary orbit $\{W^*UW : W^*W = I\}$ of U.

Corollary 3.5 *Let U be an $n \times n$ unitary matrix. If its spectrum $\sigma(U)$ lies in an arc of the unit circle with length less than π, then*

$$\max_{W^*W=I} \|U - W^*UW\| = \max\{|u-v| : u, v \in \sigma(U)\};$$

otherwise,

$$\max_{W^*W=I} \|U - W^*UW\| = 2.$$

In particular,

$$\max_{W^*W=I} \|U - W^*UW\| < 2$$

if and only if $\sigma(U)$ lies in an arc of the unit circle with length less than π.

In practice, there are different conditions ensuring that there exists an arc Γ of the unit circle $\mathbf{T}$ satisfying (3.3) so that (3.4) holds. For example, if there is an arc Γ of the unit circle $\mathbf{T}$ with length less than π containing $\sigma(U) \cup \sigma(V)$, then condition (3.3) holds. In particular, if U, V satisfy

$$\|U - I\| < \sqrt{2} \quad \textit{and} \quad \|V - I\| < \sqrt{2},$$

then all the eigenvalues of U and V lie in an open semi-circular arc of the unit circle symmetric about the point 1. The above condition is particularly useful in studying the distance between unitary matrices in a small neighborhood of I.

Another condition implying the existence of an arc Γ of the unit circle satisfying (3.3) is that

$$\max\{|u-v| : u \in \sigma(U),\ v \in \sigma(V)\} < \sqrt{2}.$$

To see this, assume that the hypothesis of Corollary 3.4 does not hold. Then there exists $e^{is_1}, e^{is_2} \in \sigma(U)$ and $e^{it_1}, e^{it_2} \in \sigma(V)$ such that $s_1 \le t_1 \le s_2 \le t_2 \le s_1 + 2\pi$, i.e., the four points divide the unit circle into 4 arcs. Thus, the largest arc must have length larger than or equal to $\pi/2$, and the distance between the two end points of this arc is at least $\sqrt{2}$. Again, this result is best possible as shown in the following example.

Example 3.6 *Let*

$$U = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \textit{and} \quad V = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Then

$$\max\{|u-v| : u \in \sigma(U),\ v \in \sigma(V)\} = \sqrt{2} \quad \textit{and} \quad \|U-V\| = 2.$$

Yet, another condition implying the hypothesis of the Corollary 3.4 is that

$$\max\{|\mu-\eta| : \mu,\eta \in \sigma(U)\cup\sigma(V)\} < \sqrt{3}.$$

In fact, if the above inequality holds, then for any $e^{it_0} \in \sigma(U)\cup\sigma(V)$, all the other elements in $\sigma(U)\cup\sigma(V)$ can be written in the form $e^{is_1},\dots,e^{is_p}$ and $e^{it_1},\dots,e^{it_q}$ so that $p+q+1 = 2n$,

$$s_1 \le \cdots \le s_p \le t_0 \le t_1 \le \cdots \le t_q,$$

$$t_0 < s_1 + 2\pi/3, \quad \textit{and} \quad t_q < t_0 + 2\pi/3.$$

Since $|e^{is_1} - e^{it_q}| < \sqrt{3}$, we see that $t_q < s_1 + 2\pi/3$. Thus, $\sigma(U)\cup\sigma(V)$ lies in an open arc of the unit circle with length less than $2\pi/3$. The above result is best possible as shown by the following example.

Example 3.7 *Let*

$$U = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \quad \textit{and} \quad V = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 1 & 0 & 0 \end{pmatrix}.$$

Then U and V are unitarily similar, $\sigma(U) = \sigma(V) = \{1, \omega, \omega^2\}$ with $\omega = e^{i2\pi/3}$, and

$$\|U-V\| = 2 > \sqrt{3} = \max\{|\mu-\eta| : \mu,\eta \in \sigma(U)\}.$$

Note that the discussion in this section reveals that there are some sorts of discontinuity phenomena concerning the spectral variation of unitary matrices. We summarize them in the following.

Proposition 3.8 *For $t \in [0,2]$, let*

$$\Phi(t) = \max\{\|U - V\| : U^*U = I_n = V^*V,\ |u - v| \le t \text{ whenever } u \in \sigma(U),\ v \in \sigma(V)\},$$

and

$$\Psi(t) = \max\{\|U - WUW^*\| : U^*U = I_n = W^*W,\ |u - v| \le t \text{ whenever } u, v \in \sigma(U)\}.$$

(i) *If $n = 1$, then $\Phi(t) = t$ for all $t \in [0,2]$; if $n \ge 2$, then*

$$\Phi(t) = \begin{cases} t & \text{for } t < \sqrt{2}, \\ 2 & \text{for } t \in [\sqrt{2}, 2]. \end{cases}$$

(ii) *If $n \le 2$, then $\Psi(t) = t$ for all $t \in [0,2]$; if $n \ge 3$, then*

$$\Psi(t) = \begin{cases} t & \text{for } t < \sqrt{3}, \\ 2 & \text{for } t \in [\sqrt{3}, 2]. \end{cases}$$

Remark 3.9 It is helpful to use harmonic analysis (alias, the geometry of the circle) to explain the underlying truth (or myth) associated with the discontinuity of $\Phi(t)$ in Proposition 3.8. Let $(\alpha_1(t), \alpha_2(t); \beta_1(t), \beta_2(t))$ be four points on the unit circle moving continuously with respect to time t. Assume further that the quadruple is $(-1, 1; , i, -i)$ at $t = 0$ and the quadruple is $(1, 1; 1, 1)$ at $t = 1$. To measure the separation between α's and β's, we introduce a function

$$g(t) = \max\{|\alpha_j(t) - \beta_k(t)| :\ 1 \le j \le 2 \text{ and } 1 \le k \le 2\}$$

with $g(0) = \sqrt{2}$ and $g(1) = 0$. Then it is clear from the structure of the unit circle, there exists $t_0 \in (0,1)$ such that $g(t_0) = 2$. Henceforth, we can describe such an astonishing phenomenon as a paradox in harmonic analysis:

Paradox (for the continuous movement of α's and β's on the unit circle). In order to come *closer*, they should go *farther apart.* Before coming *altogether*, they should have already gone *farthest apart.*

The discontinuity of $\Psi(t)$ at $t = \sqrt{3}$ is also associated with a fascinating phenomenon about the continuous movement of three points on the unit circle. Let $(\alpha_1(t), \alpha_2(t), \alpha_3(t))$ be three points on the unit circle moving continuously with respect to time t. Assume further that the triple is $(1, e^{i2\pi/3}, e^{i4\pi/3})$ at $t = 0$ and the triple is $(1, 1, 1)$ at $t = 1$. To measure the scattering of α's, we introduce a continuous function

$$h(t) = \max\{|\alpha_j(t) - \alpha_k(t)| :\ 1 \le j < k \le 3\} \text{ with } h(0) = \sqrt{3} \text{ and } h(1) = 0.$$

Then from the special feature of the unit circle, there exists $t_0 \in (0,1)$ such that $h(t_0) = 2$. In other words, the initial position $(1, e^{i2\pi/3}, e^{i4\pi/3})$ is a critical situation that prevents h from decreasing. Indeed, starting from its local minimum $h(0) = \sqrt{3}$, the function h must climb, all the way up to reach its absolute maximum $h(t_0) = 2$, before falling down to its absolute minimum $h(1) = 0$.

As a supplement to Theorem 3.2, we provide below a detailed description of the situations when the norm bounds in Theorem 3.2 are attained.

Lemma 3.10 *Let U and V be 2×2 unitary matrices with spectra $\sigma(U)=\{u,\bar{u}\}$ and $\sigma(V)=\{v,\bar{v}\}$. Assume further that* $\mathrm{Im}\, u>0$ *and* $\mathrm{Im}\, v>0$. *Then*

$$\|U+V\|=\max\{\|U+W^*VW\| : W\in M_2,\ W^*W=I_2\}$$

if and only if $U+V$ is unitarily similar to $\begin{pmatrix} u+v & 0 \\ 0 & \bar{u}+\bar{v}\end{pmatrix}$.

Proof. We may assume that $U=\begin{pmatrix} u & 0 \\ 0 & \bar{u}\end{pmatrix}$. Then by unitary similarity (without changing U), each unitary matrix with spectrum $\{v,\bar{v}\}$ is of the form

$$V_w=\begin{pmatrix} w & -\sqrt{1-|w|^2} \\ \sqrt{1-|w|^2} & \bar{w}\end{pmatrix} \quad \text{subject to} \quad |w|\le 1 \quad \text{and} \quad w+\bar{w}=v+\bar{v}.$$

Equivalently, w is subjected to the condition $w=v-is\mathrm{Im}\, v$ with $s\in[0,2]$. We proceed to find w that maximizes the norm of $T_w=U+V_w$. Since $T_w^*T_w=d_wI$ with

$$d_w=|u+w|^2+1-|w|^2=2+2\mathrm{Re}\,\bar{u}w=2+2\mathrm{Re}\,\bar{u}v-2s\mathrm{Im}\,u\mathrm{Im}\,v,$$

it follows that $\max\|T_w\|$ occurs at $s=0$ and $w=v$ and $V_w=\begin{pmatrix} v & 0 \\ 0 & \bar{v}\end{pmatrix}$ as desired. □

Proposition 3.11 *Let U and V be $n\times n$ unitary matrices with spectra $\sigma(U)$ and $\sigma(V)$.*

(i) *Suppose there is an arc Γ of the unit circle* $\mathbf{T}$ *such that (3.2) holds. Then*

$$\|U+V\|=\max\{|u+v| : u\in\sigma(U), v\in\sigma(V)\}$$

if and only if U and V have a common eigenvector with respect to $u_0\in\sigma(U)$ and $v_0\in\sigma(V)$ satisfying

$$|u_0+v_0|=\max\{|u+v| : u\in\sigma(U),\ v\in\sigma(V)\}.$$

(ii) *Suppose there does not exist an arc Γ of the unit circle* $\mathbf{T}$ *satisfying (3.2). Then $\|U+V\|=2$ if and only $U-V$ is a singular matrix.*

Proof. Case (i) As in the proof of Theorem 3.2, we may assume, without loss of generality, that there exist $u_0\in\sigma(U)$ and $v_0\in\sigma(V)$ such that

$$|u_0+v_0|=\max\{|u+v| : u\in\sigma(U),\ v\in\sigma(V)\}$$

and $\operatorname{Im} u_0 \geq 0$, $\operatorname{Im} v_0 \geq 0$, and $\operatorname{Re} u \leq \operatorname{Re} u_0 < \operatorname{Re} v_0 \leq \operatorname{Re} v$ for all $u \in \sigma(U)$ and $v \in \sigma(V)$. Now suppose further that

$$\|U+V\| = \max\{\|U + WVW^*\| : W \in M_n,\ W^*W = I_n\} = |u_0 + v_0|.$$

From the proof of Theorem 3.2 again, there is a positive real number λ_0 so that $\|U+\lambda_0 I\| = |u_0 + \lambda_0|$, $\|V - \lambda_0 I\| = |v_0 - \lambda_0|$, and

$$\|U+V\| = \|U+\lambda_0 I\| + \|V - \lambda_0 I\|.$$

Assume $u_0 + \lambda_0 \neq 0$ and $u_0 - \lambda_0 \neq 0$ to avoid trivial cases. Let $x \in \mathbb{C}^n$ be a unit vector so that

$$\begin{aligned} \|U+V\| &= \|(U+V)x\| = \|(U+\lambda_0 I)x + (V - \lambda_0 I)x\| \\ &\leq \|(U+\lambda_0 I)x\| + \|(V-\lambda_0 I)x\| \leq \|U+\lambda_0 I\| + \|V - \lambda_0 I\| = \|U+V\|. \end{aligned}$$

Then $\|(U+\lambda_0 I)x\| = \|U+\lambda_0 I\| = |u_0+\lambda_0|$, $\|(V-\lambda_0 I)x\| = \|V - \lambda_0 I\| = |v_0 - \lambda_0|$, and $(U+\lambda_0 I)x$ is a positive multiple of $(V-\lambda_0 I)x$; thus span $\{x, Ux\} =$ span $\{x, Vx\} = \mathcal{S}$, say. Since $(U+\lambda_0 I)^*(U+\lambda_0 I)x = |u_0+\lambda_0|^2 x$, we see that $U^2 x = -x + (u_0 + \bar{u}_0)Ux$; thus $\mathcal{S}$ is an invariant subspace, and hence a reducing subspace for U. Similarly, $(V-\lambda_0 I)^*(V-\lambda_0 I)x = |v_0 - \lambda_0|^2 x$ induces that $\mathcal{S}$ is a reducing subspace for V, too. Therefore $\mathcal{S}$ is a common reducing subspace of U and V. There are two possible situations.

(a) Suppose $\mathcal{S}$ is of dimension 1. Then x is a common eigenvector for U and V corresponding to the eigenvalues $u' \in \sigma(U)$ and $v' \in \sigma(V)$ with $|u' + v'| = |u_0 + v_0|$ as desired.

(b) Suppose $\mathcal{S}$ is of dimension 2. Let $U_0, V_0 \in M_2$ be the restrictions of $(U+\lambda_0 I)/|u_0+\lambda_0|$ and $(V-\lambda_0 I)/|v_0 - \lambda_0|$ to the common reducing subspace $\mathcal{S}$. As U_0 is a normal matrix of norm 1 and $\|U_0 x\| = 1$ and x is not an eigenvector of u_0, it follows that U_0 must be a unitary matrix with two distinct eigenvalues. Since $\sigma(U_0) \subseteq \{(u+\lambda_0)/|u_0+\lambda_0| : u \in \sigma(U)\}$, and $\operatorname{Re} u \leq \operatorname{Re} u_0$ and $\lambda_0 > 0$, we deduce that $\sigma(U_0) = \{(u_0+\lambda_0)/|u_0+\lambda_0|, (\bar{u}_0+\lambda_0)/|u_0+\lambda_0|\}$; hence, the 2×2 matrix formed by U restrcted to $\mathcal{S}$ is a unitary matrix with spectrum $\{u_0, \bar{u}_0\}$ such that $\operatorname{Im} u_0 > 0$. Similarly, V restriced to $\mathcal{S}$ is a unitary matrix with spectrum $\{v_0, \bar{v}_0\}$ such that $\operatorname{Im} v_0 > 0$. Therefore, we can apply Lemma 3.10 to get the conclusion.

Note that $\|U+V\| = 2$ means that there exists a unit vector x satisfying $2 = \|U+V\| = \|(U+V)x\| \leq \|Ux\| + \|Vx\| \leq 2$, which yields $Ux = Vx$, and thus that $U - V$ is singular. □

4 The Sum of two Normal Matrices

In this section, we prove that the quantity

$$\min_{\lambda \in \mathbb{C}}\{\max_{\alpha \in \sigma(A)} |\alpha + \lambda| + \max_{\beta \in \sigma(B)} |\beta - \lambda|\}$$

is the optimal norm bound for $\|A+B\|$ when A and B are normal matrices. It turns out that the major structure theory is based on the solution of the following geometrical combinatorial problem: Given two compact subsets $\mathcal{A}$ and $\mathcal{B}$ of the complex plane $\mathbb{C}$, determine

$$\min_{\lambda\in\mathbb{C}}\{\max_{\alpha\in\mathcal{A}}|\alpha-\lambda|+\max_{\beta\in\mathcal{B}}|\beta-\lambda|\},$$

and the triples $(\alpha_0,\beta_0,\lambda_0)\in\mathcal{A}\times\mathcal{B}\times\mathbb{C}$ that attain the min-max value. In other words, we are trying to find a point in $\mathbb{C}$ that minimizes the combined maximum distances to points in the sets $\mathcal{A}$ and $\mathcal{B}$.

In order to get a direct application to our setting of the summation of two normal matrices, we replace $(\mathcal{A},\mathcal{B})$ by $(-\mathcal{A},\mathcal{B})$ and consider

$$\min_{\lambda\in\mathbb{C}}\{\max_{\alpha\in\mathcal{A}}|\alpha+\lambda|+\max_{\beta\in\mathcal{B}}|\beta-\lambda|\}.$$

Denote by $\mathbf{T}$ the unit circle in $\mathbb{C}$. Each non-empty compact set $S\subset\mathbb{C}$ determines a compact subset of elements of maximal modulus,

$$M(S)=\{\mu\in S:|\mu|\geq|\nu|\text{ for all }\nu\in S\}.$$

Let $N(S)$ be the normalization of $M(S)$, i.e., $N(S)$ is a compact subset of $\mathbf{T}$ defined as

$$N(S)=\begin{cases}\mathbf{T} & \text{if } S=\{0\},\\ \{\alpha/|\alpha|:\alpha\in M(S)\} & \text{if } S\neq\{0\}.\end{cases}$$

Proposition 4.1 *Suppose $\mathcal{A}$ and $\mathcal{B}$ are non-empty compact subsets of the complex plane $\mathbb{C}$. Then*

$$\min_{\lambda\in\mathbb{C}}\{\max_{\alpha\in\mathcal{A}}|\alpha+\lambda|+\max_{\beta\in\mathcal{B}}|\beta-\lambda|\}=\max_{\alpha\in\mathcal{A}}|\alpha|+\max_{\beta\in\mathcal{B}}|\beta|. \tag{4.1}$$

if and only if the following condition holds.

(I) *There exist $u_1,u_2\in N(\mathcal{A})$ and $v_1,v_2\in N(\mathcal{B})$ such that the line segment joining u_1 and u_2 meets the line segment joining v_1 and v_2 at some point w with $|w|\leq 1$; equivalently, there does not exist an arc Γ of the unit circle $\mathbf{T}$ such that*

$$N(\mathcal{A})\subseteq\Gamma \quad\text{and}\quad N(\mathcal{B})\subseteq\mathbf{T}\setminus\Gamma.$$

Note that if $u_1=u_2$ or $v_1=v_2$ in (I), then $|w|=1$ and

$$N(\mathcal{A})\cap N(\mathcal{B})\neq\emptyset. \tag{4.2}$$

Proof. Let

$$a=\max\{|\alpha|:\alpha\in\mathcal{A}\}\qquad\text{and}\qquad b=\max\{|\beta|:\beta\in\mathcal{B}\}.$$

We assume that $a,b>0$ to avoid trivial consideration.

Suppose (4.1) holds. If there is an arc Γ of the unit circle $\mathbf{T}$ such that $N(\mathcal{A}) \subseteq \Gamma$ and $N(\mathcal{B}) \subseteq \mathbf{T} \setminus \Gamma$. We may apply a rotation to $\mathbb{C}$ and assume that $\mathbf{T} \setminus \Gamma = \{e^{it} : t \in (-t_1, t_1)\}$ with $t_1 \in (0, \pi)$. Choose two real numbers $s \in (-1, 1)$ and $\varepsilon \in (0, 1)$ such that

$$\cos\theta_1 \le \cos t_1 < s - \varepsilon \quad \text{and } s + \varepsilon < \cos\theta_2$$

for all $u = a(\cos\theta_1 + i\sin\theta_1) \in N(\mathcal{A})$ and $v = b(\cos\theta_2 + i\sin\theta_2) \in N(\mathcal{B})$. Let

$$\mathcal{A}_0 = \{\alpha \in \mathcal{A} : \mathrm{Re}\,(\alpha)/a \ge s - \varepsilon\} \quad \text{and} \quad \mathcal{B}_0 = \{\beta \in \mathcal{B} : \mathrm{Re}\,(\beta)/b \le s + \varepsilon\}.$$

Since $\mathcal{A}_0$ and $\mathcal{B}_0$ are compact and $aN(\mathcal{A}) \cap \mathcal{A}_0 = \phi = bN(\mathcal{B}) \cap \mathcal{B}_0$, there exists $\delta > 0$ such that $|\alpha| < a(1-\delta)$ and $|\beta| < b(1-\delta)$ for all $\alpha \in \mathcal{A}_0$ and $\beta \in \mathcal{B}_0$. Now for any $\alpha = a_1(\cos\theta_1 + i\sin\theta_1) \in \mathcal{A} \setminus \mathcal{A}_0$ and $\beta = b_1(\cos\theta_2 + i\sin\theta_2) \in \mathcal{B} \setminus \mathcal{B}_0$ with $a_1 \in (0, a]$ and $b_1 \in (0, b]$, and $\cos\theta_1 < s - \varepsilon < s + \varepsilon < \cos\theta_2$, we have

$$\begin{aligned}
|\alpha + t|^2 &= a_1^2 + 2a_1 t\cos\theta_1 + t^2 \\
&\le a_1^2 + 2a_1 t(s - \varepsilon) + t^2 \\
&= (a + ts)^2 - t^2 s^2 - (a - a_1)(a + a_1 + 2t(s - \varepsilon)) - t(2a\varepsilon - t) \\
&< (a + ts)^2,
\end{aligned}$$

when t is a small positive real number. Hence we get $|\alpha + t| < a + ts$ and, similarly, $|\beta - t| < b - ts$. Thus, for a sufficiently small $t > 0$, we have $|\alpha+t| + |\beta-t| < (a+ts) + (b-ts) = a + b$ for every $\alpha \in \mathcal{A} \setminus \mathcal{A}_0$ and $\beta \in \mathcal{B} \setminus \mathcal{B}_0$. Also, for any $\alpha \in \mathcal{A}$ and $\beta \in \mathcal{B}_0$, we have $|\alpha + t| + |\beta - t| < a + t + b(1 - \delta) + t < a + b$ if $2t < b\delta$; for any $\alpha \in \mathcal{A}_0$ and $\beta \in \mathcal{B}$, we have $|\alpha + t| + |\beta - t| < a(1 - \delta) + t + b + t < a + b$ if $2t < a\delta$. Consequently, if $t > 0$ is small enough, we have $|\alpha + t| + |\beta - t| < a + b$ for any $\alpha \in \mathcal{A}$ and $\beta \in \mathcal{B}$. It will then follow that

$$\max_{\alpha\in\mathcal{A}} |\alpha + t| + \max_{\beta\in\mathcal{B}} |\beta - t| < a + b,$$

which is a contradiction.

To prove the converse, suppose condition (I) holds. Let

$$w = ru_1 + (1 - r)u_2 = sv_1 + (1 - s)v_2$$

with $r, s \in [0, 1]$, u_1, u_2 in $N(\mathcal{A})$ and v_1, v_2 in $N(\mathcal{B})$. Then

$$\begin{aligned}
\max_{\alpha\in\mathcal{A}} |\alpha + \lambda| &\ge \max_{j=1,2} |au_j + \lambda| \\
&= \max_{j=1,2} |a + \lambda\bar{u}_j| \\
&\ge r|a + \lambda\bar{u}_1| + (1 - r)|a + \lambda\bar{u}_2| \\
&\ge |r(a + \lambda\bar{u}_1) + (1 - r)(a + \lambda\bar{u}_2)| \\
&= |a + \lambda\bar{w}|.
\end{aligned}$$

Similarly,

$$\max_{\beta\in\mathcal{B}}|\beta-\lambda| \geq |b-\lambda\bar{w}|.$$

Consequently, we have

$$\max_{\alpha\in\mathcal{A}}|\alpha+\lambda| + \max_{\beta\in\mathcal{B}}|\beta-\lambda| \geq |a+\lambda\bar{w}| + |b-\lambda\bar{w}| \geq a+b.$$

Thus (4.1) holds. □

Corollary 4.2 *Suppose $\mathcal{A}$ and $\mathcal{B}$ are non-empty compact subsets of the complex plane $\mathbb{C}$. Then*

$$\max\{|\alpha+\beta| : \alpha\in\mathcal{A},\ \beta\in\mathcal{B}\} \leq \min_{\lambda\in\mathbb{C}}\{\max_{\alpha\in\mathcal{A}}|\alpha+\lambda| + \max_{\beta\in\mathcal{B}}|\beta-\lambda|\}. \tag{4.3}$$

The inequality becomes an equality if and only if there exists $(\alpha_0,\beta_0,\lambda_0)\in\mathcal{A}\times\mathcal{B}\times\mathbb{C}$ such that $\alpha_0+\lambda_0$ and $\beta_0-\lambda_0$ are two complex numbers of the same argument, and

$$|\alpha_0+\lambda_0| = \max_{\alpha\in\mathcal{A}}|\alpha+\lambda_0| \qquad \textit{and} \qquad |\beta_0-\lambda_0| = \max_{\beta\in\mathcal{B}}|\beta-\lambda_0|. \tag{4.4}$$

Proof. Evidently, for any $\alpha,\beta,\lambda\in\mathbb{C}$,

$$|\alpha+\beta| \leq |\alpha+\lambda| + |\beta-\lambda|.$$

Thus (4.3) holds. Moreover, it is easy to check that inequality (4.3) becomes an equality with

$$|\alpha_0+\beta_0| = \max_{\alpha\in\mathcal{A}}|\alpha+\lambda_0| + \max_{\beta\in\mathcal{B}}|\beta-\lambda_0| = |\alpha_0+\lambda_0| + |\beta_0-\lambda_0|$$

if and only if the asserted condition holds for $(\alpha_0,\beta_0,\lambda_0)\in\mathcal{A}\times\mathcal{B}\times\mathbb{C}$. □

Theorem 4.3 *Let A and B be $n\times n$ normal matrices with spectra $\mathcal{A}$ and $\mathcal{B}$. Then*

$$\max_{W^*W=I_n}\|A+W^*BW\| = \min_{\lambda\in\mathbb{C}}\{\|A+\lambda I\| + \|B-\lambda I\|\} = \min_{\lambda\in\mathbb{C}}\{\max_{\alpha\in\mathcal{A}}|\alpha+\lambda| + \max_{\beta\in\mathcal{B}}|\beta-\lambda|\}.$$

Proof. It is clear that for each unitary W and each complex number λ, we have

$$\|A+W^*BW\| = \|A+\lambda I + W^*(B-\lambda I)W\| \leq \|A+\lambda I\| + \|B-\lambda I\|\}.$$

So,

$$\max\{\|A+W^*BW\| : W \text{ is unitary }\} \leq \min_{\lambda\in\mathbb{C}}\{\|A+\lambda I\| + \|B-\lambda I\|\}.$$

To prove the theorem, we need only to show that there exists a unitary matrix W satisfying

$$\|A+W^*BW\| = \min_{\lambda\in\mathbb{C}}\{\max_{\alpha\in\mathcal{A}}|\alpha+\lambda| + \max_{\beta\in\mathcal{B}}|\beta-\lambda|\}.$$

We may assume that the expression on the right side attains its minimum at $\lambda_0 = 0$; otherwise, replace (A, B) by $(A + \lambda_0 I, B - \lambda_0 I)$. Applying Proposition 4.1 to the spectrum of A and that of B, we have two possibilities.

If condition (4.2) holds, let W_1 and W_2 be unitary so that $W_1^* A W_1$ has α_0 as the $(1,1)$ entry, and $W_2^* B W_2$ has β_0 as the $(1,1)$ entry, where $\alpha_0 \in \sigma(A)$, $\beta_0 \in \sigma(B)$ and

$$|\alpha_0 + \beta_0| = |\alpha_0| + |\beta_0| = \|A\| + \|B\|.$$

Then for $W = W_2 W_1^*$, we have

$$\begin{aligned} & \|A + W^* B W\| = \|W_1^* A W_1 + W_2^* B W_2\| \\ \geq \quad & |\alpha_0 + \beta_0| = \min_{\lambda \in \mathbb{C}} \{\max_{\alpha \in \mathcal{A}} |\alpha + \lambda| + \max_{\beta \in \mathcal{B}} |\beta - \lambda|\}. \end{aligned}$$

Suppose condition (I) in Proposition 4.1 holds. Let $u_k = \alpha_k / |\alpha_k|$ and $v_k = \beta_k / |\beta_k|$ for $k = 1, 2$, where $\alpha_1, \alpha_2 \in \sigma(A)$ with $|\alpha_1| = |\alpha_2| = \|A\|$, and $\beta_1, \beta_2 \in \sigma(B)$ with $|\beta_1| = |\beta_2| = \|B\|$. Suppose the line segment joining u_1 and u_2 meets the line segment joining v_1 and v_2 at w. Let W_1 and W_2 be unitary so that $W_1^* A W_1 = \|A\| U \oplus A_0$ and $W_2^* B W_2 = \|B\| V \oplus B_0$, where U and V satisfy the conclusion of Lemma 3.1. Then for $W = W_2 W_1^*$, we have

$$\begin{aligned} & \|A + W^* B W\| = \|(W_1^* A W_1 + W_2^* B W_2) e_1\| \\ = \quad & \|A\| + \|B\| = \min_{\lambda \in \mathbb{C}} \{\max_{\alpha \in \mathcal{A}} |\alpha + \lambda| + \max_{\beta \in \mathcal{B}} |\beta - \lambda|\}. \end{aligned}$$

□

We note that the equality

$$\max_{W^* W = I_n} \|A + W^* B W\| = \min_{\lambda \in \mathbb{C}} \{\|A + \lambda I\| + \|B - \lambda I\|\} \tag{4.5}$$

is actually valid for general matrices without the normality assumption as explored in details in [6].

5 Extension to infinite dimensional cases

Let $B(H)$ be the algebra of bounded linear operators acting on an infinite dimensional Hilbert space H. Although the results in the previous sections are stated and proved for matrices only, we can extend them to bounded linear operators in $B(H)$. However, since the left hand side of (4.5) may not be attainable as shown in the next example, we need to make some adjustments in the statements of the results.

Example 5.1 Consider $A = \operatorname{diag}(0, 1/2, 2/3, 3/4, \ldots)$ and $B = \operatorname{diag}(1, 0, 0, \ldots)$ acting on $H = \ell_2$. Then

$$\begin{aligned} 2 &= \|A\| + \|B\| \\ &= \min\{\|A + \lambda I\| + \|B - \lambda I\| : \lambda \in \mathbb{C}\} \\ &= \sup\{\|A + W^* B W\| : W \text{ is unitary }\}, \end{aligned}$$

and the supremum is not attainable.

We proceed to show that $\|T\| < 2$ if $T = A + W^*BW$ for some unitary $W \in B(H)$. Note that A has norm 1 but it is strictly contractive in the sense $\|Ax\| < 1$ for all unit vectors $x \in H$; thus $\|AC\| < 1$ for all rank-1 norm-1 operator C. Now suppose $T = A + W^*BW$ is of norm 2, then T^2 is of norm 4. As the expansion of T^2 is a sum of four operators where each is of norm ≤ 1, it follows that each of these four operators is of norm 1. But one of these four operators is $S = AW^*BW$, so $1 = \|S\| = \|AW^*BW\| = \|AC\| < 1$ where $C = W^*BW$ is of rank-1 and norm 1. This leads to a contradiction.

To extend our results on M_n to $B(H)$, we need the following lemma.

Lemma 5.2 *Suppose $A \in B(H)$ is normal and has spectrum $\sigma(A)$. If $\alpha_1, \alpha_2 \in \sigma(A)$, then for any $\varepsilon > 0$, there is a normal operator $\tilde{A}$ such that α_1, α_2 are eigenvalues of $\tilde{A}$, $\sigma(\tilde{A}) = \sigma(A)$, and*

$$\|A - \tilde{A}\| < \varepsilon.$$

Proof. If $\alpha \in \sigma(A)$ is not an eigenvalue for a normal operator A, then by the spectral theorem, A can be written as $A_1 \oplus A_0$ where A_0 is acting on an infinite-dimensional Hilbert space and $\|A_0 - \alpha I\| < \varepsilon/2$. Rewrite A_0 as $\oplus_{n=1}^{\infty} C_n$, where C_n's are acting on a common identical Hilbert space. Let $\tilde{A} = A_1 \oplus \oplus_{n=1}^{\infty} \tilde{C}_n$, with $\tilde{C}_1 = \alpha I$, and $\tilde{C}_n = C_{n-1}$ for $n > 0$. Then α is an eigenvalue for $\tilde{A}$ while $\sigma(A) = \sigma(\tilde{A})$ and $\|A - \tilde{A}\| < \varepsilon$. The argument above can be extended to get $\tilde{A}$ with two prescribed eigenvalues α_1 and $\alpha_2 \in \sigma(A)$. □

We illustrate how to use Lemma 5.2 to prove the infinite dimensional version of Theorem 4.3 in the following.

Theorem 5.3 *Let $A, B \in B(H)$ be normal operators with spectra $\mathcal{A}$ and $\mathcal{B}$. Then*

$$\sup_{W^*W=I} \|A + W^*BW\| = \min_{\lambda \in \mathbb{C}} \{\|A + \lambda I\| + \|B - \lambda I\|\} = \min_{\lambda \in \mathbb{C}} \{\max_{\alpha \in \mathcal{A}} |\alpha + \lambda| + \max_{\beta \in \mathcal{B}} |\beta - \lambda|\}.$$

Proof. It is clear that

$$\sup\{\|A + W^*BW\| : W \text{ is unitary }\} \leq \min_{\lambda \in \mathbb{C}} \{\|A + \lambda I\| + \|B - \lambda I\|\}.$$

To prove the theorem, we need only to show that for any $\varepsilon > 0$ there is a unitary operator $W \in B(H)$ such that

$$\|A + W^*BW\| + \varepsilon \geq \min_{\lambda \in \mathbb{C}} \{\max_{\alpha \in \mathcal{A}} |\alpha + \lambda| + \max_{\beta \in \mathcal{B}} |\beta - \lambda|\}.$$

We may assume that the expression on the right side attains its minimum at $\lambda_0 = 0$; otherwise, replace (A, B) by $(A + \lambda_0 I, B - \lambda_0 I)$. Applying Proposition 4.1 to $\sigma(A)$ and $\sigma(B)$, we see that condition (I) in Proposition 4.1 holds. If (4.2) holds and α_0 (respectively, β_0) is an eigenvalue of A (respectively, B), then we can use the same arguments in the proof of Theorem 4.3 to conclude that there exists a unitary W such that

$$\|A + W^*BW\| = \min_{\lambda \in \mathbb{C}} \{\max_{\alpha \in \mathcal{A}} |\alpha + \lambda| + \max_{\beta \in \mathcal{B}} |\beta - \lambda|\}.$$

Similarly, we can prove this equality if (I) in Proposition 4.1 holds with $N(\mathcal{A}) \cap N(\mathcal{B}) = \emptyset$ and α_1, α_2 (respectively, β_1, β_2) are eigenvalues of A (respectively, B).

Suppose A and B do not have the desired eigenvalues. By Lemma 5.2, for any $\varepsilon > 0$, there exist normal operators $\tilde{A}$ and $\tilde{B}$ in $B(H)$ such that $\tilde{A}$ and $\tilde{B}$ have the desired eigenvalues, $\sigma(\tilde{A}) = \sigma(A)$, $\|A - \tilde{A}\| < \varepsilon/2$. $\sigma(\tilde{B}) = \sigma(B)$, and $\|B - \tilde{B}\| < \varepsilon/2$. Then there exists a unitary W such that

$$\|A + W^*BW\| + \varepsilon \geq \|\tilde{A} + W^*\tilde{B}W\| = \min_{\lambda \in \mathbb{C}} \{\max_{\alpha \in \mathcal{A}} |\alpha + \lambda| + \max_{\beta \in \mathcal{B}} |\beta - \lambda|\}.$$

□

We can apply similar arguments to extend other results to $B(H)$. Very often, we have to replace "maximum" by "supremum" in the statements of results as done in Theorem 5.3. For example, Theorem 2.1 can be extended to the following.

Theorem 5.4 *Suppose $A, B \in B(H)$ are Hermitian operators satisfying $a_1 I \leq A \leq a_2 I$ and $b_1 I \leq B \leq b_2 I$. Assume further that $a_2 \geq |a_1|$ and $b_2 \geq |b_1|$.*

(i) *If $a_1 b_2 + a_2 b_1 \geq 0$, then*

$$\|A + iB\| \leq |a_2 + ib_2| = \sqrt{a_2^2 + b_2^2}.$$

(ii) *If $a_1 b_2 + a_2 b_1 \leq 0$, then*

$$\|A + iB\| \leq \tau + \tau',$$

where

$$\tau = |a_1 - z_0| = |a_2 - z_0| = \frac{1}{2}\sqrt{(a_1 - a_2)^2 + (b_1 + b_2)^2}$$

and

$$\tau' = |ib_1 - z_0| = |ib_2 - z_0| = \frac{1}{2}\sqrt{(a_1 + a_2)^2 + (b_1 - b_2)^2}$$

with $z_0 = \{(a_1 + a_2) + i(b_1 + b_2)\}/2$.

(iii) *The bounds in* (i) *and* (ii) *are sharp in the following sense: If $\{a_1, a_2\} \subseteq \sigma(A)$ and $\{b_1, b_2\} \subseteq \sigma(B)$, then* $\sup\{\|A + iWBW^*\| : W \in B(H)$ *is unitary* $\}$ *attains the bound in each case.*

The extensions of other results to the infinite dimensional case can be done in a similar fashion. We omit their discussion.

References

[1] R. Bhatia, *Matrix Analysis,* Springer, New York, 1997.

[2] R. Bhatia, L. Elsner, P. Šemrl, Distance between commuting tuples of normal operators, Archiv Math. 71 (1998), 229-232.

[3] R. Bhatia and F. Kittaneh, Cartesian decompositions and Schatten norms. Linear Algebra Appl. 318 (2000), 109–116.

[4] R. Bhatia and X. Zhan, Compact operators whose real and imaginary parts are positive. Proc. Amer. Math. Soc. 129 (2001), 2277–2281

[5] R. Bhatia and X. Zhan, Norm inequalities for operators with positive real part, J. Operator Theory, to appear.

[6] M.D. Choi and C.K. Li, The ultimate estimate of the upper norm bound for the summation of matrices, preprint.

[7] B.A. Mirman, Numerical range and norm of a linear operator, *Voronež. Gos. Univ. Trudy Sem. Funkcional Anal.,* 10 (1968), 51-55.

[8] Y. Nakamura, Numerical range and norm, *Math. Japonica* 27 (1982), 149-150.