MATHEMATICS OF COMPUTATION, VOLUME 31, NUMBER 139
JULY 1977, PAGES 726–732

Inverse Iteration on Defective Matrices*

By Nai-fu Chen

Abstract. Very often, inverse iteration is used with shifts to accelerate convergence to an eigenvector. In this paper, it is shown that, if the eigenvalue belongs to a nonlinear elementary divisor, the vector sequences may diverge even when the shift sequences converge to the eigenvalue. The local behavior is discussed through a 2 × 2 example, and a sufficient condition for the convergence of the vector sequence is given.

Introduction. If an accurate approximation σ to an eigenvalue λ of a matrix B is available, then inverse iteration is an attractive technique for computing the associated eigenvector. We choose an arbitrary unit vector v_0 and a fixed shift σ. Then for $j = 1, 2, \ldots$ we solve

$$(B - \sigma I)w_j = v_{j-1}, \qquad v_j = w_j/\|w_j\|, \tag{1}$$

where $\| \cdot \|$ is the users' preferred vector norm.

Semisimple case. If λ is a simple eigenvalue with unit eigenvector x, if σ is close enough to λ, and if v_0 is not an unfortunate choice, then the vector sequence $\{v_j\}$ converges linearly to x and the convergence factor is very favorable. This well-known result holds also for multiple eigenvalue λ provided that:

(i) the dimension of λ's eigenspace is equal to λ's algebraic multiplicity (i.e. linear elementary divisors),

(ii) the spectral projection of v_0 into λ's eigenspace is not zero (i.e. the starting vector is not deficient in x).

If σ is known to equal λ to within working precision of the computer, then only one or two steps of the iteration are necessary.

Defective Case. Wilkinson [3] and Varah [2] pointed out that the situation is not so nice if λ has generalized eigenvectors of grade higher than one, i.e., when λ belongs to a nonlinear elementary divisor. In exact arithmetic the iteration converges not linearly, but harmonically like $1/j$ as $j \longrightarrow \infty$. Even worse is the fact that except for very special choices of v_0, the vectors v_2 and v_3 will be poorer approximations than v_1!

Variable Shifts. Inverse iteration can also be used with variable shifts. If the sequence of shifts $\{\sigma_j\}$ converges to λ as $j \longrightarrow \infty$, then the vector sequence generated

Received January 20, 1976; revised June 21, 1976.

AMS (MOS) subject classifications (1970). Primary 65F15; Secondary 15A18.

Key words and phrases. Eigenvector, eigenvalue, nonlinear divisors, convergence, inverse iterations, shifts.

*This work was partially supported by the Office of Naval Research under Contract No. N00014-69-A-0200-1017.

This work was performed under the auspices of the Energy Research and Development Administration while the author was visiting Lawrence Berkeley Laboratory.

by

$$(2)\qquad \begin{aligned}(B - \sigma_j I)w_j &= - v_{j-1} \quad \text{(the } - \text{ sign is for convenience)}\\ v_j &= w_j/\|w_j\|\end{aligned}$$

converges to x whenever λ is a simple eigenvalue. However, when λ has eigenvector of grade higher than one, the situation is again complicated; and the shifts can make things worse. In fact, we shall prove the following surprising result. *The sequence* $\{v_j\}$ *generated by* (2) *may fail to converge to* x *even though the sequence* $\{|\sigma_j - \lambda|\}$ *converges monotonically to* 0 *as* $j \longrightarrow \infty$.

The Construction. There is no loss of generality in studying

$$N = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

(for analogous results on $n \times n$ matrices, see Chen [1]). Note that $e_1 = \binom{1}{0}$ is the eigenvector, and any other linearly independent vector is an eigenvector of grade 2. We observe that

$$(N - \sigma I)w = - e_1 \quad \text{yields } w = \begin{cases} \tau e_1 - e_2 & \text{for any } \tau, \text{ if } \sigma = 0, \\ \sigma^{-1} e_1, & \text{if } \sigma \neq 0. \end{cases}$$

This shows that e_1 is a fixed point of the iteration (2) provided that $\sigma \neq 0$.

We are now ready to make our perverse construction. In fact, there are many ways to construct such a perverse sequence of shifts. We first observe the following facts:

Let σ_j, $j = 1, 2, \ldots$, be the sequence of shifts, $\xi_j = \sigma_j^{-1}$, and let τ_j, τ'_j be normalizing factors to keep v_j of unit length; then the vectors generated by inverse iteration with the above shifts satisfy

$$\begin{aligned} v_j &= \tau_j(N - \sigma_j I)^{-1}(N - \sigma_{j-1} I)^{-1} \cdots (N - \sigma_1 I)^{-1} v_0 \\ &= \tau'_j(I + \xi_j N)(I + \xi_{j-1} N) \cdots (I + \xi_1 N)v_0 = \tau'_j(I + (\xi_1 + \xi_2 + \cdots + \xi_j)N)v_0. \end{aligned}$$

So if we write $s_j = \Sigma_{k=1}^{j} \xi_k$, then

$$v_j = \tau'_j \begin{pmatrix} 1 & s_j \\ 0 & 1 \end{pmatrix} v_0.$$

Our goal to construct a perverse sequence is accomplished if there exists a finite number z such that z is a limit point of the infinite set $\{s_j\}$. This is so because $u = u'/\|u'\|$, where

$$u' = \begin{pmatrix} 1 & z \\ 0 & 1 \end{pmatrix} v_0 = \begin{pmatrix} 1 & z \\ 0 & 1 \end{pmatrix}\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \eta_1 + \eta_2 z \\ \eta_2 \end{pmatrix}, \qquad v_0 = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix},$$

would be a limit vector of the vector sequence $\{v_j\}$. If $v_0 \neq e_1$, then $u \neq e_1$. Hence the vector sequence cannot converge to the eigenvector.

So, we define a sequence of complex shifts $\sigma_j = r_j e^{i\theta_j}$ with $\pi \geqslant \theta_j > -\pi$, $r_j \geqslant 0$, by the rule below:

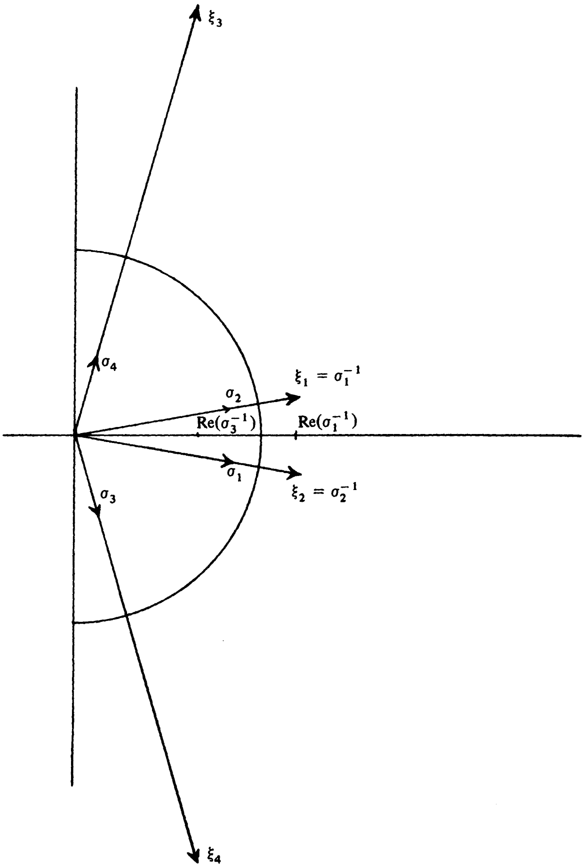

FIGURE 1

The sequence of shifts on the complex plane

(i) $\sigma_1 = r_1 e^{i\theta_1}$ so that $\mathrm{Re}(\sigma_1) \neq 0$,

(ii) for $j = 3, 5, 7, \ldots$,

$$r_j = r_{j-2}/2, \qquad \theta_j \text{ is such that } \mathrm{Re}(\sigma_j^{-1}) = \mathrm{Re}(\sigma_{j-2}^{-1})/2,$$

(equivalently, solve

$$\cos\theta_{j-2} = 4\cos\theta_j$$

for θ_j in the appropriate quadrant),

(iii) $\sigma_j = \bar{\sigma}_{j-1}$ for $j = 2, 4, 6, \ldots$.

It is clear that the sequence thus generated is monotone decreasing to zero, the eigenvalue, because $|\sigma_j| = |\sigma_{j-1}| = |\sigma_1|/2^{m-1}$ where $j = 2m$, $m = 1, 2, \ldots$. Also, $z = 4\,\mathrm{Re}(\sigma_1^{-1})$ is a finite limit point of the set $\{s_j | s_j = \Sigma_{k=1}^{j} \xi_j\}$ because

$$\begin{aligned}\lim_{m\to\infty} \sum_{k=1}^{2m} \xi_k &= \lim_{m\to\infty} (\xi_1 + \bar{\xi}_1) + (\xi_3 + \bar{\xi}_3) + \cdots + (\xi_{2m-1} + \bar{\xi}_{2m-1}) \\ &= \lim_{m\to\infty} 2\{\mathrm{Re}(\sigma_1^{-1}) + \tfrac{1}{2}\mathrm{Re}(\sigma_1^{-1}) + \cdots + (\tfrac{1}{2})^{m-1}\mathrm{Re}(\sigma_1^{-1})\} \\ &= 4\,\mathrm{Re}(\sigma_1^{-1}).\end{aligned}$$

The sequence we have constructed accomplishes our goal.

Local Behavior. This peculiar behavior of inverse iteration on defective matrices can be better understood through the local picture of our 2 × 2 matrix N. For simplicity, we keep all quantities real and let $s = \sin\theta \neq 0$.

$$(N - \sigma I)w = -\begin{pmatrix} c \\ s \end{pmatrix} \quad \text{yields } w = \begin{cases} \infty, & \sigma = 0, \\ c\sigma^{-1}[(1 - t/\sigma)e_1 + te_2], & \sigma \neq 0, \end{cases}$$

where $t = \tan\gamma$, $c = \cos\gamma$ (thus γ is the angle between $\binom{c}{s}$ and e_1).

Let γ' be the angle between w and e_1, and $t' = \tan\gamma'$. Then the iteration function for a typical step is given by

$$t' \equiv \Phi_\sigma(t) = t/(1 + t/\sigma). \tag{3}$$

The fact that $\Phi'_\sigma(0) = 1$, $\sigma \neq 0$, corresponds to the harmonic convergence of the fixed shift sequence.

If we study inverse iteration with shifts $\{\sigma_j\}$ yielding vectors $\{v_j\}$, then in applying (3), we have the following correspondence:

$\{\sigma_j\}$ close to λ corresponds to σ small,

$\{v_j\}$ close to e_1 corresponds to t small.

In the following diagram, we demonstrate how w compares with v as an approximation to the eigenvector. Each point (σ, t) on the diagram represents one step of inverse iteration with a shift σ and the vector v whose component ratio is t. We say that w is a better approximation if $|t'| < |t|$ and worse if $|t'| > |t|$:

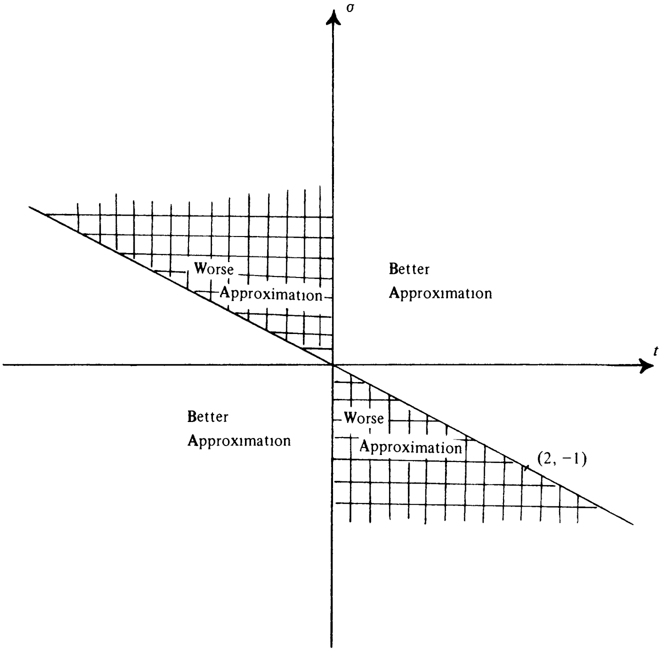

FIGURE 2

Behavior of inverse iteration applied to N

The above diagram clearly demonstrates that no matter how good our approximation to the eigenvector (when t is small) or to the eigenvalue (when σ is small) or both, inverse iteration can still give a much worse approximation in exact arithmetic. This local behavior of inverse iteration on defective matrices is precisely the source of difficulty.

A Sufficient Condition. Experience with various traditional shift strategies, even in the presence of multiple eigenvalues, indicates that inverse iteration usually converges, though very slowly. The following theorem relates how the shifts converge to the eigenvalue and gives a sufficient condition that the generated vectors converge to the eigenvector:

THEOREM. *Let C be a general complex matrix and let the shifts σ_j, $j = 1, 2, \ldots$, converge to λ, an eigenvalue of C. Let n be the dimension of the largest Jordan block corresponding to λ. Write $\sigma_j - \lambda = r_j e^{-i\theta_j}$, $r_j \geqslant 0$. If there exists ϕ and an integer k such that* (*see Figure* 3)

$$|\phi - \theta_j| \leqslant \frac{\pi}{4(n-1)} \quad \textit{for } j \geqslant k, \tag{4}$$

then $\{v_j\}$ converge to an eigenvector.

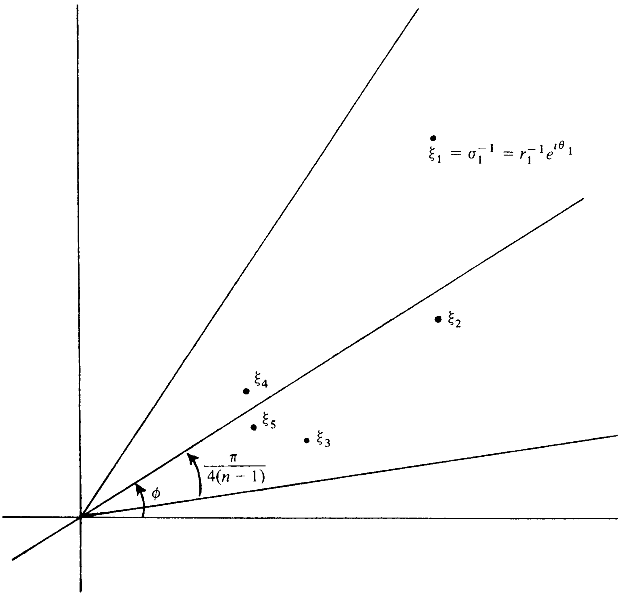

FIGURE 3

The angle condition for $\xi_j = r_j^{-1}e^{i\theta_j} = \sigma_j^{-1}$
in the complex plane where $\lambda = 0$

In other words, if the σ_j (and thus $\xi_j = \sigma_j^{-1}$) ultimately fall into a cone of size specified in (4) *in the complex plane, then the shifts will accelerate convergence.*

Proof. Without loss of generality, we can consider C to be the $n \times n$ Jordan block

$$J = \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & & \vdots \\ 0 & 0 & 0 & 1 & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ \vdots & & & & \ddots & 0 \\ \vdots & & & & & 1 \\ \cdot & \cdot & \cdot & \cdot & \cdots & 0 \end{pmatrix},$$

and we study the vector sequence from v_k on. With $\xi_j = \sigma_j^{-1}$, and τ_j, τ_j' normalizing factors, we have

$$\begin{aligned} v_{k+j} &= \tau_j(C - \sigma_{k+j}I)^{-1} \cdots (C - \sigma_{k+1}I)^{-1} v_k \\ &= \tau_j'(I - \xi_{k+j}J)^{-1} \cdots (I - \xi_{k+1}J)^{-1} v_k \\ &= \tau_j'(I + \xi_{k+j}J + \xi_{k+j}^2 J^2 + \cdots + \xi_{k+j}^{n-1} J^{n-1}) \\ &\qquad \cdots (I + \xi_{k+1}J + \cdots + \xi_{k+1}^{n-1} J^{n-1}) v_k \\ &= \tau_j'(I + d_1(j)J + d_2(j)J^2 + \cdots + d_{n-1}(j)J^{n-1}) v_k = \tau_j' S_k(j) v_k, \end{aligned}$$

where

$$S_k(j) = \begin{pmatrix} 1 & d_1(j) & d_2(j) & \cdots & d_{n-1}(j) \\ 0 & 1 & d_1(j) & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & d_1(j) \\ 0 & \cdots & & 0 & 1 \end{pmatrix}, \tag{5}$$

where $d_l(j)$ = sum of $\binom{j+l-1}{l}$ terms of the form $(\xi_{i_1}\xi_{i_2} \cdots \xi_{i_l})$, $k \leqslant i_1, i_2, \ldots, i_l \leqslant k + j$. Hence v_{k+j} is just a linear combination of the columns of the matrix $S_k(j)$. To show that $\{v_{k+j}\}$ converges to e_1, it is sufficient to show that $d_{l-1}(j)/d_l(j) \longrightarrow 0$ as $j \longrightarrow \infty$ for $l = 1, \ldots, n-1$ (with $d_0(j) \equiv 1$).

LEMMA. $|d_l(j)| > |\xi_{k+j} d_{l-1}(j)|$.

Proof. We first make the following observations: let

$$a_x = \xi_{x_1}\xi_{x_2} \cdots \xi_{x_l} = (r_{x_1}e^{i\theta_{x_1}}) \cdots (r_{x_l}e^{i\theta_{x_l}}) = r_x e^{i\theta_x}$$

(i.e., $r_x = r_{x_1}r_{x_2} \cdots r_{x_l}$, $\theta_x = \theta_{x_1} + \theta_{x_2} + \cdots + \theta_{x_l}$) and

$$a_y = \xi_{y_1}\xi_{y_2} \cdots \xi_{y_l} = r_y e^{i\theta_y}$$

be two of the $\binom{j+l-1}{l}$ terms in $d_l(j)$; then by (4)

$$|l \cdot \phi - \theta_x| \leqslant \frac{l\pi}{4(n-1)} \leqslant \frac{\pi}{4}, \qquad |l \cdot \phi - \theta_y| \leqslant \frac{\pi}{4}.$$

Further

(i) $|a_x + a_y| > |a_x|$ because a_x, a_y make an angle less than $\pi/2$ in the complex plane, and

(ii) If $a_x + a_y = re^{i\theta}$, then $|l\phi - \theta| \leqslant \pi/4$.

Now, $\xi_{k+j}d_{l-1}(j)$ is one of the terms of $d_l(j)$, and by repeated application of (ii) above, we know

$$|\arg(\xi_{k+j}d_{l-1}(j)) - l\phi| \leqslant \pi/4.$$

Similarly, the rest of $d_l(j)$, $d_l(j) - \xi_{k+j}d_{l-1}(j)$, satisfies

$$|\arg(d_l(j) - \xi_{k+j}d_{l-1}(j)) - l\phi| \leqslant \pi/4.$$

Hence

$$\begin{aligned} |d_l(j)| &= |d_l(j) - \xi_{k+j}d_{l-1}(j) + \xi_{k+j}d_{l-1}(j)| \\ &> |\xi_{k+j}d_{l-1}(j)| \quad \text{by (i).} \quad \square \end{aligned}$$

We can now complete the proof of the theorem:

$$\begin{aligned} \left|\frac{d_{l-1}(j)}{d_l(j)}\right| &< \left|\frac{d_{l-1}(j)}{\xi_{k+j}d_{l-1}(j)}\right| \quad \text{(by lemma)} \\ &= \left|\frac{1}{\xi_{k+j}}\right| = |\sigma_{k+j}| \longrightarrow 0 \quad \text{as } j \longrightarrow \infty. \quad \square \end{aligned}$$

COROLLARY. *In the real case, if* $\{\sigma_j\}$ *eventually converge to the eigenvalue from one side, then* v_j *converges to an eigenvector as* $j \longrightarrow \infty$.

Remark. Experience with Rayleigh quotient iteration shows that in fact the Rayleigh quotients do eventually converge to the eigenvalue from one side in the real case. As for the complex case, the Rayleigh quotients also eventually fall into a cone with angle less than $\pi/4(n - 1)$ as required in (4). We examine the behavior of Rayleigh quotients in another communication.

Acknowledgment. The author is very grateful to Professor Beresford Parlett for his many suggestions and comments.

Department of Mathematics
University of Southern California
University Park
Los Angeles, California 90007

1. N. CHEN, *The Rayleigh Quotient Iteration for Non-Normal Matrices,* Ph. D. Dissertation, Electronic Research Laboratory Memorandum No. ERL-M548, University of California, Berkeley, 1975.

2. J. M. VARAH, *The Computation of Bounds for the Invariant Subspaces of a General Matrix Operator,* Ph. D. Dissertation, Stanford University, 1967.

3. J. H. WILKINSON, "Inverse iteration in theory and in practice," *Symposia Mathematica,* vol. 10, Academic Press, London, 1972, pp. 361–379. MR 51 #2268.