Fields of Values and Iterative Methods

Michael Eiermann*
IBM Scientific Center
Tiergartenstrasse 15
D-6900 Heidelberg, Germany

Submitted by Michael Neumann

ABSTRACT

The performance of an iterative scheme to solve $\mathbf{x} = T\mathbf{x} + \mathbf{c}$, $T \in \mathbb{C}^{n\times n}$, $\mathbf{c} \in \mathbb{C}^n$, is often judged by spectral properties of T. If T is not normal, it is however well known that only conclusions about the *assymptotic* behavior of an iterative method can be drawn from spectral information. To anticipate the progress of the iteration after a finite number of steps, the knowledge of the eigenvalues alone is often useless. In addition, the spectrum of T may be highly sensitive to perturbations if T is not normal. An iterative method which—on the basis of some spectral information—is predicted to converge rapidly for T may well diverge if T is slightly perturbed. In practice, the convergence of the iteration $\mathbf{x}_m = T\mathbf{x}_{m-1} + \mathbf{c}$ is therefore frequently measured by some norm $\|T\|$, rather than by the spectral radius $\rho(T)$. But apart from the fact that norms lead to error estimates which are often too pessimistic, they cannot be used to analyze more general schemes such as, e.g., the Chebyshev iterative methods. Here, we discuss another tool to analyze the behavior of an iterative method, namely the field of values $W(T)$, the collection of all Rayleigh quotients of T. $W(T)$ contains the eigenvalues of T, and the numerical radius $\mu(T) = \max_{z\in W(T)}|z|$ defines a norm on $\mathbb{C}^{n\times n}$. The field of values represents therefore an "intermediate concept" to judge an iterative scheme by—it is related to the spectral approach but has also certain norm properties.

*On leave from the Institut für Praktische Mathematik, Universität Karlsruhe, Englerstr. 2, D-7500 Karlsruhe.E-mail: af07@dkauni2.bitnet.

655 Avenue of the Americas, New York, NY 10010 0024-3795/93/$6.00

1. INTRODUCTION

A standard way to solve a linear system $\mathbf{x} = T\mathbf{x} + \mathbf{c}$ iteratively, where $T \in \mathbb{C}^{n\times n}$ (for the sake of simplicity, we always suppose that $I_n - T$ is invertible) and $\mathbf{c} \in \mathbb{C}^n$, is to apply the basic iterative method

$$\mathbf{x}_m = T\mathbf{x}_{m-1} + \mathbf{c} \quad (m = 1, 2, \dots), \qquad \mathbf{x}_0 \in \mathbb{C}^n. \tag{1}$$

Judgments about the efficiency of a scheme like (1) are usually based either upon spectral properties of the iteration matrix T or upon some norm of T.

Each of these two concepts has its merits but also its drawbacks. The asymptotic behavior of (1) depends only on $\sigma(T)$, the spectrum of T. It is for instance well known that the vectors $\mathbf{x}_m$ of (1) converge to the solution $\mathbf{x} := (I_n - T)^{-1}\mathbf{c}$, for any choice of $\mathbf{x}_0$, if and only if $\sigma(T)$ is contained in the open unit disk, and that $\rho(T)$, the spectral radius of T, measures the asymptotic decay of the associated error norms $\|\mathbf{x} - \mathbf{x}_m\|$. It is, however, also well known that $\sigma(T)$ and $\rho(T)$ can give quite misleading information about the performance of (1) for a *finite number* of iteration steps if T is not a normal matrix (cf. e.g., the example in Varga's book [32, p. 67]). Another striking example of this phenomenon originates from a model equation for convection dominated flow (cf. Farrell [6]).

Consider the boundary value problem

$$-\varepsilon u''(t) + u'(t) = f(t) \qquad \text{on } (0, 1)$$

with $u(0) = \alpha$, $u(1) = \beta$. The discretization of this problem by an upwinded scheme on a uniform mesh with step size $h = 1/(n+1)$ leads to a linear system with a nonsymmetric tridiagonal Toeplitz matrix

$$A = \operatorname{tridiag}\left(-\frac{\varepsilon}{h^2} - \frac{1}{h}, \frac{2\varepsilon}{h^2} + \frac{1}{h}, -\frac{\varepsilon}{h^2}\right) \in \mathbb{R}^{n\times n}. \tag{2}$$

The standard splitting of A, $A = D - L - U$, where D is diagonal and L and U are strictly lower and upper triangular, gives rise to two iterative schemes of the form (1), namely, the *forward Gauss-Seidel method* with iteration matrix $G_f := (D - L)^{-1}U$ and the *backward Gauss-Seidel method* with iteration matrix $G_b := (D - U)^{-1}L$. It is easy to see that both matrices, G_f and G_b, have the same spectrum and that

$$\rho(G_f) = \rho(G_b) = \frac{4\varepsilon(\varepsilon + h)}{(2\varepsilon + h)^2}\cos^2 \pi h < 1.$$

But if $\varepsilon \ll h$, for instance $\varepsilon = 10^{-6}$ and $h = 0.05$, the nonasymptotic properties of these two Gauss-Seidel methods are quite different, as can be seen from Table 1, where error norms $\|\mathbf{x} - \mathbf{x}_m\|_2$ for both methods are shown. [We chose $f(x) = 1$, $\alpha = 0$, $\beta = 1$, i.e., $u(x) = x$, and $\mathbf{x}_0 = \mathbf{0}$.]

The better performance of the forward Gauss-Seidel method in this example is usually explained by the fact that it solves the linear system in the "natural" direction, i.e., in the direction given by the characteristics of the underlying differential equation. But how can one decide whether G_f or G_b is better suited as an iteration matrix if one does not know anything about the origin of those matrices? As we have seen, the spectral properties of G_f and G_b are of no help in answering this question.

Next we assume that the basic iterative method (1) converges, i.e., $\rho(T) < 1$(T is then usually called a *convergent matrix*), and we want to know how close T is to a divergent matrix M, i.e., to one satisfying $\rho(M) \geqslant 1$. Clearly, if $\rho(T)$ is very close to 1, a small perturbation of T can lead to a divergent matrix. There are, however, convergent matrices T with $\rho(T) \ll 1$ which are "nearly divergent" (see van Loan [31] and Higham [14] for the closely related question "How near is a stable matrix to an unstable matrix?").

The $n \times n$ shift matrix

$$J_n = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ & & & & 0 \end{bmatrix} \in \mathbb{R}^{n \times n} \tag{3}$$

evidently has the spectral radius $\rho(J_n) = 0$ and therefore seems to be a "perfect" iteration matrix. But it will be shown in Section 2 that there exists a

TABLE 1

	$\|\mathbf{x} - \mathbf{x}_m\|_2$ for Gauss-Seidel	
m	Forward	Backward
0	2.5	2.5
2	6.4×10^{-8}	2.1
4	9.5×10^{-16}	1.8
6	1.0×10^{-23}	1.4
8	9.8×10^{-32}	1.1
10	8.1×10^{-40}	8.4×10^{-1}

perturbation $\Delta J_n \in \mathbb{R}^{n\times n}$,

$$\|\Delta J_n\|_2 = \sqrt{2\left[1 - \cos\left(\frac{\pi}{2n+1}\right)\right]} \sim \frac{\pi}{2n+1} \qquad (n \to \infty),$$

such that $\rho(J_n + \Delta J_n) = 1$. It is easy to construct even more dramatic examples. Let $\lambda \in \mathbb{R}$, $\lambda \neq 0$. For the matrix $T_n := \lambda J_n$, we have $\rho(T_n) = 0$. But if we change the $(n,1)$ entry of T_n to λ^{1-n} (which represents a perturbation of T_n whose norm is $|\lambda|^{1-n}$), then there results a divergent matrix (cf. Reichel and Trefethen [26], Wilkinson [34, Chapter 2]).

To summarize, spectral properties of the iteration matrix T allow conclusions only about the asymptotic behavior of the scheme (1), and they are highly sensitive to perturbations of T. None of the problems described above would have occurred if we had based our analysis on some norm $\|T\|$ of the iteration matrix T. But the use of norms has another disadvantage. $\|T\|$ gives no indication how to accelerate (1), say, by a Chebyshev method.

We here discuss the advantages and disadvantages of an "intermediate" concept to judge the performance of (1), namely the *field of values*

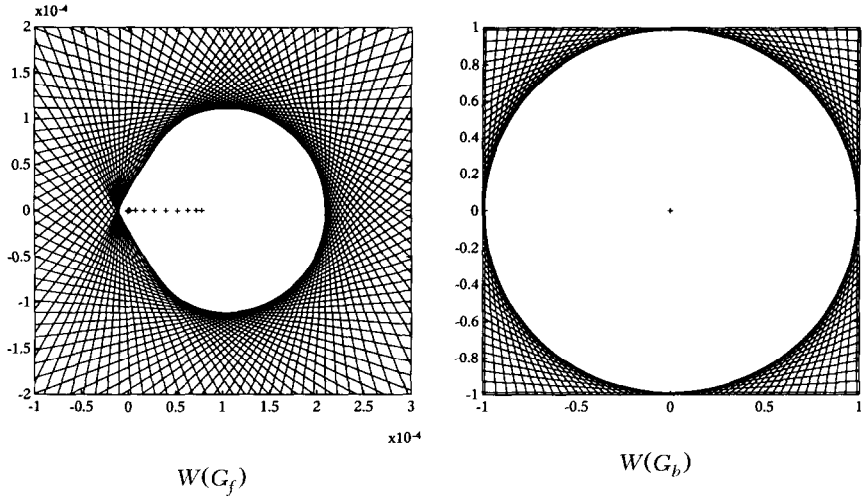
$$W(T) := \left\{\frac{\mathbf{x}^*T\mathbf{x}}{\mathbf{x}^*\mathbf{x}} : \mathbf{x} \in \mathbb{C} \setminus \{0\}\right\}$$

of T (often also called the *numerical range* of T) and the *numerical radius*

$$\mu(T) := \max\{|z| : z \in W(T)\}$$

of T. For a comprehensive discussion of the properties of $W(T)$ and $\mu(T)$, we refer to the monograph of Horn and Johnson [15, Chapter 1].

Let us come back to our introductory examples. The fields of values of the forward Gauss-Seidel matrix G_f and the backward Gauss-Seidel matrix G_b associated with A of (2) are plotted in Figure 1. (In this and also in the subsequent figures, fields of values are represented as the intersections of half planes; cf. Hausdorff [12] and Johnson [17]. An alternative method to determine the field of values numerically has been described by Marcus and Pesce [22].) $W(G_b)$ equals approximately the unit disk [$\mu(G_b) = 0.987\ldots$], suggesting that the backward Gauss-Seidel method is not well suited for this specific problem, whereas forward sweeps [$\mu(G_f) = 2.08\ldots \times 10^{-4}$] converge rapidly. For the shift matrix J_n of (3), the field of values is a disk with center 0 and radius $\cos[\pi/(n+1)]$ (cf. Lemma 3 in Section 3). Here again, $W(J_n)$ is much larger than the spectrum $\sigma(J_n)$.



$W(G_f)$ $W(G_b)$

(a) (b)

FIG. 1. Fields of values and eigenvalues (+) of the forward and backward Gauss-Seidel matrices G_f and G_b associated with A of (2) ($h = 0.05$). Note that the axes are scaled differently in the two plots.

In Section 2, we shall discuss to which extent the field of values and the numerical radius are useful tools for analyzing iterative schemes. In this paper, we restrict our analysis to Chebyshev methods or, more generally, to asymptotically stationary k-step methods (for an investigation of the alternating direction implicit (ADI) method, which is based on the field of values, see Starke [27]). There are three well-known properties of $W(T)$ which make it attractive for our purposes:

$$W(T) \text{ is always compact and convex,} \tag{4}$$

which simplifies the numerical determination of $W(T)$ considerably [although —as we shall see—the convexity is responsible for some limitations of $W(T)$];

$$\sigma(T) \subseteq W(T) \tag{5}$$

(and thus $\mathrm{Co}[\sigma(T)] \subseteq W(T)$, where $\mathrm{Co}[\Omega]$ denotes the convex hull of $\Omega \subseteq \mathbb{C}$), which relates $W(T)$ to the spectrum of T; and finally (see Lenferink and Spijker [19]),

$$\left\| (zI_n - T)^{-1} \right\|_2 \leqslant \frac{1}{\operatorname{dist}(z, W(T))} \qquad \text{for each} \quad z \notin W(T) \tag{6}$$

[where $\operatorname{dist}(\Omega, \Lambda) := \min_{z \in \Omega,\, w \in \Lambda} |z - w|$ for compact sets $\Omega, \Lambda \subset \mathbb{C}$], connecting the field of values to the growth of the resolvent.

Estimates for the fields of values are usually obtained from the Bendixson-Hirsch theorem, which requires the computation of the extreme eigenvalues of Hermitian and skew-Hermitian matrices. Section 3 is devoted to Toeplitz matrices, a class of matrices for which rather precise information about the location of the field of values can be extracted from the matrix entries alone.

Finally, Section 4 is motivated by a recent paper of Golub and de Pillis [8] on certain successive overrelaxation (SOR) methods arising from problems satisfying "Property A." We analytically determine the field of values of the SOR iteration matrices $\mathscr{L}_\omega$, which are standard examples for highly nonnormal matrices, and we are thus in the position to compute the value ω_w of the relaxation parameter which minimizes the numerical radius of $\mathscr{L}_\omega$ as a function of ω—in contrast to the classical optimal relaxation parameter ω_b, which minimizes the spectral radius $\rho(\mathscr{L}_\omega)$. In view of $\|\mathscr{L}_\omega\|_2/2 \leqslant \mu(\mathscr{L}_\omega) \leqslant \|\mathscr{L}_\omega\|_2$ (cf. Goldberg and Tadmor [7, (1.6)]), the choice of ω_w is more appropriate *at the beginning of the iteration*. Typically, there holds $\|\mathscr{L}_{\omega_b}^m\|_2 \geqslant \|\mathscr{L}_{\omega_w}^m\|_2$ for small m (cf. the example in Section 4).

2. ERROR ESTIMATES AND DISTANCE TO DIVERGENCE

Assume that the iterates $\{\mathbf{x}_m\}_{m \geqslant 0}$ of (1) converge, for any $\mathbf{x}_0$, to the solution of $\mathbf{x} = T\mathbf{x} + \mathbf{c}$, i.e., that $\rho(T) < 1$ holds. What can be said about the norms of the errors $\mathbf{e}_m = \mathbf{x} - \mathbf{x}_m$? Well-known answers to this question are, for instance, that for every $\varepsilon > 0$ there exists an integer m_ε with

$$\|\mathbf{e}_m\|_2 \leqslant [\, \rho(T) + \varepsilon \,]^m \|\mathbf{e}_0\|_2 \qquad (m = m_\varepsilon, m_\varepsilon + 1, \ldots)$$

(cf. [32, Section 3.2]), or—if T is diagonalizable—

$$\|\mathbf{e}_m\|_2 \leqslant \operatorname{cond}(R)\, \rho^m(T) \|\mathbf{e}_0\|_2 \qquad (m = 0, 1, \ldots),$$

where $\text{cond}(R) = \|R\|_2\|R^{-1}\|_2$ denotes the condition number of a matrix R such that RTR^{-1} has diagonal form. If T is not normal, neither m_ε nor $\text{cond}(R)$ is known in general, and thus the above estimates describe the behavior of $\{\mathbf{x}_m\}_{m\geq 0}$ only qualitatively. Under the stronger assumption $\|T\|_2 < 1$, we obtain the quantitative result that $\|\mathbf{e}_m\|_2 \leq \|T\|_2^m\|\mathbf{e}_0\|_2$ $(m = 0, 1, \dots)$. Since the numerical radius $\mu(T)$ has normlike properties (especially since Berger's power inequality $\mu(T^m) \leq \mu^m(T)$ is valid; cf. [25]), a similar estimate holds if $\mu(T) < 1$, i.e., if $W(T)$ is contained in the open unit disk,

$$\|\mathbf{e}_m\|_2 \leq \|T^m\|_2\|\mathbf{e}_0\|_2 \leq 2\mu(T^m)\|\mathbf{e}_0\|_2 \leq 2\mu^m(T)\|\mathbf{e}_0\|_2 \qquad (m = 0, 1, \dots).$$

The assumption $\mu(T) < 1$ is certainly less restrictive than $\|T\|_2 < 1$, but this alone would not justify the introduction of the numerical radius as a tool to analyze iterative methods. However, norms are rather useless for the investigation of many iterative schemes different from (1), whereas the numerical radius and the field of values still provide valuable information. As an example, we consider the *Chebyshev semiiterative methods* (cf. Golub and Varga [9], Manteuffel [20]).

Let $\gamma, \delta \in \mathbb{C}$ such that $1 \notin [\delta - \gamma, \delta + \gamma]$. The corresponding Chebyshev iterates are defined by

$$\begin{aligned} \mathbf{x}_1 &= \mu_{1,0}(\mathbf{c} + T\mathbf{x}_0) + (1 - \mu_{1,0})\mathbf{x}_0, \qquad \mathbf{x}_0 \in \mathbb{C}^n, \\ \mathbf{x}_m &= \mu_{m,0}(\mathbf{c} + T\mathbf{x}_{m-1}) + \mu_{m,1}\mathbf{x}_{m-1} + \mu_{m,2}\mathbf{x}_{m-2} \qquad (m = 2, 3, \dots), \end{aligned} \tag{7}$$

where the coefficients are given by

$$\mu_{1,0} = \frac{1}{1 - \delta}, \qquad \mu_{2,0} = \frac{2(1 - \delta)}{2(1 - \delta)^2 - \gamma^2},$$

$$\mu_{m,0} = \left[(1 - \delta) - \left(\frac{\gamma}{2}\right)^2 \mu_{m-1,0}\right]^{-1} \qquad (m = 3, 4, \dots), \tag{8}$$

$\mu_{m,1} = -\delta\mu_{m,0}$, and $\mu_{m,2} = 1 - (1 - \delta)\mu_{m,0}$. These procedures are called Chebyshev methods because the resulting errors $\mathbf{e}_m = \mathbf{x} - \mathbf{x}_m$ can be expressed in terms of Chebyshev polynomials $t_m(z) = \cos(m \arccos z)$, $z \in [-1, 1]$:

$$\mathbf{e}_m = p_m(T)\mathbf{e}_0, \quad \text{where } p_m(z) = \frac{t_m((z - \delta)/\gamma)}{t_m((1 - \delta)/\gamma)} \qquad (m = 0, 1, \dots).$$

The first two parts of the following theorem are merely repetitions of well-known results (cf. Manteuffel [20]). They are included here to contrast error estimates based on spectral information with error estimates derived from the field of values.

THEOREM 1. *With the Chebyshev semiiterative method defined by* (7) *and* (8), *we associate the number*

$$\kappa := \left| \frac{1 - \delta - \sqrt{(1-\delta)^2 - \gamma^2}}{\gamma} \right|$$

(*the branch of the square root has to be chosen such that* $\kappa < 1$) *and for* $\rho > 1$, *a family of elliptic regions*

$$\mathscr{E}_\rho := \left\{ z \in \mathbb{C} : |z - \delta + \gamma| + |z - \delta - \gamma| \leqslant |\gamma|(\rho + \rho^{-1}) \right\}$$

with foci $\delta \pm \gamma$ *and semiaxes* $|\gamma|(\rho \pm \rho^{-1})$. *Then*:

1. *The sequence* $\{\mathbf{x}_m\}_{m \geqslant 0}$ *of* (7) *converges, for any* $\mathbf{x}_0$, *to the solution of* $\mathbf{x} = T\mathbf{x} + \mathbf{c}$ *iff* $\sigma(T)$ *is contained in the interior of* $\partial\mathscr{E}_{1/\kappa}$ (*note that* $\partial\mathscr{E}_{1/\kappa}$ *is the unique ellipse with foci* $\delta \pm \gamma$ *passing through* $z = 1$).
2. *If in addition* T *is diagonalizable* (*assume that* R *transforms* T *into diagonal form*), *then*

$$\|\mathbf{e}_m\|_2 \leqslant 2\,\mathrm{cond}(R)\,\frac{\kappa^m}{1 - \kappa^{2m}}\|\mathbf{e}_0\|_2$$

provided that $\sigma(T) \subseteq [\delta - \gamma, \delta + \gamma]$, *and*

$$\|\mathbf{e}_m\|_2 \leqslant \mathrm{cond}(R)\,(\rho^m + \rho^{-m})\frac{\kappa^m}{1 - \kappa^{2m}}\|\mathbf{e}_0\|_2$$

provided that $\sigma(T) \subseteq \mathscr{E}_\rho$ *for some* $\rho < 1/\kappa$.
3. *For arbitrary* (*not necessarily diagonalizable*) T, *we have*

$$\|\mathbf{e}_m\|_2 \leqslant 2\frac{\kappa^m}{1 - \kappa^{2m}}\|\mathbf{e}_0\|_2$$

provided that $W(T) \subseteq [\delta - \gamma, \delta + \gamma]$, *and*

$$\|\mathbf{e}_m\|_2 \leqslant 2(\rho^m + \rho^{-m}) \frac{\kappa^m}{1 - \kappa^{2m}} \|\mathbf{e}_0\|_2$$

provided that $W(T) \subseteq \mathscr{E}_\rho$ *for some* $\rho < 1/\kappa$.

Proof. Only the third part of this theorem needs to be shown.

Assume first that $W(T) \subseteq [\delta - \gamma, \delta + \gamma]$. Then T is a normal matrix (cf. [15, Corollary 1.6.7]) and thus unitarily diagonalizable. We therefore have $\|\mathbf{e}_m\|_2 \leqslant \|p_m\|_{[\delta-\gamma, \delta+\gamma]} \|\mathbf{e}_0\|_2$, where $\|p_m\|_\Omega$ denotes the maximum norm of p_m on a compact set $\Omega \subset \mathbb{C}$. Now the assertion follows from well known estimates for Chebyshev polynomials.

Next, let $W(T) \subseteq \mathscr{E}_\rho$ for some $\rho < 1/\kappa$. The image $p_m(\partial\mathscr{E}_\rho)$ of $\partial\mathscr{E}_\rho$ under p_m is $\partial\tilde{\mathscr{E}}_{\rho^m}$ (covered exactly m times), where

$$\tilde{\mathscr{E}}_{\rho^m} := \frac{1}{t_m((1 - \delta)/\gamma)} \{z \in \mathbb{C} : |z + 1| + |z - 1| \leqslant \rho^m + \rho^{-m}\}.$$

In other words, $p_m^{-1}(\tilde{\mathscr{E}}_{\rho^m}) = \mathscr{E}_\rho$. Since both sets, $\tilde{\mathscr{E}}_{\rho^m}$ and $\mathscr{E}_\rho$, are compact and convex, it follows from a result of Kato [18, Theorem 1] that $W(p_m(T)) \subseteq \tilde{\mathscr{E}}_{\rho^m}$ and consequently $\mu(p_m(T)) \leqslant \max\{|p_m(z)| : z \in \mathscr{E}_\rho\}$. Now,

$$\|\mathbf{e}_m\|_2 \leqslant \|p_m(T)\|_2 \|\mathbf{e}_0\|_2 \leqslant 2\mu(p_m(T)) \|\mathbf{e}_0\|_2 \leqslant 2 \max_{z \in \mathscr{E}_\rho} |p_m(z)| \|\mathbf{e}_0\|_2.$$

The desired estimate follows again from well known properties of Chebyshev polynomials (cf. [4, Section 3.2]).

We finally remark that under additional assumptions on γ and δ, e.g., if γ and δ are both real, the factor $\kappa^m/(1 - \kappa^{2m})$ appearing in the above estimates can be replaced by the smaller number $\kappa^m/(1 + \kappa^{2m})$. ■

Whether an iterative scheme of the form (1) or (7) (theoretically) converges or diverges depends only on spectral properties of the matrix T. But if T is highly nonnormal, even small perturbations ΔT—which are unavoidable in practical computations—can change the spectrum dramatically.[1] An iterative method which is predicted to converge rapidly for T may well diverge if

[1] In contrast to $\sigma(T)$, the field of values is "perfectly stable", since $\max_{z \in W(T+\Delta T)} \operatorname{dist}(z, W(T)) \leqslant \|\Delta T\|_2$.

it is applied to $T + \Delta T$ (cf. Trefethen [28] for a beautiful example concerning the first order Richardson method). The question is, how close is a matrix T for which a specific iterative process converges to the set of matrices for which this process fails to converge?

To go beyond the schemes (1) and (7), we need some additional terminology. A scheme of the form

$$\mathbf{x}_m = \mu_{m,0}(T\mathbf{x}_{m-1} + \mathbf{c}) + \mu_{m,1}\mathbf{x}_{m-1} + \cdots + \mu_{m,k}\mathbf{x}_{m-k}, \tag{9}$$

where

$$\mu_{m,0} \neq 0, \quad \sum_{j=0}^{k} \mu_{m,k} = 1 \qquad (m = k, k+1, \ldots)$$

and $\mathbf{x}_0, \ldots \mathbf{x}_{k+1}$ are suitably chosen starting vectors, is called a *k-step iterative method* for the solution of $(I_n - T)\mathbf{x} = \mathbf{c}$. Here, we concentrate on *asymptotically stationary k-step methods*, i.e., schemes of the form (9) with

$$\lim_{m\to\infty} \mu_{m,j} = \mu_j \quad (j = 0, 1, \ldots, k), \qquad \mu_0 \neq 0. \tag{10}$$

With such a method, we associate a rational function (cf. Niethammer and Varga [24])

$$h(w) := \frac{1 - \mu_1 w - \cdots - \mu_k w^k}{\mu_0 w} \tag{11}$$

[note that h has a simple pole at infinity, and that $h(1) = 1$] and a family of subsets of the complex plane

$$\begin{aligned} U(h) &:= \mathbb{C}_\infty \setminus h\big(\overline{\mathbb{D}}(0;1)\big), \\ U_\eta(h) &:= \mathbb{C}_\infty \setminus h(\mathbb{D}(0;\eta)) \qquad (\eta > 1). \end{aligned} \tag{12}$$

Here, $\mathbb{C}_\infty$ denotes the complex plane together with the point at infinity, and $\mathbb{D}(\alpha;\rho)$ is the (open) disk with center α and radius ρ. $U(h)$ is an open set with $1 \in \partial U(h)$, while the $U_\eta(h)$'s are closed (for $\eta > 1$).

These definitions allow an elegant description of the convergence behavior of an asymptotically stationary k-step method (cf. [24]): A k-step method

given by (9) and (10) converges, for any choice of the initial vectors $\mathbf{x}_0, \ldots \mathbf{x}_{k-1}$, to $\mathbf{x} = (I_n - T)^{-1}\mathbf{c}$ if and only if $\sigma(T) \subseteq U(h)$. Moreover, with $\mathbf{e}_m := \mathbf{x} - \mathbf{x}_m$,

$$\kappa(h, T) := \limsup_{m \to \infty} \left[\sup_{\mathbf{e}_0 \neq \mathbf{0}} \frac{\|\mathbf{e}_m\|}{\mathbf{e}_0\|} \right]^{1/m} = \min \left\{ \frac{1}{\eta} : \eta > 1 \text{ and } \sigma(T) \subseteq U_\eta(h) \right\}.$$

Note that $\kappa(h, T) < 1$ and $\sigma(T) \subseteq U(h)$ are equivalent for every matrix $T \in \mathbb{C}^{n \times n}$ with $1 \notin \sigma(T)$. In addition, $\kappa(h, T) \leqslant 1/\eta$ if and only if $\sigma(T) \subseteq U_\eta(h)$.

For the basic iteration (1), we have $h(w) = 1/w$ and thus $U(h) = \mathbb{D}(0; 1)$, $U_\eta(h) = \overline{\mathbb{D}}(0; 1/\eta)$ $(\eta > 1)$. We therefore regain the classical result that (1) converges, for every $\mathbf{x}_0$, iff $\sigma(T) \subseteq \mathbb{D}(0; 1)$ [with the asymptotic convergence factor $\kappa(1/w, T) = \rho(T)$]. The Chebyshev method defined by (7) and (8) is an asymptotically stationary two step method (cf. [9]). With the notation of Theorem 1, $U(h)$ is the interior of $\mathscr{E}_{1/\kappa}$, and $U_\eta(h) = \mathscr{E}_{1/(\kappa\eta)}$ (if $1 < \eta < \kappa^{-1}$), $U_\eta(h) = [\delta - \gamma, \delta + \gamma]$ (if $\eta = \kappa^{-1}$), $U_\eta(h) = \varnothing$ (if $\eta > \kappa^{-1}$).

Let now h be a rational function of the form (11), and assume that $T \in \mathbb{C}^{n \times n}$ satisfies $\kappa(h, T) < 1$, i.e., the k-step method given by (9) and (10) converges. We seek the smallest perturbation ΔT of T such that $\kappa(h, T + \Delta T) \geqslant 1$, i.e.,

$$\delta_2(h, T) := \inf\{\|T - M\|_2 : M \in \mathbb{C}^{n \times n},\ \kappa(h, M) \geqslant 1\}. \tag{13}$$

A standard continuity argument [applied to $\alpha T + (1 - \alpha)M$, $0 \leqslant \alpha \leqslant 1$] yields

$$\delta_2(h, T) = \min\{\|T - M\|_2 : M \in \mathbb{C}^{n \times n},\ \kappa(h, M) = 1\}.$$

Further, $\delta_2(h, T)$ is unitarily invariant, i.e., $\delta_2(h, T) = \delta_2(h, U^*TU)$ for every unitary $U \in \mathbb{C}^{n \times n}$.

THEOREM 2. *Let h be the rational function defined by* (11), *and set $U(h) = \mathbb{C}_\infty \setminus h(\overline{\mathbb{D}}(0; 1))$. For $T \in \mathbb{C}^{n \times n}$ with $\sigma(T) \subset U(h)$, i.e., $\kappa(h, T) < 1$,*

$$\delta_2(h, T) = \min_{z \in \partial U(h)} s_{\min}(zI_n - T),$$

where $s_{\min}(M)$ denotes the smallest singular value of $M \in \mathbb{C}^{n \times n}$. Moreover, $\delta_2(h, T)$ can be estimated by

$$\operatorname{dist}(W(T), \partial U(h)) \leqslant \delta_2(h, T) \leqslant \operatorname{dist}(\sigma(T), \partial U(h)).$$

Proof. Let $\delta_2(h, T) = \|T - M\|_2$, where $\kappa(h, M) = 1$, i.e., there exists a $z_0 \in \sigma(M) \cap \partial U(h)$. Therefore $z_0 I_n - M$ is singular, and $s_{\min}(z_0 I_n - T)$, the distance of the nonsingular matrix $z_0 I_n - T$ [note that $\sigma(T) \subseteq U(h)$] to the collection of all singular matrices, is dominated by $\delta_2(h, T)$:

$$s_{\min}(z_0 I_n - T) \leqslant \|(z_0 I_n - T) - (z_0 I_n - M)\|_2 = \|T - M\|_2 = \delta_2(h, T).$$

On the other hand, for each $z \in \partial U(h)$, there is a singular matrix $S_z \in \mathbb{C}^{n\times n}$ with $s_{\min}(zI_n - T) = \|zI_n - T - S_z\|_2$. Since $z \in \partial U(h) \cap \sigma(zI_n - S_z)$ and thus $\kappa(h, zI_n - S_z) \geqslant 1$, we conclude that $s_{\min}(zI_n - T) = \|T - (zI_n - S_z)\|_2 \geqslant \delta_2(h, T)$. Now, $\delta_2(h, T) = \min_{z\in\partial U(h)} s_{\min}(zI_n - T)$ is shown.

We next prove that $\operatorname{dist}(W(T), \partial U(h)) \leqslant \delta_2(h, T)$. If $\operatorname{dist}(W(T), \partial U(h)) = 0$, then there is nothing to show. If $\operatorname{dist}(W(T), \partial U(h)) > 0$, we choose ε with $0 < \varepsilon < \operatorname{dist}(W(T), \partial U(h))$, and $M \in \mathbb{C}^{n\times n}$ with $\|M - T\|_2 \leqslant \varepsilon$. Since $W(M) \subseteq W(T) + W(M - T)$ and $W(M - T) \subseteq \overline{\mathbb{D}}(0; \varepsilon)$, it follows that $W(M) \subseteq U(h)$ and thus $\sigma(M) \subseteq U(h)$. In other words, $\kappa(h, M) < 1$ for all M satisfying $\|T - M\|_2 \leqslant \varepsilon$.

Finally, we come to $\delta_2(h, T) \leqslant \operatorname{dist}(\sigma(T), \partial U(h))$. Since $\delta_2(h, \cdot)$ is unitarily invariant we may assume that T is given in Schur form, i.e., $T = D + N$, where $D = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ is a diagonal matrix and N is strictly upper triangular. We further assume that

$$\operatorname{dist}(\sigma(T), \partial U(h)) = \operatorname{dist}(\lambda_1, \partial U(h)) = |\lambda_1 - z|$$

with $z \in \partial U(h)$. For $M := \operatorname{diag}(z, \lambda_2, \ldots, \lambda_n) + N$, we obtain $\kappa(h, M) \geqslant 1$ and $\|T - M\|_2 = |\lambda_1 - z| = \operatorname{dist}(\sigma(T), \partial U(h))$. ∎

As an example, we consider the $n \times n$ shift matrix [cf. (3)] and the basic iterative method (1), i.e., $h(w) = 1/w$, $\partial U(h) = \{|z| = 1\}$. The singular values of $(e^{i\theta} I_n - J_n)$, $0 \leqslant \theta < 2\pi$, are independent of θ. We therefore have

$$\delta_2\left(\frac{1}{w}, J_n\right) = s_{\min}(I_n - J_n) = \lambda_{\min}^{1/2}\left((I_n - J_n)^T (I_n - J_n)\right).$$

But $(I_n - J_n)^T (I_n - J_n)$ is the inverse of Frank's matrix (cf. [33, Appendix C]), and its eigenvalues are known to be

$$\lambda_j = 2\left[1 - \cos\left(\frac{(2j-1)\pi}{2n+1}\right)\right] \qquad (j = 1, 2, \ldots, n).$$

This implies $\delta_2(1/w, J_n) = \sqrt{2\{1 - \cos[\pi/(2n+1)]\}}$. From Theorem 2, it follows that

$$1 - \cos\left(\frac{\pi}{n+1}\right) = 1 - \mu(J_n) \leqslant \delta_2\left(\frac{1}{w}, J_n\right) \leqslant 1 - \rho(J_n) = 1$$

(cf. Lemma 3); thus both bounds differ from the correct value of $\delta_2(1/w, J_n)$ by an order of magnitude [e.g., $1 - \mu(J_{10}) = 0.040\ldots$ and $\delta_2(1/w, J_{10}) = 0.149\ldots$].

There is no doubt that the concept of pseudospectra leads to more satisfactory theoretical results. Trefethen (cf. [28], [29], and [30]) defined $\Lambda_\varepsilon(T)$, the ε-pseudospectrum of T, by

$$\Lambda_\varepsilon(T) := \{\lambda \in \mathbb{C} : \lambda \in \sigma(T + \Delta T) \text{ for some } \Delta T \text{ with } \|\Delta T\|_2 \leqslant \varepsilon\}.$$

As an immediate consequence of this definition,

$$\delta_2(h, T) = \sup\{\varepsilon > 0 : \Lambda_\varepsilon(T) \subset U(h)\}.$$

There are natural relationships between the field of values and the pseudospectra of a matrix T (cf. Trefethen [30]). As in the above question (what effect does a perturbation of T have on the convergence of an iterative method?), pseudospectra often provide the exact answer, whereas the field of values leads merely to upper or lower bounds. However, fields of values are in general much easier to compute than pseudospectra.

3. FIELDS OF VALUES OF TOEPLITZ MATRICES

The field of values of a nonnormal matrix is in general much larger than the convex hull of its spectrum. This well-known fact can be easily illustrated within the class of Toeplitz matrices

$$T_n = \begin{bmatrix} \tau_0 & \tau_1 & \cdots & \tau_{n-1} \\ \tau_{-1} & \tau_0 & & \vdots \\ \vdots & & & \tau_1 \\ \tau_{1-n} & \cdots & \tau_{-1} & \tau_0 \end{bmatrix} \in \mathbb{C}^{n\times n}. \tag{14}$$

As a first example, we consider powers of the $n \times n$ shift matrix J_n [cf. (3)] whose spectrum is the singleton $\{0\}$, whereas for large dimensions n, its field of values is approximately the unit disk.

LEMMA 3. *Let $J_n \in \mathbb{R}^{n\times n}$ denote the $n \times n$ shift matrix (cf. (3)), and let k be an integer with $1 \leqslant k \leqslant n-1$. Then*

$$W(J_n^k) = \overline{\mathbb{D}}\left(0; \cos\left(\frac{\pi}{[(n-1)/k]+2}\right)\right),$$

where $[r]$ denotes the largest integer m with $m \leqslant r$.

Proof. Let $w \in W(J_n^k)$, i.e., there is a vector $\mathbf{x} \in \mathbb{C}^n$, $\|\mathbf{x}\|_2 = 1$, with $w = \mathbf{x}^* J_n^k \mathbf{x}$. If x_j $(j = 1, 2, \ldots, n)$ are the components of $\mathbf{x}$, we define a vector $\mathbf{y}$ by $y_j := e^{ij\varphi} x_j$ $(j = 1, 2, \ldots, n)$, where $0 \leqslant \varphi < 2\pi$ is an arbitrary angle. Since $\|\mathbf{y}\|_2 = 1$, it follows that

$$e^{ik\varphi} w = e^{ik\varphi} \mathbf{x}^* J_n^k \mathbf{x} = e^{ik\varphi} \sum_{j=1}^{n-k} \bar{x}_j x_{j+k} = \sum_{j=1}^{n-k} e^{-ij\varphi} \bar{x}_j e^{i(j+k)\varphi} x_{j+k}$$

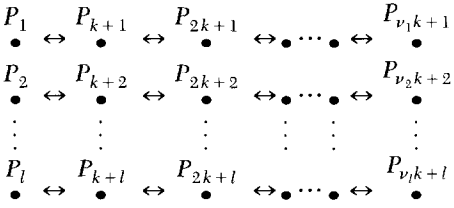
$$= \sum_{j=1}^{n-k} \bar{y}_j y_{j+k} = \mathbf{y}^* J_n^k \mathbf{y} \in W(J_n^k),$$

and thus, $W(J_n^k)$ is a disk centered at the origin.

According to the Bendixson-Hirsch theorem (cf. [21, 5.2.7]), its radius $\rho_{n,k}$ is equal to the spectral radius of $H_{n,k} := (J_n^k + (J_n^k)^T)/2 \in \mathbb{R}^{n\times n}$, the symmetric part of J_n^k. For $k = 1$,

$$\rho_{n,1} = \rho(H_{n,1}) = \cos\left(\frac{\pi}{n+1}\right)$$

(cf. [5]). For $k > 1$, the directed graph (cf. [32, p. 19]) of $H_{n,k}$ has the structure



with $l := \min\{k, n-k\}$ and $\nu_j := [(n-j)/k]$ $(j = 1, 2, \ldots, l)$. There therefore exists a permutation matrix P such that

$$P^* H_{n,k} P = \begin{bmatrix} H_{\nu_1+1,1} & & & \\ & H_{\nu_2+1,1} & & \\ & & \ddots & \\ & & & H_{\nu_l+1,1} \end{bmatrix}$$

is block diagonal. Since $\rho_{\nu,1}$ is a monotonically increasing function of ν and since $\nu_1 = \max\{\nu_j : j = 1, 2, \ldots, l\}$, we conclude

$$\rho_{n,k} = \rho(H_{n,k}) = \rho(H_{\nu_1+1,1}) = \rho_{\nu_1+1,1}$$

$$= \cos\left(\frac{\pi}{\nu_1 + 2}\right) = \cos\left(\frac{\pi}{[(n-1)/k] + 2}\right). \quad \blacksquare$$

As a direct consequence of Lemma 3, we obtain the field of values of a Toeplitz tridiagonal matrix.

COROLLARY 4. *The field of values of* $\operatorname{tridiag}(\alpha, 0, \beta) \in \mathbb{C}^{n\times n}$ *is the closed interior of the ellipse*

$$\mathscr{E}_n(\alpha, \beta) := \left\{ z \in \mathbb{C} : z = \cos\left(\frac{\pi}{n+1}\right)(\alpha e^{-i\theta} + \beta e^{i\theta}),\ 0 \leqslant \theta < 2\pi \right\}.$$

Proof. For $w \in W(\operatorname{tridiag}(\alpha, 0, \beta))$, there exists a vector $\mathbf{x} \in \mathbb{C}^n$, $\|\mathbf{x}\|_2 = 1$, such that

$$w = \mathbf{x}^* \operatorname{tridiag}(\alpha, 0, \beta)\, \mathbf{x} = \alpha \mathbf{x}^* J_n^* \mathbf{x} + \beta \mathbf{x}^* J_n \mathbf{x}$$

$$= \alpha \overline{\mathbf{x}^* J_n \mathbf{x}} + \beta \mathbf{x}^* J_n \mathbf{x} = \alpha \bar{z} + \beta z$$

with $z \in W(J_n)$— and vice versa, $\alpha\bar{z} + \beta z$ belongs to $W(\operatorname{tridiag}(\alpha, 0, \beta))$ for every $z \in W(J_n)$. Now the assertion follows from Lemma 3. $\blacksquare$

We next derive inclusion sets for the field of values of an arbitrary Toeplitz matrix T_n which do not require any eigenvalue computations. A first estimate for $W(T_n)$ can be derived by "embedding" $T_n \in \mathbb{C}^{n\times n}$ into a circulant matrix of order $2n-1$ (cf. [1, Satz 9.2]), or more generally, into a

$\{z\}$-circulant matrix $Z_{2n-1} \in \mathbb{C}^{(2n-1)\times(2n-1)}$. Recall that a matrix $C = \{z\}\text{-circ}(\gamma_0, \gamma_1, \ldots, \gamma_{n-1}) \in \mathbb{C}^{n\times n}$ is $\{z\}$-circulant if it has the form

$$C = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{n-2} & \gamma_{n-1} \\ z\gamma_{n-1} & \gamma_0 & \gamma_1 & \cdots & \gamma_{n-3} & \gamma_{n-2} \\ z\gamma_{n-2} & z\gamma_{n-1} & \gamma_0 & \cdots & \gamma_{n-4} & \gamma_{n-3} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ z\gamma_2 & z\gamma_3 & z\gamma_4 & \cdots & \gamma_0 & \gamma_1 \\ z\gamma_1 & z\gamma_2 & z\gamma_3 & \cdots & z\gamma_{n-1} & \gamma_0 \end{bmatrix}$$

(cf. Davis [3, p. 84]).

Let f denote the *symbol* of the Toeplitz matrix T_n of (14),

$$f(z) := \sum_{j=-n+1}^{n-1} \tau_j z^j. \tag{15}$$

As we shall see, there is an intimate relationship between $W(T_n)$ and the image of the unit circle under f. The following estimate implies that $W(T_n)$ is always a subset of $\text{Co}[f(|z| = 1)]$.

LEMMA 5. *For an arbitrary* $0 \leqslant \theta < 2\pi$, *let* $\zeta_1, \zeta_2, \ldots, \zeta_{2n-1}$ *denote the* $(2n-1)$*th roots of* $z := e^{i\theta}$. *Then*

$$W(T_n) \subseteq \mathscr{P}_z(T_n) := \text{Co}\big[\{f(\zeta_1), f(\zeta_2), \ldots, f(\zeta_{2n-1})\}\big].$$

Proof. We augment T_n of (14) to the $\{z\}$-circulant matrix

$$Z_{2n-1} = \{z\}\text{-circ}(\tau_0, \ldots, \tau_{n-1}, \bar{z}\tau_{1-n}, \ldots, \bar{z}\tau_{-1}) \in \mathbb{C}^{(2n-1)\times(2n-1)}.$$

Since T_n is a principal submatrix of Z_{2n-1}, we have $W(T_n) \subseteq W(Z_{2n-1})$. Now Z_{2n-1} is a normal matrix ($|z| = 1$), and the eigenvalues of Z_{2n-1} are (note that $\bar{z}\zeta_j^{2n-1} = 1$)

$$\begin{aligned} &\tau_0 + \tau_1\zeta_j + \cdots + \tau_{n-1}\zeta_j^{n-1} + \bar{z}\tau_{1-n}\zeta_j^{n} + \bar{z}\tau_{2-n}\zeta_j^{n+1} + \cdots + \bar{z}\tau_{-1}\zeta_j^{2n-2} \\ &\quad = \tau_0 + \tau_1\zeta_j + \cdots + \tau_{n-1}\zeta_j^{n-1} + \tau_{1-n}\zeta_j^{1-n} \\ &\qquad + \tau_{2-n}\zeta_j^{2-n} + \cdots + \tau_{-1}\zeta_j^{-1} \\ &\quad = f(\zeta_j) \qquad (j = 1, 2, \ldots, 2n-1). \end{aligned}$$

■

For $T_n = J_n$ [cf. (3)], an elementary geometric consideration shows that

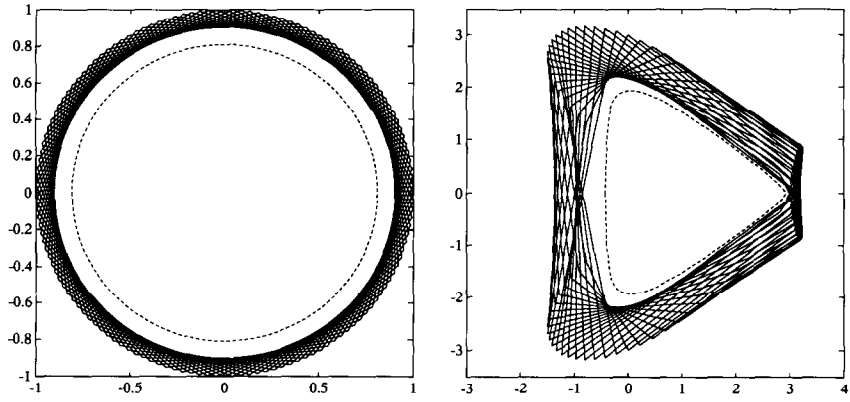
$$\bigcap_{|z|=1} \mathscr{P}_z(J_n) = \overline{\mathbb{D}}\left(0; \cos\left(\frac{\pi}{2n+1}\right)\right),$$

which overestimates $W(J_n)$ by a quantity of order n^{-2} $(n \to \infty)$ (cf. Lemma 3 and Figure 2). In Figure 2, also inclusion sets $\mathscr{P}_z$ for the field of values of

$$R_4 = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 2 & 1 & 0 & 1 \\ 0 & 2 & 1 & 0 \\ -1 & 0 & 2 & 1 \end{bmatrix} \in \mathbb{R}^{4\times 4} \tag{16}$$

are shown.

For a general matrix $A \in \mathbb{C}^{n\times n}$, $W(A)$ can be estimated by the Bendixson-Hirsch theorem, which is based on a splitting of A into a sum of two normal matrices, namely its Hermitian part A_H and its skew-Hermitian part A_S. To determine $W(A_H)$ and $W(A_S)$, the extremal eigenvalues of these matrices have to be computed, and finally, $W(A) \subseteq W(A_H) + W(A_S)$. For Toeplitz matrices T_n, other additive decompositions into normal matrices are possible. It is well known that T_n can be split into its circulant and



$W(J_4)$ (a) $W(R_4)$ (b)

FIG. 2. Field of values (bounded by dashed curve) and inclusion set $\bigcap \mathscr{P}_z$, where $z = \exp(2\pi ik/20)$ $(k = 1, 2, \ldots, 20)$ (cf. Lemma 5) for the matrices J_4 [cf. (3)] and R_4 [cf. (16)].

skew-circulant parts, or more generally, T_n can be represented as the sum of a $\{z\}$-circulant and a $\{w\}$-circulant matrix as long as $z \neq w$.

LEMMA 6. *Let the symbol f (cf. (15)) of the Toeplitz matrix $T_n \in \mathbb{C}^{n\times n}$ (cf. (14)) be split according to $f = f_- + f_+$ with*

$$f_-(z) := \sum_{j=-n+1}^{-1} \tau_j z^j \quad \text{and} \quad f_+(z) := \sum_{j=0}^{n-1} \tau_j z^j.$$

Then, for every pair of numbers $z, w \in \mathbb{C}$ with $z \neq w$ and $|z| = |w| = 1$, one has $W(T_n) \subseteq \mathscr{E}_{z,w}(T_n)$, where

$$\mathscr{E}_{z,w}(T_n) := \operatorname{Co}\left[\left\{\frac{z(f_-(\zeta_k) + f_+(\xi_l)) - w(f_-(\xi_l) + f_+(\zeta_k))}{z - w} : 1 \leqslant k, l \leqslant n\right\}\right],$$

where $\zeta_1, \zeta_2, \ldots, \zeta_n$ are the nth roots of z, and where $\xi_1, \xi_2, \ldots, \xi_n$ are the nth roots of w.

Proof. Upon setting $\gamma_0 = -w\tau_i/(z - w)$, $\sigma_0 = z\tau_i/(z - w)$,

$$\gamma_i = \frac{\tau_{i-n} - w\tau_i}{z - w}, \quad \sigma_i = \frac{z\tau_i - \tau_{i-n}}{z - w} \qquad (i = 1, 2, \ldots, n - 1),$$

and

$$C_n = \{z\}\text{-circ}(\gamma_0, \gamma_1, \ldots, \gamma_{n-1}), \qquad S_n = \{w\}\text{-circ}(\sigma_0, \sigma_1, \ldots, \sigma_{n-1}),$$

we see that $T_n = C_n + S_n$. The fields of values of the circulant components C_n and S_n are known explicitly:

$$W(C_n) = \frac{1}{z - w} \operatorname{Co}\left[\{zf_-(\zeta_k) - wf_+(\zeta_k) : 1 \leqslant k \leqslant n\}\right]$$

(cf. Davis [3, p. 84]), and—analogously—

$$W(S_n) = \frac{1}{z - w} \operatorname{Co}\left[\{zf_+(\xi_l) - wf_-(\xi_l) : 1 \leqslant l \leqslant n\}\right].$$

Thus $\mathscr{E}_{z,w}(T_n) = W(C_n) + W(S_n)$ is another enclosure of $W(T_n)$ which can be constructed directly from the entries of T_n. ∎

For the matrices J_4 [cf. (3)] and R_4 [cf. (16)], enclosures of the field of values resulting from Lemma 6 are shown in Figure 3.

With the Toeplitz matrix $T_n = (\tau_{k-l})_{1 \leqslant k, l \leqslant n}$ of (14), we associate a sequence $\{T_m\}_{m \geqslant n}$ of banded Toeplitz matrices,

$$T_m = (\tau_{k-l})_{1 \leqslant k, l \leqslant m} \in \mathbb{C}^{m \times m} \quad \text{with} \quad \tau_j = 0 \quad \text{for } |j| \geqslant n. \tag{17}$$

Obviously (cf. Lemma 5),

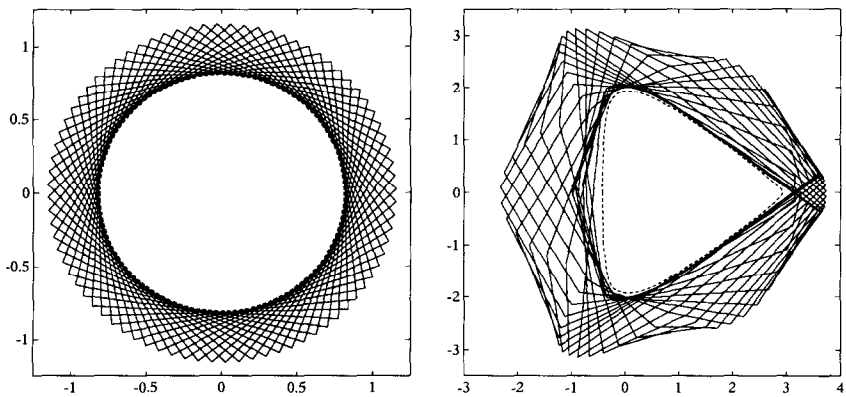
$$W(T_n) \subseteq W(T_{n+1}) \subseteq W(T_{n+2}) \subseteq \cdots \subseteq \mathrm{Co}[\Gamma],$$

where $\Gamma := f(|z| = 1)$ [cf. (15)]. Moreover, it is easy to see that

$$d(\mathrm{Co}[\Gamma], W(T_m)) \to 0 \qquad (\text{for} \quad m \to \infty), \tag{18}$$

where $d(\cdot, \cdot)$ denotes the *Hausdorff distance* of two compact subsets of the plane (cf. [11, p. 115]).

The question we want to address here is how fast the convergence in (18) is. As an example we first consider $T_m = \mathrm{tridiag}(\alpha, 0, \beta)$ $(m = 2, 3, \ldots)$.



$W(J_4)$ (a) $W(R_4)$ (b)

FIG. 3. Field of values (bounded by dashed curve) and inclusion set $\bigcap \mathscr{E}_{z \neq w}$, where $z = \exp(2\pi ik/20)$ and $w = \exp[2\pi i(k/20 + \frac{1}{3}]$ $(k = 1, 2, \ldots, 20)$ (cf. Lemma 6) for the matrices J_4 [cf. (3)] and R_4 [cf. (16)].

Here, $\mathrm{Co}[\Gamma]$ is the closed interior of the ellipse

$$\mathscr{E}_\infty(\alpha, \beta) := \{z \in \mathbb{C} : z = (\alpha e^{i\theta} + \beta e^{-i\theta}), 0 \leqslant \theta < 2\pi\}.$$

This, together with Corollary 4, implies

$$d(\mathrm{Co}[\Gamma], W(T_m)) = (|\alpha| + |\beta|)\left[1 - \cos\left(\frac{\pi}{m+1}\right)\right]$$

$$= (|\alpha| + |\beta|)\frac{\pi^2}{2}\left(\frac{1}{m+1}\right)^2 + O(m^{-3}) \qquad (m \to \infty). \tag{19}$$

The following theorem states that this asymptotic behavior is valid for every banded Toeplitz matrix.

THEOREM 7. *Let the sequence $\{T_m\}_{m \geqslant n}$ of banded Toeplitz matrices be given by* (14) *and* (17), *and further, denote by Γ the image of the unit circle under the symbol f (cf.* (15)). *Then*

$$d(\mathrm{Co}[\Gamma], W(T_m)) \leqslant C_n\left(\frac{1}{m+1}\right)^2 + O(m^{-3}) \qquad (m \to \infty)$$

with

$$C_n := \pi^2 \max_{1 \leqslant j \leqslant n-1} \{|\tau_j| + |\tau_{-j}|\}\frac{(n-1)n(2n-1)}{12}$$

(*recall that n is the bandwidth of T_m*). *Equation* (19) *implies that in general, the constant C_n cannot be replaced by a smaller one.*

Proof. For $0 \leqslant \varphi < 2\pi$, we consider the vector $\mathbf{x}(\varphi) = (x_1, \ldots, x_m)^T \in \mathbb{C}^m$ whose components are

$$x_j := \sin\left(\frac{\pi j}{m+1}\right)e^{ij\varphi} \qquad (j = 1, 2, \ldots, m), \tag{20}$$

and the closed curve

$$\Gamma_m := \left\{ \frac{\mathbf{x}^*(\varphi) T_m \mathbf{x}(\varphi)}{\mathbf{x}^*(\varphi)\mathbf{x}(\varphi)} : 0 \leqslant \varphi < 2\pi \right\} \subseteq W(T_m)$$

$(m = n, n+1, \ldots)$. It is sufficient to show that, for each $z \in \Gamma$,

$$\operatorname{dist}(z, \Gamma_m) \leqslant C_n \left(\frac{1}{m+1} \right)^2 + O(m^{-3}) \qquad (m \to \infty).$$

To this end, we fix $\varphi \in [0, 2\pi)$ and $z = f(e^{i\varphi}) \in \Gamma$. Then

$$\left| z - \frac{\mathbf{x}^*(\varphi) T_m \mathbf{x}(\varphi)}{\mathbf{x}^*\mathbf{x}} \right| = \left| \sum_{j=1}^{n-1} \left(\tau_{-j} e^{-ij\varphi} + \tau_j e^{ij\varphi} \right) \left[1 - \frac{\mathbf{x}^*(0) J_m^j \mathbf{x}(0)}{\mathbf{x}^*(0)\mathbf{x}(0)} \right] \right|$$

$$\leqslant \max_{1 \leqslant j \leqslant n-1} \left\{ |\tau_j| + |\tau_{-j}| \right\} \sum_{j=1}^{n-1} \left| 1 - \frac{\mathbf{x}^*(0) J_m^j \mathbf{x}(0)}{\mathbf{x}^*(0)\mathbf{x}(0)} \right|. \tag{21}$$

Using trigonometric identities (cf. [10, 1.351.1]), we obtain for $\mathbf{x} := \mathbf{x}(0)$

$$\mathbf{x}^* J_m^j \mathbf{x} = \sum_{l=1}^{m-j} \bar{x}_l x_{l+j} = \frac{1}{2} \left[(m-j) \cos\left(\frac{j\pi}{m+1} \right) + \frac{\sin[(j+1)\pi/(m+1)]}{\sin[\pi/(m+1)]} \right],$$

$$\mathbf{x}^*\mathbf{x} = \sum_{l=1}^{m} \bar{x}_l x_l = \frac{m+1}{2},$$

and thus

$$1 - \frac{\mathbf{x}^* J_m^j \mathbf{x}}{\mathbf{x}^*\mathbf{x}} = \frac{m-j}{m+1} \left[1 - \cos\left(\frac{j\pi}{m+1} \right) \right]$$

$$+ \left[\frac{j+1}{m+1} - \frac{\sin[(j+1)\pi/(m+1)]}{(m+1)\sin[\pi/(m+1)]} \right]$$

$$= \frac{j^2\pi^2}{2} \left(\frac{1}{m+1} \right)^2 + O(m^{-3}) \qquad (m \to \infty).$$

Inserting into (21) leads to the desired conclusion

$$\left| z - \frac{\mathbf{x}^*(\varphi) T_m \mathbf{x}(\varphi)}{\mathbf{x}^*(\varphi)\mathbf{x}(\varphi)} \right| \leqslant C_n \left(\frac{1}{m+1} \right)^2 + O(m^{-3}). \qquad \blacksquare$$

Theorem 7 can be interpreted as a statement concerning the speed with which the fields of values $W(T_m)$ of the finite Toeplitz matrices T_m approaches $W(T_\infty)$, the field of values of the associated semiinfinite banded Toeplitz matrix $T_\infty = (\tau_{k-l})_{1 \leqslant k, l < \infty}$, which is a linear bounded operator on l^2. The spectrum $\sigma(T_\infty)$, of T_∞ has an elegant characterization (cf. Calderón, Spitzer, and Widom [2]), namely

$$\sigma(T_\infty) = \{ z \in \mathbb{C} : z \in \Gamma \quad \text{or} \quad n(\Gamma, z) \neq 0 \},$$

where $n(\Gamma, z)$ denotes the winding number of z with respect to Γ. The matrix T_∞ is in general not normal; however, there holds (cf. Halmos [11, Chapter 20])

$$\overline{W(T_\infty)} = \mathrm{Co}[\sigma(T_\infty)] = \mathrm{Co}[\Gamma].$$

Thus, Theorem 7 reads as

$$d\left(\overline{W(T_\infty)}, W(T_m) \right) \leqslant C_n \left(\frac{1}{m+1} \right)^2 + O(m^{-3}) \qquad (m \to \infty).$$

To show that a statement like that is no longer valid if T_∞ is not banded, we consider $T_m = (\tau_{k-l})_{1 \leqslant k, l \leqslant m}$ $(m = 1, 2, \ldots)$ together with $T_\infty = (\tau_{k-l})_{1 \leqslant k, l < \infty}$, where

$$\tau_j = \begin{cases} 0 & \text{if} \quad j < 0, \\ 1 & \text{if} \quad j = 0, \\ 2 & \text{if} \quad j > 0, \end{cases} \tag{22}$$

Toeplitz matrices which have been investigated previously by Reichel and Trefethen [26]. Here, $\overline{W(T_\infty)}$ is the right half plane $\{\mathrm{Re}\, z \geqslant 0\}$, and the Bendixson-Hirsch theorem implies that $W(T_m) \subset R_m$, where R_m denotes the rectangle with vertices $\pm i[\sin(\pi/m)]/[1 - \cos(\pi/m)]$ and $m \pm i[\sin(\pi/m)]/[1 - \cos(\pi/m)]$. On the other hand, $\{0, m\} \subset W(T_m)$, and for

$$\mathbf{x} = \frac{1}{\sqrt{m}} (1, \omega, \omega^2, \ldots, \omega^{m-1}), \qquad \text{where} \quad \omega := \exp\left(i \frac{\pi}{m} \right),$$

there holds

$$\mathbf{x}^* T_m \mathbf{x} = \frac{2}{m} \frac{1}{1 - \cos(\pi/m)} + i \frac{\sin(\pi/m)}{1 - \cos(\pi/m)} \in W(T_m),$$

because $\mathbf{x}^*\mathbf{x} = 1$. Therefore $W(T_m)$ contains the quadrilateral Q_m with vertices 0, m, $\mathbf{x}^* T_m \mathbf{x}$, and $\overline{\mathbf{x}^* T_m \mathbf{x}}$. It is now more appropriate to consider these sets on the Riemann sphere $\mathbb{C}_\infty$ rather than in the complex plane. Let $d_\chi(\cdot, \cdot)$ denote the Hausdorff metric based on the chordal distance $\chi(\cdot, \cdot)$ (cf. [13, p. 311]). Simple geometric considerations lead us to

$$d_\chi\left(\overline{W(T_\infty)}, R_m\right) = \chi(\infty, m) = \frac{2}{\sqrt{m^2 + 1}} \sim \frac{2}{m}$$

and

$$d_\chi\left(\overline{W(T_\infty)}, Q_m\right) \geqslant d_\chi\left(\overline{W(T_\infty)}, \mathbf{x}^* T_m \mathbf{x}\right) > \frac{1}{m},$$

implying that $d_\chi(\overline{W(T_\infty)}, W(T_m))$ tends to zero like $1/m$, which is an order of magnitude slower than the speed of convergence in Theorem 7 (cf. Figure 4).

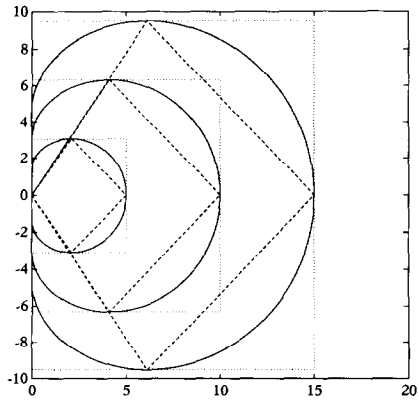


FIG. 4. $\partial W(T_m)$ (solid curve), ∂R_m (dotted curve), and ∂Q_m (dashed curve) for the Toeplitz matrix T_m defined by (22) ($m = 5, 10, 15$).

4. FIELDS OF VALUES OF CERTAIN SOR ITERATION MATRICES

We first consider symmetric matrices $A = I_n - B \in \mathbb{R}^{n\times n}$ with *property A*, i.e.,

$$B = \begin{bmatrix} 0 & M \\ M^T & 0 \end{bmatrix} \qquad (\text{with} \quad M \in \mathbb{R}^{p\times q}, \quad p \geqslant q, \quad p + q = n), \tag{23}$$

and the associated SOR iteration matrices

$$\mathscr{L}_\omega = \begin{bmatrix} (1-\omega) I_p & \omega M \\ \omega(1-\omega) M^T & (1-\omega) I_q + \omega^2 M^T M \end{bmatrix}, \tag{24}$$

where ω, $0 < \omega < 2$, denotes the relaxation parameter.

D. M. Young's identity [35],

$$(\lambda + \omega - 1)^2 = \lambda \omega^2 \mu^2, \tag{25}$$

relates the eigenvalues λ of $\mathscr{L}_\omega$ to the eigenvalues μ of the block Jacobi matrix B [cf. (23)]. Simple examples show that (25) is not valid for the field of values: There is not always, for a given $\mu \in W(B)$, a $\lambda \in W(\mathscr{L}_\omega)$ fulfilling (25); nor is it generally possible to determine, for $\lambda \in W(\mathscr{L}_\omega)$, a number $\mu \in W(B)$ such that (25) is satisfied. The following theorem, however, shows that $W(\mathscr{L}_\omega)$ is uniquely determined by $W(B) = [-\|B\|_2, \|B\|_2]$ and by the relaxation parameter ω.

THEOREM 8. *For the matrix $\mathscr{L}_\omega$ of (24), there holds $W(\mathscr{L}_\omega) = \mathscr{E}(\omega, \|B\|_2)$. Here, $\mathscr{E}(\omega, \|B\|_2)$ denotes the closed interior of the ellipse with center*

$$c(\omega, \|B\|_2) := 1 - \omega + \tfrac{1}{2}\omega^2 \|B\|_2^2$$

and semiaxes

$$a(\omega, \|B\|_2) := \tfrac{1}{2}\omega \|B\|_2 \left[\omega^2 \|B\|_2^2 + (2-\omega)^2\right]^{1/2} \qquad (\textit{on the real axis}),$$

$$b(\omega, \|B\|_2) := \tfrac{1}{2}\omega^2 \|B\|_2 \qquad (\textit{perpendicular to the real axis}).$$

Proof. Let $\|B\|_2 = s_1 \geqslant s_2 \geqslant \ldots \geqslant s_q$ be the singular values of M [cf. (23)]. Golub and de Pillis showed in [8] that $\mathscr{L}_\omega$ of (24) is unitarily similar to the block diagonal matrix

$$\begin{bmatrix} M(\omega, s_1) & & & & \\ & M(\omega, s_2) & & & \\ & & \ddots & & \\ & & & M(\omega, s_q) & \\ & & & & (1-\omega) I_{p-q} \end{bmatrix},$$

where—for $0 < \omega < 2$ and $s > 0$—

$$M(\omega, s) = \begin{bmatrix} 1-\omega & \omega s \\ w(1-\omega)s & 1 - \omega + \omega^2 s^2 \end{bmatrix} \in \mathbb{R}^{2\times 2}.$$

Since the field of values is unitarily invariant, and since the field of values of a block diagonal matrix is the convex hull of the union of the fields of values of the diagonal blocks, there follows

$$W(\mathscr{L}_\omega) = \operatorname{Co}\left[\bigcup_{j=1}^{q} W\big(M(\omega, s_j)\big) \cup \{1-\omega\}\right] = \operatorname{Co}\left[\bigcup_{j=1}^{q} W\big(M(\omega, s_j)\big)\right], \tag{26}$$

because $1 - \omega$ is a diagonal entry of $M(\omega, s)$, and thus $1 - \omega \in W(M(\omega, s))$ for every $s > 0$.

We intend to show that, for each $0 < \omega < 2$,

$$s < t \quad \text{implies} \quad W(M(\omega, s)) \subset W(M(\omega, t)). \tag{27}$$

To this end, we introduce the matrix

$$N(\omega, s) = s\begin{bmatrix} 0 & 1 \\ 1-\omega & \omega s \end{bmatrix} \in \mathbb{R}^{2\times 2}.$$

Since $M(\omega, s) = (1-\omega)I_2 + \omega N(\omega, s)$, it is sufficient to prove that

$$s < t \quad \text{implies} \quad W(N(\omega, s)) \subset W(N(\omega, t)).$$

For $0 \leqslant \theta < 2\pi$, we consider the matrix $N_\theta(\omega, s) = e^{i\theta}N(\omega, s) \in \mathbb{C}^{2\times 2}$, its Hermitian part $H_\theta(\omega, s) = [N_\theta(\omega, s) + N_\theta^*(\omega, s)]/2$, and $\lambda_\theta(\omega, s)$, the largest eigenvalue of $H_\theta(\omega, s)$. It is well known (cf. Hausdorff [12]) that the field of values of any matrix can be represented as the intersection of certain half planes; here

$$W(N(\omega, s)) = \bigcap_{0\leqslant\theta<2\pi} \{z \in \mathbb{C} : \operatorname{Re}(e^{i\theta}z) \leqslant \lambda_\theta(\omega, s)\}.$$

But

$$\lambda_\theta(\omega, s) = \frac{s}{2}\left[\omega s \cos\theta + s\sqrt{\left[\omega^2 s^2 + (2-\omega)^2\right]\cos^2\theta + \omega^2 \sin^2\theta}\right]$$

is an increasing function of s, because

$$\frac{\partial\lambda_\theta(\omega, s)}{\partial s} = \frac{1}{2}\frac{\left(\omega s\cos\theta + \sqrt{\left[\omega^2 s^2 + (2-\omega)^2\right]\cos^2\theta + \omega^2\sin^2\theta}\right)^2}{\sqrt{\left[\omega^2 s^2 + (2-\omega)^2\right]\cos^2\theta + \omega^2\sin^2\theta}} > 0.$$

This implies $W(N(\omega, s)) \subset W(N(\omega, t))$ (for $s < t$) and thus the desired conclusion of (27). From (26), we now deduce

$$W(\mathscr{L}_\omega) = W(M(\omega, s_1)) = W(M(\omega, \|B\|_2)).$$

Finally, we use the fact that the field of values of any 2×2 matrix is known (e.g., Johnson [16]) to conclude that

$$W(\mathscr{L}_\omega) = W(M(\omega, \|B\|_2)) = \mathscr{E}(\omega, \|B\|_2). \qquad \blacksquare$$

According to Young's theory [35], the optimal relaxation parameter which minimizes $\rho(\mathscr{L}_\omega)$ as a function of ω is given by

$$\omega_b = \frac{2}{1 + \sqrt{1 - \|B\|_2^2}}$$

(if $\|B\|_2 < 1$). Using Theorem 8, it is easy to determine ω_w, the value of ω which minimizes $\mu(\mathscr{L}_\omega)$ (cf. Figure 5). Note that in contrast to the choice of ω_b, which always yields an *over*relaxation scheme, the use of ω_w leads to an *under*relaxation scheme if $\|B\|_2 > 0.786\ldots$.

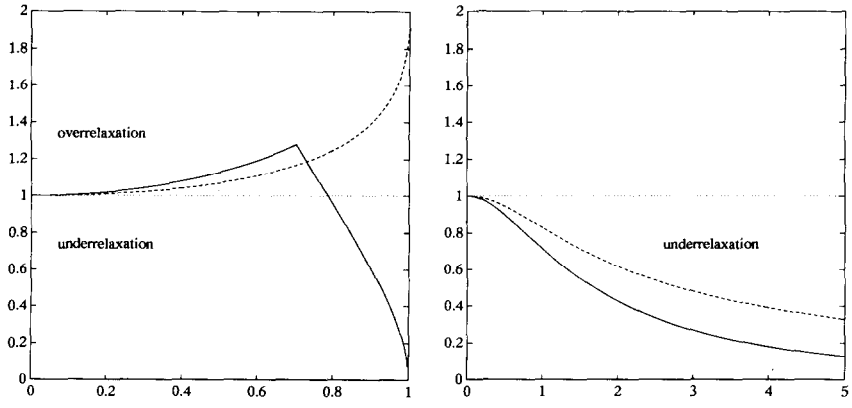


FIG. 5. ω_b (dashed curve), ω_w (solid curve) and functions of $\|B\|_2$ for the symmetric (cf. Theorem 8) and the skew-symmetric case (cf. Theorem 9).

Figure 6 suggests that there is a great difference between $\|\mathscr{L}_{\omega_b}^m\|_2$ and $\|\mathscr{L}_{\omega_w}^m\|_2$ for small m if $\|B\|_2$ is close to 1. The values which are shown in Table 2 originate from the one-dimensional model problem, i.e., the Jacobi matrix B results from the red-black ordering of $\frac{1}{2}\,\mathrm{tridiag}(-1, 0, -1) \in \mathbb{R}^{100\times 100}$ ($\|B\|_2 = 0.9995\ldots$). Here, $\lim_{m\to\infty} \|\mathscr{L}_{\omega_b}^m\|_2^{1/m} = 0.97\ldots$ and $\lim_{m\to\infty} \|\mathscr{L}_{\omega_w}^m\|_2^{1/m} = 0.99998\ldots$, but it requires a rather large exponent m for the asymptotic optimality of ω_b to be observed.

A result analogous to Theorem 8 holds for the skew-symmetric case $A = I_n - B \in \mathbb{R}^{n\times n}$, where

$$B = \begin{bmatrix} 0 & M \\ -M^T & 0 \end{bmatrix} \quad (\text{with} \quad M \in \mathbb{R}^{p\times q}, \quad p \geqslant q, \quad p + q = n). \tag{28}$$

Now the induced SOR iteration matrix has the form

$$\mathscr{L}_\omega = \begin{bmatrix} (1-\omega) I_p & \omega M \\ -\omega(1-\omega) M^T & (1-\omega) I_q - \omega^2 M^T M \end{bmatrix}. \tag{29}$$

THEOREM 9. *For the matrix* $\mathscr{L}_\omega$ *of* (29) *there holds* $W(\mathscr{L}_\omega) = \tilde{\mathscr{E}}(\omega, \|B\|_2)$. *Here,* $\tilde{\mathscr{E}}(\omega, \|B\|_2)$ *denotes the closed interior of the ellipse with*

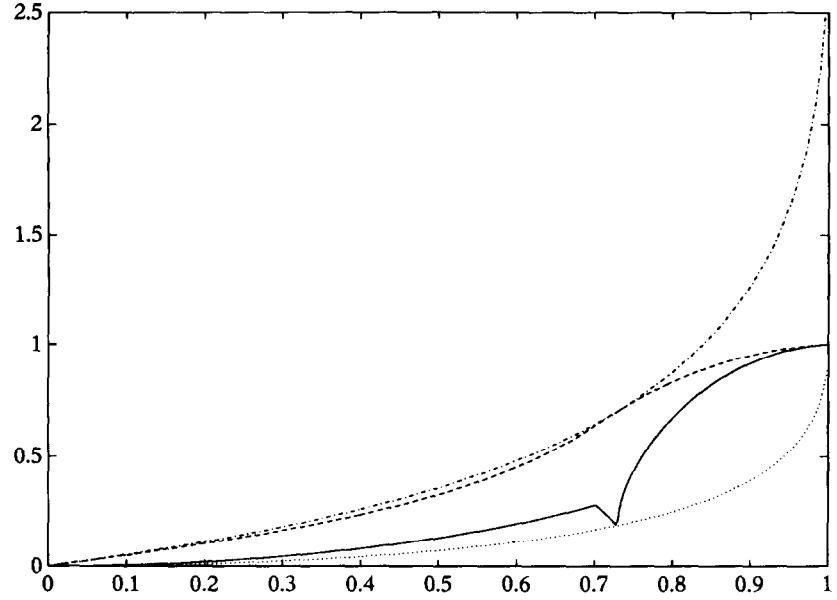


FIG. 6. $\rho(\mathscr{L}_{\omega_b})$ (dotted curve), $\rho(\mathscr{L}_{\omega_w})$ (solid curve), $\mu(\mathscr{L}_{\omega_w})$ (dashed curve), and $\mu(\mathscr{L}_{\omega_b})$ (dash-dot curve) as functions of $\|B\|_2$ in the symmetric case.

center

$$\tilde{c}(\omega, \|B\|_2) := 1 - \omega - \tfrac{1}{2}\omega^2\|B\|_2^2$$

and semiaxes

$$\tilde{a}(\omega, \|B\|_2) := \tfrac{1}{2}\omega^2\|B\|_2\left[1 + \|B\|_2^2\right]^{1/2} \qquad (\textit{on the real axis}),$$

$$\tilde{b}(\omega, \|B\|_2) := \tfrac{1}{2}\omega(2 - \omega)\|B\|_2 \qquad (\textit{perpendicular to the real axis}).$$

In Figure 5, we compare—now for the skew-symmetric case—

$$\omega_b = \frac{2}{1 + \sqrt{1 + \|B\|_2^2}},$$

TABLE 2

m	$\|\mathscr{L}_{\omega_b}^m\|_2$	$1 - \|\mathscr{L}_{\omega_w}^m\|_2$
1	$3.9\ldots$	$1.1\ldots \times 10^{-5}$
5	$1.4\ldots \times 10^{1}$	$5.8\ldots \times 10^{-5}$
10	$2.1\ldots \times 10^{1}$	$1.1\ldots \times 10^{-4}$
20	$2.4\ldots \times 10^{1}$	$2.5\ldots \times 10^{-4}$
50	$1.4\ldots \times 10^{1}$	$7.2\ldots \times 10^{-4}$
100	$5.0\ldots$	$1.6\ldots \times 10^{-3}$

which minimizes $\rho(\mathscr{L}_\omega)$ (cf. Niethammer [23]), and ω_w, which minimizes $\mu(\mathscr{L}_\omega)$ (and which is easily determined from Theorem 9). Their difference is not as dramatic as in the symmetric case.

I would like to thank G. Starke for many stimulating discussions, and L. N. Trefethen, whose suggestions improved this paper considerably. All figures were produced by Matlab.

REFERENCES

1 L. Berg, *Lineare Gleichungssysteme mit Bandstruktur und ihr asymptotisches Verhalten*, Carl Hanser, München, 1986.

2 A. Calderón, F. Spitzer, and H. Widom, The inversion of Toeplitz matrices, *Illinois J. Math.* 3:490–498 (1959).

3 P. J. Davis, *Circulant Matrices*, Wiley, New York, 1979.

4 M. Eiermann, *Semiiterative Verfahren für nichtsymmetrische lineare Gleichungssysteme*, Habilitationsschrift, Univ. Karlsruhe, 1989.

5 H. C. Elman and G. H. Golub, Iterative methods for cyclically reduced non-selfadjoint problems, *Math. Comp.* 54:671–700 (1990).

6 P. A. Farrell, Flow conforming iterative methods for convection dominated flows, in *Numerical and Applied Mathematics* (C. Brezinski, Ed.), J. C. Baltzer AG, Scientific Publishing, 1989, pp. 681–686.

7 M. Goldberg and E. Tadmor, On the numerical radius and its applications, *Linear Algebra Appl.* 42:263–284 (1982).

8 G. H. Golub and J. E. de Pillis, Toward an effective two-parameter SOR method, in *Iterative Methods for Large Linear Systems* (D. R. Kincaid and L. J. Hayes, Eds.), Academic Press, Boston, 1989, pp. 107–119.

9 G. H. Golub and R. S. Varga, Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second order Richardson iterative methods, parts I, II, *Numer. Math.* 3:147–168 (1961).

10 I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, New York, 1965.
11 P. R. Halmos, *A Hilbert Space Problem Book*, Van Nostrand, New York, 1967.
12 F. Hausdorff, Der Wertevorrat einer Bilinearform, *Math. Z.* 3:314–316 (1919).
13 P. Henrici, *Applied and Computational Complex Analysis. Volume I*, Wiley, New York, 1974.
14 N. J. Higham, Matrix nearness problems and applications, in *Applications of Matrix Theory* (M. J. C. Gover and S. Barnett, Eds.), Clarendon, Oxford, 1989, pp. 1–27.
15 R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge U.P., New York, 1991.
16 C. R. Johnson, Computation of the field of values of a 2×2 matrix, *J. Res. Nat. Bur. Standards* 78B:105–107 (1974).
17 C. R. Johnson, Numerical determination of the field of values of a general complex matrix, *SIAM J. Numer. Anal.* 15:595–602 (1978).
18 T. Kato, Some mapping theorems for the numerical range, *Proc. Japan Acad.* 41:652–655 (1965).
19 H. W. J. Lenferink and M. N. Spijker, A generalization of the numerical range of a matrix, *Linear Algebra Appl.* 140:251–266 (1990).
20 T. A. Manteuffel, The Tchebychev iteration for nonsymmetric linear systems, *Numer. Math.* 28:307–327 (1977).
21 M. Marcus and H. Minc, *A Survey of Matrix Theory and Matrix Inequalities*, Allyn and Bacon, Boston, 1964.
22 M. Marcus and C. Pesce, Computer generated numerical ranges and some resulting theorems, *Linear and Multilinear Algebra* 20:121–157 (1987).
23 W. Niethammer, Relaxation bei Matrizen mit der Eigenschaft "A," *Z. Angew. Math. Mech.* 44:T49–T52 (1964).
24 W. Niethammer and R. S. Varga, The analysis of k-step iterative methods for linear systems from summability theory, *Numer. Math.* 41:177–206 (1983).
25 C. Pearcy, An elementary proof of the power inequality for the numerical radius, *Michigan Math. J.* 13:289–291 (1966).
26 L. Reichel and L. N. Trefethen, Eigenvalues and pseudoeigenvalues of Toeplitz and block-Toeplitz matrices, *Linear Algebra Appl.*, 162–164:153–185 (1992).
27 G. Starke, Field of values and the ADI method for nonnormal matrices, *Linear Algebra Appl.*, 180:199–218 (1993).
28 L. N. Trefethen, Approximation theory and numerical linear algebra, in *Algorithms for Approximation II* (J. C. Mason and M. G. Cox, Eds.), Chapman and Hall, New York, 1990, pp. 336–360.
29 L. N. Trefethen, Pseudospectra of matrices, in *Proceedings of the 14th Dundee Biennial Conference on Numerical Analysis* (D. F. Griffiths and G. A. Watson, Eds.), to appear.
30 L. N. Trefethen, *Non-normal Matrices and Pseudospectra*, in preparation.
31 C. F. Van Loan, How near is a stable matrix to an unstable matrix?, *Contemp. Mathematics* 47:465–477 (1985).

32 R. S. Varga, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, N.J., 1962.
33 J. R. Westlake, *A Handbook of Numerical Matrix Inversion and Solution of Linear Equations*, Wiley, New York, 1968.
34 J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford U.P., Oxford, 1965.
35 D. M. Young, Iterative methods for solving partial differential equations of elliptic type, *Trans. Amer. Math. Soc.* 76:92–111 (1954).

Received 31 July 1991; final manuscript accepted 6 January 1992