PROCEEDINGS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 39, Number 2, July 1973

EVERY $n \times n$ MATRIX Z WITH REAL SPECTRUM SATISFIES $\|Z - Z^*\| \leqq \|Z + Z^*\| (\log_2 n + 0.038)$

W. KAHAN

ABSTRACT. The title's inequality is proved for the operator bound-norm in a unitary space. An example is exhibited to show that the inequality cannot be improved by more than about 8% when n is large. The numerical range, of an $n \times n$ matrix Z with real spectrum, is then shown to be not arbitrarily different in shape from the spectrum.

The norm in question is the matrix bound-norm in a unitary space; $\|B\| \equiv \max_{v \neq 0} \|Bv\|/\|v\|$ where the vector norm is $\|v\| \equiv (v^*v)^{1/2}$.

Publication of the title's inequality was stimulated by work of Alan McIntosh ([**1971**], [**1972**]) on questions posed by Tosio Kato, but the inequality has some interesting aspects of its own. First, the surprising appearance of the logarithm function is unavoidable because for every $n > 1$ an example exists of an $n \times n$ matrix Z, with real spectrum, which satisfies

$$\|Z - Z^*\|/\|Z + Z^*\| > (2/\pi)(\log n + \tfrac{1}{4} - \tfrac{1}{2} \log 2 + 1/2n);$$

and $(2/\pi) \log n \doteqdot 0.92 \log_2 n$. Secondly, the inequality implies that the *numerical range*—the range of values taken by v^*Zv/v^*v as v runs through all nonzero vectors—cannot differ arbitrarily in shape from the spectrum of Z when that spectrum is real. These assertions are elaborated and proved in §§1 and 2 below.

0. **Proof of the title's inequality.** Schur's theorem allows any $n \times n$ matrix Z to be transformed into an upper-triangular matrix by a unitary similarity without changing $\|Z-Z^*\|$ nor $\|Z+Z^*\|$ nor Z's spectrum, which appears on the triangular matrix's diagonal. Therefore let us restrict attention to $n \times n$ upper-triangular matrices Z with real diagonal. Then $Z \neq 0$ if and only if $Z+Z^* \neq 0$, so we might as well normalize the nonzero matrices Z to satisfy $\|Z+Z^*\| = 1$, from which it follows that no element of Z can exceed 1 in magnitude and hence $\|Z-Z^*\| \leqq n-1$. Now we seek an estimate for $\beta_n \equiv \max \|Z-Z^*\|$ over $n \times n$ upper-triangular $Z \neq 0$ with

Received by the editors August 11, 1972.
AMS (MOS) subject classifications (1970). Primary 15A45, 15A60, 47A10, 47A30, 47B35, 47B99.
Key words and phrases. Matrix with real spectrum, numerical range.

real diagonal and $\|Z+Z^*\|=1$, from which we hope to deduce $\beta_n \leqq \log_2 n + 0.038$. Observe that the maximum sought is the maximum of a continuous function over a compact set, so the maximum is achieved at some Z_n (not uniquely determined) for each n, and Z_n must satisfy: Z_n is $n \times n$ upper-triangular with real diagonal, $\|Z_n+Z_n^*\|=1$, $\|Z_n-Z_n^*\|=\beta_n$.

For example, $Z_1=(1/2)$ and $\beta_1=0$, and $Z_2=\left(\begin{smallmatrix}0 & 1\\ 0 & 0\end{smallmatrix}\right)$ and $\beta_2=1$, as may be verified by elementary computation. We might as well assume further that β_n is the largest eigenvalue of $\iota(Z_n-Z_n^*)$; otherwise replace Z_n by $-Z_n$; and we shall let x_n denote a corresponding eigenvector,

$$\iota(Z_n - Z_n^*)x_n = \beta_n x_n \qquad (\iota \equiv \sqrt{(-1)})$$

normalized so that $\|x_n\|=1$.

For any $n>2$ choose $k \equiv \lfloor n/2 \rfloor$ (=the greatest integer in $n/2$) and $m \equiv n-k$, and partition Z_n and x_n conformally thus:

$$Z_n = \begin{matrix}\begin{pmatrix} P & Q \\ O & R \end{pmatrix} \begin{matrix} m \\ k \end{matrix} \\ \begin{matrix} m & k \end{matrix}\end{matrix} \qquad x_n = \begin{pmatrix} q \\ r \end{pmatrix} \begin{matrix} m \\ k \end{matrix}$$

where P is $m \times m$ upper-triangular with real diagonal, and R is $k \times k$ upper-triangular with real diagonal. We may now estimate

$$\|Q\| = \left\| \begin{pmatrix} 0 & Q \\ 0 & 0 \end{pmatrix} \right\| \leqq \left\| \begin{pmatrix} P^* + P & Q \\ Q^* & R^* + R \end{pmatrix} \right\| = \|Z_n^* + Z_n\| = 1,$$

and similarly $\|P^*+P\| \leqq 1$ and $\|R^*+R\| \leqq 1$, whence it follows that $\|P-P^*\| \leqq \beta_m$ and $\|R-R^*\| \leqq \beta_k$. And since $\|q\|^2+\|r\|^2=\|x_n\|^2=1$,

$$\begin{aligned} \beta_n &= \iota x_n^*(Z_n - Z_n^*)x_n \\ &= \iota(q^*(P - P^*)q + q^*Qr - r^*Q^*q + r^*(R - R^*)r) \\ &\leqq \beta_m\|q\|^2 + 2\|q\| \cdot \|r\| + \beta_k\|r\|^2 \\ &\leqq \left\| \begin{pmatrix} \beta_m & 1 \\ 1 & \beta_k \end{pmatrix} \right\| = (\beta_m + \beta_k)/2 + (1 + (\beta_m - \beta_k)^2/4)^{1/2}. \end{aligned}$$

We must now exploit the properties of the function $\|\left(\begin{smallmatrix}\xi & 1\\ 1 & \eta\end{smallmatrix}\right)\|$; e.g. it is a monotonic increasing function of ξ and η for all real ξ and η with $\xi+\eta>0$. Therefore an argument by induction proves that $\beta_n \leqq \mu_n$ for all n where the sequence $\{\mu_n\}$ is defined recursively thus:

$$\mu_1 \equiv 0, \quad \mu_2 \equiv 1, \quad \mu_{2n} \equiv \mu_n + 1,$$

$$\mu_{2n+1} \equiv \left\| \begin{pmatrix} \mu_{n+1} & 1 \\ 1 & \mu_n \end{pmatrix} \right\| = (\mu_{n+1} + \mu_n)/2 + (1 + (\mu_{n+1} - \mu_n)^2/4)^{1/2}.$$

Another argument by induction proves that the sequence $\{\mu_n\}$ is monotonic; $\mu_{2n}<\mu_{2n+1}<\mu_{2n+2}$. What remains to be proved is that $\mu_n<\log_2 n+0.038$.

In fact more is true. To any positive integer n correspond integers i and j defined uniquely by $2^i \leqq n=2^i+j<2^{i+1}$, so $0 \leqq 2^{-i}j<1$; it turns out that

$$\log_2 n = i + \log_2(1 + 2^{-i}j)$$
$$\leqq \mu_n \leqq i + \log(1 + (e-1)2^{-i}j) < \log_2 n + 0.038.^1$$

The first of these inequalities will not be proved here (it is presented only to show how closely μ_n approximates $\log_2 n$); and the last inequality is a consequence of the elementary observation that the function $\log(1+(e-1)\xi)-\log_2(1+\xi)$ is positive for $0<\xi<1$ and takes its maximum value there when $\xi=(\log 2-1/(e-1))/(1-\log 2)$, and that maximum value is about $0.0379\cdots$. The nontrivial inequality is the middle one, and it will be deduced from the following elementary inequality:

LEMMA 0. $1/\log(1+\xi)+1/\log(1-\xi)>1$ *whenever* $0<\xi<1$.

PROOF. Define $\phi_1(\xi) \equiv -\log(1-\xi)-\log(1+\xi)=-\log(1-\xi^2)>0$ for $0<\xi<1$. Then in turn

$$\phi_2(\xi) \equiv \int_0^\xi \phi_1(\eta)\,d\eta = (1-\xi)\log(1-\xi) - (1+\xi)\log(1+\xi) + 2\xi > 0;$$

$$\phi_3(\xi) \equiv \phi_2(\xi)/(1-\xi^2) = \frac{\log(1-\xi)}{1+\xi} - \frac{\log(1+\xi)}{1-\xi} - \frac{1}{1+\xi} + \frac{1}{1-\xi} > 0;$$

$$\phi_4(\xi) \equiv \int_0^\xi \phi_3(\eta)\,d\eta = \log(1+\xi)\log(1-\xi) - \log(1+\xi) - \log(1-\xi) > 0;$$

$$\frac{1}{\log(1+\xi)} + \frac{1}{\log(1-\xi)} - 1 = \frac{\phi_4(\xi)}{-\log(1-\xi)\log(1+\xi)} > 0 \quad \text{as claimed.}$$

Now for some serious work. We begin with the induction hypothesis that $\mu_n \leqq i+\log(1+(e-1)2^{-i}j)$, where i and j are derived from n as described above, for each $n=1, 2, \cdots, 2^m-1, 2^m$ and some $m \geqq 1$. The hypothesis is obviously true for $m=1$. Since $2n$ corresponds respectively to $i+1$ and $2j$ (i.e. $2n=2^{i+1}+2j$), and $\mu_{2n}=\mu_n+1 \leqq i+1+\log(1+(e-1)2^{-i-1}(2j))$, we see how the induction hypothesis is conveyed from n to $2n$ and hence to all the even integers $2n=2^m+2, 2^m+4, \cdots, 2^{m+1}$. As for the odd integers $2n+1$, we observe that $n+1$ corresponds either to i and $j+1$ (i.e. $n+1=2^i+j+1$) or to $i+1$ and 0 (i.e. $n+1=2^{i+1}$), so in either case, provided

---

[1] The referee claims that the inequality $\mu_n<\log_2 n+1$ is trivial.

$n+1 \leqq 2^m$,

$$\mu_{2n+1} = \left\| \begin{pmatrix} \mu_{n+1} & 1 \\ 1 & \mu_n \end{pmatrix} \right\|$$
$$\leqq \left\| \begin{pmatrix} i + \log(1 + (e-1)2^{-i}(j+1)) & -1 \\ -1 & i + \log(1 + (e-1)2^{-i}j) \end{pmatrix} \right\|;$$

and we wish to show that the last expression is less than

$$i + 1 + \log(1 + (e-1)2^{-i-1}(2j+1)).$$

But that is soon seen to be tantamount to showing

$$\det \begin{pmatrix} 1 + \log(1 + (e-1)2^{-i-1}(2j+1)) & 1 \\ -\log(1 + (e-1)2^{-i}(j+1)) & \\ & 1 + \log(1 + (e-1)2^{-i-1}(2j+1)) \\ 1 & -\log(1 + (e-1)2^{-i}j) \end{pmatrix} > 0$$

and this is tantamount to showing for $\xi \equiv 1/(2j+1+2^{i+1}/(e-1))$ that the inequality of Lemma 0 is true. The proof of the title's inequality is soon completed.

1. **An example** Z. We shall exhibit an $n \times n$ matrix Z, with real spectrum, which satisfies

$$\|Z - Z^*\|/\|Z + Z^*\| > (2/\pi)(\log n + \tfrac{1}{4} - \tfrac{1}{2}\log 2 + 1/2n);$$

since $(2/\pi)\log n \doteq 0.92 \log_2 n$ we must conclude that the title's inequality, $\|Z-Z^*\|/\|Z+Z^*\| < \log_2 n + 0.038$, cannot be improved by more than about 8% when n is large.

The example is

$$Z \equiv \iota \begin{pmatrix} 0 & 1 & \frac{1}{2} & \frac{1}{3} & \cdot & \cdot & \cdot & 1/(n-1) \\ & 0 & 1 & \frac{1}{2} & \frac{1}{3} & \cdot & \cdot & \cdot \\ & & 0 & 1 & \frac{1}{2} & \frac{1}{3} & \cdot & \cdot \\ & & & & \cdot & \cdot & \cdot & \cdot \\ & & & & & & 0 & 1 \\ & & & & & & & 0 \end{pmatrix}_{n \times n};$$

i.e.

$$z_{ij} \equiv 0 \quad \text{if } i \geqq j,$$
$$\equiv \iota/(j-i) \quad \text{if } i < j.$$

Our object is to obtain estimates $\|Z+Z^*\| < \pi$ and $\|Z-Z^*\| > 2\log n + \frac{1}{2} - \log 2 + 1/n$.

The matrix $Z+Z^*$ is the $n \times n$ Toeplitz matrix belonging to the function

$\phi(\theta) \equiv \pi - \theta$ on $0 < \theta < 2\pi$ in so far as to any n-vector x with components $\xi_0, \xi_1, \cdots, \xi_{n-1}$ correspond the quadratic forms

$$x^*(Z + Z^*)x = \frac{1}{2\pi}\int_0^{2\pi} \left|\sum_0^{n-1} \xi_j e^{ij\theta}\right|^2 \phi(\theta)\, d\theta,$$

$$x^*x = \frac{1}{2\pi}\int_0^{2\pi} \left|\sum_0^{n-1} \xi_j e^{ij\theta}\right|^2 d\theta.$$

Evidently $-\pi < Z + Z^* < \pi$, so $\|Z + Z^*\| < \pi$. Moreover, the constant π in the last inequality cannot be replaced by any smaller constant for all n without contradicting theorems in Grenander and Szegö [**1958**, pp. 19, 64].

The matrix $\iota(Z^* - Z)$ is another Toeplitz matrix, this time belonging to

$$\psi(\theta) \equiv -2 \log(2 \sin \tfrac{1}{2}\theta) = 2\sum_1^\infty \frac{\cos n\theta}{n},$$

but the fact that $\psi(\theta) \to \infty$ as $\theta \to 0$ places that matrix out of reach of the theory in Grenander and Szegö [**1958**, pp. 72–75], so we shall resort to elementary methods. Specifically, we invoke the fact that

$$\|Z^* - Z\| = \max |x^*(Z^* - Z)x / x^*x| \quad \text{over } x \neq 0,$$

and consider a trial vector x with all components equal. Thus we find

$$\begin{aligned}
\|Z^* - Z\| &\geqq \frac{1}{n}\sum\sum \frac{1}{|i - j|} \quad (\text{over } 1 \leqq i \leqq n, 1 \leqq j \leqq n \text{ and } i \neq j) \\
&= 2(\tfrac{1}{2} + \tfrac{1}{3} + \tfrac{1}{4} + \cdots + 1/(n-1) + 1/n) \\
&= 2(\tfrac{1}{2} \cdot \tfrac{1}{2} + \tfrac{1}{3} + \tfrac{1}{4} + \cdots + 1/(n-1) + 1/2n) + \tfrac{1}{2} + 1/n \\
&\geqq 2\int_2^n d\xi/\xi + \tfrac{1}{2} + 1/n \quad (\text{cf. trapezoidal rule}) \\
&= 2(\log n - \log 2 + \tfrac{1}{4} + 1/2n), \quad \text{as claimed.}
\end{aligned}$$

This inequality cannot much overestimate $\|Z^* - Z\|$ when n is large since

$$\begin{aligned}
\|Z^* - Z\| &\leqq \text{largest row-sum of magnitudes of elements of } \iota(Z^* - Z) \\
&= (1 + \tfrac{1}{2} + \cdots + 1/\lfloor (n-1)/2 \rfloor) + (1 + \tfrac{1}{2} + \cdots + 1/\lfloor n/2 \rfloor) \\
&\leqq \int_{1/2}^{\lfloor (n-1)/2 \rfloor + 1/2} \frac{d\xi}{\xi} + \int_{1/2}^{\lfloor n/2 \rfloor + 1/2} \frac{d\xi}{\xi} \quad (\text{cf. midpoint rule}) \\
&\leqq 2 \log n.
\end{aligned}$$

Thus we conclude $\|Z - Z^*\| / (\|Z + Z^*\| \log n)$ approaches $2/\pi$ from below as $n \to \infty$, so the title's inequality overestimates $\|Z - Z^*\| / \|Z + Z^*\|$ by about 8% when n is very large.

Can this example be improved to provide a larger limit for

$$\|Z - Z^*\|/(\|Z + Z^*\| \log n)$$

as $n \to \infty$? Can the title's inequality be improved to provide a smaller bound for that quotient?

The title's inequality is very much an artifact of the chosen norm. A different norm, say $\|B\|_2 \equiv (\text{trace}(B^*B))^{1/2}$, would lead to a different result:

Every $n \times n$ matrix Z with real spectrum satisfies $\|Z - Z^*\|_2 \leqq \|Z + Z^*\|_2$. The proof is immediate after Z has been transformed to an upper triangle by a unitary similarity, and shows that the inequality becomes equality just when Z is nilpotent.

2. **The shape of Z's numerical range when its spectrum is real.** Z's numerical range $\mathcal{N}(Z)$ is the set of all complex numbers $\xi + \iota\eta = v^*Zv/v^*v$ obtained as v runs through all nonzero n-vectors. The Toeplitz-Hausdorff theorem, for which C. Davis [**1971**] has recently provided a brief proof, asserts that $\mathcal{N}(Z)$ is a convex set containing Z's spectrum. When Z's spectrum is real, how different can the shape of $\mathcal{N}(Z)$ be from that of a horizontal line segment?

Let us denote the height and width of $\mathcal{N}(Z)$ by

$$\mathcal{H}(Z) \equiv \max_{\xi+\iota\eta \in \mathcal{N}(Z)} \eta - \min_{\xi+\iota\eta \in \mathcal{N}(Z)} \eta \quad \text{and}$$

$$\mathcal{W}(Z) \equiv \max_{\xi+\iota\eta \in \mathcal{N}(Z)} \xi - \min_{\xi+\iota\eta \in \mathcal{N}(Z)} \xi.$$

We claim that every $n \times n$ matrix Z with real spectrum satisfies

$$\mathcal{H}(Z) \leqq \mathcal{W}(Z)(\log_2 n + 0.038).$$

First observe that $\mathcal{N}(Z-\alpha) = \mathcal{N}(Z) - \alpha$, $\mathcal{H}(Z-\alpha) = \mathcal{H}(Z)$ and $\mathcal{W}(Z-\alpha) = \mathcal{W}(Z)$ for every scalar α and, in particular, for every real scalar α. Therefore we lose no generality by setting

$$\alpha \equiv \frac{1}{2}\left(\max_{\xi+\iota\eta \in \mathcal{N}(Z)} \xi + \min_{\xi+\iota\eta \in \mathcal{N}(Z)} \xi \right)$$

and considering $Z - \alpha$ in place of Z, i.e. assume $\alpha = 0$. Then

$$\max_{\xi+\iota\eta \in \mathcal{N}(Z)} \xi = - \min_{\xi+\iota\eta \in \mathcal{N}(Z)} \xi = \max_{v \neq 0} \frac{|v^*(Z + Z^*)v|}{2v^*v} = \tfrac{1}{2}\|Z + Z^*\|,$$

so $\mathcal{W}(Z) = \|Z + Z^*\|$. On the other hand, a similar calculation shows $\mathcal{H}(Z) \leqq \|Z - Z^*\|$. Consequently the title's inequality proves the claim.

Does that claim always grossly overestimate $\mathscr{H}(Z)/\mathscr{W}(Z)$? No. Consider for example a $2n \times 2n$ matrix Z which is the *diagonal* sum of the previous section's example and its conjugate transpose; for this new example $\mathscr{H}(Z)/(\mathscr{W}(Z)\log n) \to 2/\pi$ as $n \to \infty$.

ACKNOWLEDGEMENTS. This work was supported in part by a grant from the U.S. Office of Naval Research, contract no. N00014-69-A-0200-1017. The results were first presented at the Fifth Gatlinburg Symposium on Numerical Linear Algebra held at Los Alamos on June 5–10, 1972. I am indebted to T. Kato and A. McIntosh for a prepublication view of the latter's results.

REFERENCES

1971. Chandler Davis, *The Toeplitz-Hausdorff theorem explained*, Canad. Math. Bull. **14** (1971), 245–246.

1958. Ulf Grenander and Gabor Szegö, *Toeplitz forms and their applications*, Univ. of California Press, Berkeley, Calif., 1958. MR **20** #1349.

1971. Alan McIntosh, *Counterexample to a question on commutators*, Proc. Amer. Math. Soc. **29** (1971), 337–340. MR **43** #2538.

1972. ———, *On the comparability of* $A^{1/2}$ *and* $(A^*)^{1/2}$, Proc. Amer. Math. Soc. **32** (1972), 430–434. MR **44** #7354.

COMPUTER SCIENCE DEPARTMENT, UNIVERSITY OF CALIFORNIA, BERKELEY, CALIFORNIA 94720