Filomat 37:27 (2023), 9183–9195
https://doi.org/10.2298/FIL2327183C

Published by Faculty of Sciences and Mathematics,
University of Niš, Serbia
Available at: http://www.pmf.ni.ac.rs/filomat

Sharp bounds for the operator norm of commutator

Che-Man Cheng[a,*], Ka Leong Hoi[a]

[a] *Department of Mathematics, University of Macau, Macao, China*

Abstract. Let $\|\cdot\|_p$ denote the Schatten p-norm, $1 \le p \le \infty$. The smallest constant $C^{\mathbb{F}}_{\infty,q,r}$ such that

$$\|XY - YX\|_\infty \le C^{\mathbb{F}}_{\infty,q,r}\|X\|_q\|Y\|_r$$

for all real ($\mathbb{F} = \mathbb{R}$) or complex ($\mathbb{F} = \mathbb{C}$) matrices X and Y is determined.

1. Introduction

A general problem is the determination of the best (i.e., smallest) constant $C^{\mathbb{F}}_{p,q,r}$ such that

$$\|XY - YX\|_p \le C^{\mathbb{F}}_{p,q,r}\|X\|_q\|Y\|_r, \quad X, Y \in M_n(\mathbb{F}), \tag{1}$$

where $M_n(\mathbb{F})$ denotes the set of $n \times n$ matrices with entries in $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, and $\|\cdot\|_p$ denotes the Schatten p-norm, $1 \le p \le \infty$. In [2], Böttcher and Wenzel first raised the problem for real matrices and Frobenius norm (i.e., for $\mathbb{F} = \mathbb{R}$ and $p = q = r = 2$). It was conjectured that $C^{\mathbb{R}}_{2,2,2} = \sqrt{2}$. The conjecture was proved by László [11] for $n = 3$, and in general by Vong and Jin [16] and by Lu [12]. The result was extended to complex matrices in [1] and [3], giving $C^{\mathbb{C}}_{2,2,2} = \sqrt{2}$. For other interesting results about commutator norm inequalities, see the surveys [9] and [13].

The generalization (1) was raised in [18]. Simple formulas for $C^{\mathbb{F}}_{p,q,r}$ are obtained in [6] and [18]. In particular, for the operator norm of commutators, i.e., $p = \infty$, we have (see [18]) $C^{\mathbb{F}}_{\infty,1,1} = \frac{\sqrt{27}}{4}$ and $C^{\mathbb{F}}_{\infty,q,r} = 2^{1-\frac{1}{\max\{q,r\}}}$ if $q \ge 2$ or $r \ge 2$. Difficulty arises when coming to the remaining situation. In [8], it is shown that $C^{\mathbb{R}}_{\infty,q,1}$, $1 < q < 2$, can be found by solving a polynomial-like equation. This hinted that simple expressions for $C^{\mathbb{R}}_{\infty,q,r}$ are not likely.

For $1 \le p \le \infty$, let $\tilde{p}$ satisfy $\frac{1}{p} + \frac{1}{\tilde{p}} = 1$. As usual, take $\frac{1}{\infty} = 0$. It is shown in [18] that the commutator bounds $C^{\mathbb{F}}_{p,q,r}$ satisfy

$$C^{\mathbb{F}}_{p,q,r} = C^{\mathbb{F}}_{p,r,q} \quad \text{and} \quad C^{\mathbb{F}}_{p,q,r} = C^{\mathbb{F}}_{\tilde{q},\tilde{p},r}. \tag{2}$$

2020 *Mathematics Subject Classification.* 15A60
Keywords. Commutator, Schatten norm, operator norm, singular values.
Received: 21 January 2023; Revised: 30 April 2023; Accepted: 09 May 2023
Communicated by Mohammad Sal Moslehian
This research is supported by the research project 0013/2021/A from the Science and Technology Development Fund, Macao S.A.R. (FDCT).

* Corresponding author: Che-Man Cheng
Email addresses: fstcmc@um.edu.mo (Che-Man Cheng), mb95454@connect.um.edu.mo (Ka Leong Hoi)

Thus, $C^{\mathbb{F}}_{\infty,q,r} = C^{\mathbb{F}}_{\tilde{q},1,r} = C^{\mathbb{F}}_{\tilde{q},r,1}$. Recently, the determination of $C^{\mathbb{F}}_{p,1,1}$, $2 < p < \infty$, is solved in [4] with a new form of solution: the constants $C^{\mathbb{F}}_{p,1,1}$ are expressed in terms of a parametrization of p. To determine $C^{\mathbb{F}}_{\infty,q,r}$, we are going to determine $C^{\mathbb{F}}_{p,1,r}$ for all the unsolved situations of p and r, i.e., $2 < p < \infty$ and $1 < r < 2$. This problem is solved in [4] for 2×2 real matrices. Here, in Theorem 2.7, we solve the problem for general $n \times n$ matrices.

Let us introduce some notations. Let $\mathbf{e}_j \in \mathbb{F}^n$ denote the column vector with 1 at the j-th component and 0 otherwise. The identity and zero matrices (of appropriate orders) are denoted by I and 0, respectively. For $X, Y \in M_n(\mathbb{F})$, the commutator $XY - YX$ is denoted by $[X, Y]$. Let $E_{ij} \in M_n(\mathbb{F})$ denote the matrix with 1 at the (i, j) entry and 0 otherwise. Let $s_1(X) \geq \cdots \geq s_n(X)$ denote the singular values of X arranged in non-increasing order, and $\mathbf{s}(X) = (s_1(X), \ldots, s_n(X))^T$. Let $\mathbf{s}_{1,2}(X) = (s_1(X), s_2(X))^T$. Let $O(n)$ and $U(n)$ denote the sets of $n \times n$ orthogonal and unitary matrices, respectively. The set of the extreme points of a convex set X is denoted by $\operatorname{ext}(X)$. For simplicity, we use $\|\cdot\|$ to denote the Euclidean norm on $\mathbb{F}^n$ and the Frobenius norm on $M_n(\mathbb{F})$. We also use the following notations:

$$\begin{aligned}
\Sigma_{1,r}(n,\mathbb{F}) &= \{(X,Y) : X, Y \in M_n(\mathbb{F}), \|X\|_1 = 1, \|Y\|_r = 1\},\\
\Sigma^0_{1,r}(n,\mathbb{F}) &= \{(X,Y) : X, Y \in M_n(\mathbb{F}), \|X\|_1 = 1, \|Y\|_r = 1, \operatorname{tr} X = \operatorname{tr} Y = 0\},\\
\Sigma_{(1),r}(n,\mathbb{F}) &= \{(X,Y) : X, Y \in M_n(\mathbb{F}), \mathbf{s}(X) = (1,0,...,0)^T, \|Y\|_r = 1\},\\
\Sigma^0_{(1),r}(n,\mathbb{F}) &= \{(X,Y) : X, Y \in M_n(\mathbb{F}), \mathbf{s}(X) = (1,0,...,0)^T, \|Y\|_r = 1, \operatorname{tr} X = \operatorname{tr} Y = 0\},\\
S_{(1),r}(n,\mathbb{F}) &= \{(s_1([X,Y]), s_2([X,Y])) : (X,Y) \in \Sigma_{(1),r}(n,\mathbb{F})\},\\
S^0_{(1),r}(n,\mathbb{F}) &= \{(s_1([X,Y]), s_2([X,Y])) : (X,Y) \in \Sigma^0_{(1),r}(n,\mathbb{F})\}.
\end{aligned}$$

2. Results

2.1. Some background and lemmas

A norm is a convex function and a closed unit ball is a convex set. For the determination of $C^{\mathbb{F}}_{p,q,r}$, one may focus on the extreme points of the unit balls. When $q = 1$, $\operatorname{ext}\{X : X \in M_n(\mathbb{F}), \|X\|_1 \leq 1\}$ is the set of normalized rank one matrices. Note that when $\operatorname{rank} X = 1$, $\operatorname{rank}[X, Y] \leq 2$. It is known (see [4, Introduction]) that (with $\|\mathbf{x}\|_p$ is the l_p norm of $\mathbf{x}$),

$$C^{\mathbb{F}}_{p,1,r} = \max\{\|[X,Y]\|_p : (X,Y) \in \Sigma_{(1),r}(n,\mathbb{F})\} = \max\{\|\mathbf{x}\|_p : \mathbf{x}^T \in S_{(1),r}(n,\mathbb{F})\}. \tag{3}$$

A simple fact about the set $S_{(1),r}(n, \mathbb{F})$ is given in the following lemma.

Lemma 2.1. *Suppose $n \geq 2$, $1 \leq r \leq \infty$ and $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$. The set $S_{(1),r}(n, \mathbb{F})$ is star-shaped with the origin as a star center.*

Proof. The proof is similar to that given in [4, Lemma 3.1]. □

The set $S_{(1),r}(2, \mathbb{R})$ is characterized in [4]. Let

- C_1 be the segment joining the points $(1, 0)$ and $\left(\frac{\sqrt{2}+1}{2}, \frac{\sqrt{2}-1}{2}\right)$ together with the curve

$$C_{(1)} : \frac{4\sqrt{\sin\phi\cos\phi}}{(\sin\phi + \cos\phi)^2}(\cos\phi, \sin\phi), \quad \tan^{-1}\left(\frac{\sqrt{2}-1}{\sqrt{2}+1}\right) \leq \phi \leq \frac{\pi}{4},$$

- C_r, $1 < r < \infty$, be the curve

$$C_r : \Big(x(t), y(t)\Big) = \Big(f(t) + g(t), f(t) - g(t)\Big), \quad 0 \leq t \leq 1, \tag{4}$$

where

$$f(t) = \frac{\sqrt{(t^{r+1}+t^{r-1})^2+(1-t^{2r})^2}}{(t^{2r-2}+1)(t^{2r}+1)^{\frac{1}{r}}} \quad \text{and} \quad g(t) = \frac{(t^{r+1}+t^{r-1})}{(t^{2r-2}+1)(t^{2r}+1)^{\frac{1}{r}}}.$$

(Note that the above formulas are obtained with $t = s^{\frac{1}{r-1}}$ and so cannot be used for $r = 1$.)

- C_∞ be the segment joining the points $(1,1)$ and $(2,0)$.

For $1 \le r \le \infty$, the curve C_r, the x-axis and the line $x = y$ bounded a region. We denote this region by $\mathcal{R}_r$. For illustration, Figure 2.1, which can be found in [4], shows the curve C_r for $r = 1, 1.3, 1.7, 2, 2.5, 5, \infty$. The broken line shows the segment joining the origin and the point $(1,1)$.

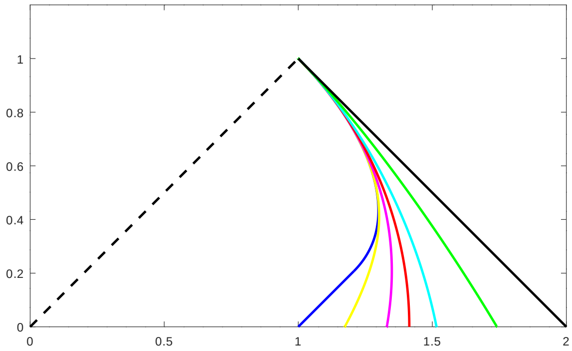

Figure 2.1. The curve C_r for $r = 1$ (blue), 1.3 (yellow), 1.7 (magenta), 2 (red), 2.5 (cyan), 5 (green), ∞ (black).

Theorem 2.2. *[4, Theorem 3.3] Suppose* $1 \le r \le \infty$. *Then* $S_{(1),r}(2,\mathbb{R}) = S^0_{(1),r}(2,\mathbb{R}) = \mathcal{R}_r$.

Lemma 2.3. *Suppose* $1 \le r \le \infty$. *The curve* C_r *is a convex curve and the slope* m *of any of its non-vertical tangent line satisfies*

$$m \in \begin{cases} (-\infty,-1] \cup [\frac{r}{2-r},\infty) & \text{if } 1 \le r < 2, \\ (-\infty,-1] & \text{if } r = 2, \\ [\frac{r}{2-r},-1] & \text{if } 2 < r < \infty, \\ \{-1\} & \text{if } r = \infty. \end{cases} \tag{5}$$

Proof. We first consider $1 < r < \infty$. When t goes from 0 to 1, the curve C_r goes from $(1,1)$ to $(2^{1-\frac{1}{r}},0)$ in the first quadrant. When $r = 2$, the curve C_2 is the arc (in the first quadrant) of the circle with radius $\sqrt{2}$ and centered at the origin, joining $(1,1)$ and $(\sqrt{2},0)$. When $r \in (1,\infty)\setminus\{2\}$, referring to the second paragraph of [4, Section 4], it is noted that at the points $(1,1)$ and $(2^{1-\frac{1}{r}},0)$, the slopes of the tangent lines of C_r are -1 and $\frac{r}{2-r}$, respectively. In [4, Lemma 3.5], by showing that $\frac{dy}{dx}\cdot\frac{d^2y}{dx^2} \ge 0$ (except possibly at the point where the curve has a vertical tangent), it is proved that C_r is a convex curve. The result follows readily.

When $r = 1$, the fact that $C_{(1)}$ is a convex curve is proved in the proof of [4, Theorem 2.1]. At the points $(1,1)$ and $(\frac{\sqrt{2}+1}{2}, \frac{\sqrt{2}-1}{2})$, the slopes of the tangent lines are -1 and 1, respectively. The result follows readily. Finally, the result is trivial when $r = \infty$. □

For $\mathbf{x} \in \mathbb{R}^n$, let $x_{[i]}$ denote the i-th largest component of $\mathbf{x}$, i.e., $x_{[1]} \ge \cdots \ge x_{[n]}$. For $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$, we say that $\mathbf{x}$ is weakly majorized by $\mathbf{y}$, and write $\mathbf{x} \prec_w \mathbf{y}$, if $\sum_{i=1}^k x_{[i]} \le \sum_{i=1}^k y_{[i]}, k = 1,2,\ldots,n$. See [14] for more information about majorization.

Lemma 2.4. *Let* $n \ge 1$, $\mathbb{F} \in \{\mathbb{R},\mathbb{C}\}$ *and* $X, Y \in M_n(\mathbb{F})$.

(a) *Suppose* $\mathbf{d} = (|x_{11}|, \dots, |x_{nn}|)^T$. *Then* $\mathbf{d} \prec_w \mathbf{s}(X)$.

(b) *For any* $1 \le r \le \infty$, *if* $\mathbf{s}(X) \prec_w \mathbf{s}(Y)$ *then* $\|X\|_r \le \|Y\|_r$.

Proof. For (a), see [14, 9.D.1]. For (b), see [14, 10.A.2]. □

Lemma 2.5. *Let* $1 \le r \le \infty$, $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ *and* $X \in M_2(\mathbb{F})$. *Then* $\|X - \frac{\operatorname{tr} X}{2} I\|_r \le \|X\|_r$.

Proof. Suppose $X \in M_2(\mathbb{C})$. By [10, Lemma 1.3.1], X is unitarily similar to a matrix with equal diagonal entries. In fact, when X is real, it is orthogonally similar to a matrix with equal diagonal entries. The argument is as follows. Let $X = S + K$ where $S = \frac{1}{2}(X + X^T)$ is symmetric and $K = \frac{1}{2}(X - X^T)$ is skew-symmetric. Let $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^2$ be orthonormal eigenvectors of S corresponding to its real eigenvalues λ_1 and λ_2. Let $\mathbf{v}_1 = \frac{1}{\sqrt{2}}\mathbf{u}_1 + \frac{1}{\sqrt{2}}\mathbf{u}_2$, $\mathbf{v}_2 = \frac{1}{\sqrt{2}}\mathbf{u}_1 - \frac{1}{\sqrt{2}}\mathbf{u}_2$. Then, $V = [\mathbf{v}_1\, \mathbf{v}_2]$ is orthogonal. By direct verification, both diagonal entries of $V^T S V$ are $\frac{1}{2}(\lambda_1 + \lambda_2)$. As K is skew-symmetric, $V^T K V$ is also skew-symmetric and hence its diagonal entries are zero. Hence $V^T X V$ has equal diagonal entries.

Write $\hat{X} = X - \frac{\operatorname{tr} X}{2} I$. We now show that $\mathbf{s}(\hat{X}) \prec_w \mathbf{s}(X)$. Under unitary (orthogonal if $\mathbb{F} = \mathbb{R}$) similarity, we may assume $X = \begin{bmatrix} a & b \\ c & a \end{bmatrix}$ and so $\hat{X} = \begin{bmatrix} 0 & b \\ c & 0 \end{bmatrix}$. Then,

$$s_1(X) \ge \max\left\{ \sqrt{|a|^2 + |b|^2}, \sqrt{|a|^2 + |c|^2} \right\} \ge \max\{|b|, |c|\} = s_1(\hat{X}).$$

Furthermore, we have

$$\begin{aligned} (s_1(X) + s_2(X))^2 &= \|X\|^2 + 2|\det X| \\ &= 2|a|^2 + |b|^2 + |c|^2 + 2|a^2 - bc| \\ &\ge 2|a|^2 + |b|^2 + |c|^2 + 2|bc| - 2|a|^2 \\ &= (|b| + |c|)^2 \\ &= (s_1(\hat{X}) + s_2(\hat{X}))^2. \end{aligned}$$

The required weak majorization result follows. The result then follows from Lemma 2.4(b). □

Lemma 2.6. *Suppose* $X \in M_2(\mathbb{C})$. *Then* $\mathbf{s}(\mathrm{Re}(X)) \prec_w \mathbf{s}(X)$.

Proof. Suppose $U, V \in O(2)$ such that $U\mathrm{Re}(X)V = \operatorname{diag}\big(s_1(\mathrm{Re}(X)), s_2(\mathrm{Re}(X))\big)$ by the singular value decomposition. Then, using Lemma 2.4(a),

$$\mathbf{s}(\mathrm{Re}(X)) = \mathbf{s}(U\mathrm{Re}(X)V) \le (|(UXV)_{11}|, |(UXV)_{22}|)^T \prec_w \mathbf{s}(UXV) = \mathbf{s}(X).$$

Here, '≤' denotes the component-wise comparison. The result follows readily. □

2.2. The main theorem

To state our main theorem, we first note that when $1 \le r \le 2$, referring to (5) and Figure 2.1, there is a unique point $(x(t_r), y(t_r))$ on the curve C_r at which the tangent line is vertical.

Theorem 2.7.

(a) *Let* x *and* y *be defined as in (4), and let (with* x' *denoting the derivative of* x*)*

$$p(t) = 1 + \frac{\ln\left(-\frac{y'(t)}{x'(t)}\right)}{\ln \frac{x(t)}{y(t)}}, \quad 0 < t < t_r.$$

Suppose $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, $n \ge 2$ *and* $1 < r < 2$. *Then,*

$$C^{\mathbb{F}}_{\infty, \frac{p(t)}{p(t)-1}, r} = C^{\mathbb{F}}_{p(t),1,r} = \left[(x(t))^{p(t)} + (y(t))^{p(t)} \right]^{1/p(t)} = y(t) \left(1 - \frac{x(t) y'(t)}{y(t) x'(t)} \right)^{1/p(t)}, \quad 0 < t < t_r.$$

(b) The function p (resp. $\frac{p}{p-1}$) is strictly increasing (resp. decreasing) and its range is $(2,\infty)$ (resp. $(1,2)$).

Utilizing the set $S_{(1),r}(2,\mathbb{R})$, in particular C_r, Theorem 2.7 is proved when $n=2$ and $\mathbb{F}=\mathbb{R}$ in [4, Theorem 4.1]. Though $C^{\mathbb{R}}_{p(t),1,r}$ in Theorem 2.7(a) looks complicated, its derivation from $S_{(1),r}(2,\mathbb{R})$ is not difficulty. Let Γ_k denote the curve $x^p+y^p=k^p$, $k\geq 0$. Then, it is obvious that

$$C^{\mathbb{R}}_{p,1,r}=\max\{\|\mathbf{x}\|_p:\mathbf{x}^T\in S_{(1),r}(2,\mathbb{R})\}=\max\{k:\Gamma_k\cap S_{(1),r}(2,\mathbb{R})\text{ is non-empty}\}.$$

Let $\mathbf{x}_0^T\in S_{(1),r}(2,\mathbb{R})$ such that $\max\{\|\mathbf{x}\|_p:\mathbf{x}^T\in S_{(1),r}(2,\mathbb{R})\}$ is achieved. Obviously, $\mathbf{x}_0^T\in C_r$. The two endpoints of C_r are $(1,1)$ and $(2^{1-\frac{1}{r}},0)$. From [18, Theorem 3]), we have $\sqrt{2}=C^{\mathbb{R}}_{2,1,1}$. As $p>2$, $\|(1,1)\|_p=2^{\frac{1}{p}}<\sqrt{2}=C^{\mathbb{R}}_{2,1,1}\leq C^{\mathbb{R}}_{p,1,r}$ and so $\mathbf{x}_0^T\neq(1,1)$. Similarily, as $1<r<2$, $\mathbf{x}_0^T\neq(2^{1-\frac{1}{r}},0)$. Any increase in k from $C^{\mathbb{R}}_{p,1,r}$ will make $\Gamma_k\cap S_{(1),r}(2,\mathbb{R})$ empty. Thus, we see that the curves $\Gamma_{C^{\mathbb{R}}_{p,1,r}}$ and C_r touch each other at $\mathbf{x}_0^T=(x(t),y(t))$ with $0<t<1$, and so have a common tangent line. The slope of the tangent line of C_r at $\mathbf{x}_0^T$ is $y'(t)/x'(t)$, while that of Γ_k is $-(x(t)/y(t))^{p-1}$. Equating the two slopes, we get $y'(t)/x'(t)=-(x(t)/y(t))^{p-1}$, from which we get the function p and consequently $C^{\mathbb{R}}_{p(t),1,r}$.

Our main objective is to show that Theorem 2.7 is true for general $n\times n$ complex matrices. Of course, if we can show that $S_{(1),r}(n,\mathbb{C})=S_{(1),r}(2,\mathbb{R})$, the result follows immediately (see (3)). However, this result is stronger than what we need, and we tried in vain to prove it. Nevertheless, we see that if we can show that the maximum

$$\max\{\|[X,Y]\|_p:(X,Y)\in\Sigma_{(1),r}(n,\mathbb{C})\}=C^{\mathbb{C}}_{p,1,r}$$

is attained with some $(X,Y)\in\Sigma_{(1),r}(n,\mathbb{C})$ such that $\mathbf{s}([X,Y])^T\in C_r$, then $C^{\mathbb{C}}_{p,1,r}$ is the same as the one given in [4, Theorem 4.1] for 2×2 real matrices. This is our approach to prove the result.

Let C_r^+ (resp. C_r^-) denote the part of C_r joining the points $\big(x(t_r),y(t_r)\big)$ and $\big(2^{1-\frac{1}{r}},0\big)$ (resp. $(1,1)$). This is the part of C_r where, except for the point $\big(x(t_r),y(t_r)\big)$ at which the tangent line is vertical, the tangent lines have positive (resp. negative) slopes. Let Q_r be the region bounded by the x-axis, the curve C_r^+ and the vertical tangent line of C_r. It has nonempty interior if and only if $1\leq r<2$ (when $r=2$, $Q_2=\{(\sqrt{2},0)\}$). For $1\leq r<2$, define the set $\mathcal{T}_r=\mathcal{R}_r\cup Q_r$. For an illustration, see Figure 2.2 for $\mathcal{T}_{1.3}$. Our main task is to show that $S_{(1),r}(n,\mathbb{C})\subset\mathcal{T}_r$ (see Theorem 2.12). With this result, we can readily prove Theorem 2.7, as follows.

Proof of Theorem 2.7. Let $2<p<\infty$. It is clear that

$$\max\{k:x^p+y^p=k^p,(x,y)\in\mathcal{T}_r\}$$

is attained at a point lying on C_r^-. Hence, as $C_r^-\subset S_{(1),r}(2,\mathbb{R})$, we get

$$\begin{aligned}C^{\mathbb{C}}_{p,1,r}&=\max\{k:x^p+y^p=k^p,(x,y)\in S_{(1),r}(n,\mathbb{C})\}\\&\leq\max\{k:x^p+y^p=k^p,(x,y)\in\mathcal{T}_r\}\qquad(6)\\&\leq\max\{k:x^p+y^p=k^p,(x,y)\in S_{(1),r}(2,\mathbb{R})\}\\&=C^{\mathbb{R}}_{p,1,r}\ \text{ for }2\times 2\text{ real matrices,}\end{aligned}$$

where (6) follows from Theorem 2.12. The result follows readily from [4, Theorem 4.1]. □

The proof of $S_{(1),r}(n,\mathbb{C})\subset\mathcal{T}_r$ is divided into two main parts. We prove the result for $n=2$ in Section 2.3. Then, extending the idea used in [5], we prove result for $n>2$ in Section 2.4.

2.3. *The inclusion $S_{(1),r}(2,\mathbb{C}) \subset \mathcal{T}_r$*

We first consider $n = 2$ in this section. We introduce a few more notations.

$$\begin{aligned}
\mathcal{D}_r(\mathbb{F}) &= \{X : X \in M_2(\mathbb{F}), \|X\|_r \le 1\},\\
\partial\mathcal{D}_r(\mathbb{F}) &= \{X : X \in M_2(\mathbb{F}), \|X\|_r = 1\},\\
\mathcal{D}_r^0(\mathbb{F}) &= \{X : X \in M_2(\mathbb{F}), \|X\|_r \le 1, \operatorname{tr} X = 0\},\\
\partial\mathcal{D}_r^0(\mathbb{F}) &= \{X : X \in M_2(\mathbb{F}), \|X\|_r = 1, \operatorname{tr} X = 0\}.
\end{aligned}$$

Suppose $\mathbf{c} = (c_1, c_2)^T$ with $c_1 \ge c_2 \ge 0$. Let $D_{\mathbf{c}} = \operatorname{diag}(c_1, c_2)$. Define $\nu_{\mathbf{c}}$ on $M_2(\mathbb{C})$ by

$$\nu_{\mathbf{c}}(X) = \max\{\operatorname{Re}\big(\operatorname{tr}(D_{\mathbf{c}}UXV)\big) : U, V \in U(2)\} = c_1 s_1(X) + c_2 s_2(X). \tag{7}$$

The second equality is due to von Neumann [15] (see, for example, [14, 20.B.1]). It gives a variational characterization of the inner product of $(c_1, c_2)^T$ and $\mathbf{s}(X)$. The following lemma, giving two very nice properties of $\nu_{\mathbf{c}}$ that we need, can be verified readily.

Lemma 2.8. *Suppose $\mathbf{c} = (c_1, c_2)^T \in \mathbb{R}^2$ with $c_1 \ge c_2 \ge 0$. Then*

(a) The function $\nu_{\mathbf{c}}$ is unitary similarity invariant.

(b) For each fixed $Y \in M_2(\mathbb{C})$, $\nu_{\mathbf{c}}([X, Y])$ is a convex function in X.

Lemma 2.9. *Let $1 \le r \le \infty$. For each $\mathbf{c} = (c_1, c_2)^T \in \mathbb{R}^2$ with $c_1 \ge c_2 \ge 0$,*

$$\max\{\mathbf{c}^T\mathbf{x} : \mathbf{x}^T \in S_{(1),r}(2,\mathbb{C})\} = \max\{\mathbf{c}^T\mathbf{x} : \mathbf{x}^T \in S^0_{(1),r}(2,\mathbb{R})\}.$$

Proof. We first show that for any $Y \in \partial\mathcal{D}_r^0(\mathbb{C})$,

$$\mathbf{c}^T\mathbf{s}([E_{12}, Y]) \le \max\{\mathbf{c}^T\mathbf{x} : \mathbf{x} \in S^0_{(1),r}(\mathbb{R})\}. \tag{8}$$

As $\operatorname{tr} Y = 0$, there exist a unit $\theta \in \mathbb{C}$ and a diagonal unitary matrix D such that $\theta D^*YD = \begin{bmatrix} |y_{11}| & * \\ |y_{21}| & -|y_{11}| \end{bmatrix}$. By Lemmas 2.6 and 2.4(b), $\|\operatorname{Re}(\theta D^*YD)\|_r \le \|\theta D^*YD\|_r = 1$. We may assume $\operatorname{Re}(\theta D^*YD) \ne 0$. Otherwise, $y_{11} = y_{21} = 0$ and this implies $[E_{12}, Y] = 0$. Then, (8) is trivial. Let $Z = \frac{\operatorname{Re}(\theta D^*YD)}{\|\operatorname{Re}(\theta D^*YD)\|_r} \in \partial\mathcal{D}_r^0(\mathbb{R})$. Then,

$$\mathbf{s}([E_{12}, Y]) = \mathbf{s}\left(\begin{bmatrix} y_{21} & -2y_{11} \\ 0 & -y_{21} \end{bmatrix}\right) = \mathbf{s}\left(\begin{bmatrix} |y_{21}| & -2|y_{11}| \\ 0 & -|y_{21}| \end{bmatrix}\right) = \mathbf{s}([E_{12}, \operatorname{Re}(\theta D^*YD)]) = t\mathbf{s}([E_{12}, Z]),$$

where $0 < t = \|\operatorname{Re}(\theta D^*YD)\|_r \le 1$. So, (8) holds because

$$\mathbf{c}^T\mathbf{s}([E_{12}, Y]) = t\mathbf{c}^T\mathbf{s}([E_{12}, Z]) \le \mathbf{c}^T\mathbf{s}([E_{12}, Z]).$$

To prove the lemma, it suffices to prove the inequality ($\le$). The other part is trivial. Suppose

$$\max\{\mathbf{c}^T\mathbf{x} : \mathbf{x}^T \in S_{(1),r}(2,\mathbb{C})\} = \mathbf{c}^T\mathbf{s}([X, Y]),$$

where $(X, Y) \in \Sigma_{(1),r}(2,\mathbb{C})$. We first show that

$$\nu_{\mathbf{c}}([X, Y]) \le \max\{\nu_{\mathbf{c}}([X, Y]) : (X, Y) \in \Sigma^0_{1,r}(2,\mathbb{C})\}. \tag{9}$$

Obviously, X and Y are not multiples of I. Thus, $\|X - \frac{\operatorname{tr} X}{2}I\|_1$ and $\|Y - \frac{\operatorname{tr} Y}{2}I\|_r$ are positive. By Lemma 2.5, $\|X - \frac{\operatorname{tr} X}{2}I\|_1 \le 1$ and $\|Y - \frac{\operatorname{tr} Y}{2}I\|_r \le 1$. Let

$$\tilde{X} = \frac{X - \frac{\operatorname{tr} X}{2}I}{\|X - \frac{\operatorname{tr} X}{2}I\|_1} \quad \text{and} \quad \tilde{Y} = \frac{Y - \frac{\operatorname{tr} Y}{2}I}{\|Y - \frac{\operatorname{tr} Y}{2}I\|_r},$$

so that $(\tilde{X}, \tilde{Y}) \in \Sigma^0_{1,r}(2, \mathbb{C})$. Then,

$$[\tilde{X}, \tilde{Y}] = \frac{1}{\|X - \frac{\operatorname{tr} X}{2} I\|_1 \cdot \|Y - \frac{\operatorname{tr} Y}{2} I\|_r} [X, Y],$$

from which we easily get $\nu_{\mathbf{c}}([X, Y]) \leq \nu_{\mathbf{c}}([\tilde{X}, \tilde{Y}])$. Thus (9) holds. Then,

$$\begin{aligned} \mathbf{c}^T \mathbf{s}([X, Y]) &\leq \max\{\nu_{\mathbf{c}}([X, Y]) : (X, Y) \in \Sigma^0_{1,r}(2, \mathbb{C})\} && (10) \\ &\leq \max\{\nu_{\mathbf{c}}([X, Y]) : X \in \mathcal{D}^0_1(\mathbb{C}), Y \in \partial\mathcal{D}^0_r(\mathbb{C})\} \\ &= \max\{\nu_{\mathbf{c}}([X, Y]) : X \in \operatorname{ext}(\mathcal{D}^0_1(\mathbb{C})), Y \in \partial\mathcal{D}^0_r(\mathbb{C})\} && (11) \\ &= \max\{\nu_{\mathbf{c}}([E_{12}, Y]) : Y \in \partial\mathcal{D}^0_r(\mathbb{C})\} && (12) \\ &\leq \max\{\mathbf{c}^T \mathbf{x} : \mathbf{x} \in S^0_{(1),r}(2, \mathbb{R})\}. && (13) \end{aligned}$$

Note that (10) follows from (7) and (9); (11) follows from Lemma 2.8(b); (12) follows from 2.8(a) and the fact that $\operatorname{ext}(\mathcal{D}^0_1(\mathbb{C})) = \{U^* E_{12} U : U \in U(2)\}$; and (13) follows from (8). □

In the following theorem, though we just need the result for $1 \leq r < 2$, we include the result for $2 \leq r \leq \infty$ which gives the precise description of the set $S_{(1),r}(2, \mathbb{C})$.

Theorem 2.10. *Suppose $n = 2$ and $1 \leq r \leq \infty$. Then,*

(a) $S_{(1),r}(2, \mathbb{C}) \subset \mathcal{T}_r$, $1 \leq r < 2$;

(b) $S_{(1),r}(2, \mathbb{C}) = S^0_{(1),r}(2, \mathbb{R}) = \mathcal{R}_r$, $2 \leq r \leq \infty$.

Proof. (a) We first note that the constant

$$\max\{s_1([X, Y]) : (X, Y) \in \Sigma_{(1),r}(n, \mathbb{F})\} = C^{\mathbb{F}}_{\infty,1,r} = C^{\mathbb{R}}_{\tilde{r},1,1} \qquad (14)$$

is independent of n (≥ 2) and $\mathbb{F}$, see [5]. For the last equality, see (2). As $\mathcal{R}_r = S_{(1),r}(2, \mathbb{R})$, we know that the vertical tangent line of C_r is $x = C^{\mathbb{R}}_{\tilde{r},1,1}$. Also, from (14), we deduce that $S_{(1),r}(2, \mathbb{C})$ is contain in $T(C^{\mathbb{R}}_{\tilde{r},1,1})$, the triangular region with the origin, $(C^{\mathbb{R}}_{\tilde{r},1,1}, 0)$ and $(C^{\mathbb{R}}_{\tilde{r},1,1}, C^{\mathbb{R}}_{\tilde{r},1,1})$ as vertices. For illustration, Figure 2.2 shows the regions $T(C^{\mathbb{R}}_{\tilde{r},1,1})$, $\mathcal{R}_r$ (green) and $\mathcal{Q}_r$ (blue), with $r = 1.3$.

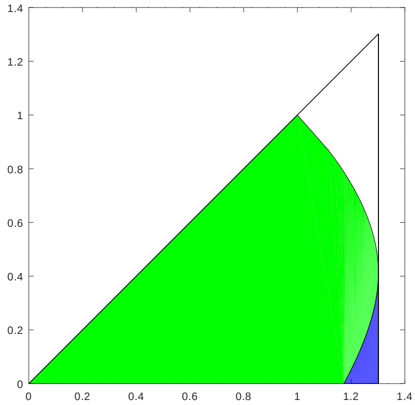

Figure 2.2. The triangular region $T(C^{\mathbb{R}}_{\tilde{r},1,1})$, the region $\mathcal{R}_r$ (green), and the region $\mathcal{Q}_r$ (blue), where $r = 1.3$.

Assume to the contrary that $S_{(1),r}(2, \mathbb{C}) \not\subset \mathcal{T}_r$ $(= \mathcal{R}_r \cup \mathcal{Q}_r)$. Then, there exists $\mathbf{u}^T \in S_{(1),r}(2, \mathbb{C})$ such that $\mathbf{u}^T \in T(C^{\mathbb{F}}_{\tilde{r},1,1}) \setminus \mathcal{T}_r$. In other words, $\mathbf{u}^T \in T(C^{\mathbb{F}}_{\tilde{r},1,1})$ and is above the curve C^-_r (i.e., in the white bounded region in Figure 2.2). So, there exists a non-vertical tangent line L of C^-_r with negative slope which separates

$\mathbf{u}^T$ from $\mathcal{T}_r$. From Lemma 2.3, the slope m of L must satisfy $m \le -1$. So, we may take $\mathbf{c} = (c_1, c_2)^T$ with $c_1 \ge c_2 > 0$ to be a normal vector of L. Then, as $S^0_{(1),r}(2, \mathbb{R}) \subset \mathcal{T}_r$,

$$\max\{\mathbf{c}^T\mathbf{x} : \mathbf{x}^T \in S^0_{(1),r}(2, \mathbb{R})\} \le \max\{\mathbf{c}^T\mathbf{x} : \mathbf{x}^T \in \mathcal{T}_r\} < \mathbf{c}^T\mathbf{u} \le \max\{\mathbf{c}^T\mathbf{x} : \mathbf{x}^T \in S_{(1),r}(2, \mathbb{C})\}.$$

This contradicts Lemma 2.9. The result follows.

(b) When $2 \le r \le \infty$, by Lemma 2.3, the slopes of tangent lines of C_r are always negative, except when $r = 2$ and at the point $(\sqrt{2}, 0)$ where the tangent line is vertical. The proof is similar to that of part (a), with $\mathcal{T}_r$ replaced by $\mathcal{R}_r = S^0_{(1),r}(2, \mathbb{R})$. □

2.4. *The inclusion* $S_{(1),r}(n, \mathbb{C}) \subset \mathcal{T}_r$

Lemma 2.11. *Suppose* $n \ge 5$ *and* $1 \le r \le \infty$. *Then,* $S_{(1),r}(n, \mathbb{C}) = S_{(1),r}(4, \mathbb{C})$.

Proof. Suppose $n \ge 5$. It suffices to show $S_{(1),r}(n, \mathbb{C}) \subseteq S_{(1),r}(4, \mathbb{C})$. Let $(X, Y) \in \Sigma_{(1),r}(n, \mathbb{C})$. As $\operatorname{rank} X = 1$, let $X = \mathbf{ab}^*$, where $\mathbf{a}$ and $\mathbf{b}$ are unit vectors in $\mathbb{C}^n$. Then,

$$[X, Y] = \mathbf{a}(\mathbf{b}^*Y) - (Y\mathbf{a})\mathbf{b}^* = \mathbf{a}(Y^*\mathbf{b})^* - (Y\mathbf{a})\mathbf{b}^*.$$

Let $U = [\mathbf{u}_1\, \mathbf{u}_2 \cdots \mathbf{u}_n] \in U(n)$ with $\{\mathbf{a}, \mathbf{b}, Y\mathbf{a}, Y^*\mathbf{b}\} \subseteq \operatorname{span}\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4\}$. Readily, we have

$$U^*XU = \hat{X} \oplus 0 \quad \text{and} \quad U^*[X, Y]U = Z \oplus 0,$$

where $\hat{X}, Z \in M_4(\mathbb{C})$ with $\mathbf{s}(\hat{X}) = (1, 0, 0, 0)^T$. Let $U^*YU = [U_i^*YU_j]_{i,j=1,2} = [\hat{Y}_{i,j}]_{i,j=1,2}$. Then,

$$Z \oplus 0 = U^*[X, Y]U = [U^*XU, U^*YU] = \begin{bmatrix} [\hat{X}, \hat{Y}_{11}] & * \\ * & 0 \end{bmatrix},$$

and hence $Z = [\hat{X}, \hat{Y}_{11}]$. Thus, $\mathbf{s}_{1,2}([X, Y]) = \mathbf{s}_{1,2}(Z) = \mathbf{s}_{1,2}([\hat{X}, \hat{Y}_{11}])$. If $\hat{Y}_{11} = 0$, the result is trivial. Suppose $\hat{Y}_{11} \neq 0$. Then

$$\mathbf{s}_{1,2}([X, Y]) = \|\hat{Y}_{11}\|_r \; \mathbf{s}_{1,2}\left(\left[\hat{X}, \hat{Y}_{11}/\|\hat{Y}_{11}\|_r\right]\right).$$

Note that $\hat{Y}_{11}$ is a submatrix of U^*YU and so $\|\hat{Y}_{11}\|_r \le \|U^*YU\|_r = \|Y\|_r = 1$. By Lemma 2.1, the result follows. □

Theorem 2.12. *Suppose* $n \ge 2$ *and* $1 \le r < 2$. *Then* $S_{(1),r}(n, \mathbb{C}) \subset \mathcal{T}_r$.

Proof. We first note that when $\operatorname{rank} X \le 1$, $\operatorname{rank}([X, Y]) \le 2$ and so $s_1^2([X, Y]) + s_2^2([X, Y]) = \|\|[X, Y]\|\|^2$. Thus, $\|\|[X, Y]\|\|$ is also the distance of the point $\mathbf{s}_{1,2}([X, Y])^T$ from the origin. For notation simplicity, abbreviate $s_1([X, Y]) \cdot s_2([X, Y])$ as $s_{1\cdot 2}([X, Y])$, and let $s^2_{1\cdot 2}([X, Y]) = (s_{1\cdot 2}([X, Y]))^2$.

Theorem 2.10(a) gives the result for $n = 2$. As $S_{(1),r}(3, \mathbb{C}) \subset S_{(1),r}(4, \mathbb{C})$, by Lemma 2.11, it suffices to prove the result for $n = 4$. The idea of our proof comes from [5] and [7] in which, instead of $\mathbf{s}([X, Y])$, the coefficients of the characteristic polynomials of $[X, Y]^*[X, Y]$ are studied. Here, we carry out our proof with reference to $\mathcal{T}_r$. In [5], $Y = \mathbf{cd}^*$ is a rank one matrix. Perturbation can be done by varying the unit vectors $\mathbf{c}$ and $\mathbf{d}$. Without the rank one condition on Y here, we need different arguement and the techniques are more involved. We divide the proof into three steps.

Step 1. Problem formulation. Suppose to the contrary that $S_{(1),r}(4, \mathbb{C}) \not\subset \mathcal{T}_r$. Then there exists $(X_0, Y_0) \in \Sigma_{(1),r}(4, \mathbb{C})$ such that $\mathbf{s}_{1,2}([X_0, Y_0])^T \in S_{(1),r}(4, \mathbb{C}) \setminus \mathcal{T}_r$. As explained in (14), the vertical tangent line of C_r is $x = C^{\mathbb{R}}_{\tilde{r},1,1}$ and $C^{\mathbb{R}}_{\tilde{r},1,1}$ is independent of n and $\mathbb{F}$. So, $\mathbf{s}_{1,2}([X_0, Y_0])^T$ lies in the left closed open half plane determined by $x = C^{\mathbb{R}}_{\tilde{r},1,1}$. On the other hand, from (20) below, we know that for any $(X, Y) \in \Sigma_{(1),r}(4, \mathbb{C})$, $s^2_{1\cdot 2}([X, Y]) \le 1$. Thus, $\mathbf{s}_{1,2}([X_0, Y_0])^T$ lies on or below the branch of the hyperbola $xy = 1$ in the first quadrant. Hence, $\mathbf{s}_{1,2}([X_0, Y_0])^T$ lies in the region $\mathcal{P}_r$ bounded by the vertical line $x = C^{\mathbb{R}}_{\tilde{r},1,1}$, the curves C_r^- and $xy = 1$, but not on C_r^-. For illustration, Figure 2.3 shows the regions $\mathcal{R}_r$ (green), $\mathcal{Q}_r$ (blue) and $\mathcal{P}_r$ (red), with $r = 1.3$.

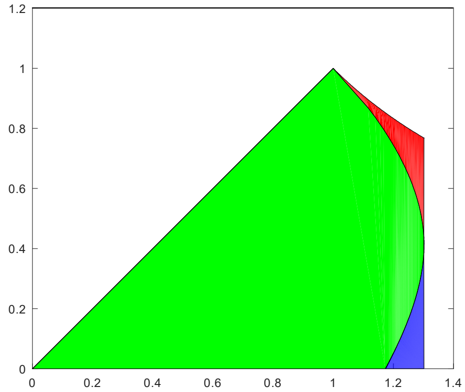

Figure 2.3. The regions $\mathcal{R}_r$ (green), $\mathcal{Q}_r$ (blue) and $\mathcal{P}_r$ (red), with $r = 1.3$.

The point at which C_r has a vertical tangent is $(x(t_r), y(t_r))$ (where $x(t_r) = C^{\mathbb{R}}_{\tilde{r},1,1}$). Let $\beta_r = x^2(t_r)y^2(t_r)$. Obviously, the point $\mathbf{s}_{1,2}([X_0, Y_0])^T$ lies above the hyperbola $xy = \sqrt{\beta_r}$ and so $\beta_r < s^2_{1\cdot 2}([X_0, Y_0]) \leq 1$. Suppose, for some $\beta_r < \beta_0 \leq 1$, $\mathbf{s}_{1,2}([X_0, Y_0])^T$ lies on the hyperbola $xy = \sqrt{\beta_0}$. It is easy to check that when a point on the part of the hyperbola $\{(x, y) : xy = \sqrt{\beta_0}, x \geq y > 0\}$ goes along the hyperbola to the right, the distance of the point from the origin increases. Consequently,

$$\begin{aligned}
& \max\{\|[X,Y]\|^2 : X, Y \in \Sigma_{(1),r}(4,\mathbb{C}),\ s^2_{1\cdot 2}([X,Y]) = \beta_0\} \\
\geq\ & s_1^2([X_0,Y_0]) + s_2^2([X_0,Y_0]) \\
>\ & \max\{x^2 + y^2 : (x,y) \in \mathcal{T}_r, xy = \sqrt{\beta_0}\} && (15) \\
\geq\ & \max\{\|[X,Y]\|^2 : X, Y \in \Sigma_{(1),r}(2,\mathbb{C}),\ s^2_{1\cdot 2}([X,Y]) = \beta_0\}. && (16)
\end{aligned}$$

Inequality (16) follows from Theorem 2.10(a).

The inequality (15) is strict. To prove the result, we are going to obtain a contradiction by showing that for all $\beta_r < \beta \leq 1$, the first and the last terms in the above inequalities are equal, i.e.,

$$\begin{aligned}
& \max\{\|[X,Y]\|^2 : X, Y \in \Sigma_{(1),r}(4,\mathbb{C}),\ s^2_{1\cdot 2}([X,Y]) = \beta\} \\
=\ & \max\{\|[X,Y]\|^2 : X, Y \in \Sigma_{(1),r}(2,\mathbb{C}),\ s^2_{1\cdot 2}([X,Y]) = \beta\}.
\end{aligned}$$

The result then follows. Our strategy is to establish (17) and (18) below for all $\beta_r < \beta \leq 1$:

$$\begin{aligned}
& \max\{\|[X,Y]\|^2 : X, Y \in \Sigma_{(1),r}(4,\mathbb{C}),\ s^2_{1\cdot 2}([X,Y]) = \beta\} \\
\leq\ & \max\{\|[X,Y]\|^2 : X, Y \in \Sigma_{(1),r}(4,\mathbb{C}),\ s^2_{1\cdot 2}([X,Y]) \leq \beta\} \\
\leq\ & \max\{\|[X,Y]\|^2 : X, Y \in \Sigma_{(1),r}(2,\mathbb{C}),\ s^2_{1\cdot 2}([X,Y]) \leq \beta\} && (17) \\
=\ & \max\{\|[X,Y]\|^2 : X, Y \in \Sigma_{(1),r}(2,\mathbb{C}),\ s^2_{1\cdot 2}([X,Y]) = \beta\} && (18) \\
\leq\ & \max\{\|[X,Y]\|^2 : X, Y \in \Sigma_{(1),r}(4,\mathbb{C}),\ s^2_{1\cdot 2}([X,Y]) = \beta\}.
\end{aligned}$$

Step 2. We establish (18). For each $\beta_r < \xi \leq 1$, let

$$\begin{aligned}
M(\xi) &= \max\{\|[X,Y]\|^2 : (X,Y) \in \Sigma_{(1),r}(2,\mathbb{C}),\ s^2_{1\cdot 2}([X,Y]) = \xi\} \\
&= \max\{\|\mathbf{x}\|^2 : \mathbf{x}^T \in S_{(1),r}(2,\mathbb{C}),\ x_1^2 x_2^2 = \xi\}.
\end{aligned}$$

To show (18), it suffices to show that M is an increasing function in ξ. As $\xi > \beta_r$, by Theorem 2.10(a), it is obvious that $\max\{\|\mathbf{x}\|^2 : \mathbf{x}^T \in S_{(1),r}(2,\mathbb{C}),\ x_1^2x_2^2 = \xi\}$ has it maximum achieved at the unique intersection point of $xy = \sqrt{\xi}$ and C_r^-. Thus, it suffices to show that when ξ increases, the square of the distance of the intersection point from the origin increases. In terms of the parametrization of C_r in (4), let

$$F(t) = x^2(t) + y^2(t) = 2(f^2(t) + g^2(t)) \quad \text{and} \quad \xi(t) = x^2(t)y^2(t) = (f^2(t) - g^2(t))^2, \quad 0 \leq t \leq 1.$$

Note that we have $f^2(t) - g^2(t) > 0$ if $t \neq 1$, and it is shown in the proof of [4, Lemma 3.5] that $f' < 0$ and $g' > 0$ on the open interval $(0, 1)$. Thus, $\frac{d\xi}{dt} = 4(f^2(t) - g^2(t))(f(t)f'(t) - g(t)g'(t)) < 0$ on $(0, 1)$. Hence, ξ is one-to-one and F may be regarded as a function in ξ. Using symbolic calculation with MATLAB, we get, for $0 < t < 1$ and $1 < r < 2$,

$$\frac{dF}{d\xi} = \frac{\frac{d}{dt}[2(f^2(t) + g^2(t))]}{\frac{d}{dt}(f^2(t) - g^2(t))^2} = \frac{(t^{2r} - t^4)(t^{2r} + 1)^{\frac{2}{r}}(t^{2r} + t^2)^2}{t^6(1 - t^{2r})^3} > 0.$$

Hence, F is an increasing function in ξ (for all $t \in (0, 1)$). Consequently, (18) is valid.

Step 3. We establish (17). Suppose $(X, Y) \in \Sigma_{(1),r}(4, \mathbb{C})$ with $X = \mathbf{a}\mathbf{b}^*$ where $\mathbf{a}, \mathbf{b} \in \mathbb{C}^4$ are unit vectors. Let $U = [\mathbf{u}_1 \cdots \mathbf{u}_4] \in U(4)$ with $\mathbf{u}_1 = \mathbf{a}$ and $\mathbf{u}_2 = \frac{Y\mathbf{a} - (\mathbf{a}^*Y\mathbf{a})\mathbf{a}}{\|Y\mathbf{a} - (\mathbf{a}^*Y\mathbf{a})\mathbf{a}\|}$ if $Y\mathbf{a}$ is not a multiple of $\mathbf{a}$. Then $\mathbf{a}, Y\mathbf{a} \in \operatorname{span}\{\mathbf{u}_1, \mathbf{u}_2\}$. Let $V = [\mathbf{v}_1 \cdots \mathbf{v}_4] \in U(4)$ such that $\mathbf{v}_1 = \mathbf{b}$ and $\mathbf{v}_2 = \frac{Y^*\mathbf{b} - (\mathbf{b}^*Y^*\mathbf{b})\mathbf{b}}{\|Y^*\mathbf{b} - (\mathbf{b}^*Y^*\mathbf{b})\mathbf{b}\|}$ if $Y^*\mathbf{b}$ is not a multiple of $\mathbf{b}$. We have

$$U^*([X, Y])V = U^*\Big(\mathbf{a}(Y^*\mathbf{b})^* - (Y\mathbf{a})\mathbf{b}^*\Big)V = \mathbf{e}_1(Y^*\mathbf{b})^*V - U^*(Y\mathbf{a})\mathbf{e}_1^* = Z \oplus 0_2,$$

where

$$Z = \begin{bmatrix} \mathbf{b}^*Y\mathbf{b} - \mathbf{a}^*Y\mathbf{a} & \sqrt{\|Y^*\mathbf{b}\|^2 - |\mathbf{b}^*Y^*\mathbf{b}|^2} \\ -\sqrt{\|Y\mathbf{a}\|^2 - |\mathbf{a}^*Y\mathbf{a}|^2} & 0 \end{bmatrix}.$$

Hence,

$$\|[X, Y]\|^2 = \|Z\|^2 = \|Y\mathbf{a}\|^2 - 2\mathrm{Re}\Big((\mathbf{a}^*Y\mathbf{a})(\mathbf{b}^*Y^*\mathbf{b})\Big) + \|Y^*\mathbf{b}\|^2 \tag{19}$$

and

$$s_{1\cdot 2}^2([X, Y]) = |\det Z|^2 = \Big(\|Y\mathbf{a}\|^2 - |\mathbf{a}^*Y\mathbf{a}|^2\Big)\Big(\|Y^*\mathbf{b}\|^2 - |\mathbf{b}^*Y^*\mathbf{b}|^2\Big). \tag{20}$$

Suppose $\beta_r < \beta \leq 1$ and

$$\max\{\|[X, Y]\|^2 : X, Y \in \Sigma_{(1),r}(4, \mathbb{C}),\ s_{1\cdot 2}^2([X, Y]) \leq \beta\} = \|[X, Y]\|^2 > 0,$$

where $(X, Y) \in \Sigma_{(1),r}(4, \mathbb{C})$ with $X = \mathbf{a}\mathbf{b}^*$ where $\mathbf{a}, \mathbf{b} \in \mathbb{C}^4$ are unit vectors. We first note that we must have $s_{1\cdot 2}^2([X, Y]) > \beta_r$. Otherwise, $\mathbf{s}_{1,2}([X, Y])^T$ lies in the region bounded by the x-axis, the line $x = y$, the hyperbola $xy = \sqrt{\beta_r}$ and the vertical line $x = C_{\tilde{r},1,1}^{\mathbb{R}}$. Obviously, the point in this region that is furthest from the origin is the point $(x(t_r), y(t_r))$. So, $x^2(t_r) + y^2(t_r) \geq \|[X, Y]\|^2$. Suppose $(H, K) \in \Sigma_{(1),r}(2, \mathbb{R})$ such that $\mathbf{s}([H, K])^T \in C_r^-$ and $s_{1\cdot 2}^2([H, K]) = \beta$. From the proof in Step 2, we know that $\|[H, K]\|^2 > x^2(t_r) + y^2(t_r)$. Then, $(H \oplus 0, K \oplus 0) \in \Sigma_{(1),r}(4, \mathbb{C})$, $s_{1\cdot 2}^2([H \oplus 0, K \oplus 0]) = \beta$ and $\|[H \oplus 0, K \oplus 0]\|^2 > \|[X, Y]\|^2$. This contradicts the maximality of $\|[X, Y]\|^2$.

We now have different arguement, depending on whether **a** and **b** are linearly dependent or not.

Suppose **a** and **b** are linearly dependent. Under simultaneous unitary similarity on X and Y and the multiplication of a suitable unit scalar to X, we may assume $X = E_{11}$. Let $U_i = [1] \oplus W_i$ where $W_i \in U(3)$, $i = 1, 2$. By direct calculation, we have $U_1^*[E_{11}, U_1YU_2]U_2^* = [E_{11}, Y]$ so that $\mathbf{s}([E_{11}, U_1YU_2]) = \mathbf{s}([E_{11}, Y])$. Thus, with U_1 and U_2 suitably chosen, we may assume $y_{13} = y_{14} = y_{31} = y_{41} = 0$. Then, $[X, Y] = \begin{bmatrix} 0 & y_{12} \\ -y_{21} & 0 \end{bmatrix} \oplus 0$. On the other hand, as $\begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix}$ is a submatrix of Y, $\left\|\begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix}\right\|_r \leq 1$. Let $\tilde{Y} = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & t \end{bmatrix}$ where t is chosen such that $\|\tilde{Y}\|_r = 1$. Then, with $E_{11} \in M_2(\mathbb{C})$, $[E_{11}, \tilde{Y}] = \begin{bmatrix} 0 & y_{12} \\ -y_{21} & 0 \end{bmatrix}$. Readily, (17) follows.

Suppose now **a** and **b** are linearly independent.

Case 1. $\mathbf{a}^*Y\mathbf{a} \neq 0$ or $\mathbf{b}^*Y^*\mathbf{b} \neq 0$. Since $\mathbf{s}([X,Y]) = \mathbf{s}([X^*,Y^*])$, by replacing $[X,Y]$ by $[X^*,Y^*]$ if necessary, we may assume $\mathbf{a}^*Y\mathbf{a} \neq 0$. In addition, under simultaneous unitary similarity on X and Y and the multiplication of a suitable unit scalar to X, we may assume that $\mathbf{a} = \mathbf{e}_1$, $\mathbf{b} = (b_1, b_2, 0, 0)^T$, where $b_1 \geq 0$ and $b_2 > 0$. Write $Y^*\mathbf{b} = \mathbf{d} = (d_1, d_2, d_3, d_4)^T$.

Claim. $(b_2, 0, 0)^T$ and $(d_2, d_3, d_4)^T$ are linearly dependent.

Before we prove the claim, let us first see how the claim works. Let $Y = [Y_{ij}]_{i,j=1,2}$ where $Y_{ij} \in M_2(\mathbb{C})$. For $\mathbf{x} = (x_1, x_2, x_3, x_4)^T \in \mathbb{C}^4$, let $\hat{\mathbf{x}} = (x_1, x_2)^T \in \mathbb{C}^2$. With the claim, we see that (since $b_2 \neq 0$) $Y_{12}^*\hat{\mathbf{b}} = (d_3, d_4)^T = 0$. So,

$$[X,Y] = \mathbf{ab}^*Y - Y\mathbf{ab}^* = \begin{bmatrix} \hat{\mathbf{a}}\hat{\mathbf{b}}^*Y_{11} - Y_{11}\hat{\mathbf{a}}\hat{\mathbf{b}}^* & 0 \\ -Y_{21}\hat{\mathbf{a}}\hat{\mathbf{b}}^* & 0 \end{bmatrix}. \tag{21}$$

We now show that $Y_{21}\hat{\mathbf{a}} = 0$. As $\hat{\mathbf{a}}, \hat{\mathbf{b}} \in \mathbb{R}^2$ are unit vectors, $\hat{\mathbf{a}} + \hat{\mathbf{b}}$ and $\hat{\mathbf{a}} - \hat{\mathbf{b}}$ are orthogonal. Suppose $\hat{\mathbf{a}} + \hat{\mathbf{b}} = \|\hat{\mathbf{a}} + \hat{\mathbf{b}}\|\mathbf{u}_1$ and $\hat{\mathbf{a}} - \hat{\mathbf{b}} = \|\hat{\mathbf{a}} - \hat{\mathbf{b}}\|\mathbf{u}_2$. Then,

$$2\hat{\mathbf{a}} = \|\mathbf{a} + \mathbf{b}\|\mathbf{u}_1 + \|\mathbf{a} - \mathbf{b}\|\mathbf{u}_2 \quad \text{and} \quad 2\hat{\mathbf{b}} = \|\mathbf{a} + \mathbf{b}\|\mathbf{u}_1 - \|\mathbf{a} - \mathbf{b}\|\mathbf{u}_2.$$

Take $W \in O(2)$ such that $W\mathbf{u}_1 = \mathbf{u}_1$ and $W\mathbf{u}_2 = -\mathbf{u}_2$. We see that $W\hat{\mathbf{a}} = \hat{\mathbf{b}}$ and $W\hat{\mathbf{b}} = \hat{\mathbf{a}}$. Let $U = W \oplus I$. Then $UX^*U^* = X$. As

$$\mathbf{s}([X, UY^*U^*]) = \mathbf{s}([UX^*U^*, UY^*U^*]) = \mathbf{s}([X^*,Y^*]) = \mathbf{s}(-[X,Y]^*) = \mathbf{s}([X,Y]),$$

we may apply the claim to $[X, UY^*U^*]$ to have $(WY_{21}^*)^*\hat{\mathbf{b}} = 0$, i.e., $Y_{21}\hat{\mathbf{a}} = 0$. Thus, from (21),

$$\mathbf{s}_{1,2}([X,Y]) = \mathbf{s}([\hat{\mathbf{a}}\hat{\mathbf{b}}^*, Y_{11}]).$$

If $\|Y_{11}\|_r = 1$, (17) follows. Otherwise, as Y_{11} is a nonzero submatrix of Y, $0 < \|Y_{11}\|_r < 1$. Then,

$$\mathbf{s}([\hat{\mathbf{a}}\hat{\mathbf{b}}^*, Y_{11}]) = \|Y_{11}\|_r\, \mathbf{s}\left(\left[\hat{\mathbf{a}}\hat{\mathbf{b}}^*, \frac{Y_{11}}{\|Y_{11}\|_r}\right]\right).$$

As $S_{(1),r}(2,\mathbb{C})$ is star-shaped by Lemma 2.1, we deduce that $\mathbf{s}([\hat{\mathbf{a}}\hat{\mathbf{b}}^*, Y_{11}])^T \in S_{(1),r}(2,\mathbb{C})$ but not on the boundary curve C_r^-. The hyperbola $xy = s_{1\cdot 2}([\hat{\mathbf{a}}\hat{\mathbf{b}}^*, Y_{11}])$ $(= s_{1\cdot 2}([X,Y]) > \sqrt{\beta_r})$ intersects with C_r^- at some point $\mathbf{s}([H,K])^T$, where $(H,K) \in \Sigma^0_{(1),r}(2,\mathbb{R})$. The points $\mathbf{s}([\hat{\mathbf{a}}\hat{\mathbf{b}}^*, Y_{11}])^T$ and $\mathbf{s}([H,K])^T$ lie on the same hyperbola, and $\mathbf{s}([H,K])^T$ is on the right of $\mathbf{s}([\hat{\mathbf{a}}\hat{\mathbf{b}}^*, Y_{11}])^T$. Thus, for $(H \oplus 0, K \oplus 0) \in \Sigma_{(1),r}(4,\mathbb{C})$, we easily get $\|[H \oplus 0, K \oplus 0]\|^2 > \|[\hat{\mathbf{a}}\hat{\mathbf{b}}^*, Y_{11}]\|^2 = \|[X,Y]\|^2$. Also, we have $s^2_{1\cdot 2}([H \oplus 0, K \oplus 0]) = s^2_{1\cdot 2}([\hat{\mathbf{a}}\hat{\mathbf{b}}^*, Y_{11}]) = s^2_{1\cdot 2}([X,Y]) \in [\beta_r, \beta]$. This contradicts the maximality of $\|[X,Y]\|^2$. The result follows.

Proof of the claim. The idea of the proof can be found in [5]. We include the necessary modification here. Assume to the contrary that $(b_2, 0, 0)^T$ and $(d_2, d_3, d_4)^T$ are linearly independent. Then d_3 or d_4 is nonzero. By multiplying a suitable scalar to Y, we may assume $d_1 \geq 0$. Let $W \in U(3)$ such that

$$([1] \oplus W)Y^*\mathbf{b} = ([1] \oplus W)\mathbf{d} = (d_1, \|(d_2, d_3, d_4)^T\|, 0, 0)^T,$$

and let, for $\alpha \in [-\pi, \pi]$,

$$Y_\alpha = Y\Big([1] \oplus (e^{-\mathrm{i}\alpha}W^*)\Big),$$

and

$$\mathbf{d}_\alpha = Y_\alpha^*\mathbf{b} = \Big([1] \oplus (e^{\mathrm{i}\alpha}W)\Big)\mathbf{d} = (d_1, e^{\mathrm{i}\alpha}\|(d_2, d_3, d_4)^T\|, 0, 0)^T.$$

Define closed unit disks

$$D_1 = \{z : z \in \mathbb{C}, |z| \leq |\mathbf{b}^*\mathbf{d}|\} \quad \text{and} \quad D_2 = \{z : z \in \mathbb{C}, |z - b_1 d_1| \leq |b_2| \cdot \|(d_2, d_3, d_4)^T\|\}.$$

Let D_i^o and ∂D_i, $i = 1, 2$, denote the interior and boundary of D_i, respectively. Note that

$$\partial D_2 = \{\mathbf{b}^*\mathbf{d}_\alpha : -\pi \le \alpha \le \pi\}$$

is the circle centered at $b_1 d_1 \ge 0$ with radius $|b_2| \cdot \|(d_2, d_3, d_4)^T\| > |b_2 d_2|$. The arguement is now similar to those given in [5, page 171-172]. We can find an $Y_{\alpha_0} \in M_4(\mathbb{C})$ with $\|Y_{\alpha_0}\|_r = 1$ such that $|||[X, Y_{\alpha_0}]|||^2 > |||[X, Y]|||^2$ and $s^2_{1\cdot 2}([X, Y_{\alpha_0}]) \le \beta$. This gives a contradiction.

The proof of the following is different from the corresponding part in [5, Case 2] which, in Subcase 2.2, makes use of the result from [17] that $\{\mathbf{s}_{1,2}([X, Y])^T : X, Y \in M_3(\mathbb{R}), \mathbf{s}(X) = \mathbf{s}(Y) = \mathbf{e}_1\} = \mathcal{R}_1$. For our purpose, however, result on $S_{(1),r}(n, \mathbb{R})$, $n \ge 3$, is not available.

Case 2. $\mathbf{a}^*Y\mathbf{a} = \mathbf{b}^*Y^*\mathbf{b} = 0$. If $Y\mathbf{a} = Y^*\mathbf{b} = 0$, then $[X, Y] = 0$ by (19), and we have a contradiction. Again, by replacing (X, Y) by (X^*, Y^*) if necessary, we may assume $Y\mathbf{a} \ne 0$.

Subcase 2.1. $Y\mathbf{a} \notin \operatorname{span}\{\mathbf{b}\}$. With $\mathbf{a}$ and $\mathbf{b}$ are linearly independent, we now have $\mathbf{a}, Y\mathbf{a} \notin \operatorname{span}\{\mathbf{b}\}$. The orthogonal projections of $\mathbf{a}$ and $Y\mathbf{a}$ on $\{\mathbf{x} : \mathbf{x} \in \mathbb{C}^4, \mathbf{x}^*\mathbf{b} = 0\}$ are nonzero, and thus there exists $U \in U(4)$ such that $U\mathbf{b} = \mathbf{b}$ (hence $U^*\mathbf{b} = \mathbf{b}$), and $\mathbf{a}^*(UY\mathbf{a}) \ne 0$. Then, using (19) and (20), we check that (as $\mathbf{b}^*Y^*\mathbf{b} = 0$),

$$|||[X, UY]|||^2 = \|(UY)\mathbf{a}\|^2 - 2\mathrm{Re}\Big((\mathbf{b}^*(UY)^*\mathbf{b})(\mathbf{a}^*(UY)\mathbf{a})\Big) + \|(UY)^*\mathbf{b}\|^2 = |||[X, Y]|||^2$$

and

$$\begin{aligned} s^2_{1\cdot 2}([X, UY]) &= \Big(\|(UY)\mathbf{a}\|^2 - |\mathbf{a}^*(UY)\mathbf{a}|^2\Big)\Big(\|(UY)^*\mathbf{b}\|^2 - |\mathbf{b}^*(UY)^*\mathbf{b}|^2\Big) \\ &< \|Y\mathbf{a}\|^2 \cdot \|Y^*\mathbf{b}\|^2 \\ &= s^2_{1\cdot 2}([X, Y]) \\ &\le \beta. \end{aligned}$$

Thus, with Y replaced by UY, we are back to Case 1. The result follows.

Subcase 2.2. $0 \ne Y\mathbf{a} \in \operatorname{span}\{\mathbf{b}\}$. In this case, as $\mathbf{a}^*Y\mathbf{a} = 0$, we have $\mathbf{a}^*\mathbf{b} = 0$. Let $U \in U(4)$ such that $U\mathbf{a} = \mathbf{e}_1$ and $U\mathbf{b} = \mathbf{e}_2$. Replacing (X, Y) by (UXU^*, UYU^*), we may assume $X = E_{12} = \mathbf{e}_1\mathbf{e}_2^*$. Then, the assumptions $\mathbf{a}^*Y\mathbf{a} = \mathbf{b}^*Y^*\mathbf{b} = 0$ mean $y_{11} = y_{22} = 0$. Let $U_i = I_2 \oplus W_i$ where $W_i \in U(2)$, $i = 1, 2$. By direct calculation, we have $U_1^*[E_{12}, U_1YU_2]U_2^* = [E_{12}, Y]$ so that $\mathbf{s}([E_{12}, U_1YU_2]) = \mathbf{s}([E_{12}, Y])$. Thus, with U_1 and U_2 suitably chosen, we may assume $y_{41} = y_{24} = 0$, $y_{23} \le 0$ and $y_{31} \ge 0$. Then,

$$[X, Y] = \begin{bmatrix} y_{21} & 0 & y_{23} \\ 0 & -y_{21} & 0 \\ 0 & -y_{31} & 0 \end{bmatrix} \oplus [0].$$

Note that $|||[X, Y]|||^2 = 2|y_{21}|^2 + |y_{23}|^2 + |y_{31}|^2$ and

$$(|y_{21}|^2 + |y_{23}|^2)(|y_{21}|^2 + |y_{31}|^2) = \Big(\|Y^*\mathbf{b}\|^2 - |\mathbf{b}^*Y^*\mathbf{b}|^2\Big)\Big(\|Y\mathbf{a}\|^2 - |\mathbf{a}^*Y\mathbf{a}|^2\Big) \le \beta.$$

As $\begin{bmatrix} y_{21} & y_{23} \\ y_{31} & y_{33} \end{bmatrix}$ is a submatrix of Y, we deduce that $\left\|\begin{bmatrix} y_{23} & y_{33} \\ y_{21} & y_{31} \end{bmatrix}\right\|_r \le 1$. Let $E_{12} = \mathbf{e}_1\mathbf{e}_2^* \in M_2(\mathbb{C})$ and $K = \begin{bmatrix} y_{23} & t \\ y_{21} & y_{31} \end{bmatrix}$, where t is chosen such that $\|K\|_r = 1$. Then, $[E_{12}, K] = \begin{bmatrix} y_{21} & y_{31} - y_{23} \\ 0 & -y_{21} \end{bmatrix}$, from which we have (since $y_{23} \le 0$ and $y_{31} \ge 0$)

$$|||[E_{12}, K]|||^2 = 2|y_{21}|^2 + (y_{31} - y_{23})^2 \ge 2|y_{21}|^2 + |y_{23}|^2 + |y_{31}|^2 = |||[X, Y]|||^2$$

and

$$s^2_{1\cdot 2}([E_{12}, K]) = |\det[E_{12}, K]|^2 = |y_{21}|^4 \le (|y_{21}|^2 + |y_{23}|^2)(|y_{21}|^2 + |y_{31}|^2) \le \beta.$$

Thus, (17) follows. □

References

[1] K. Audenaert, *Variance bounds, with an application to norm bounds for commutators*, Linear Algebra Appl. **432** (2010), 1126–1143.

[2] A. Böttcher and D. Wenzel, *How big can the commutator of two matrices be and how big is it typically?* Linear Algebra Appl. **403** (2005), 216–228.

[3] A. Böttcher, D. Wenzel, *The Frobenius norm and the commutator*, Linear Algebra Appl. **429** (2008), 1864-1885.

[4] C.-M. Cheng, D. O. Akintoye, R. Jiao, *Commutator bounds and region of singular values of the commutator with a rank one matrix*, Linear Algebra Appl. **613** (2021), 347-376.

[5] C.-M. Cheng, R. Jiao, *Proof of Wenzel's conjecture concerning singular values of the commutator of rank one matrices*, Linear Algebra Appl. **592** (2020), 165-174.

[6] C.-M. Cheng, C. Lei, *On Schatten p-norms of commutators*, Linear Algebra Appl. **484** (2015), 409-434.

[7] C.-M. Cheng, Y. Liang, *Singular values, eigenvalues and diagonal elements of the commutator of 2×2 rank one matrices*, Electron. J. Linear Algebra **36** (2020), 1-20.

[8] C.-M. Cheng, Y. Liang, *Some sharp bounds for the commutator of real matrices*, Linear Algebra Appl. **521** (2017), 263-282.

[9] C.-M. Cheng, X. Jin, S. Vong, *A Survey on the Böttcher-Wenzel conjecture and related problems*, Oper. Matrices **9** (2015), 659-673.

[10] R.A. Horn, C.R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, 1991.

[11] L. László, *Proof of Böttcher and Wenzel's conjecture on commutator norms for* 3*-by-*3 *matrices*, Linear Algebra Appl. **422** (2007), 659-663.

[12] Z. Lu, *Normal scalar curvature conjecture and its applications*, J. Funct. Anal. **261** (2011), 1284-1308.

[13] Z. Lu, D. Wenzel, *Commutator estimates comprising the Frobenius norm - Looking back and forth*, In: Bini D., Ehrhardt T., Karlovich A., Spitkovsky I. (eds) Large Truncated Toeplitz Matrices, Toeplitz Operators, and Related Topics. Operator Theory: Advances and Applications, **259** (2017), 533-559, Birkhäuser, Cham.

[14] A. W. Marshall, I. Olkin, B. C. Arnold, *Inequalities: Theory of Majorization and Its Applications*, (2nd edition), Springer, New York, 2011.

[15] J. von Neumann, *Some matrix inequalities and metrization of matrix space*, Tomsk. Univ. Rev. **1** (1937), 286-300.

[16] S. Vong and X. Jin, *Proof of Böttcher and Wenzel's conjecture*, Oper. Matrices **2** (2008), 435-442.

[17] D. Wenzel, *A strange phenomenon for the singular values of commutators with rank one matrices*, Electron. J. Linear Algebra **30** (2015), 649-669.

[18] D. Wenzel and K. Audenaert, *Impressions of convexity: an illustration for commutator bounds*, Linear Algebra Appl. **433** (2010), 1726-1759.