BIT 31 (1991), 551–555.

ON A CONJECTURE BY LE VEQUE AND TREFETHEN RELATED TO THE KREISS MATRIX THEOREM

M. N. SPIJKER

Department of Mathematics and Computer Science, University of Leiden, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands.

Abstract.

In the Kreiss matrix theorem the power boundedness of $N \times N$ matrices is related to a resolvent condition on these matrices. LeVeque and Trefethen proved that the ratio of the constants in these two conditions can be bounded by $2eN$. They conjectured that this bound can be improved to eN.

In this note the conjecture is proved to be true. The proof relies on a lemma which provides an upper bound for the arc length of the image of the unit circle in the complex plane under a rational function. This lemma may be of independent interest.

1980 AMS Subject Classification (1985 revision): primary 39A11, 65M10; secondary 15A45, 30A10.

1. Introduction

In the stability analysis of numerical methods one is often faced with the question whether given matrices A have powers that are uniformly bounded. The Kreiss matrix theorem (see e.g. [1], [3], [5]) provides an important tool for answering this question. One of the assertions of the theorem relates the inequality

$$\|A^n\| \leq C_0 \qquad \text{for } n = 1, 2, 3, \ldots \tag{1}$$

to the resolvent condition

$$(zI - A) \text{ is regular with } \|(zI - A)^{-1}\| \leq C_1 \cdot (|z| - 1)^{-1} \text{ for all complex } z \text{ with } |z| > 1. \tag{2}$$

Here A denotes an arbitrary complex $N \times N$ matrix, I is the $N \times N$ identity matrix and $\|\cdot\| = \|\cdot\|_2$ is the spectral norm. By power series expansion it is easily seen that (1) implies (2) with $C_1 = C_0$. The Kreiss theorem asserts that, conversely, (2) implies (1) with C_0 depending on C_1 and N only. Various authors (e.g. [5], [6]) studied the size of (the optimal) C_0 as a function of C_1, N. Eventually it was proved

Received November, 1990. Revised March, 1991.

by LeVeque and Trefethen [2] that (2) implies (1) with

$$C_0 = 2eN \cdot C_1. \tag{3a}$$

Moreover, these authors showed by means of a counterexample that, for arbitrary C_1, the factor 2 in (3a) cannot be replaced by any factor smaller than 1. They conjectured that (3a) can be strengthened to the optimal value

$$C_0 = eN \cdot C_1. \tag{3b}$$

In this note we prove (3b) as well as a related conjecture to be stated below.

2. The proof by LeVeque and Trefethen of (3a).

Let S denote the unit circle $\{z | z \in \mathbb{C} \text{ with } |z| = 1\}$. By $\mathscr{R}_N$ we denote the class of all rational functions

$$R(z) = P(z)/Q(z)$$

where $P(z)$, $Q(z)$ are polynomials, with complex coefficients, of a degree not exceeding N with $Q(z) \neq 0$ on S.

Suppose γ is any constant such that

$$\int_S |R'(z)||dz| \leq \gamma \cdot 2\pi N \cdot \max_S |R(z)| \text{ for all } R \in \mathscr{R}_N, \qquad N \geq 1. \tag{4}$$

LeVeque and Trefethen [2] proved that (4) is valid with

$$\gamma = 2 \tag{5a}$$

and they conjectured that

$$\gamma = 1 \tag{5b}$$

is possible. The proof in [2] of (3a) can be viewed as consisting in a combination of (5a) and the following lemma (implicitly proved in [2]),

LEMMA 1. *Let γ be such that* (4) *is valid. Then* (2) *implies* (1) *with* $C_0 = \gamma \cdot eN \cdot C_1$.

In an attempt to prove conjecture (5b), Smith [4] obtained the value $\gamma = 1 + 2/\pi$, which is better than (5a) but still larger than (5b).

3. Proof of the conjectures (3b), (5b).

We shall prove

LEMMA 2. *Relation* (4) *is valid with* $\gamma = 1$.

Clearly, in view of lemma 1, our lemma proves both of the above conjectures (3b), (5b).

PROOF OF LEMMA 2.

1. Let $N \geq 1$ and $R \in \mathscr{R}_N$ be given. We denote the integral appearing in the left-hand member of inequality (4) by L, and we use the notation

$$f(t) = g(t) + i \cdot h(t) = R(\exp[it]) \qquad \text{for real } t.$$

Clearly

$$L = \int_0^{2\pi} |f'(t)| dt.$$

For each t there is a real ω such that $g'(t) = |f'(t)| \cos \omega$, $h'(t) = |f'(t)| \sin \omega$, which implies

$$\int_0^{2\pi} |g'(t) \cos \theta + h'(t) \sin \theta| \, d\theta = \int_0^{2\pi} |\cos(\omega - \theta)| \cdot |f'(t)| \, d\theta = 4 |f'(t)|.$$

We thus arrive at the following representation,

$$L = \tfrac{1}{4} \int_0^{2\pi} \left\{ \int_0^{2\pi} |g'(t) \cos \theta + h'(t) \sin \theta| \, dt \right\} d\theta. \tag{6}$$

Let $\theta \in [0, 2\pi]$ be given, and write

$$F(t) = g(t) \cos \theta + h(t) \sin \theta.$$

Below (in parts 2, 3) we shall establish the inequality

$$\int_0^{2\pi} |F'(t)| \, dt \leq 4N \cdot \max_{0 \leq t \leq 2\pi} |F(t)|. \tag{7}$$

Since

$$\max_{0 \leq t \leq 2\pi} |F(t)| \leq \max_{z \in S} |R(z)|$$

the lemma follows by combining (6) and (7).

2. In order to prove (7) we assume, without loss of generality, that $F'(t)$ does not vanish identically on $[0, 2\pi]$. Consequently, the integral in (7) is equal to a sum,

$$\sum_{j=1}^{k} \left| \int_{t_{j-1}}^{t_j} F'(t) dt \right|$$

where $t_0 = 0 < t_1 < \ldots < t_k = 2\pi$ and $F'(t) \neq 0$ on each open interval (t_{j-1}, t_j). Hence

$$\int_0^{2\pi} |F'(t)| dt = \sum_{j=1}^{k} |F(t_j) - F(t_{j-1})|.$$

We denote the range of values $F(t)$ obtained when t runs through the interval (t_{j-1}, t_j) by A_j. Further we define

$$a = \max_{0 \le t \le 2\pi} |F(t)|,$$

and $\varphi_j(x) = 1$ (for $x \in A_j$), $\varphi_j(x) = 0$ (for $x \in \mathbb{R} \setminus A_j$). Since $A_j \subset [-a, a]$ we have

$$\sum_{j=1}^{k} |F(t_j) - F(t_{j-1})| = \sum_{j=1}^{k} \int_{-a}^{a} \varphi_j(x)dx = \int_{-a}^{a} \left\{ \sum_{j=1}^{k} \varphi_j(x) \right\} dx.$$

Consequently

$$\int_0^{2\pi} |F'(t)|\, dt \le 2ab \tag{8}$$

where $b = \sup_x \{\varphi_1(x) + \varphi_2(x) + \ldots + \varphi_k(x)\}$.

3. In order to bound b, we assume $x \in \mathbb{R}$ to be given and we write $z = e^{it}$ for $t \in (0, 2\pi)$. A straightforward calculation shows that

$$F(t) = x \tag{9}$$

is equivalent to

$$e^{-i\theta} P(z)\overline{Q(z)} + e^{i\theta} \overline{P(z)} Q(z) - 2xQ(z)\overline{Q(z)} = 0.$$

Multiplying this equality by z^N we arrive, in view of $\bar{z} = z^{-1}$, at a relation

$$p(z) = 0,$$

where $p(z)$ is a polynomial of a degree not exceeding $2N$. Moreover $p(z)$ does not vanish identically (since $F'(t)$ does not). Therefore, there exist at most $2N$ different values $t \in (0, 2\pi)$ satisfying (9). This implies that $x \in A_j$ for at most $2N$ different values j. Hence $\{\varphi_1(x) + \varphi_2(x) + \ldots + \varphi_k(x)\} \le 2N$, so that

$$b \le 2N.$$

In view of (8) we have proved (7), which completes the proof of the lemma.

REMARK 1. Lemma 2 is equivalent to the following remarkable fact. Denote for any rational function R the maximum distance of $R(z)$ to the origin when $|z| = 1$ by M_R, and the arc length of the image of the unit circle under the mapping R by L_R. Then for all $R \in \mathscr{R}_N$ the ratio L_R/M_R will never exceed the ratio L_P/M_P, where $P(z) = z^N$.

REMARK 2. The above proof of lemma 2 allows the following geometrical interpretation.

Consider the projection of the curve $\zeta = R(e^{it})$ (where $0 \leq t \leq 2\pi$) onto the straight line passing through the origin with angle θ to the real axis. Denote by $L_R(\theta)$ the length of this projection, and by $M_R(\theta)$ the maximum value of the distances between the origin and the projections of the points $\zeta = R(e^{it})$.

The length L_R of the original curve can be computed by finding the average value of $L_R(\theta)$ (when $0 \leq \theta \leq 2\pi$), and then multiplying by $\pi/2$. This is equation (6). Clearly, the theorem is proved if we can show that $L_R(\theta) \leq 4N \cdot M_R(\theta)$. This is inequality (7).

In part 2 of the proof it is shown that this inequality holds if we can prove that the projection of $\zeta = R(e^{it})$ passes any given point on the θ-line at most $2N$ times as t ranges over $(0, 2\pi)$.

Since the value $x = F(t)$ specifies the position on the θ-line of the projection of $\zeta = R(e^{it})$, one sees that part 3 of the proof amounts to proving the above upper-bound $2N$.

ACKNOWLEDGEMENT. I am grateful to L. N. Trefethen and J. F. B. M. Kraaijevanger for drawing my attention to reference [4], and to the referee for a remark concerning the presentation of the proof.

REFERENCES

[1] Kreiss, H.-O., *Über die Stabilitätsdefinition für Differenzengleichungen die partielle Differentialgleichungen approximieren*, BIT 2, 153–181 (1962).

[2] LeVeque, R. J., L. N. Trefethen, *On the resolvent condition in the Kreiss matrix theorem*, BIT 24, 584–591 (1984).

[3] Richtmyer, R. D., K. W. Morton, *Difference Methods for Initial-value Problems*, 2nd Ed., J. Wiley & Sons. New York, London, Sydney, 1967.

[4] Smith, J. C., *An inequality for rational functions*, The American Mathem. Monthly 92, 740–741 (1985).

[5] Sod, G. A., *Numerical Methods in Fluid Dynamics*, Cambridge University Press. Cambridge, 1985.

[6] Tadmor, E., *The equivalence of L_2-stability, the resolvent condition, and strict H-stability*. Linear Algebra Appl. 41, 151–159 (1981).