Contemporary Mathematics
Volume 47, 1985

HOW NEAR IS A STABLE MATRIX TO AN UNSTABLE MATRIX?

Charles Van Loan[1]

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

1. INTRODUCTION. Suppose $A \in C^{nxn}$ 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 $\dot{x} = Ax$. 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 stability which is typically based upon $\alpha(A)$.

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

$$A = \begin{bmatrix} J_9(-.1) & 0 \\ 0 & -.001 \end{bmatrix} \in R^{10x10}$$

where $J_9(-.1)$ is a 9-by-9 Jordan block associated with the eigenvalue -.1 . If

---

1980 Mathematics Subject Classification. 65L07, 93D20

[1]Supported by the Office of Naval Research Contract N00014-83-K-0640.

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 measure of 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 n-by-n upper triangular matrix that has 1's on the diagonal and -1's above the diagonal. 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 $A \in C^{nxn}$, which we denote by $\sigma_{min}(A)$, satisfies

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

Recall that $\|E\|_F^2 = \sum\sum |e_{ij}|^2$. Moreover, if

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

is the singular value decomposition (SVD) with unitary

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

and

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

and ordered singular values

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

then the minimizing E in (1.1) is given by $E_{min} = -\sigma_n u_n v_n^H$. We mention that if A is real then U and V may be chosen to be real orthogonal in the SVD. A thorough discussion of SVD 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:

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

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 Frobenius 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

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

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 $\psi(X) = AX + XA^H$ is singular. (See Golub and Van Loan (1983, p.194).) Let $sep(A)$ denote the smallest singular value of ψ, i.e.,

$$sep(A) = \min \{ \| AX + XA^H \|_F \mid X \in C^{nxn} , \| X \|_F = 1 \}$$

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 connections between these four quantities.

THEOREM 1.1. If $A \in C^{nxn}$ is stable then

$$\tfrac{1}{2} sep(A) \leq \beta(A) \leq \sigma_{min}(A)$$

and

$$\beta(A) \leq |\alpha(A)|$$

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

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

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 a matrix representation of the linear transformation ψ.

Now let E be such that $\beta(A) = \|E\|_F$ and $\alpha(A + E) = 0$. Note from Lyapunov theory that the linear transformation $X \rightarrow (A + E)X + X(A + E)^H$ 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

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

Here, we used the easily derived inequality $\| B \otimes C \|_2 \leq \| B \|_2 \| C \|_2$.

To establish the relationship between $\alpha(A)$ and $\beta(A)$, let $Q^H A Q = T$ be the Scur 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 $\beta(A) \leq |\alpha(A)|$. Q.E.D.

If $\text{sep}(A) = 0$ implies that A is unstable, then shouldn't $\text{sep}(A) << 1$ 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 $\beta(A) \leq \text{constant} \cdot \text{sep}(A)$. The problem concerns the behavior of the matrix M in (1.2) under perturbation. If we allow arbitrary perturbations, then M has distance $\text{sep}(A)$ from the set of singular matrices. If we only allow perturbations of the form $I \otimes E + E \otimes I$, then its distance to the set of singular matrices is of order $\beta(A)$. Hence, we are not able to claim that $\text{sep}(A)$ is a reliable indicator of nearness-to-instability. Nevertheless, the reader should be aware that an efficient method for estimating $\text{sep}(A)$ is detailed in Byers (1983).

Our plan in the remainder of this paper is to focus on $\beta(A)$. In §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 §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.

2. ESTIMATING $\beta(A)$. Note that if A is stable and E is a destabilizing perturbation, then $(A + E + i\mu I)z = 0$ for some $\mu \in R$ and some nonzero $z \in C^n$. 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 the minimum singular value of $(A - \mu i I)$ and so

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

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

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

Because it does not involve 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 minimum of f in the interval.

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

Fortunately the volume of work can be reduced by an order of magnitude if A is first transformed into upper triangular form and condition estimation ideas are used. Suppose $Q^H A Q = T$ is the Schur decomposition of A. Since $Q^H(A - \mu iI)Q = T - \mu iI$, it follows that $A - \mu iI$ and $T - \mu iI$ have the same minimum singular value. (Q is unitary and singular values are preserved under unitary transformations.) Thus,

$$\beta(A) = \min \{ \sigma_{min}(T - \mu iI) \mid \mu \in R \} \tag{2.2}$$

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

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

be the estimate of $\sigma_{min}(T - \mu iI)$ 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 $f(\mu) \leq \hat{f}(\mu) \leq 1.1\, f(\mu)$. 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 [a,b] 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 $\mu_{opt} \in R$, 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.

The lemma suggests that FMIN be applied with the initial interval $[-\gamma,\gamma]$ where $\gamma = 2\|A\|_2$. ($\|A\|_2 = \|T\|_2$ can also be estimated in $O(n^2)$ operations using techniques that are described in Cline, Conn, and Van Loan(1982).) Unfortunately, FMIN returns only local minima and so a plan must be devised for finding the global minimum of f.

To this end let Π denote the imaginary part of A's spectrum,

$$\Pi = \{\text{imag}(\lambda) \mid \lambda \in \lambda(A)\} = \{\mu_1, \ldots, \mu_k\}. \tag{2.4}$$

Our computational experience has shown that the local minima of $f(\mu)$ occur in the vicinity of $\mu_1, \ldots, \mu_k$. Thus, if $\mu_1 \leq \mu_2 \leq \ldots \leq \mu_k$ and

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

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

We implemented this procedure but were pleasntly surprised to learn the following from numerous numerical experiments:

(2.5) 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 six digits, $\beta(A) = .316224\ 10^{-4}$ 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

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

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

These computations were performed using double precision VAX 780 arithmetic, unit roundoff $\cong 10^{-16}$. The singular values reported were obtained via the LINPACK SVD algorithm. In all cases the corresponding σ_{min} estimate was 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 confirm heuristic (2.5) to such an extent 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 $\mu_1,\ldots,\mu_k$ denote the imaginary part of the spectrum.
2. For $j = 1,\ldots,k$ compute $\beta_j = \hat{f}(\mu_j)$. (See (2.3).)
3. Set $\beta = \min \{ \beta_1,..,\beta_k \} \cong \beta(A)$

There are several 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 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 reliability 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 $\| E \|_2 \cong$ (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 and E in C^{nxn} 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 $(A + E + E_1 - E) = 0$ and $\alpha(A + E + E_2) = 0$ it follows that

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

and

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

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 real orthogonal and T is upper quasi-triangular, i.e., block triangular with 1-by-1 and 2-by-2 diagonal blocks. Let $\Pi_+ = \{\text{Imag}(\lambda) \mid \lambda \in \lambda(A) , \text{Imag}(\lambda) \geq 0 \} = \{ \mu_1,\ldots,\mu_k \}$.
2. For $j = 1,\ldots,k$ compute $\beta_j = \hat{f}(\mu_j) = \hat{\sigma}_{min}(T - \mu_j iI)$.
3. Set $\beta = \min\{\beta_1 ,\ldots,\beta_k \}$.

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

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

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

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

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

It is easy to show that $\{ E \mid \alpha(A + E) = 0 , E \in R^{nxn} \}$ has an element of minimal Frobenius norm. Moreover, if $A = U \Sigma V^T$ is the (real) SVD of A then the matrix $A - \sigma_n u_n v_n^T$ is singular where u_n and v_n are the n-th columns of U and V respectively. (See §1.) It follows that

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

It is possible to have strict inequality in this expression. Indeed, if we set

$A = (-\lambda + i\mu)I_n$ with $\lambda,\mu > 0$, then $\sigma_{min} = \sqrt{\lambda^2 + \mu^2}$. 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 $A \in R^{nxn}$ is stable and that $\beta_R(A) = \|E\|_F$ where $E \in R^{nxn}$ is destabilizing. If

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

then

$$\beta_R(A)^2 = \min_{\substack{r,t \in R^n \\ r^Tr + t^Tt = 1 \\ r^Tt = 0 \\ (r^TA\,t)(t^TA\,r) < 0}} \| Ar \|^2 + \| At \|^2 - (r^TA\,t)^2 - (t^TA\,r)^2$$

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

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

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

Without loss of generality we can assume that $x^Ty = 0$. This follows because if $x + iy$ is an eigenvector 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 the equations

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

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 all 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 and Y are in $R^{nx(n-2)}$ and $Y^Tx = Y^Ty = 0$. Since

$$\| E \|_F^2 = \frac{\| u \|^2}{\| x \|^2} + \frac{\| v \|^2}{\| y \|^2} + \| W Y^T \|_F^2$$

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

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

and

$$\beta_R(A)^2 = \frac{\| u \|^2}{\| x \|^2} + \frac{\| v \|^2}{\| y \|^2} = \frac{\| Ax + \mu y \|^2}{\| x \|^2} + \frac{\| Ay - \mu x \|^2}{\| y \|^2}$$

$$= \frac{\| Ax \|^2}{\| x \|} + \frac{\| Ay \|^2}{\| y \|} + \mu^2 \left[\frac{\| x \|^2}{\| y \|^2} + \frac{\| y \|^2}{\| x \|^2} \right]$$

$$-2\mu \left[\frac{x^TAy}{y^Ty} - \frac{y^TAx}{x^Tx} \right] .$$

As a function of μ this expression is minimized by setting

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

Thus, because E is a minimizer, we must have

$$\| E \|_F^2 = \frac{x^TA^TAx}{x^Tx} + \frac{y^TA^TAy}{y^Ty} - \frac{[x^TAy / y^Ty - y^TAx / x^Tx]^2}{y^Ty / x^Tx \quad x^Tx / y^Ty} .$$

Since both x and y are nonzero, we may assume that

$$x = cr \qquad \| r \| = 1 \quad , \quad c = \cos(\theta) \neq 0$$

$$y = st \qquad \| t \| = 1 \quad , \quad s = \sin(\theta) \neq 0 \quad .$$

Setting

$$a = t^TAr \quad \text{and} \quad b = r^TAt$$

we find that

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

As a function of θ, this expression is minimized if

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

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

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

contradicting the hypothesis. Thus, $0 \neq 4cs(c^2b - s^2a)$ can be divided out in (3.11) giving

$$(3.12) \quad 0 = (c^4 + s^4)(a + b) - (c^2b - s^2a)(c^2 - s^2)$$
$$= ac^2(c^2 + s^2) + bs^2(c^2 + s^2) = ac^2 + bs^2 .$$

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

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

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

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

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

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

Thus, we see that $\beta_R(A)$ is achieved by minimizing the right hand side of this expression over all unit vectors r and t that are orthogonal to each other and satisfy $(r^T A t)(t^T A r) \leq 0$. 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 R^{nxn}, and the pure imaginary eigenvalue $i\mu_{opt}$ of $A + E_{opt}$:

Algorithm 3.1

1. Find vectors r and t in R^n that minimize

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

subject to the constraints $\|r\| = \|t\| = 1$, $r^Tt = 0$, and $(r^TAt)(t^TAr) \leq 0$. Let β denote the minimum value and set $a = t^TAr$ and $b = r^TAt$. Without loss of generality we can assume that $a \geq 0$ while $b \leq 0$.

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

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

The expresssions 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):

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

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

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

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

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

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

$$= \min_{z \in R} \| zt - Ar \|^2 + \min_{z \in R} \| zr - At \|^2$$

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 $\text{span}\{t\}$ and At is close to $\text{span}\{r\}$ in the least squares sense subject to the constraint $(r^T A t)(t^T A r) < 0$. 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 should have 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 $(r^T A t)(t^T A r) < 0$.

Unfortunately, we are 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 $u = u_1 + u_2$ and $v = v_1 + iv_2$ are the left and right singular vectors associated with $\sigma_{min}(A - \mu iI) = \sigma_{min}$. (The u_i and v_i are real.) We can assume without loss of generality that $v_1^T v_2 = 0$. If

$$E = -\sigma_{min} [(u_1 v_1^T)/v_1^T v_1 + (u_2 v_2^T)/v_2^T v_2]$$

is defined, then the equation $(A - \mu iI)(u_1 + iu_2) = \sigma_{min}(u_1 + i\,u_2)$ implies that

$$(A + E - \mu iI)(v_1 + iv_2) = 0 .$$

Since E is real we have

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

This suggests the following approach:

Algorithm 3.2

1. Compute the real Schur decomposition $Q^T A Q = T$ and let $\{\mu_1, \ldots, \mu_k\} = \{ \text{Re}(\lambda) \mid \lambda \in \lambda(T), \text{Re}(\lambda) \geq 0 \}$
2. For $j = 1,\ldots,k$, use the σ_{min} estimator to find $\sigma_{min} \cong \sigma_{min}(T - \mu iI)$ and singular vectors $u = u_1 + iu_2$ and $v = v_1 + iv_2$. Force $v_1^T v_2 = 0$ and set $\beta_k = \hat{\sigma}_{min} [\, \|u_1\|^2 / \|v_1\|^2 + \|u_2\|^2 / \|v_2\|^2]$.
3. Set $\beta_R = \min \{ \beta_1, \ldots, \beta_k \}$.

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

$$\min_{\mu \in R} \sigma_{min}(A - \mu iI) = \beta(A) \leq \beta_R(A)$$

it follows that $\beta(A)$ will be close to $\beta_R(A)$ if the quotients $\|u_1\| / \|v_1\|$ and $\|u_2\| / \|v_2\|$ encountered in the algorithm are modestly sized. We have no intuitive or rigorous understanding of these ratios, however, and are unable to comment on the quality of β_R as an approximation to $\beta_R(A)$.

ACKNOWLEDGEMENT The author would like to thank Professor Jim Demmel of NYU for his interesting comments concerning the estimation of $\sigma_{min}(A - \mu iI)$.

BIBLIOGRAPHY

1. R. Brent, Minimization Without Derivatives, Prentice-Hall, Englewood-Cliffs, New Jersey, 1973.

2. R. Byers, "Hamiltonian and Symplectic Methods for the Algebraic Ricatti Equation", Ph.D. Thesis, Center for Applied Mathematics, Cornell University, Ithaca, New York.

3. A.K. Cline, A.R.Conn, and C. Van Loan, "Generalizing the LINPACK Condition Estimator", in Numerical Analysis, J.P. Hennart (ed.), Lecture Notes in Mathematics, No. 909, Springer-Verlag, New York, pp. 73-83.

4. J. Dongarra, J.Bunch, C.B. Moler, and G.W.Stewart, LINPACK Users Guide, SIAM Publications, Philadelphia, Pa., 1978.

5. G.E. Forsythe, M.Malcolm, and C.B. Moler, Computer Methods for Mathmatical Computations, Prentice-Hall, Englewood-Cliffs, New Jersey, 1977.

6. G.H. Golub and C. Van Loan, Matrix Computations, Johns Hopkins Press, Baltimore, Maryland, 1983.

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

8. B.T. Smith et al, Matrix Eigensystem Routines: EISPACK Guide, Springer Verlag, New York, 1970.

DEPARTMENT OF COMPUTER SCIENCE
405 UPSON HALL
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853