ON THE NEAREST STABLE $2 \times 2$ MATRIX, DEDICATED TO PROF. SZE-BI HSU IN APPRECIATION OF HIS INSPIRING IDEAS

YUEH-CHENG KUO, HUAN-CHANG CHENG AND JHIIH-YOU SYU

Department of Applied Mathematics, National University of Kaohsiung
Kaohsiung, 811, Taiwan

SHIH-FENG SHIEH*

Department of Mathematics, National Taiwan Normal University
Taipei 116, Taiwan

Abstract. In this paper, we study the continuous-time nearest stable matrix problem: given a $2 \times 2$ real matrix $A$, minimize the Frobenius norm of $A - X$, where $X$ is a stable matrix. We provide an explicit formula for the global minimizer $X_*$. The uniqueness of the minimizer is also studied.

1. Introduction. We study the nearest stable matrix $X$ to a given unstable matrix $A \in \mathbb{R}^{n \times n}$. That is, we consider the nearest stable matrix problem:

$$\inf_{X \in \mathbb{S}^{n \times n}} \| A - X \|_F,$$

(1.1)

where $\| \cdot \|_F$ denotes the Frobenius norm of a matrix and $\mathbb{S}^{n \times n}$ is the set of all stable matrices of size $n \times n$. Here, a matrix $X$ is stable, denoted by $X \in \mathbb{S}^{n \times n}$, if all eigenvalues of $X$ are in the closed left half of the complex plane, denoted by $\mathbb{C}_{\leq}$, and all eigenvalues on the imaginary axis are semisimple. This kind of problem occurs in system identification or system modification. The measurements included noise or truncation may happen that the model system is unstable. Hence, we need to modify the errors of the unstable system and obtain a suitable system. Such problems have been considered and investigated in many contexts [2, 4, 5, 9]. For continuous-time systems, [9] presents an algorithm that constructs a sequence of successive stable iterates to compute a nearby stable approximation $X$. A reformulation of (1.1) is studied by using linear dissipative Hamiltonian systems in [7]. The authors of [4] use the projected gradient descent method to solve such a problem. In paper [1], a block coordinate descent method for this optimization problem (1.1) is proposed. The nearest stable matrix pair problems have been considered in [2]. For a discrete-time system, the nearest stable matrix problems have been investigated in [3, 9].

It is easily seen that the set $\mathbb{S}^{n \times n}$ is not open because the eigenvalues of a stable matrix are in $\mathbb{C}_{\leq}$. In addition, it is also not closed, because $A_\varepsilon = \begin{bmatrix} 0 & 1 \\ 0 & \varepsilon \end{bmatrix}$ with

2020 Mathematics Subject Classification. 15A39, 15A60.

Key words and phrases. Linear continuous-time systems, matrix approximation, stability radius, global minimizer.

* Corresponding author: Shih-Feng Shieh.
\( \varepsilon < 0 \) is stable but the imaginary eigenvalue of \( \lim_{\varepsilon \to 0} A_\varepsilon = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \) are not semisimple. Hence, the solution of problem (1.1) may not exist. Further, we have

\[
\inf_{X \in S^{n \times n}} \|A - X\|_F = \min_{X \in cl(S^{n \times n})} \|A - X\|_F,
\]

where \( cl(S^{n \times n}) = \{ X \in \mathbb{R}^{n \times n} \mid \text{all eigenvalues of } X \text{ are in } \mathbb{C}_\leq \} \) is the closure of the set \( S^{n \times n} \). The set \( cl(S^{n \times n}) \) of stable matrix is nonconvex, and therefore it is in general difficult to compute a globally optimal solution to problem (1.2).

The optimization problem (1.1) is difficult and possibly can only be solved by efficient numerical schemes. To simplify this problem, we consider that the unstable matrix \( A \) is a real \( 2 \times 2 \) matrix, i.e., \( n = 2 \). To be more precise, we consider the optimization problem

\[
\min_{X \in cl(S^{2 \times 2})} \|A - X\|_F,
\]

where \( A \in \mathbb{R}^{2 \times 2} \) is an unstable matrix. Denote by \( X_* \in \mathbb{R}^{2 \times 2} \) the minimizer of (1.3). In this paper, we give an explicit formula of \( X_* \) for the nearest stable matrix problem (1.3). Let \( A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathbb{R}^{2 \times 2} \) be unstable. The stable matrix \( X_* \) can be explicitly formulated as follows.

(i) If \( A \) with \( \tau = \text{tr}(A)/2 > 0 \) has complex conjugate eigenvalues. Then (1.3) has a unique minimizer \( X_* = A - \tau I \).

(ii) Suppose that \( \text{tr}(A) \leq 0 \). If \( \hat{R} = \sqrt{(a + d)^2 + (b - c)^2} > 0 \), then (1.3) has a unique minimizer

\[
X_* = R + \frac{\hat{R}}{4} \begin{bmatrix} -\sin \omega & -\sin \phi \\ \cos \omega + \cos \phi & -\sin \omega + \sin \phi \end{bmatrix},
\]

where \( R > 0 \) and \( \phi \) are in (3.8b), and \( \omega \in [0, \pi] \) satisfies \( \sin \omega = -(a + d)/\hat{R} \), \( \cos \omega = (b - c)/\hat{R} \). In addition, if \( \hat{R} = 0 \), then (1.3) has infinitely many minimizers.

(iii) Suppose that \( \text{tr}(A) > 0 \) and \( A \) has real eigenvalues. If \( b \neq c \), then (1.3) has a unique minimizer

\[
X_* = \frac{R + |b - c|}{4} \begin{bmatrix} -\sin \phi & \frac{b - c}{|b - c|} + \cos \phi \\ -\frac{-b + c}{|b - c|} + \cos \phi & \sin \phi \end{bmatrix},
\]

where \( R \) and \( \phi \) are in (3.8b). In addition, if \( b = c \), then (1.3) has exactly two minimizers.

The organization of this paper is as follows. In section 2, some preliminaries are given. In section 3, we provide an explicit formula of the minimizer of (1.3). We conclude this paper in Section 4.

2. Preliminaries. In this section, we introduce notations, definitions and some preliminary results. For a matrix \( A \in \mathbb{R}^{n \times m} \), \( A^\top \) denotes the transpose of \( A \), and the trace of \( A \), denoted by \( \text{tr}(A) \), is the sum of diagonal entries of \( A \). We use capital letters to denote matrices and lowercase (bold) letters to denote scalars (vectors).

In this paper, we consider the nearest stable matrix problem with \( n = 2 \). In the following, we quote some preliminary results for the size of matrix being 2. It is
well-known that a $2 \times 2$ orthogonal matrix is either
\[
Q_\theta = \begin{bmatrix}
\cos \theta & -\sin \theta \\
\sin \theta & \cos \theta
\end{bmatrix}
\text{ or }
\begin{bmatrix}
\cos \theta & \sin \theta \\
\sin \theta & -\cos \theta
\end{bmatrix} = Q_\theta \begin{bmatrix}
1 & 0 \\
0 & -1
\end{bmatrix}
\]  \hspace{1cm} (2.1)
for some $\theta \in [0, 2\pi)$, and the Frobenius norm $\| \cdot \|_F$ is orthogonal invariant. In the following, we quote two orthogonal transformation theorems.

**Theorem 2.1.** [6, Schur’s Theorem] Given $A \in \mathbb{R}^{2 \times 2}$ with two real eigenvalues $\lambda_1$ and $\lambda_2$, there is an orthogonal matrix $Q_\theta$ with $\theta \in [0, \pi)$ such that
\[
Q_\theta^\top AQ_\theta = \begin{bmatrix}
\lambda_1 & \alpha \\
0 & \lambda_2
\end{bmatrix} \in \mathbb{R}^{2 \times 2}.
\]

**Theorem 2.2.** Given $A \in \mathbb{R}^{2 \times 2}$, there is $\theta \in [0, \pi/2)$ such that
\[
Q_\theta^\top AQ_\theta = \begin{bmatrix}
\tau & \alpha \\
\beta & \tau
\end{bmatrix} \in \mathbb{R}^{2 \times 2},
\]  \hspace{1cm} (2.2)
where $\tau = \text{tr}(A)/2$ and $Q_\theta$ is defined in (2.1). In addition, there exists $\theta \in [0, \pi)$ such that we can require $|\alpha| \geq |\beta|$ in (2.2).

**Proof.** Let $A = \begin{bmatrix}
a_{11} & a_{12} \\
a_{21} & a_{22}
\end{bmatrix}$ and $Q_\theta = \begin{bmatrix}
\cos \theta & -\sin \theta \\
\sin \theta & \cos \theta
\end{bmatrix}$ be orthogonal. Denote $\hat{A} \equiv [\hat{a}_{ij}(\theta)] = Q_\theta^\top AQ_\theta$, and let $f(\theta) = \hat{a}_{11}(\theta) - \hat{a}_{22}(\theta)$. Then
\[
f(\theta) = (a_{11} - a_{22}) \cos^2 \theta + (a_{22} - a_{11}) \sin^2 \theta + 2(a_{12} + a_{21}) \cos \theta \sin \theta
\]
is a continuation function. Since $f(0)$ and $f(\pi/2)$ have opposite sign, there exists a $\theta \in (0, \pi/2)$ such that $f(\theta) = 0$. Hence, we have $\hat{a}_{11}(\theta) = \hat{a}_{22}(\theta) = \tau$ and (2.2) holds.

Suppose that $|\alpha| < |\beta|$ in (2.2) for some $\theta \in [0, \pi/2)$. Replacing $\theta \in [0, \pi/2)$ by $\hat{\theta} = \theta + \pi/2 \in [0, \pi)$ in (2.2), we have
\[
Q_{\hat{\theta}}^\top AQ_{\hat{\theta}} = \begin{bmatrix}
0 & 1 \\
-1 & 0
\end{bmatrix} \begin{bmatrix}
\tau & \alpha \\
\beta & \tau
\end{bmatrix} \begin{bmatrix}
0 & -1 \\
1 & 0
\end{bmatrix} = \begin{bmatrix}
\tau & -\beta \\
-\alpha & \tau
\end{bmatrix} \equiv \begin{bmatrix}
\tau & \hat{\alpha} \\
\hat{\beta} & \tau
\end{bmatrix}.
\]
Then $|\hat{\alpha}| = |\beta| > |\alpha| = |\hat{\beta}|$. This completes the proof. \hfill \Box

Let
\[
\mathcal{J}^{2 \times 2} = \{ A \in \mathbb{R}^{2 \times 2} \mid \text{A has one real eigenvalue with multiplicity 2} \}.
\]  \hspace{1cm} (2.3)
For each $A \in \mathcal{J}^{2 \times 2}$, the eigenvalue of $A$ is $\lambda = \text{tr}(A)/2$ and $(A - \lambda I)^2 = 0$. It follows from Theorem 2.2 that there exists an orthogonal matrix $Q_\theta$ such that
\[
Q_\theta^\top AQ_\theta = \begin{bmatrix}
\lambda & \alpha \\
\beta & \lambda
\end{bmatrix}.
\]
Using the fact that $(A - \lambda I)^2 = 0$, we can require $\beta = 0$, i.e.,
\[
Q^\top AQ = \begin{bmatrix}
\lambda & \alpha \\
0 & \lambda
\end{bmatrix}.
\]
This is the result of Theorem 2.1. The following lemma is useful to investigate the nearest stable matrix problem (1.2).

**Lemma 2.3.** Suppose that $A \in \mathbb{R}^{2 \times 2}$ has two real eigenvalues. If the eigenvalues of $\tilde{X} \in \mathbb{R}^{2 \times 2}$ are non-real, then
\[
\min_{X \in \mathcal{J}^{2 \times 2}} \| A - X \|_F < \| A - \tilde{X} \|_F.
\]
Assume $A(t) = A + tP$ for $t \in [0,1]$. Then $A(0) = A$ and $A(1) = \hat{A}$. Let $\lambda_1(t)$ and $\lambda_2(t)$ denote the eigenvalues of $A(t)$. Since $A(t)$ is a real $2 \times 2$ matrix, we have either $\lambda_1(t)$ and $\lambda_2(t)$ are real or $\lambda_1(t) = \bar{\lambda}_2(t)$. Now, let

$$E = \{ t \in [0,1] | \text{Im} \lambda_1(t) = -\text{Im} \lambda_2(t) \neq 0 \}.$$  

Let $t_* = \inf E$ and write $\lambda_{1,2}(t) = a(t) \pm b(t)i$ for $t \in E$. Since $\text{Im} \lambda_{1,2}(t)$ is continuous, this implies $b(t_*) = 0$. In addition, let $\{t_k\} \subset E$ such that $t_k \to t_*$. By the continuity of $\text{Re} \lambda_{1,2}(t)$, we have

$$\text{Re} \lambda_{1,2}(t_*) = \lim_{k \to \infty} a(t_k) \pm b(t_k)i = \lim_{k \to \infty} a(t_k).$$

This means that $A(t_*) \in \mathbb{J}^{2 \times 2}$. Hence, $\min_{X \in \mathbb{J}^{2 \times 2}} \|A - X\|_F \leq \|t_* P\|_F < \|P\|_F = \|A - \hat{X}\|_F.$

3. The nearest stable matrix problem. In this section, we first show that if $A \in \mathbb{R}^{2 \times 2}$ has real eigenvalues (or complex conjugate eigenvalues), then so is the minimizer $X$ of (1.3). Given a matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathbb{R}^{2 \times 2}$. Let $X \in cl(\mathbb{S}^{2 \times 2})$. It follows from Theorem 2.2 that there exists an orthogonal matrix $Q_\theta$ such that such that

$$Q_\theta^T X Q_\theta = \begin{bmatrix} \tau_X & \gamma \\ \delta & \tau_X \end{bmatrix} \in \mathbb{R}^{2 \times 2},$$

where $\tau_X = \text{tr}(X)/2 \leq 0$. Let $\tau = \text{tr}(A)/2 = (a + d)/2$. we have

$$\|A - X\|_F^2 = \|Q_\theta^T A Q_\theta - \begin{bmatrix} \tau_X & \gamma \\ \delta & \tau_X \end{bmatrix}\|_F^2 = \|(\tau - \tau_X)I + Q_\theta^T (A - \tau I) Q_\theta - \begin{bmatrix} 0 & \gamma \\ \delta & 0 \end{bmatrix}\|_F^2. \quad (3.1)$$

Denote $A_\tau = A - \tau I$. Then $\text{tr}(A_\tau) = 0$. Using the fact that $\langle A, B \rangle_F \equiv \text{tr}(B^T A)$ is an inner product and $\|A\|_F^2 = \langle A, A \rangle_F$, if $B \in \mathbb{R}^{2 \times 2}$ with $\text{tr}(B) = 0$ then $\langle I, B \rangle_F = \text{tr}(B^T) = 0$, i.e., $I$ is orthogonal to $B$ in the sense of the inner product $\langle \cdot, \cdot \rangle_F$. Since

$$\text{tr} \left( Q_\theta^T A_\tau Q_\theta - \begin{bmatrix} 0 & \gamma \\ \delta & 0 \end{bmatrix} \right) = 0,$$

it follows from (3.1) that

$$\|A - X\|_F^2 = \|(\tau - \tau_X)I\|_F^2 + \|Q_\theta^T A_\tau Q_\theta - \begin{bmatrix} 0 & \gamma \\ \delta & 0 \end{bmatrix}\|_F^2. \quad (3.2)$$

First, we consider the matrix $A$ is unstable and has complex conjugate eigenvalues. In this case, we have $\tau = \text{tr}(A)/2 > 0$. The following theorem gives the form of the minimizer of (1.3). The minimizer has complex conjugate eigenvalues.

**Theorem 3.1.** Let $A \in \mathbb{R}^{2 \times 2}$ with $\tau = \text{tr}(A)/2 > 0$ have complex conjugate eigenvalues. Then (1.3) has a unique minimizer $X_* = A_\tau \equiv A - \tau I$ and the optimal value of (1.3) is $\sqrt{2}\tau$.

**Proof.** Assume $X_*$ is the minimizer of (1.3). Since $X_* \in cl(\mathbb{S}^{2 \times 2})$, $\tau_{X_*} = \text{tr}(X_*)/2 \leq 0$. From (3.2), we conclude that

$$\|A - X_*\|_F^2 \geq \|\tau I\|_F^2 = 2\tau^2. \quad (3.3)$$
Since the eigenvalues of $A$ are complex conjugate, so is $A_\tau = A - \tau I$. Furthermore, the real part of eigenvalues of $A_\tau$ is 0, hence, $A_\tau \in cl(S^{2 \times 2})$. Since $\|A - A_\tau\|^2_F = 2\tau^2$, it follows from (3.3) that $X = A_\tau$ is the unique minimizer of (1.3). \( \square \)

Next, we consider the unstable matrix $A$ has real eigenvalues and show that the minimizer of (1.3) also has real eigenvalues. To prove this, we consider the optimization problem:

$$\min_{X \in \mathbb{J}^{2 \times 2}} \|A_\tau - X\|^2_F, \quad (3.4)$$

where $\mathbb{J}^{2 \times 2}$ is defined in (2.3), $A_\tau = A - \tau I$, and $\tau = \text{tr}(A)/2$. Then $\text{tr}(A_\tau) = 0$ and $A_\tau$ has real eigenvalues. Using the fact that $\| \cdot \|_F$ is orthogonal invariant, it follows from Theorem 2.2 that we can first assume the matrix $A_\tau$ in (3.4) has the form

$$A_\tau = \begin{bmatrix} 0 & \alpha \\ \beta & 0 \end{bmatrix} \quad \text{with } |\alpha| \geq |\beta| \text{ and } \alpha \beta \geq 0. \quad (3.5)$$

The following lemma gives the minimizer of (3.4).

**Lemma 3.2.** Let $A_\tau \in \mathbb{R}^{2 \times 2}$ have the form in (3.5). Then $\tilde{X} = \begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix} \in \mathbb{J}^{2 \times 2}$ is a minimizer of (3.4).

**Proof.** Let $X \in \mathbb{J}^{2 \times 2}$. Using the definition of $\mathbb{J}^{2 \times 2}$ and Theorem 2.1, there exists $\theta$ such that $X = Q_\theta^T \begin{bmatrix} \lambda & \gamma \\ 0 & \lambda \end{bmatrix} Q_\theta$. Then

$$\|A_\tau - X\|^2_F = \|\lambda I\|^2_F + \|A_\tau - Q_\theta^T \begin{bmatrix} 0 & \gamma \\ 0 & 0 \end{bmatrix} Q_\theta\|^2_F \geq \sigma^2_{\min}(A_\tau), \quad (3.6)$$

where $\sigma_{\min}(A_\tau) = |\beta|$ is the smallest singular value of $A_\tau$. Here, the last inequality of (3.6) holds because $\|\lambda I\|^2_F \geq 0$ and $Q_\theta^T \begin{bmatrix} 0 & \gamma \\ 0 & 0 \end{bmatrix} Q_\theta$ is a rank-1 matrix. Let $\tilde{X} = \begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix} \in \mathbb{J}^{2 \times 2}$. Then $\|A_\tau - \tilde{X}\|^2_F = |\beta|^2 = \sigma^2_{\min}(A_\tau)$, i.e., the equality of (3.6) holds. This completes the proof. \( \square \)

Now we are ready to prove that if the unstable matrix $A$ has real eigenvalues, then so is the minimizer $X$ of (1.3).

**Theorem 3.3.** Let $A \in \mathbb{R}^{2 \times 2}$ be an unstable matrix and have real eigenvalues. Then the minimizer of (1.3) has real eigenvalues.

**Proof.** We prove it by contradiction. Suppose that the minimizer $X_* \in cl(S^{2 \times 2})$ has two complex conjugate eigenvalues $\lambda \pm i\varepsilon$, where $\lambda \leq 0$ and $\varepsilon > 0$. First, we claim that $\lambda = 0$. Suppose that $\lambda < 0$. Then there exists $\rho > 0$ such that

$$B_\rho(X_*) = \{ B \in \mathbb{R}^{2 \times 2} \mid \|B - X_*\|_F < \rho \} \subseteq S^{2 \times 2}.$$

Because $A \notin S^{2 \times 2}$, this implies that the minimizer $X_*$ should be in the boundary of the $cl(S^{2 \times 2})$. This contradicts to the assumption $\lambda < 0$. Hence, $\lambda = 0$ and $\text{tr}(X_*) = 0$. Let $\tau = \text{tr}(A)/2$ and $A_\tau = A - \tau I$. Then we have

$$\|A - X_*\|^2_F = \|\tau I\|^2_F + \|A_\tau - X_*\|^2_F.$$
Since $X_*$ is a minimizer of (1.3), we obtain

$$X_* = \arg \min_{X \in \mathcal{S}(S^{2 \times 2})} \| A_r - X \|_F.$$  

From Lemma 2.3, we have $\min_{X \in \mathcal{S}(S^{2 \times 2})} \| A_r - X \|_F < \| A_r - X_* \|_F$. Let $\tilde{X} \in \mathcal{S}(S^{2 \times 2})$ be the minimizer of $\min_{X \in \mathcal{S}(S^{2 \times 2})} \| A_r - X \|_F$. We have

$$\| A - \tilde{X} \|_F^2 = \| \tau I \|_F^2 + \| A_r - \tilde{X} \|_F^2 < \| \tau I \|_F^2 + \| A_r - X_* \|_F^2 = \| A - X_* \|_F^2.$$  

From Lemma 3.2, we obtain $\tilde{X} \in \mathcal{S}(S^{2 \times 2})$. This is a contradiction because $X_*$ is a minimizer of (1.3).

Theorem 3.3 shows that the nearest stable matrix $X_*$ of the unstable matrix $A$ has real eigenvalues when the eigenvalues of $A$ are real. In the following, we will give an explicit formula for $X_*$. Let $A \in \mathbb{R}^{2 \times 2}$ be unstable and have real eigenvalues. It follows from Theorem 3.3 that the nearest stable matrix $X_*$ has real eigenvalues. Hence, there exists an orthogonal matrix $Q_{\theta_*}$ such that $Q_{\theta_*} X Q_{\theta_*}$ is upper triangular. In order to investigate the minimizer, we define a function $\Delta : \mathbb{R}^{2 \times 2} \to \mathbb{R}^{2 \times 2}$ by

$$\Delta \left( \begin{bmatrix} a & b \\ c & d \end{bmatrix} \right) = \begin{bmatrix} \chi(a) & b \\ 0 & \chi(d) \end{bmatrix},$$

where $\chi(x) = x$ if $x \leq 0$ and $\chi(x) = 0$ if $x > 0$. It is easily seen that $\Delta(A)$ is the upper triangular matrix which is closet to $A$ in $\mathcal{S}(S^{2 \times 2})$.

Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathbb{R}^{2 \times 2}$. Suppose that $Q_{\theta}$ in (2.1) is an orthogonal matrix. Then

$$Q_{\theta}^T A Q_{\theta} = \begin{bmatrix} (a + d) - (d - a) \cos 2 \theta + (b + c) \sin 2 \theta & (b - c) - (b + c) \cos 2 \theta + (d - a) \sin 2 \theta \\ (b + c) - (b - d) \cos 2 \theta + (d + a) \sin 2 \theta & (a - d) - (a + c) \cos 2 \theta + (b - c) \sin 2 \theta \end{bmatrix}$$

and

$$R = \sqrt{(a - d)^2 + (b + c)^2}, \quad u = 2 \theta - \phi, \quad (3.8b)$$

where $\phi \in [0, 2\pi)$ satisfies $R \sin \phi = d - a, R \cos \phi = b + c$. By Theorem 3.3, we assume that eigenvalues of $X$ are real and nonpositive. It follows from Theorem 2.1 that there are a rotation matrix $Q_{\theta}$ and an upper triangular matrix $U$ such that $Q_{\theta}^T A Q_{\theta} = U$. Then

$$\| A - X \|_F = \| Q_{\theta}^T (A - X) Q_{\theta} \|_F = \| Q_{\theta}^T A Q_{\theta} - U \|_F,$$

and hence, (1.3) is equivalent

$$\min_{Q_{\theta}, U} \| Q_{\theta}^T A Q_{\theta} - U \|_F = \min_{Q_{\theta}} \| Q_{\theta}^T A Q_{\theta} - U \|_F.$$  

Using the definition of $\Delta(\cdot)$ in (3.7), we let

$$\Delta_{\theta} = \Delta(Q_{\theta}^T A Q_{\theta}) \in \mathcal{S}(S^{2 \times 2}) \text{ for } \theta \in \mathbb{R}. \quad (3.9)$$

Then $\Delta_{\theta}$ is a continuous periodic function. For each rotation matrix $Q_{\theta}$, the minimizer of $\min_U \| Q_{\theta}^T A Q_{\theta} - U \|_F$ attains at $U = \Delta_{\theta}$. Since $\| Q_{\theta}^T A Q_{\theta} - \Delta_{\theta} \|_F^2 = \| A \|_F^2 - \| \Delta_{\theta} \|_F^2$, let $\theta_*$ be the minimizer of

$$\min_{\theta \in [0, \pi]} \| Q_{\theta}^T A Q_{\theta} - \Delta_{\theta} \|_F^2 = \| A \|_F^2 - \max_{\theta \in [0, \pi]} \| \Delta_{\theta} \|_F^2. \quad (3.10)$$

Then we have $X_* = Q_{\theta_*}^T \Delta_{\theta_*} Q_{\theta_*} \in \mathcal{S}(S^{2 \times 2})$ is a minimizer of (1.3).
Remark 3.1. If \( R = 0 \), then \( a = d \) and \( b = -c \). The matrix \( A \) has complex conjugate eigenvalues \( a \pm ib \). The nearest stable matrix \( X_* \) has been shown in Theorem 3.1. Hence, we require \( R > 0 \) in the following.

We study the problem \((1.3)\) for two different cases \( \text{tr}(A) \leq 0 \) and \( \text{tr}(A) > 0 \). In the following theorem, we show the results of the optimization problem \((1.3)\) when \( \text{tr}(A) \leq 0 \).

**Theorem 3.4.** Let \( A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathbb{R}^{2 \times 2} \) with \( \text{tr}(A) \leq 0 \) be an unstable matrix. Let \( R = \sqrt{(a - d)^2 + (b + c)^2} > 0, \phi \in [0, 2\pi) \) in \((3.8b)\) and \( \hat{R} = \sqrt{(a + d)^2 + (b - c)^2} \).

(i) If \( \hat{R} > 0 \), then \((1.3)\) has a unique minimizer

\[
X_* = \frac{R + \hat{R}}{4} \begin{bmatrix} -\sin \omega - \sin \phi & \cos \omega + \cos \phi \\ -\cos \omega + \cos \phi & -\sin \omega + \sin \phi \end{bmatrix},
\]

where \( \omega \in [0, \pi) \) satisfies \( \sin \omega = -(a + d)/\hat{R}, \cos \omega = (b - c)/\hat{R} \). In addition, the optimal value of \((1.3)\) is \((R - \hat{R})/2\).

(ii) If \( \hat{R} = 0 \), the \((1.3)\) has infinitely many minimizers. The solutions are

\[
X_* = \frac{R}{4} \begin{bmatrix} -\sin u - \sin \phi & \cos u + \cos \phi \\ -\cos u + \cos \phi & -\sin u + \sin \phi \end{bmatrix}
\]

for \( u \in [0, \pi) \)

and the optimal value of \((1.3)\) is \( R/2 \).

**Proof.** Since \( A \) is unstable and \( \text{tr}(A) \leq 0 \), we obtain that the eigenvalues of \( A \) are real. Suppose that \( X_* \in \text{cl}(S^{2 \times 2}) \) is a minimizer of \((1.3)\). It follows from Theorem 3.3 that \( X_* \) has real eigenvalues. From Theorem 2.1, there exists \( \theta_* \in [0, \pi) \) such that \( \Delta_* = Q_{\theta_*}^T X_* Q_{\theta_*} \) is upper triangular, and the diagonal entries of \( \Delta_* \) are nonpositive, where \( Q_{\theta_*} \) in \((2.1)\) is orthogonal. Next, we will find the angular \( \theta_* \).

The angular \( \theta_* \in [0, \pi) \) is the minimizer of \((3.10)\) and \( \Delta_{\theta_*} \) in \((3.9)\) is the Schur form of \( X_* \). Since \( A \) is unstable, the minimizer \( X_* \) can not stay in the interior of \( S^{2 \times 2} \), i.e., \( \Delta_{\theta_*} \) has at least one zero diagonal entry. From \((3.7)\) and \((3.8a)\), we obtain that

\[
\theta_* \in \{ \theta \in [0, \pi) \mid |a + d| \leq R|\sin u| \text{ and } u = 2\theta - \phi \},
\]

where \( \phi \) is given in \((3.8b)\). Let \( u = 2\theta - \phi \). Since \( \text{tr}(A) = a + d \leq 0 \), \( \Delta_{\theta} \) can have the following two possibilities:

\[
\Delta_{\theta} = \begin{cases} \frac{1}{2} \begin{bmatrix} 0 & (b - c) + R \cos u \\ (a + d) - R \sin u \\ (a + d) + R \sin u & (b - c) + R \cos u \\ 0 & 0 \end{bmatrix}, & \text{if } \sin u \geq \frac{-(a + d)}{R}, \\ \frac{1}{2} \begin{bmatrix} 0 & (b - c) \\ (a + d) - R \sin u \\ (a + d) + R \sin u & (b - c) + R \cos u \\ 0 & 0 \end{bmatrix}, & \text{if } \sin u \leq \frac{a + d}{R}. \end{cases}
\]

From \((3.10)\), the angular \( \theta_* \) is the maximizer of \( \max_{\theta \in [0, \pi)} \|\Delta_{\theta}\|_F \).

For assertion (i), suppose that \( \hat{R} = \sqrt{(a + d)^2 + (b - c)^2} > 0 \). Since \( \text{tr}(A) = a + d \leq 0 \) and \( A \) is unstable, we obtain that \( \det(A) < 0 \) and there exists \( \omega \in [0, \pi) \) such that

\[
\begin{bmatrix} a + d \\ b - c \end{bmatrix} = \hat{R} \begin{bmatrix} -\sin \omega \\ \cos \omega \end{bmatrix}.
\]
Using $R^2 - \hat{R}^2 = 4(bc - ad) = -4\det(A) > 0$ leads to $R > \hat{R}$. Let $u_* = \omega$ (or $u_* = -\omega$), then $u_*$ satisfies
\[
\sin u_* = \sin \omega = \frac{-(a + d)}{R} > \frac{-(a + d)}{\hat{R}} \quad \left( \text{or } \sin \omega = \frac{a + d}{R} < \frac{a + d}{\hat{R}} \right).
\]
From (3.10), (3.13), and (3.14), and $u = 2\theta - \phi$, we obtain that $\theta_{*1} = (\omega + \phi)/2$ and $\theta_{*2} = (-\omega + \phi)/2$ are two maximizers of $\max_{\theta \in [0, \pi)} \|\Delta_\theta\|_F$. Hence,
\[
X_{*1} = Q_{\theta_{*1}} \Delta_{\theta_{*1}} Q_{\theta_{*1}}^T \quad \text{and} \quad X_{*2} = Q_{\theta_{*2}} \Delta_{\theta_{*2}} Q_{\theta_{*2}}^T.
\]
are two minimizers of (1.3), where
\[
\Delta_{\theta_{*1}} = \frac{R + \hat{R}}{2} \begin{bmatrix} 0 & \cos \omega \\ \sin \omega & 0 \end{bmatrix} \quad \text{and} \quad \Delta_{\theta_{*2}} = \frac{R + \hat{R}}{2} \begin{bmatrix} -\sin \omega & \cos \omega \\ 0 & 0 \end{bmatrix}. \tag{3.15b}
\]

Now, we claim that $X_{*1} = X_{*2}$. From (3.15) and using $\theta_{*1} - \theta_{*2} = \omega$, we have
\[
X_{*2} = Q_{\theta_{*2}} X_{*1} Q_{\theta_{*1}}^T = Q_{\theta_{*2}} (Q_{\theta_{*1}} \Delta_{\theta_{*1}} Q_{\theta_{*1}}^T) Q_{\theta_{*2}}^T
\]
\[
= \frac{R + \hat{R}}{2} Q_{\theta_{*1}} \begin{bmatrix} \cos \omega & \sin \omega \\ -\sin \omega & \cos \omega \end{bmatrix} \begin{bmatrix} -\sin \omega & \cos \omega \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \cos \omega & -\sin \omega \\ \sin \omega & \cos \omega \end{bmatrix} Q_{\theta_{*2}}^T
\]
\[
= \frac{R + \hat{R}}{2} Q_{\theta_{*1}} \begin{bmatrix} 0 & \cos \omega \\ \sin \omega & 0 \end{bmatrix} Q_{\theta_{*2}}^T = X_{*1}.
\]
This shows that (1.3) has a unique solution $X_* = X_{*1} = X_{*2}$. Using product-to-sum identities of trigonometric functions, we obtain (3.11). In addition, it follows from (3.10) and (3.15b) that the square of the optimal value is
\[
\|A - X_*\|^2_F = (a^2 + b^2 + c^2 + d^2) - \frac{(R + \hat{R})^2}{4} = \frac{(R - \hat{R})^2}{4}.
\]
Hence, assertion (i) holds.

For assertion (ii), if $\hat{R} = \sqrt{(a + d)^2 + (b - c)^2} = 0$, then $A = \begin{bmatrix} a & b \\ b & -a \end{bmatrix}$ and
\[
\Delta_\theta = \begin{cases} \frac{R}{2} \begin{bmatrix} \sin u & \cos u \\ 0 & 0 \end{bmatrix}, & \text{if } u \in [\pi, 2\pi), \\ \frac{R}{2} \begin{bmatrix} 0 & \cos u \\ \sin u & 0 \end{bmatrix}, & \text{if } u \in [0, \pi), \end{cases} \tag{3.16}
\]
where $u = 2\theta - \phi \in [0, 2\pi)$. It is easily seen that $\|\Delta_\theta\|_F = R/2$ for each $\theta \in [0, \pi)$. Let
\[
X_*(u) = Q_{\frac{u + \pi}{2}} \Delta_{\frac{u + \pi}{2}} Q_{\frac{u + \pi}{2}}^T, \quad \text{for } u \in [0, 2\pi).
\]
Then $X_*(u)$ are minimizers of (1.3). For each $u \in [0, \pi)$, let $\tilde{u} = 2\pi - u \in [\pi, 2\pi)$, then
\[
X_*(u) = Q_{\frac{u + \pi}{2}} \Delta_{\frac{u + \pi}{2}} Q_{\frac{u + \pi}{2}}^T = Q_{\frac{u + \pi}{2}} (Q_{u - \pi} \Delta_{u - \pi} Q_{u - \pi}) Q_{\frac{u + \pi}{2}}^T
\]
\[
= \frac{R + \hat{R}}{2} Q_{\frac{u + \pi}{2}} \begin{bmatrix} -\cos u & \sin \tilde{u} \\ \sin u & -\cos \tilde{u} \end{bmatrix} \begin{bmatrix} \sin \tilde{u} & \cos \tilde{u} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} -\cos u & \sin \tilde{u} \\ \sin u & -\cos \tilde{u} \end{bmatrix} Q_{\frac{u + \pi}{2}}^T
\]
\[
= \frac{R + \hat{R}}{2} Q_{\frac{u + \pi}{2}} \begin{bmatrix} \cos u & \sin \tilde{u} \\ -\sin u & \cos \tilde{u} \end{bmatrix} \begin{bmatrix} -\sin \tilde{u} & \cos \tilde{u} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \cos u & -\sin \tilde{u} \\ \sin u & \cos \tilde{u} \end{bmatrix} Q_{\frac{u + \pi}{2}}^T
\]
\[
= \frac{R + \hat{R}}{2} Q_{\frac{u + \pi}{2}} \begin{bmatrix} 0 & \cos \tilde{u} \\ \sin \tilde{u} & 0 \end{bmatrix} Q_{\frac{u + \pi}{2}}^T = X_*(u).
\]

Similarly, using (3.16) and product-to-sum identities of trigonometric functions, we obtain (3.12) and the optimal value is $R/2$. This completes the proof. $\Box$
In the following, we show the minimizer of (1.3) when the unstable $A$ has $\text{tr}(A) > 0$.

**Theorem 3.5.** Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathbb{R}^{2 \times 2}$ with $\text{tr}(A) > 0$ have real eigenvalues. Let $R = \sqrt{(a-d)^2 + (b+c)^2} > 0$ and $\phi \in [0,2\pi)$ in (3.8b).

(i) If $b \neq c$, then (1.3) has a unique minimizer

$$X_\ast = \frac{R + |b-c|}{4} \begin{bmatrix} -\sin \phi & \frac{b-c}{|b-c|} + \cos \phi \\ \frac{-b+c}{|b-c|} + \cos \phi & \sin \phi \end{bmatrix}.$$  

(3.17)

In addition, the optimal value of (1.3) is $\frac{1}{2} \sqrt{4||A||_F^2 - ((b-c) + R)^2}$.

(ii) If $b = c$, then (1.3) has two minimizers

$$X_{\ast \pm} = \frac{R}{4} \begin{bmatrix} -\sin \phi & \pm 1 + \cos \phi \\ \mp 1 + \cos \phi & \sin \phi \end{bmatrix}.$$  

(3.18)

In addition, the optimal value of (1.3) is $\frac{1}{2} \sqrt{4||A||_F^2 - R^2}$.

**Proof.** Suppose that $A$ with $\text{tr}(A) = a + d > 0$ has real eigenvalues. Let $X_\ast \in \text{cl}(S^{2 \times 2})$ is a minimizer of (1.3). It follows from Theorem 3.3 that $X_\ast$ has real eigenvalues. Then there exists an angular $\theta_\ast \in [0,\pi)$ such that $\Delta_{\theta_\ast}$ in (3.9) is the Schur form of $X_\ast$ and $\theta_\ast$ is the minimizer of (3.10). Next, we will find the angular $\theta_\ast$.

(i) Suppose that $b \neq c$. First, we consider $b > c$. For each $\theta \in [0,\pi)$, let $u = 2\theta - \phi$. Using $a + d > 0$ and (3.8a), we know that

$$\Delta_\theta = \begin{cases}
\frac{1}{4} \begin{bmatrix} (a+d) + R\sin u & (b-c) + R\cos u \\
0 & 0 \end{bmatrix}, & \text{if } \sin u \leq \frac{-(a+d)}{R}, \\
\frac{1}{4} \begin{bmatrix} 0 & (b-c) + R\cos u \\
0 & (a+d) - R\sin u \end{bmatrix}, & \text{if } \sin u \geq \frac{a+d}{R}, \\
\frac{1}{4} \begin{bmatrix} 0 & (b-c) + R\cos u \\
0 & 0 \end{bmatrix}, & \text{if } \frac{-(a+d)}{R} \leq \sin u \leq \frac{a+d}{R}.
\end{cases}$$

For the case $\frac{-(a+d)}{R} \leq \sin u \leq \frac{a+d}{R}$, we have the maximal value of $||\Delta_\theta||_F$ is $(|b-c| + R)/2$ when $u = u_\ast = 0$. For the cases $\sin u \leq \frac{-(a+d)}{R} < 0$ or $\sin u \geq \frac{a+d}{R} > 0$, since $a + d > 0$ and $(a+d) - R|\sin u| < 0$, then

$$||\Delta_\theta||_F = \frac{1}{2} \left\| \begin{bmatrix} (b-c) + R\cos u \\
(a+d) - R|\sin u| \end{bmatrix} \right\| < \frac{1}{2} \left\| \begin{bmatrix} (b-c) + R\cos u \\
-R|\sin u| \end{bmatrix} \right\| \leq \frac{|b-c| + R}{2}.$$  

(3.19)

Hence, the maximizer of $\max_{\theta \in [0,\pi]} ||\Delta_\theta||_F$ attains at the case $\frac{-(a+d)}{R} \leq \sin u \leq \frac{a+d}{R}$ and $\theta_\ast = (u_\ast + \phi)/2 = \phi/2$ is the unique maximizer. In this case,

$$X_\ast = Q_{\theta_\ast} \Delta_{\theta_\ast} Q_{\theta_\ast}^T = \frac{|b-c| + R}{2} \begin{bmatrix} \cos \frac{\phi}{2} & -\sin \frac{\phi}{2} \\
\sin \frac{\phi}{2} & \cos \frac{\phi}{2} \end{bmatrix} \begin{bmatrix} 0 & 1 \\
0 & 0 \end{bmatrix} \begin{bmatrix} \cos \frac{\phi}{2} & \sin \frac{\phi}{2} \\
-\sin \frac{\phi}{2} & \cos \frac{\phi}{2} \end{bmatrix}$$

$$= \frac{R + |b-c|}{4} \begin{bmatrix} -\sin \phi & 1 + \cos \phi \\
-1 + \cos \phi & \sin \phi \end{bmatrix}.$$  

is the unique minimizer of (1.3). The matrix $X_\ast$ is the same as (3.17) for $b > c$.

Next, we consider the case $b < c$ in $A$. Let $\hat{A} \equiv \begin{bmatrix} \hat{a} & \hat{b} \\ \hat{c} & \hat{d} \end{bmatrix} = \begin{bmatrix} a & c \\ b & d \end{bmatrix} = A^T$.

From (3.8b), we have $\hat{R} \equiv \sqrt{(\hat{a} - \hat{d})^2 + (\hat{b} + \hat{c})^2} = R$ and $\phi = \phi$. Since $b = c > b = \ldots$
Hence, for the case $b \leq c$ in $A$, the minimizer of (1.3) is
\[
X_\ast = \begin{cases}
\frac{R + |b - c|}{4} & \begin{bmatrix} -\sin \phi & 1 + \cos \phi \\ -1 + \cos \phi & \sin \phi \end{bmatrix} \\
\frac{R + |b - c|}{4} & \begin{bmatrix} -\sin \phi & 1 + \cos \phi \\ -1 + \cos \phi & \sin \phi \end{bmatrix}
\end{cases}
\]
Hence, for the case $b < c$ in $A$, the minimizer of (1.3) is
\[
X_\ast = \hat{X}_\ast^\top = \frac{R + |b - c|}{4} \begin{bmatrix} -\sin \phi & -1 + \cos \phi \\ 1 + \cos \phi & \sin \phi \end{bmatrix}.
\]
The matrix $X_\ast$ coincides with the formula (3.17) with $b < c$.

In addition, it follows from (3.10) and (3.19) that the optimal value of (1.3) is
\[
\frac{1}{2} \sqrt{4|A| - (|b - c| + R)^2}.
\]

(ii) Assume that $b = c$. A similar argument to (i) shows that the maximal value of $\|\Delta \phi\|_F$ is $R/2$ when $u = u_1 = 0$ or $u = u_2 = \pi$. Let $\theta_\ast_1 = \phi/2$ and $\theta_\ast_2 = (\pi + \phi)/2$. Then $\theta_\ast_1$ and $\theta_\ast_2$ are two minimizers of max$_{\theta \in [0, \pi]} \|\Delta \phi\|_F$. Since
\[
X_{\ast 1} = Q_{\theta_\ast 1} \Delta_{\theta_\ast 1} Q_{\theta_\ast 1}^\top = \frac{R}{2} Q_{\theta_\ast 1} \begin{bmatrix} \sin \phi & 1 + \cos \phi \\ -1 + \cos \phi & \sin \phi \end{bmatrix},
\]
\[
X_{\ast 2} = Q_{\theta_\ast 2} \Delta_{\theta_\ast 2} Q_{\theta_\ast 2}^\top = \frac{R}{2} Q_{\theta_\ast 2} \begin{bmatrix} -\sin \phi & -1 + \cos \phi \\ 1 + \cos \phi & \sin \phi \end{bmatrix},
\]
are different, (1.3) has two minimizers in (3.18). In addition, using (3.10) again, the optimal value of (1.3) is
\[
\frac{1}{2} \sqrt{4|A| - R^2}.
\]

4. Conclusion. In this paper, we provide an explicit form of the nearest stable real matrix to a $2 \times 2$ real matrix in the sense of Frobenius norm. Recently, a similar result is obtained in [8], in which the optimizer $X_\ast$ is characterized in a set of five matrices. It is also of our interest to study the nearest stable complex matrix. The difficulty arises in two parameters of $2 \times 2$ unitary matrices. For a discrete-time system, the explicit form of the nearest stable $2 \times 2$ matrix problem is also of interest. It is also of our interest to investigate the higher degree case by using current results.

REFERENCES

[1] N. Choudhary, N. Gillis and P. Sharma, On approximating the nearest $\Omega$-stable matrix, Numer Alg. Appl., 27 (2020), e2282, 13pp.
[2] N. Gillis, V. Mehrmann and P. Sharma, Computing the nearest stable matrix pairs, Numer. Linear Alg. Appl., 25 (2018), e2153, 16pp.
[3] N. Gillis, M. Karow and P. Sharma, Approximating the nearest stable discrete-time system, Linear Alg. Appl., 573 (2019), 37–53.
[4] N. Gillis and P. Sharma, On computing the distance to stability for matrices using linear dissipative Hamiltonian systems, Automation, 85 (2017), 113–121.
[5] N. Higham, Matrix nearness problems and applications, Applications of Matrix Theory (Bradford, 1988), 1–27, Inst. Math. Appl. Conf. Ser. New Ser., 22, Oxford Univ. Press, New York, 1989.
[6] R. A. Horn and C. R. Johnson, Topics in Matrix Analysis, Cambridge University Press, 1991.
[7] V. Mehrmann and P. Van Dooren, Optimal Robustness of Port-Hamiltonian Systems, SIAM J. Matrix Anal. Appl., 41 (2020), 134–151.
[8] V. Noferini and F. Poloni, Nearest $\Omega$-stable matrix via Riemannian optimization, arXiv:2002.07052.
[9] F.-X. Orbandexivry, Y. Nesterov and P. Van Dooren, Nearest stable system using successive convex approximations, *Automatica*, 49 (2013), 1195–1203.

Received August 2020; revised October 2020.

E-mail address: yckuo@nuk.edu.tw
E-mail address: gn311311@gmail.com
E-mail address: a1064132@mail.nuk.edu.tw
E-mail address: sfshieh@ntnu.edu.tw