2014

A bound for condition numbers of matrices

Michael I. Gil
gilmi@bezeqint.net

Follow this and additional works at: http://repository.uwyo.edu/ela

Recommended Citation
DOI: https://doi.org/10.13001/1081-3810.1610
A BOUND FOR CONDITION NUMBERS OF MATRICES

MICHAEL I. GIL

Abstract. Let $A$ be a diagonalizable matrix; so there is an invertible matrix $T$ and a normal matrix $\hat{D}$, such that $T^{-1}AT = \hat{D}$. A sharp bound for the constant $\kappa_T = \|T\|\|T^{-1}\|$ is suggested. Some applications of the obtained bound are also discussed.

Key words. Matrix, Similarity, Condition number.

AMS subject classifications. 15A45, 15A42.

1. Introduction and preliminaries. Let $\mathbb{C}^n$ be the $n$-dimensional complex Euclidean space with a scalar product $(\cdot, \cdot)$, the Euclidean norm $\| \cdot \| = \sqrt{(\cdot, \cdot)}$ and the identity matrix $I$. For an $n \times n$ matrix $A$, $\sigma(A)$ denotes the spectrum of $A$, $\|A\|$ is the spectral norm; $A^*$ is the adjoint to $A$; $\|A\|_F = (\text{Trace } A^*A)^{1/2}$ is the Frobenius norm; $\lambda_k (k = 1, \ldots, n)$ are the eigenvalues of $A$. Everywhere below it is assumed that

$$\lambda_j \neq \lambda_m, \text{ whenever } j \neq m. \quad (1.1)$$

So, $A$ is a diagonalizable matrix: There is an invertible matrix $T$ and a normal matrix $\hat{D}$, such that

$$T^{-1}AT = \hat{D}. \quad (1.2)$$

The condition number $\kappa_T := \|T\|\|T^{-1}\|$ is very important for various applications, cf. [3,16]. That number is mainly numerically calculated.

In the present paper, we suggest a sharp bound for $\kappa_T$. Applications of the obtained bound to spectrum perturbations and matrix functions are also discussed.

The following quantity (departure from normality) plays an essential role hereafter:

$$g(A) := \left(\|A\|_F^2 - \sum_{k=1}^{n} |\lambda_k|^2 \right)^{1/2}.$$
\(g(A)\) enjoys the following properties:

\[
g^2(A) \leq 2\|A_I\|_F^2 \quad (A_I = (A - A^*)/2i) \quad \text{and} \quad g^2(A) \leq \|A\|_F^2 - |\text{Trace } A^2|,
\]

(1.3) cf. [10, Section 2.1]. If \(A\) is normal, then \(g(A) = 0\). Put

\[
\delta := \min_{j,k=1,\ldots,n, k \neq j} |\lambda_j - \lambda_k|.
\]

Corollary 3.6 from [11], under condition (1.1) gives us the inequality

\[
\kappa_T \leq n \sum_{k=0}^{n-1} \frac{g^k(A)2^k}{\delta^k k!}.
\]

(1.4)

That inequality is not sharp: If \(A\) is a normal matrix, then it gives \(\kappa_T \leq n\), but \(\kappa_T = 1\) in this case. Inequality (1.4) has been slightly improved in [12]. In this paper, we considerably refine (1.4) and the corresponding result from [12].

Put

\[
\tau(A) := \sum_{k=0}^{n-2} \frac{g^{k+1}(A)}{\sqrt{k!} \delta^{k+1}}
\]

and

\[
\gamma(A) := \left(1 + \frac{\tau(A)}{\sqrt{n-1}}\right)^{2(n-1)}.
\]

Now we are in a position to formulate the main result of this paper.

**Theorem 1.1.** Let condition (1.1) be fulfilled. Then there is an invertible matrix \(T\), such that (1.2) holds with

\[
\kappa_T \leq \gamma(A).
\]

(1.5)

The proof of this theorem is presented in the next two sections. Theorem 1.1 is sharp: If \(A\) is normal, then \(g(A) = 0\) and \(\gamma(A) = 1\). Thus, we obtain the equality \(\kappa_T = 1\). So Theorem 1.1 is obviously sharper than (1.4) at least for matrices “close” to normal ones. The proof of Theorem 1.1 is absolutely different from the proof of inequality (1.4) and the proof of the corresponding result from [12].

2. Auxiliary results. Let matrix $A$ have in $\mathbb{C}^n$ a chain of invariant projections $P_k$ ($k = 1, \ldots, m; m \leq n$):

$$0 \subset P_1\mathbb{C}^n \subset P_2\mathbb{C}^n \subset \cdots \subset P_m\mathbb{C}^n = \mathbb{C}^n \quad (2.1)$$

and

$$P_kAP_k = AP_k \quad (k = 1, \ldots, m). \quad (2.2)$$

Put $\Delta P_k = P_k - P_{k-1}$ ($P_0 = 0$), $A_k = \Delta P_k A \Delta P_k$,

$$Q_k = I - P_k, \quad B_k = Q_k A Q_k \quad \text{and} \quad C_k = \Delta P_k A Q_k.$$ 

It is assumed that the spectra $\sigma(A_k)$ of $A_k$ in $\Delta P_k\mathbb{C}^n$ satisfies the condition

$$\sigma(A_k) \cap \sigma(A_j) = \emptyset \quad (j \neq k). \quad (2.3)$$

**Lemma 2.1.** One has

$$\sigma(A) = \cup_{k=1}^{m} \sigma(A_k).$$

**Proof.** Put

$$S = \sum_{k=1}^{m} A_k \quad \text{and} \quad W = A - S.$$ 

Due to (2.2), we have $WP_k = P_{k-1}WP_k$. Hence,

$$W^m = W^mP_m = W^{m-1}P_{m-1}WP_m = W^{m-2}P_{m-2}WP_{m-1}WP_m = \cdots = P_0W^m = 0.$$ 

So, $W$ is nilpotent. Similarly, taking into account that

$$(S - \lambda I)^{-1} WP_k = P_{k-1}(S - \lambda I)^{-1} WP_k,$$

we prove that $((S - \lambda I)^{-1} W)^m = 0$ ($\lambda \notin \sigma(S)$). Thus,

$$(A - \lambda I)^{-1} = (S + W - \lambda I)^{-1} = (I + (S - \lambda I)^{-1} W)^{-1}(S - \lambda I)^{-1} = \sum_{k=0}^{m-1} (-1)^k((S - \lambda I)^{-1} W)^k(S - \lambda I)^{-1}.$$ 

Hence, it easily follows that $\sigma(S) = \sigma(A)$. This proves the lemma. $\square$
Since $B_j$ is a block triangular matrix, according to the previous lemma, we have

$$\sigma(B_j) = \cup_{k=j+1}^m \sigma(A_k) \quad (j = 0, \ldots, m - 1).$$

So, due to (2.3),

$$\sigma(B_j) \cap \sigma(A_j) = \emptyset.$$

Under this condition, the equation

$$A_jX_j - X_jB_j = -C_j \quad (j = 1, \ldots, m - 1). \quad (2.4)$$

has a unique solution, e.g., [1, Section VII.2] or [2].

**Lemma 2.2.** Let condition (2.3) hold and $X_j$ be a solution to (2.4). Then

$$(I - X_{m-1})(I - X_{m-2}) \cdots (I - X_1) A(I + X_1)(I + X_2) \cdots (I + X_{m-1})$$

$$= A_1 + A_2 + \cdots + A_m. \quad (2.5)$$

**Proof.** Since $X_j = \Delta P_jX_jQ_j$, we have $X_jA_j = B_jX_j = X_jC_j = C_jX_j = 0$. Due to (2.2), $Q_jAP_j = 0$. Thus, $A = A_1 + B_1 + C_1$, and consequently,

$$(I - X_1) A(I + X_1) = (I - X_1)(A_1 + B_1 + C_1)(I + X_1) =$$

$$A_1 + B_1 + C_1 - X_1B_1 + A_1X_1 = A_1 + B_1.$$

Furthermore, $B_1 = A_2 + B_2 + C_2$. Hence,

$$(Q_1 - X_2)B_1(Q_1 + X_2) = (Q_1 - X_1)(A_2 + B_2 + C_2)(Q_1 + X_1) =$$

$$A_2 + B_2 + C_2 - X_2B_2 + A_2X_2 = A_2 + B_2.$$

Therefore,

$$(I - X_2)(A_1 + B_1)(I + X_2) = (P_1 + Q_1 - X_2)(A_1 + B_1)(P_1 + Q_1 + X_2) =$$

$$A_1 + (Q_1 - X_2)(A_1 + B_1)(Q_1 + X_2) = A_1 + A_2 + B_2.$$

Consequently,

$$(I - X_2)(A_1 + B_1)(I + X_2) = (I - X_2)(I - X_1) A(I + X_1)(I + X_2) = A_1 + A_2 + B_2.$$
Continuing this process and taking into account that $B_{m-1} = A_m$, we obtain the required result. □

Take
$$T = (I + X_1)(I + X_2)\cdots(I + X_{m-1}).$$
(2.6)

It is simple to see that the inverse to $I + X_j$ is the matrix $I - X_j$. Thus,
$$T^{-1} = (I - X_{m-1})(I - X_{m-2})\cdots(I - X_1)$$
(2.7)
and (2.5) can be written as
$$T^{-1}AT = \text{diag}(A_{kk})_{k=1}^m.$$ (2.8)

By the inequalities between the arithmetic and geometric means, we get
$$\|T\| \leq \prod_{k=1}^{m-1} (1 + \|X_k\|) \leq \left(1 + \frac{1}{m-1} \sum_{k=1}^{m-1} \|X_k\| \right)^{m-1}$$
(2.9)
and
$$\|T^{-1}\| \leq \left(1 + \frac{1}{m-1} \sum_{k=1}^{m-1} \|X_k\| \right)^{m-1}.$$ (2.10)

3. Proof of Theorem 1.1. Let $\{e_k\}$ be the Schur basis (the orthogonal normal basis of the triangular representation) of matrix $A$:

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & a_{22} & a_{23} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & a_{nn} \end{pmatrix}$$

with $a_{jj} = \lambda_j$ in that basis. Besides,
$$\sum_{k=2}^{n} \sum_{i=1}^{k-1} |a_{ik}|^2 = g^2(A).$$

Take $P_j = \sum_{k=1}^{i} (\cdot, e_k)e_k$. Then one can apply Lemma 2.2 with $m = n$, $\Delta P_k = (\cdot, e_k)e_k$,
$$Q_j = \sum_{k=j+1}^{n} (\cdot, e_k)e_k, A_k = \Delta P_k A \Delta P_k = \lambda_k \Delta P_k, \quad \text{diag}(A_{kk})_{k=1}^n = \text{diag}(\lambda_k)_{k=1}^n,$$
A Bound for Condition Numbers of Matrices

\[ B_j = Q_j AQ_j = \begin{pmatrix}
  a_{j+1,j+1} & a_{j+1,j+2} & \cdots & a_{j+1,n} \\
  0 & a_{j+2,j+2} & \cdots & a_{j+2,n} \\
  \vdots & \vdots & \ddots & \vdots \\
  0 & 0 & \cdots & a_{nn}
\end{pmatrix} \]

and

\[ C_j = \Delta P_j AQ_j = \begin{pmatrix}
  a_{j,j+1} & a_{j,j+2} & \cdots & a_{j,n}
\end{pmatrix}. \]

Besides,

\[ A = \begin{pmatrix}
  \lambda_1 & C_1 \\
  0 & B_1
\end{pmatrix}, \quad B_1 = \begin{pmatrix}
  \lambda_2 & C_2 \\
  0 & B_2
\end{pmatrix}, \quad \ldots, \quad B_j = \begin{pmatrix}
  \lambda_{j+1} & C_{j+1} \\
  0 & B_{j+1}
\end{pmatrix} \quad (j < n). \]

So, \( B_j \) is an upper-triangular \((n - j) \times (n - j)\) matrix. Equation (2.4) takes the form \( \lambda_j X_j - X_j B_j = -C_j \). Since \( X_j = X_j Q_j \), we can write \( X_j (\lambda_j Q_j - B_j) = C_j \).

Therefore,

\[ X_j = C_j (\lambda_j Q_j - B_j)^{-1} \quad (3.1) \]

The inverse matrix is understood in the sense of subspace \( Q_j C^n \). Hence,

\[ \| X_j \| \leq \| C_j \| \| (\lambda_j Q_j - B_j)^{-1} \|. \]

Besides,

\[ \| C_j \|^2 = \sum_{k=j+1}^{n} |a_{jk}|^2, \]

and due to [10] Corollary 2.2.2, we have

\[ \| (\lambda_j Q_j - B_j)^{-1} \| \leq \sum_{k=0}^{n-j-1} \frac{g^k(B_j)}{\sqrt{k!} \delta^{k+1}} \quad (j = 1, 2, \ldots, n - 1). \]

But

\[ g^2(B_j) = g^2(Q_j AQ_j) = \sum_{k=j+2}^{n} \sum_{i=j+1}^{k-1} |a_{ik}|^2 \leq g^2(A). \]

So, with the notation

\[ \tau_1(A) := \sum_{k=0}^{n-2} \frac{g^k(A)}{\sqrt{k!} \delta^{k+1}}, \]

we have

\[ \| (\lambda_j Q_j - B_j)^{-1} \| \leq \tau_1(A) \quad \text{and} \quad \| X_j \| \leq \| C_j \| \tau_1(A). \]
Take $T$ as is in (2.6) with $X_k$ defined by (3.1). Besides (2.9) and (2.10), imply

$$\|T\| \leq \left(1 + \frac{1}{n-1} \sum_{k=1}^{n-1} \|X_k\|\right)^{n-1} \leq \left(1 + \frac{\tau_1(A)}{n-1} \sum_{k=1}^{n-1} \|C_k\|\right)^{n-1}$$

and

$$\|T^{-1}\| \leq \left(1 + \frac{\tau_1(A)}{n-1} \sum_{j=1}^{n-1} \|C_j\|\right)^{n-1}.$$  

But, by the Schwarz inequality,

$$\left(\sum_{j=1}^{n-1} \|C_j\|^2\right)^2 \leq (n-1) \sum_{j=1}^{n-1} \|C_j\|^2 = (n-1) \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} |a_{jk}|^2 = (n-1)g^2(A).$$

Thus,

$$\|T\|^2 \leq \left(1 + \frac{\tau_1(A)}{\sqrt{n-1}} g(A)\right)^{2(n-1)} = \left(1 + \frac{\tau(A)}{\sqrt{n-1}}\right)^{2(n-1)} = \gamma(A)$$

and $\|T^{-1}\|^2 \leq \gamma(A)$. Now (2.8) proves the theorem. □

4. Applications of Theorem 1.1. Theorem 1.1 immediately implies the following.

**Corollary 4.1.** Let condition (1.1) hold and $f(z)$ be a scalar function defined on the spectrum of $A$. Then $\|f(A)\| \leq \gamma(A) \max_k |f(\lambda_k)|$.

Let $A$ and $\tilde{A}$ be complex $n \times n$ matrices whose eigenvalues $\lambda_k$ and $\tilde{\lambda}_k$, respectively, are taken with their algebraic multiplicities. Recall that

$$sv_A(\tilde{A}) := \max_k \min_j |\tilde{\lambda}_k - \lambda_j|$$

is the spectral variation of $\tilde{A}$ with respect to $A$.

**Corollary 4.2.** Let condition (1.1) hold. Then $sv_A(\tilde{A}) \leq \gamma(A) \|A - \tilde{A}\|$.

Indeed, the matrix $\tilde{D} = TAT^{-1}$ is normal. Put $B = T\tilde{A}T^{-1}$. Thanks to the well-known Corollary 3.4 [10], $sv_B(B) \leq \|\tilde{D} - B\|$. Now the required result is due to Theorem 1.1.

Furthermore, let us suppose that $\lambda_k$ are real and

$$\lambda_1 > \lambda_2 > \cdots > \lambda_n.$$  

(4.1)
Then (1.2) holds with a Hermitian matrix \( \hat{D} \). Again put \( B = T\hat{A}T^{-1} \). Then, due to Theorem 1.1, we have
\[
\|\hat{D} - B\|_F = \|T\hat{A}T^{-1} - T\hat{A}T^{-1}\|_F = \|T\|\|A - \hat{A}\|_F \|T^{-1}\| \leq \gamma(A)\|A - \hat{A}\|_F. \tag{4.2}
\]
The eigenvalues of \( B \) coincide with the eigenvalues \( \tilde{\lambda}_k \) of \( \tilde{A} \). Denote \( \mu_k = \text{Re} \hat{\lambda}_k \) and assume that \( \hat{\lambda}_k \) are ordered in such a way that
\[
\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n. \tag{4.3}
\]
Due to the Kahan theorem \([16, \text{Theorem IV.5.2, p. 213, inequality (5.4)}]\), we can write
\[
\left[ \sum_{k=1}^{n} |\tilde{\lambda}_k - \lambda_k|^2 \right]^{1/2} \leq \sqrt{2}\|\hat{D} - B\|_F.
\]
Hence, taking into account (4.2), we arrive at our next result.

**Corollary 4.3.** Let the inequalities (4.1) and (4.3) hold. Then
\[
\left[ \sum_{k=1}^{n} |\tilde{\lambda}_k - \lambda_k|^2 \right]^{1/2} \leq \sqrt{2}\gamma(A)\|A - \hat{A}\|_F.
\]

In addition, note that Theorem 4.5.4 in \([16, \text{p. 215}]\) and Theorem 1.1 yield the following corollary.

**Corollary 4.4.** Let \( A \) and \( \hat{A} \) be diagonalizable \( n \times n \) matrices having purely real eigenvalues:
\[
\lambda_1 < \lambda_2 < \cdots < \lambda_n \quad \text{and} \quad \hat{\lambda}_1 < \hat{\lambda}_2 < \cdots < \hat{\lambda}_n, \quad \text{respectively}.
\]
Then
\[
|\hat{\lambda}_j - \lambda_j| \leq \gamma(A)\gamma(\hat{A})\|A - \hat{A}\| \quad (j = 1, \ldots, n).
\]

To consider an additional application of Theorem 1.1, put
\[
\text{md}(A, \hat{A}) := \min_{\pi} \left[ \sum_{k=1}^{n} |\tilde{\lambda}_k - \lambda_k|^2 \right]^{1/2},
\]
where \( \pi \) ranges over all permutations of the integers 1, 2, \ldots, \( n \), cf. \([16]\). Let us use Theorem 4.5.5 \([16, \text{p. 216}]\). That theorem together with Theorem 1.1 implies our next result.

**Corollary 4.5.** Let the conditions (1.1) and
\[
\hat{\lambda}_j \neq \hat{\lambda}_m, \quad \text{whenever} \quad j \neq m \tag{4.4}
\]
be fulfilled. Then

$$\| md (A, \tilde{A}) \| \leq \gamma (A) \gamma (\tilde{A}) \| A - \tilde{A} \| _{F}.$$  

About the interesting recent publications on spectrum perturbations see for instance [6, 8].

Finally, note that Corollaries 2.2 and 2.3 from [11] and the above proved Theorem 1.1 yield the following corollary.

**Corollary 4.6.** Let conditions (1.1) and (4.4) be fulfilled, and $f$ be a function defined on $\sigma (A) \cup \sigma (\tilde{A})$. Then the inequalities

$$\| f(A) - f(\tilde{A}) \| _{F} \leq \gamma (A) \gamma (\tilde{A}) \max _{j,k} \left| \frac{f(\lambda _{k}) - f(\tilde{\lambda }_{j})}{\lambda _{k} - \tilde{\lambda }_{j}} \right| \| A - \tilde{A} \| _{F},$$

and

$$\| f(A) - f(\tilde{A}) \| _{F} \leq \gamma (A) \gamma (\tilde{A}) \max _{j,k} | f(\lambda _{k}) - f(\tilde{\lambda }_{j}) |$$

are valid.

The recent interesting results devoted to matrix-valued functions can be found in [7, 14].

5. **Example.** To illustrate Theorem 1.1, consider the simple matrix

$$A = \begin{pmatrix} 5 & 0 & -\frac{1}{3} \\ 0 & 7 & 0 \\ 0 & 0 & 3 \end{pmatrix}.$$  

It is simple to check that $\lambda _{1} = 5, \lambda _{2} = 7, \lambda _{3} = 3$. So, $\delta = 2$. In addition, due to (1.3), we have $g(A) \leq \| A - A^{\ast} \| _{F} / \sqrt{2} = \frac{1}{3}$. Thus,

$$\tau (A) = \sum _{k=0}^{1} \frac{g^{k+1}(A)}{\sqrt{k!} \delta ^{k+1}} \leq \frac{1}{6} + \frac{1}{36} \approx 0.1944$$

and

$$\gamma (A) \leq \left( \frac{1 + 0.1944}{\sqrt{2}} \right) ^{4} \approx 1.6741.$$  

On the other hand, it is not hard to check that the matrix

$$T = \begin{pmatrix} 1 & 0 & \frac{1}{6} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

has the inverse one $T^{-1} = \begin{pmatrix} 1 & 0 & -\frac{1}{6} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.  

and satisfies relation (1.2) with
\[
\hat{D} = \begin{pmatrix}
5 & 0 & 0 \\
0 & 7 & 0 \\
0 & 0 & 3
\end{pmatrix}.
\]
The direct calculations gives us \(\kappa_T \approx 1.1815\).

Acknowledgment. I am very grateful to the referee of the present paper for his/her really deep and helpful remarks.

REFERENCES