How Near is a Stable Matrix to an Unstable Matrix?

C. Van Loan
TR 84-649
October 1984

Department of Computer Science
Cornell University
Ithaca, New York 14853

HOW NEAR IS A STABLE MATRIX TO AN UNSTABLE MATRIX?

Charles Van Loan
Department of Computer Science
Cornell University
Ithaca, New York 14853

ABSTRACT

In this paper we explore how close a given stable matrix A is to being unstable. As a measure of "how stable" a stable matrix is, the spectral abscissa is shown to be flawed. A better measure of stability is the Frobenius norm of the smallest perturbation that shifts one of A's eigenvalues to the imaginary axis. This leads to a singular value minimization problem that can be approximately solved by heuristic means. However, the minimum destabilizing perturbation may be complex even when A is real. This suggests that in the real case we look for the smallest *real* perturbation that shifts one of the eigenvalues to the imaginary axis. Unfortunately, a difficult constrained minimization problem ensues and no practical estimation technique could be devised.

This paper was presented at the AMS Summer Conference on Linear Algebra and Systems Theory held at Bowdoin College, Brunswick, Maine, July 30 - August 3, 1984.

Section 1. Introduction

Suppose $A \in \mathbf{C}^{n\times n}$ and denote the set of its n eigenvalues by $\lambda(A)$. The spectral abscissa of A is defined by

$$\alpha(A) = \max \{ \mathrm{Re}(\lambda) \mid \lambda \in \lambda(A) \} .$$

If $\alpha(A)$ is negative (non-negative) then we say that A is stable (unstable). The terminology is tied to the asymptotic behavior of the system $(d/dt)x = Ax$. In this paper we address the question, "How near is a given stable matrix to an unstable matrix?" Our aim is to suggest an alternative to the control engineer's traditional way of appraising stabilty which is typically based on $\alpha(A)$.

Obviously, if $\alpha(A)$ is negative but very small, then a small perturbation in A can make it unstable. In particular, the matrix $A - \alpha(A)I_n$ has an eigenvalue on the imaginary axis. On the other hand, it is possible for A to be very nearly unstable without $\alpha(A)$ being particularly small. For example, suppose

$$A = \begin{bmatrix} J_9(-.1) & 0 \\ 0 & -.001 \end{bmatrix} \in \mathbf{R}^{10\times 10}$$

where $J_9(-.1)$ is a 9-by-9 Jordan block associated with the eigenvalue -.1 . If the (9,1) entry in A is changed to 10^{-9} then the resulting matrix is unstable. Thus, A is within 10^{-9} of being unstable even though $\alpha(A) = -10^{-3}$. From this we conclude that the spectral abscissa is flawed as a nearness-to-instability measure: it can give a false impression about how stable the matrix is.

The notion of a "flawed measure" can be clarified by considering the more familiar problem of nearness-to-singularity. In this context, the smallest eigenvalue of a matrix can give a misleading impression about how close a matrix is to being singular. To illustrate, consider the following well-known example:

$$B_n = \begin{bmatrix} 1 & -1 & -1 & \cdots & -1 \\ 0 & 1 & -1 & \cdots & -1 \\ & & \ddots & & \\ & & & \ddots & \\ 0 & 0 & \cdots & 1 & -1 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \in \mathbf{R}^{n\times n}$$

It can be shown that a perturbation of order 2^{-n} makes this matrix singular even though its smallest eigenvalue is 1 in modulus. This is why one uses singular values rather than eigenvalues when quantifying nearness-to-singularity.

The smallest singular value of $A \in \mathbf{C}^{n\times n}$, which we denote by $\sigma_{min}(A)$, satisfies

$$\sigma_{min}(A) = \min \{ \| E \|_F \mid \det(A+E) = 0 , E \in \mathbf{C}^{n\times n} \} . \tag{1.1}$$

(Recall: $\| E \|_F^2 = \Sigma \Sigma |e_{ij}|^2$) . Moreover, if

$$U^H A V = \mathrm{diag}(\sigma_1 , \ldots , \sigma_n)$$

is the singular value decomposition (SVD) with unitary

$$U = [u_1 , \ldots , u_n] \qquad u_i \in \mathbf{C}^n$$

and

$$V = [v_1 , \ldots , v_n] \qquad v_i \in \mathbf{C}^n$$

and ordered singular values

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \equiv \sigma_{min} \geq 0 ,$$

then the minimizing E in (1.1) is given by

$$E_{min} = -\sigma_{min} u_n v_n^H$$

We mention that if A is real, then U and V may be chosen to be real orthogonal.

A thorough discussion of the above SVD-related concepts is given in Golub and Van Loan (1983).

Returning to our problem, we propose to parallel (1.1) and examine the following measure of stability :

$$\beta(A) = \min \{ \| E \|_F \mid \alpha(A+E) \geq 0 , E \in \mathbf{C}^{n \times n} \}.$$

This quantity measures the size of the smallest matrix E such that A + E has an eigenvalue in the closed right half plane. We have chosen to work in the Frobenious norm for reasons of *analytic* convenience. Other norms may be more suitable in specific applications.

If A is stable, then a simple continuity-of-eigenvalue argument applied to $A + tE$, $0 \leq t \leq 1$, reveals that

$$\beta(A) = \min \{ \| E \|_F \mid \alpha(A+E) = 0 , E \in \mathbf{C}^{n \times n} \}.$$

We say that E is a *destabilizing* perturbation if A is stable and A + E has an eigenvalue on the imaginary axis. Thus, the problem of computing $\beta(A)$ involves finding the smallest destabilizing perturbation.

Note that if A has an eigenvalue on the imaginary axis, then the *Lyapunov transformation* $\phi(X) = AX + XA^H$ is singular. (See Golub and Van Loan (1983,p.194).) Let sep(A) denote the smallest singular value of ϕ, i.e.,

$$sep(A) = \min \{ \| AX + X A^H \|_F \mid X \in \mathbf{C}^{n \times n} , \| X \|_F = 1 \} .$$

Note that sep(A) = 0 if and only if A has an eigenvalue on the imaginary axis.

Thus, if A is stable, then $\alpha(A)$, $\beta(A)$, sep(A), and $\sigma_{min}(A)$ each have "something to say" about what's involved in moving one of A's eigenvalues to the imaginary axis. The following result establishes some of the connections between these four quantities

Theorem 1.1

If $A \in \mathbf{C}^{n\times n}$ is stable then

$$\tfrac{1}{2}\,\mathrm{sep}(A) \le \beta(A) \le \sigma_{min}(A)$$

and

$$\beta(A) \le |\alpha(A)| .$$

Proof.

The set of perturbations that move one of A's eigenvalues to the imaginary axis is larger than the set of perturbations that move one of A's eigenvalues to the origin. Thus,

$$\begin{aligned}\beta(A) &= \min \{ \| E \|_F \mid \alpha(A+E) = 0 ,\ E \in \mathbf{C}^{n\times n} \} \\ &\le \min \{ \| E \|_F \mid \det(A+E) = 0 ,\ E \in \mathbf{C}^{n\times n} \} = \sigma_{min}(A)\end{aligned}$$

To establish the other inequality, we use the fact that sep(A) is just the smallest singular value of the n^2-by-n^2 matrix

$$M = I \otimes A + A \otimes I \tag{1.2}$$

where $B \otimes C$ denotes the block matrix $(b_{ij}C)$, i.e., the Kronecker product of B and C . M is just a matrix representation of the linear transformation ϕ .

Now let E be such that $\beta(A) = \| E \|_F$ and $\alpha(A+E) = 0$. Note from Lyapunov theory that the linear transformation $X \to (A+E)X + X(A+E)^H$ is singular. Corresponding to (1.2), this linear mapping has the singular matrix representation $I \otimes (A+E) + (A+E) \otimes I$. But since the smallest singular value of a matrix is a lower bound on the 2-norm of any perturbation that renders it singular we have

$$\text{sep}(A) \le \| I \otimes E + E \otimes I \|_2 \le 2 \| E \|_2 \le 2 \| E \|_F = 2\beta(A).$$

Here we used the easily derived fact that $\| B \otimes C \|_2 \le \| B \|_2 \| C \|_2$.

To establish the relationship between $\alpha(A)$ and $\beta(A)$, let $Q^H A Q = T$ be the Schur decomposition of A where Q is unitary and T is upper triangular. (A discussion of the Schur decomposition may be found in Golub and Van Loan (1983, Chapter 7).) Q can be chosen so that $\alpha(A) = \text{Re}(t_{11})$. If we set $E = -\alpha(A) q_1 q_1^H$ where q_1 is the first column of Q, then it is easy to show that A + E has an eigenvalue on the imaginary axis and that $\| E \|_F = |\alpha(A)|$. It follows that $\beta(A) \le |\alpha(A)|$.

Q.E.D.

If $\text{sep}(A) = 0$ implies that A is unstable, then shouldn't $\text{sep}(A) \ll 1$ imply that A is nearly unstable? This question is not answered by the theorem. Moreover, we have not been able to produce a useful inequality of the form $\beta(A) \le \text{constant} \cdot \text{sep}(A)$. The problem concerns the behavior of the matrix M in (1.2) under perturbation. If we allow *arbitrary* perturbations, then M has distance sep(A) from the set of singular matrices. However, if we only allow perturbations of the form $I \otimes E + E \otimes I$, then it's distance to the set of singular matrices is of order $\beta(A)$. Hence, we are not able to claim that sep(A) is a reliable indicator of nearness-to-instability . Nevertheless, the reader should be aware that an efficient method for estimating sep(A) is detailed in Byers (1983).

Our plan in the remainder of this paper is to focus on $\beta(A)$. In Section 2 we give a practical characterization of $\beta(A)$ that involves singular values. Using this characterization we develop an algorithm that can be used to estimate $\beta(A)$. It assumes that A has been reduced to triangular form using the QR algorithm and it utilizes some recent condition estimation techniques. In Section 3 we study the case when A is real and only real perturbations are considered. The analysis shows that a rather complicated constrained optimization problem must be solved.

Section 2. Estimating $\beta(A)$

Note that if A is stable and E is a destabilizing perturbation, then

$$(A + E - \mu i I)z = 0$$

for some $\mu \in \mathbf{R}$ and some nonzero $z \in \mathbf{C}^n$. Thus, $A - \mu i I$ is made singular when perturbed by E. It follows from (1.1) that the minimum Frobenius norm of any such E is minimum singular value of $(A - \mu i I)$ and so

$$\beta(A) = \min \{ \| E \|_F \mid \alpha(A + E) = 0 \} = \min \{ \sigma_{min}(A - \mu i I) \mid \mu \in \mathbf{R} \}.$$

The problem of computing $\beta(A)$ is thus the problem of minimizing the smallest singular value of $A - \mu i I$.

To this end let us apply a one-dimensional minimizer to the function

$$f(\mu) = \sigma_{min}(A - \mu I) \quad . \tag{2.1}$$

Because it does not require derivatives, the subroutine FMIN in Forsythe, Malcolm, and Moler (1977, Chapter 8) is particularly well-suited. FMIN is based on golden-section search and successive parabolic interpolation and is originally described in Brent (1973). A call to FMIN requires an interval [a,b], a procedure for evaluating the function $f(\mu)$, and a termination criteria. The routine then proceeds to find a local minima of f in the interval.

The function $f(\mu)$ can be evaluated by applying the Golub-Reinsch SVD algorithm to $A - \mu i I$. A complex arithmetic implementation of this routine may be found in LINPACK (1978). However, every function call requires $O(n^3)$ arithmetic operations since each different μ forces recomputation of the SVD.

However, the volume of work can be reduced by an order of magnitude if A is first transformed to upper triangular form and condition estimation ideas are used. Suppose $Q^H A Q = T$ is the Schur decomposition of A. (See Golub and Van Loan (1983 , Chap.7) for a

discussion of this decomposition.) Since

$$Q^H(A - \mu i I)Q = T - \mu i I$$

it follows that $A - \mu i I$ and $T - \mu i I$ have the same minimum singular value regardless of μ because of unitary invariance. Thus,

$$\beta(A) = \min \{ \sigma_{min}(T - \mu i I) \mid \mu \in \mathbf{R} \}. \tag{2.2}$$

The reason for reducing A to triangular form is that the smallest singular value of $T - \mu i I$ can now be estimated in $O(n^2)$ arithmetic operations using the σ_{min} estimator described in Cline, Conn, and Van Loan (1982) . (See also Van Loan (1984).) Let

$$\hat{f}(\mu) = \hat{\sigma}_{min}(T - \mu i I) \tag{2.3}$$

be the estimate of $\sigma_{min}(T - \mu i I)$ produced by this means. Extensive tests reveal that $\hat{f}(\mu)$ is a reliable estimate of $f(\mu)$ in that for any T and any given μ we have

$$f(\mu) \leq \hat{f}(\mu) \leq 1.1 \; f(\mu). \tag{2.4}$$

Applying FMIN to $\hat{f}(\mu)$ instead of $f(\mu)$ is justified because in practice all we need in most control engineering applications is an order of magnitude estimate of $\beta(A)$. Stated another way, it is worth using the σ_{min} estimator and sacrificing fifteen-digit accuracy in order to make the FMIN approach feasible.

Another practical detail that warrants consideration is the choice of the initial interval that FMIN requires. A useful result in this regard is the following

Lemma 2.1

If $\beta(A) = \sigma_{min}(A - \mu_{opt} iI)$ with $\mu_{opt} \in \mathbf{R}$, then $|\mu_{opt}| \leq 2 \| A \|_2$.

Proof.

Using SVD perturbation theory (c.f. Golub and Van Loan (1983,p.286)) we have $\sigma_{min}(A) \geq \sigma_{min}(A - \mu_{opt} iI) \geq |\mu_{opt}| - \| A \|_2$. Thus, $|\mu_{opt}| \leq \| A \|_2 + \sigma_{min}(A) \leq 2 \| A \|_2$.

Q.E.D.

This suggests that FMIN be applied with the initial interval $[-\omega, \omega]$ where $\omega = 2\| A \|_2$. ($\| A \|_2 = \| T \|_2 \equiv \sigma_{max}(T)$ can also be estimated in $O(n^2)$ operations.) Unfortunately, FMIN returns only *local* minima and so a plan must be devised for finding the global minimum of f.

To this end, define let $\mathbf{I}$ denote the imaginary part of A's spectrum,

$$\mathbf{I} = \{ \text{Imag}(\lambda) \mid \lambda \in \lambda(A) \} = \{ \mu_1, \ldots, \mu_k \} . \tag{2.5}$$

Our computational experience has shown that the local minima of $f(\mu)$ occur in the vicinity of $\mu_1, \ldots, \mu_k$. Thus, if $\mu_1 < \cdots < \mu_k$ and

$$x_0 \equiv -2\| A \|_2 \leq \mu_1 < x_1 < \mu_2 < \cdots < x_{k-1} < \mu_k < x_k \equiv 2\| A \|_2$$

then we could apply FMIN to the intervals $[x_{j-1}, x_j]$, $j = 1,\ldots,k$. In this context is reasonable to set $x_j = (\mu_j + \mu_{j+1})/2$ for $j = 1,\ldots,k-1$.

We implemented this procedure but were then pleasantly surprised to learn the following from numerous experiments:

(2.6) *The local minima of $f(\mu)$ seem to coincide with the μ_j.*

We have been unable to rigorously establish this result. The following examples are intended to suggest its validity.

Example 2.1

$$A = \begin{bmatrix} -0.01 & 5.00 & -1.00 & -1.00 \\ -5.00 & -0.01 & 5.00 & -1.00 \\ 0.00 & 0.00 & -0.01 & 5.00 \\ 0.00 & 0.00 & -5.00 & -0.01 \end{bmatrix}$$

Here, A has a double defective eigenvalue at $-.01 \pm 5i$. To within 6 digits, $\beta(A) = .316224 \cdot 10^{-4}$ with $\mu_{opt} = \pm 5.00000$.

Example 2.2

$$A = \begin{bmatrix} -10^{-5} & 4 & -1 & -1 & -1 & -1 & -1 & -1 \\ 0 & -10 & 4 & -1 & -1 & -1 & -1 & -1 \\ 0 & 0 & -10^{-5} & 4 & -1 & -1 & -1 & -1 \\ 0 & 0 & -1 & -10^{-5} & 4 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & -10^{-5} & 4 & -1 & -1 \\ 0 & 0 & 0 & 0 & -4 & -10^{-5} & 4 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -10^{-5} & 6 \\ 0 & 0 & 0 & 0 & 0 & 0 & -6 & -10^{-5} \end{bmatrix}$$

Here, A has distinct eigenvalues -10^{-5}, -10, $-10^{-5} \pm 2i$, $-10^{-5} \pm 4i$ and $-10^{-5} \pm 6i$. To within six significant digits we find

$$\sigma_{min}(A - \mu_1 i I) = .641982 \cdot 10^{-5}$$
$$\sigma_{min}(A - \mu_2 i I) = \sigma_{min}(A + \mu_2 i I) = .308193 \cdot 10^{-5}$$
$$\sigma_{min}(A - \mu_3 i I) = \sigma_{min}(A + \mu_3 i I) = .293227 \cdot 10^{-5}$$
$$\sigma_{min}(A - \mu_4 i I) = \sigma_{min}(A + \mu_4 i I) = .518362 \cdot 10^{-5}$$

where $\mu_1 = 0$, $\mu_2 = 2$, $\mu_3 = 4$, and $\mu_4 = 6$. In this case we confirmed that $\mu_{opt} = 4$ and so to six digits we have $\beta(A) = .293227 \cdot 10^{-5}$.

These computations were performed using double precision VAX 780 arithmetic; unit roundoff $\approx 10^{-16}$. The singular values reported were obtained via the LINPACK SVD algorithm. In all cases the corresponding

$\hat{\sigma}_{min}$ estimate has correct to one significant digit. The values reported for $\beta(A)$ were checked using a systematic global search with MATLAB(1980). The global minima were always found to coincide with one of the μ_i.

Our numerical experiments so confirm the heuristic (2.6) that we recommend the following strategy for estimating $\beta(A)$:

Algorithm 2.1

1. Compute the Schur Decomposition $Q^H A Q = T$ and let $\mu_1,...,\mu_k$ denote the imaginary part of the spectrum. (See (2.5).)

2. For $j = 1,...,k$ compute $\beta_j = \hat{f}(\mu_j)$. (See (2.3).)
3. Set $\beta = \min \{ \beta_1, ..., \beta_k \}$.

There are several other reasons why we prefer this approach to one that makes serious use of FMIN. (1) We are willing to settle for an order-of-magnitude estimate of $\beta(A)$. (2) There are already several layers of approximation imbedded in our technique -- roundoff in computing T, the heuristic (2.4), etc. (3) FMIN requires several evaluations of $\hat{f}$ per call in contrast to Algorithm 2.1 which requires just one evaluation of f per μ_j.

We mention that in practice the computed eigenvalues may differ significantly from their exact counterparts. Would this not undermine the reliablity of Algorithm 2.1? After all, it makes heavy use of the imaginary parts of the computed eigenvalues. Fortunately, this is not the case for if the QR algorithm for eigenvalues is used then each computed eigenvalue is an exact eigenvalue for some $A+E$ where $\| E \|_2 \approx$ (machine precision)$\cdot \| A \|_2$. The following theorem shows that $\beta(A)$ is not sensitive to perturbations in A.

Theorem 2.2

For all $A,E \in \mathbf{C}^{n\times n}$ we have $| \beta(A+E) - \beta(A) | \leq \| E \|_2$

Proof.

Let E_1 and E_2 be such that $\alpha(A + E_1) = 0$, $\| E_1 \|_F = \beta(A)$, $\alpha(A + E + E_2) = 0$ and $\| E_2 \|_F = \beta(A + E)$. Since $\alpha(A + E + E_1 - E) = 0$ and $\alpha(A + E + E_2) = 0$ it follows that

$$\beta(A + E) = \| E_2 \|_F \leq \| E_1 - E \|_F \leq \| E_1 \|_F + \| E \|_F = \beta(A) + \| E \|_F .$$

$$\beta(A) = \| E_1 \|_F \leq \| E + E_2 \|_F \leq \| E \|_F + \| E_2 \|_F = \beta(A + E) + \| E \|_F .$$

and so $\beta(A) - \| E \|_F \leq \beta(A + E) \leq \beta(A) + \| E \|_F$.

Q.E.D.

Thus, rounding errors should not effect the reliability of Algorithm 2.1

Finally, we mention that Algorithm 2.1 should be modified as follows in the event that A is real:

Algorithm 2.2

1. Compute the Real Schur Decomposition $Q^T A Q = T$ where Q is orthogonal and T is upper quasi-triangular. Let $\mathbb{I}_+ = \{ \text{Imag}(\lambda) \mid \lambda \in \lambda(A), \text{Imag}(\lambda) \geq 0 \} = \{ \mu_1, \ldots, \mu_k \}$.
2. For $j = 1, \ldots, k$, compute $\beta_j = \hat{f}(\mu_j) = \sigma_{min}(T - \mu i I)$.
3. Set $\beta = \min \{ \beta_1, \ldots, \beta_k \}$.

The Real Schur decomposition is discussed in Golub and Van Loan (1983, Chap. 7) and can be computed using EISPACK (1970). Step 2 exploits the fact that the complex eigenvalues of a real matrix come in conjugate pairs and that $\sigma_{min}(A - \mu i I) = \sigma_{min}(A + \mu i I)$. Hence we need only examine $\sigma_{min}(A - \mu i I)$ for $\mu \in \mathbb{I}_+$ instead of $\mu \in \mathbb{I}$. Before the σ_{min} estimator can be applied the quasi-triangular matrix $T - \mu i I$ must first be reduced to triangular form. This can be accomplished with Givens rotations at minimal cost.

Section 3. The Real Perturbation Case

In many applications where nearness to instability is an issue and where the matrix in question is real, it makes sense to consider only real destabilizing perturbations . Thus, it is natural to consider

$$\beta_R(A) = \min \{ \| E \|_F \mid \alpha(A+E) \geq 0 \ , E \in \mathbf{R}^{n\times n} \}. \tag{3.1}$$

If A is unstable, then $\beta_R(A) = 0$. If A is stable, which we hereafter assume, then

$$\beta_R(A) = \min \{ \| E \|_F \mid \alpha(A+E) = 0 \ , \ E \in \mathbf{R}^{n\times n} \}. \tag{3.2}$$

It is easy to show that $\{ E \mid \alpha(A+E) = 0 \ , E \in \mathbf{R}^{n\times n} \}$ is closed and therefore has an element of minimal norm. Moreover, if $A = U\Sigma V^T$ is the (real) SVD of A then $A - \sigma_n u_n v_n^T$ is singular where u_n and v_n are the n-th columns of U and V respectively. It follows that

$$\beta_R(A) \leq \sigma_{min}(A). \tag{3.3}$$

It is possible to have strict inequality . Indeed, if we set $A = (-\lambda + i\mu)I_n$ with $\lambda , \mu > 0$, then $\sigma_{min}(A) = \sqrt{(\lambda^2 + \mu^2)}$. However, $E = \lambda I$ is real and destabilizing and so $\beta_R(A) \leq \| E \|_F = \sqrt{n}\, \lambda$. Clearly, if $\mu > \sqrt{n-1}\ \lambda$, then $\beta_R(A) < \sigma_{min}(A)$.

The following theorem characterizes $\beta_R(A)$ for the case when we have strict inequality in (3.3).

Theorem 3.1

Suppose $A \in \mathbf{R}^{n\times n}$ is stable and that $\beta_R(A) = \| E \|_F$ where $E \in \mathbf{R}^{n\times n}$ is destabilizing. If

$$\beta_R(A) < \sigma_{min}(A) \tag{3.4}$$

then

$$\begin{aligned} \beta_R(A)^2 = \min \quad & \|Ar\|^2 + \|At\|^2 - (r^T A t)^2 - (t^T A r)^2 \\ \|r\| &= 1 \\ \|t\| &= 1 \\ r^T t &= 0 \\ (r^T A t)(t^T A r) &\le 0 \end{aligned} \tag{3.5}$$

(All vector norms in this section are 2-norms.)

Proof.

Since $A + E$ has a pure imaginary eigenvalue then there exist $x,y \in \mathbf{R}^n$ (not both zero) and $\mu \in \mathbf{R}$ such that

$$(A + E - \mu i I)(x + iy) = 0. \tag{3.6}$$

Without loss of generality we can assume that $x^T y = 0$. This follows because if $x + iy$ is an eigenvector for $A + E$, then so is

$$e^{i\theta}(x + iy) = [\cos(\theta)x - \sin(\theta)y] + i[\cos(\theta)y + \sin(\theta)x].$$

Clearly, $e^{i\theta}$ can be chosen so that the real and imaginary parts of this vector are orthogonal.

From (3.6) the matrix E must satisfy

$$\begin{aligned} Ex &= -(Ax + \mu y) \equiv u \\ Ey &= -(Ay - \mu x) \equiv v. \end{aligned} \tag{3.7}$$

If $y = 0$ then $\mu = 0$. It follows from (3.6) that $A + E$ is singular and so

$$\beta_R(A) = \| E \|_F \geq \sigma_{min}(A).$$

Likewise, $x = 0$ implies that $\beta_R(A) = \| E \|_F \geq \sigma_{min}(A)$. Hence, the assumption (3.4) guarantees that x, y, and μ are nonzero.

From the constraints (3.7) the matrix E must have the form

$$E = (ux^T/x^Tx) + (vy^T / y^Ty) + WY^T$$

where $W,Y \in \mathbf{R}^{nx(n-2)}$ and $Y^Tx = Y^Ty = 0$. Since

$$\| E \|_F^2 = (\| u \|/ \| x \|)^2 + (\| v \| / \| y \|)^2 + \| WY^T \|_F^2$$

and $\| E \|_F$ is minimum, we must have $WY^T = 0$. Thus,

$$E = ux^T/(x^Tx) + vy^T/(y^Ty) \tag{3.8}$$

and

$$\begin{aligned}
\beta_R(A)^2 &= (\| u \| / \| x \|)^2 + (\| v \|/ \| y \|)^2 \\
&= (\|Ax + \mu y\| / \| x \|)^2 + (\|Ay - \mu x \| / \| y \|)^2 \\
&= (\| Ax \| / \| x \|)^2 + (\| Ay \| / \| y \|)^2 \\
&\quad + \mu^2 [(\| y \| / \| x \|)^2 + (\| x \| / \| y \|)^2] \\
&\quad - 2\mu [(x^TAy)/ y^Ty - (y^TAx)/x^Tx] .
\end{aligned}$$

As a function of μ this expression is minimized by setting

$$\mu = \mu_{opt} \equiv \frac{(x^TAy) / y^Ty - (y^TAx) / x^Tx}{y^Ty / x^Tx + x^Tx / y^Ty} \tag{3.9}$$

Thus, because E is a minimizer, we must have

$$\| E \|_F^2 = \frac{x^T A^T A x}{x^T x} + \frac{y^T A^T A y}{y^T y} - \frac{[\, x^T A y / y^T y - y^T A x / x^T x \,]^2}{x^T x / y^T y + y^T y / x^T x}$$

Since x and y are nonzero, we may assume that

$$x = cr \qquad \| r \| = 1 \;, \qquad c = \cos(\theta) \neq 0$$
$$y = st \qquad \| t \| = 1 \;, \qquad s = \sin(\theta) \neq 0 \;.$$

Setting

$$a = t^T A r \quad \text{and} \quad b = r^T A t$$

we find that

$$\| E \|_F^2 = \| Ar \|^2 + \| At \|^2 - \frac{(c^2 b - s^2 a)^2}{c^4 + s^4} \;. \tag{3.10}$$

As a function of θ, this expression is minimized if

$$0 = 2(s^4 + c^4)(c^2 b - s^2 a)(-2sca - 2scb) - (c^2 b - s^2 a)^2 (4s^3 c - 4sc^3) \;. \tag{3.11}$$

Recall from above that x and y are nonzero and so $sc \neq 0$. Moreover, we cannot have $c^2 b - s^2 a = 0$ for then (3.10) implies

$$\| E \|_F^2 = \| Ar \|^2 + \| At \|^2 \geq \min \{ \| Az \|^2 \mid \| z \| = 1 \} = \sigma_{min}(A)^2$$

contradicting the hypothesis. Thus, $0 \neq 4cs(c^2 b - s^2 a)$ can be divided into

(3.11) giving

$$(3.12) \qquad \begin{aligned} 0 &= (c^4 + s^4)(a + b) - (c^2 b - s^2 a)(c^2 - s^2) \\ &= ac^2(c^2 + s^2) + bs^2(c^2 + s^2) = ac^2 + bs^2 \quad . \end{aligned}$$

If $a = 0$ then from (3.10) we have $\| E \|_F^2 = \| Ar \|^2 + \| At \|^2 - b^2$. But

$$|b| = |r^T A t| \leq \| r \| \, \| At \| = \| At \|$$

implies

$$\| E \|_F \geq \| Ar \|^2 + \| At \|^2 - \| At \|^2 \geq \| Ar \|^2 \geq \sigma_{min}(A)^2 .$$

Thus, $a \neq 0$. Likewise, if $b = 0$ we find that $\| E \|_F \geq \sigma_{min}(A)$. Hence, we must have $a \neq 0$ and $b \neq 0$. It follows from (3.12) that

$$(3.13) \qquad (s/c)^2 = -a/b .$$

This implies that a and b must have opposite signs. Substituting into (3.10) we find after some manipulation that

$$\| E \|_F^2 = \| Ar \|^2 + \| At \|^2 - \frac{[c^2(b - (s/c)^2 a)]^2}{c^4 [1 + (s/c)^4]}$$

$$= \| Ar \|^2 + \| At \|^2 - (r^T A t)^2 - (t^T A r)^2 \quad .$$

Thus, we see that $\beta_R(A)$ is achieved by minimizing the righthand side of this expression over all unit vectors r and t that are orthogonal to each other and satisfy $(r^T At)(t^T Ar) = ab \leq 0$.

Q.E.D.

Using the theorem and the results given in its proof, we have (in principle) a procedure for calculating $\beta_R(A)$, the minimum destabilizing $E_{opt} \in \mathbf{R}^{n \times n}$, and the pure imaginary eigenvalue $i\mu_{opt}$ of $A + E_{opt}$:

Algorithm 3.1

1. Find vectors r and t that minimize

$$f(r,t) = \| Ar \|^2 + \| At \|^2 - (r^T At)^2 - (t^T A r)^2$$

subject to $\| r \| = \| t \| = 1$, $r^T t = 0$, and $(r^T At)(t^T Ar) \leq 0$. Let β denote the minimum value. Set $a = t^T Ar$ and $b = r^T At$. Without loss of generality, we may assume that $a \leq 0$ and $b \geq 0$.

2. Calculate $\sigma_{min}(A)$ and the corresponding left and right singular vectors u and v.

3. If $\beta \geq \sigma_{min}(A)$

then

$$\beta_R(A) = \sigma_{min}(A) \; ; \; \mu_{opt} = 0 \; ; \; E_{opt} = -\sigma_{min}(A) uv^T$$

else

$$\beta_R(A) = \beta \; ; \; \mu_{opt} = \sqrt{-ab} \; ; \; E_{opt} = (at - Ar)r^T + (br - At)t^T$$

The expressions for μ_{opt} and E_{opt} for the case $\beta < \sigma_{min}(A)$ follow by substituting $x = cr$ and $y = st$ into (3.8) and (3.9) and using (3.13) :

$$\mu_{opt} = [\, csb/s^2 - csa/c^2 \,] / [\, s^2/c^2 + c^2/s^2 \,] = \sqrt{-ab}$$

$$\begin{aligned} E_{opt} &= -ux^T/(x^T x) - vy^T/y^T y \\ &= -(cAr + \mu_{opt} st)(cr)^T/c^2 - (sAt - \mu_{opt} cr)(st)^T/s^2 \\ &= -(Ar + \mu_{opt}(s/c)t)r^T - (At - \mu_{opt}(c/s)r)t^T \\ &= -(Ar - at)r^T - (At - br)t^T \\ &= (at - Ar)r^T + (br - At)t^T \end{aligned}$$

It can be shown that if $\beta_R(A) = \sigma_{min}(A)$ then

$$(A + E_{opt})v = \mu_{opt}v$$

while $\beta_R(A) < \sigma_{min}(A)$ implies

$$(A + E_{opt})(cr + ist) = i\mu_{opt}(c + ist)$$

with $c = \sqrt{b/(b - a)}$ and $s = \sqrt{-a/(b - a)}$.

Note that for all orthonormal bases $\{r,t\}$ we have

$$f(r,t) = \|(t^TAr)t - Ar\|^2 + \|(r^TAt)r - At\|^2$$

$$= \min\{\|zt - Ar\|^2 \mid z \in \mathbf{R}\} + \min\{\|zr - At\|^2 \mid z \in \mathbf{R}\}.$$

Thus, the minimization problem in Step 1 of Algorithm 3.1 involves finding an orthonormal basis $\{r,t\}$ with the property that Ar is close to span$\{t\}$ and At is close to span$\{r\}$ in the least square sense subject to the constraint $(r^TAt)(t^TAr) < 0$. The resulting vector cr + ist (c and s suitably chosen) attempts to look like an eigenvector associated with a nonzero pure imaginary eigenvalue. Indeed, if A has such an eigenvalue and $A(cr + ist) = \mu i(cr + ist)$ where r and t are unit vectors, μ is real, and $c^2 + s^2 = 1$, then it is easy to show that $f(r,t) = 0$ with $(r^TAt)(t^TAr) < 0$.

Unfortunately, we were not able to devise a simple means for computing $\beta_R(A)$ based on the nice geometry of $f(r,t)$. As in the case of $\beta(A)$, we can only suggest a heuristic means of approximation. Suppose μ is real and that $u = u_1 + iu_2$ and $v = v_1 + iv_2$ are the left and right singular vectors associated with $\sigma_{min} \equiv \sigma_{min}(A - \mu i I)$. We may assume that $v_1^Tv_2 = 0$. If

$$E = -\sigma_{min}[(u_1v_1^T)/(v_1^Tv_1) + (u_2v_2^T)/(v_2^Tv_2)]$$

is defined then the equation $(A - \mu i I)(v_1 + iv_2) = \sigma_{min}(u_1 + iu_2)$ implies

that

$$(A + E - \mu i I)(v_1 + i v_2) = 0.$$

Since E is real we have

$$\beta_R(A)^2 \leq \| E \|_F^2 = \sigma_{min}(A - \mu i I)^2 \{ \| u_1 \|^2 / \| v_1 \|^2 + \| u_2 \|^2 / \| v_2 \|^2 \}$$

This suggests the following approach:

Algorithm 3.2

1. Compute the real Schur decomposition $Q^T A Q = T$ and let $\{\mu_1,...,\mu_k\} = \{ Re(\lambda) \mid \lambda \in \lambda(T), Re(\lambda) \geq 0 \}$.

2. For $j = 1,...,k$, use the σ_{min} estimator to find $\hat{\sigma}_{min} \approx \sigma_{min}(T - \mu i I)$ and singular vectors $u = u_1 + i u_2$ and $v = v_1 + i v_2$. Assume that $v_1^T v_2 = 0$. Set $\beta_k = \sigma_{min} \sqrt{ \| u_1 \|^2 / \| v_1 \|^2 + \| u_2 \|^2 / \| v_2 \|^2 }$.

3. $\beta_R = \min \{ \beta_1, ..., \beta_k \}$.

We know from our remarks in Section 2 that the minimum σ_{min} generated in this way will be very close to $\beta(A)$. Since

$$\min \{ \sigma_{min}(A - \mu i I) \mid \mu \in \mathbf{R} \} = \beta(A) \leq \beta_R(A)$$

it follows that $\beta(A)$ will be close to $\beta_R(A)$ if the quantities $\| u_1 \| / \| v_1 \|$ and $\| u_2 \| / \| v_2 \|$ encountered in the algorithm are modestly sized. We have no intuitive or rigorous understanding for these ratios, however, and we are thus unable to offer a comment concerning the quality of the approximation $\beta_R \approx \beta_R(A)$.

REFERENCES

R. Brent (1973), *Minimization Without Derivatives,* Prentice-Hall, Englewood Cliffs, New Jersey.

R. Byers (1983), " Hamiltonian and Symplectic Methods for the Algebraic Ricatti Equation", PhD Thesis, Center for Applied Mathematics, Cornell University, Ithaca, New York.

A.K. Cline, A.R. Conn, and C. Van Loan (1982), "Generalizing the LINPACK condition estimator," in *Numerical Analysis,* J.P. Hennart (ed), Lecture Note in Mathematics, No. 909, Springer-Verlag, New York.

J. Dongarra, J. Bunch, C. Moler, and G.W. Stewart (1978), *LINPACK Users Guide,* SIAM Publications, Philadelphia, Pa.

G.E. Forsythe, M. Malcolm, and C. Moler (1977), *Computer Methods for Mathematical Computations,* Prentice-Hall, Englewood Cliffs, NJ.

G.H. Golub and C. Van Loan (1983), *Matrix Computations* , Johns Hopkins University Press, Baltimore, Md.

C. Moler (1980), "MATLAB User's Guide," Technical Report CS81-1, Department of Computer Science, University of New Mexico, Albuquerque, New Mexico.

B.T. Smith et al (1970), *Matrix Eigensystem Routines: EISPACK Guide,* Springer-Verlag, New York.

C. Van Loan (1984), "Estimating the Accuracy of Computed Eigenvalues and Eigenvectors", Cornell Computer Science Technical Report TR84-777, Ithaca, New York.