The computation and sensitivity of double eigenvalues *

Ross A. Lippert and Alan Edelman

March 15, 1999

*This paper is a slightly modified form of the first chapter in the PhD thesis of Ross Lippert [11] This paper is an extension of an article of the same name by Lippert and Edelman included in [13].

Abstract

This paper explores the problem left open by Wilkinson of computing the distance of a given diagonalizable matrix to the nearest non-diagonalizable matrix. Algorithms for finding the nearest non-diagonalizable matrix to a given diagonalizable matrix are presented and analyzed. It will be shown that this problem reduces to finding the critical points of the *spectral portrait* (the graph of the pseudospectrum) and that critical points are ill-conditioned if and only iff the corresponding nearby non-diagonalizable matrix problem is ill-conditioned.

1 Introduction

A fundamental problem in numerical analysis is to understand how eigenstructure affects the numerical behavior of algorithms. Classical insights come from studying the conditioning of the eigenvalue problem. From this point of view, ill-conditioning is related to nearness to a defective matrix. More recent insights are based on pseudospectra where interesting behavior is associated with non-normality. In this paper, we revisit the classical insights with the added technology of differential geometry. Pseudospectra are windows into n^2 dimensional space. Singularities in the spectral portrait are relics of a defective matrix. "Singularities" of these singularities (to be defined later) indicate a center of curvature of the defective matrices and play a role in conditioning.

If A is an $n \times n$, diagonalizable matrix, we can find a neighborhood about A in matrix space where the eigenvalues are all analytic functions of the elements of A. Thus we have n analytic functions $\lambda_j = \mathbf{eig}_j(A)$. An eigenvalue λ_j is said to be ill-conditioned if first order variations A result in very large first order variations in λ_j. A textbook example is the 12×12 Frank matrix

$$F_{12} = \begin{pmatrix} 12 & 11 & 10 & \dots & 3 & 2 & 1 \\ 11 & 11 & 10 & \dots & 3 & 2 & 1 \\ 0 & 10 & 10 & \dots & 3 & 2 & 1 \\ 0 & 0 & 9 & \dots & 3 & 2 & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 2 & 2 & 1 \\ 0 & 0 & 0 & \dots & 0 & 1 & 1 \end{pmatrix}.$$

An ϵ perturbation of the upper right element results in a change in the smallest two eigenvalues of order $10^7\epsilon$ (a relative change of order 2×10^8).

Wilkinson [22, 23] showed that if the conditioning of some of the eigenvalues of a matrix is poor, then that matrix must lie near some defective matrix. He further presented several bounds on this distance. In particular, the Frobenius norm distance from F_{12} to the nearest matrix with a double eigenvalue is only 1.822×10^{-10}.

Nearness to a defective matrix can also sometimes explain transient behaviors in the matrix exponenital. Such transient behavior is explored by Trefethen [20], who exhibited the transient behavior of

$$A = \begin{pmatrix} -1 & 5 \\ 0 & -2 \end{pmatrix}, \tag{1}$$

a matrix with distinct negative eigenvalues. His message from this example is that defective eigenvalues are neither necessary nor sufficient for transient growth. It is curious, however, that in his example, there is a defective matrix lurking. Let U be defined by the singular value decomposition $U\Sigma V^T = A + \frac{3}{2}I$. Then

$$U^TAU = \begin{pmatrix} -1.5 & 5.0495 \\ 0.049509 & -1.5 \end{pmatrix}$$

and thus A is a distance 0.049509 ($= 9\times10^{-3}||A||_F$) from a matrix $\hat{A}$ with a double eigenvalue, -1.5. A plot of the norms of the exponentiations of A and $\hat{A}$ in Figure 1 shows that the defective matrix does play a role.

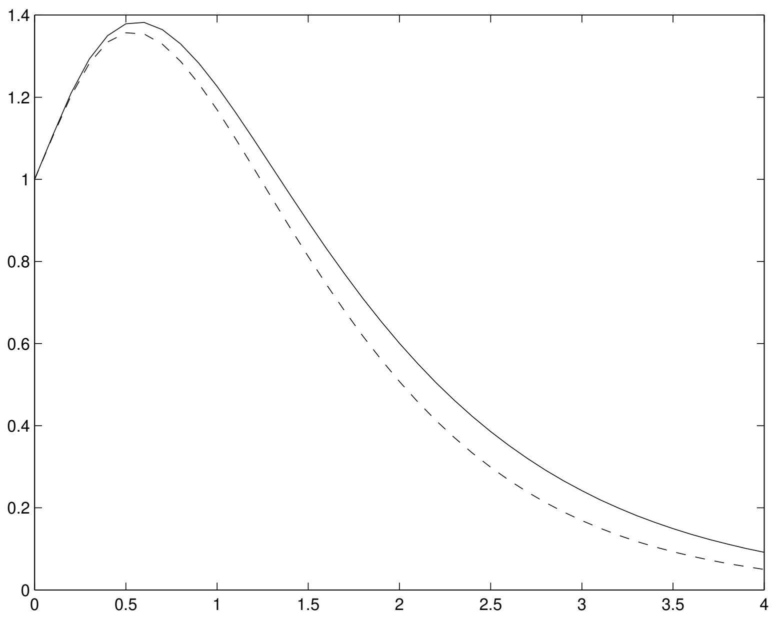

Figure 1: $||e^{tA}||_2$ (solid) and $||e^{t\hat{A}}||_2$ (dashed) vs. t for A defined by (1) and its nearest defective matrix

Given the significance of a lurking nearby defective matrix in these examples, this paper studies the computation of the shortest (least squares) distance of any matrix, A, with distinct eigenvalues, to the nearest matrix, $\hat{A}$, with at least one repeated eigenvalue. Wilkinson reviewed and critiqued several bounds on this distance, although no procedure was given for this computation. Recently, Malyshev [15] has studied the 2-norm version of this problem. There has also been recent work on the computation of the distance to more general Jordan structures in [3, 8, 6, 7, 12].

This paper closely studies the n^2-1 dimensional surface of defective matrices and presents two methods which will identify the $\hat{A}$ for a given A and give appropriate condition numbers on $\hat{A}$ and its repeated eigenvalue. One method proposed will relate the nearby $\hat{A}$ to the critical points of the spectral portrait. We shall find that the conditioning of $\hat{A}$ is directly related to the degeneracy of the associated critical point. A second method proposed relies on successive improvements to an approximation of the repeated eigenvalue. This method will have a stability which is also dependent upon conditioning of the critical points of the spectral portrait. Although we feel that there is more theoretical value in the first method, we present the second because of its easy implementation in terms of common Matlab operations.

Throughout this paper, unless otherwise stated, we will use the Frobenius matrix norm and inner product. We also assume passing familiarity with concepts from differential geometry and singularity theory. (We recommend [1] as an excellent reference for singularity theory.)

2 The geometry of eigenvalues

In this section, we will examine the level sets of the **eig** function and their envelope, the double eigenvalue variety. We will consider a matrix A with eigenvalue λ to belong to a surface M_λ in matrix space. This is a surface of co-dimension 1 with a continuous normal vector field almost everywhere. The envelope of the M_λ, $\mathcal{D}$, the double eigenvalue variety is comprised of all matrices with at least one repeated eigenvalue. This surface is also of co-dimension 1 and also has a continuous normal field almost everywhere.

2.1 The level sets of eig

Consider the function $\mathbf{eig}(A)$ returning some eigenvalue of A. If A is diagonal, then **eig** may be presumed analytic in a neighborhood of A. We define $M_\lambda = \mathbf{eig}^{-1}(\lambda)$ to be the set of all matrices over either $\mathcal{R}$ or $\mathcal{C}$ with eigenvalue λ. M_λ is a variety which is differentiable almost everywhere. M_λ may be thought of as a level set of **eig** . All $A \in M_\lambda$ can be characterized by the constraint equation $e(\lambda, A) = 0$ for λ fixed, where $e(\lambda, A) = \mathbf{eig}(A) - \lambda$ in a neighborhood of A.

It is well known (see [9], pp. 320-324) that if $\lambda = \mathbf{eig}(A)$ is a simple eigenvalue with right eigenvector x ($Ax = \lambda x$) and left eigenvector y ($y^H A = \lambda y^H$), then variations in λ relate to the variations in A by the formula

$$\dot{\lambda} = \frac{y^H \dot{A} x}{y^H x}. \tag{2}$$

This formula captures the differential behavior of $\lambda(t) = \mathbf{eig}(A(t))$ at diagonalizable matrices. More generally, if $\lambda(t)$ is independent of $A(t)$ the derivative of the constraint function, e, is

$$\frac{d}{dt} e(\lambda(t), A(t)) = \frac{y^H \dot{A} x}{y^H x} - \dot{\lambda}. \tag{3}$$

(Note, the condition $\lambda(t) = \mathbf{eig}(A(t))$ implies $e = \frac{d}{dt} e = 0$ reducing to Equation (2).)

The differential behavior of any scalar (analytic) function $z = g(w)$ can be captured by a gradient and a Riemannian (Hermitian) inner product.

$$\dot{z} = \langle \nabla g, \dot{w} \rangle.$$

Therefore, using the Frobenius inner product on matrices $\langle A, B \rangle = \mathrm{tr}(A^H B)$, we may rewrite the perturbation formula (2) as

$$\dot{\lambda} = \langle \frac{y x^H}{x^H y}, \dot{A} \rangle,$$

from which we see that the gradient of **eig** is $\nabla\mathbf{eig} = \frac{y x^H}{x^H y}$, and the differential behavior of e is

$$\frac{d}{dt} e(\lambda(t), A(t)) = \langle \nabla\mathbf{eig}\ , \dot{A} \rangle - \dot{\lambda}.$$

Geometrically, the gradient of any function is perpendicular to that function's level sets. Thus, $\frac{y x^H}{x^H y}$ is normal to the surface M_λ. If we select y and x such that $||y|| = ||x|| = 1$ then

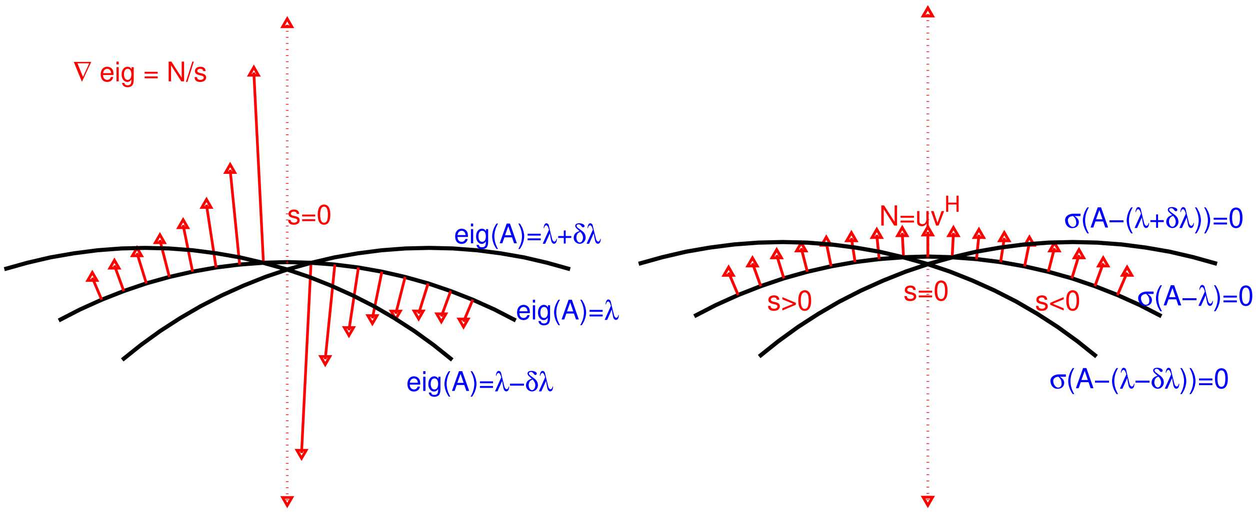

Figure 2: (Left) M_λ defined by the constraint $\mathbf{eig}(A) - \lambda = 0$ for fixed λ. The gradient of the constraint function points in the direction of increasing λ and has magnitude inversely proportional to the distance to $M_{\lambda+\delta\lambda}$. (Right) the M_λ defined by the constraint $\sigma_{\min}(A - \lambda I) = 0$ for fixed λ, giving a gradient of unit length in the direction of increasing λ, the distance to next M_λ being proportional to $s = v^H u$.

$N = yx^H$ is the unit normal (unique up to sign [or phase]). We may identify $s = x^H y$ as the inverse of the length of the gradient (also up to sign [or phase]).

There is an unfortunate characteristic of the constraint function e. At points of M_λ where the eigenvalue, λ, is ill-conditioned, the right and left eigenvectors of λ are close to being perpendicular, thus $s = x^H y$ is very small and the magnitude of $\nabla\mathbf{eig}$ approaches infinity. If A were a matrix with a nontrivial Jordan block for λ then $x^H y$ would vanish and the gradient of **eig** (as well as the differential of e) would be infinite.

To obtain a better differential behavior at A with ill-conditioned λ, we must dispense with e and use a constraint equation with a better differential behavior to characterize the M_λ surfaces. If we let $f(\lambda, A) = \sigma_{\min}(A - \lambda I)$, we may define M_λ by the constraint equation $f(\lambda, A) = 0$, λ fixed. The function f has the differential behavior

$$\frac{d}{dt} f(\lambda(t), A(t)) = u^H \dot{A} v - \dot{\lambda} u^H v, \tag{4}$$

where u and v are the right and left singular vectors of $A - \lambda I$ (when $f(\lambda, A) = 0$ the phases of u and v are such that the right hand side of (4) is real and nonnegative).

Observe that u and v are unit length left and right eigenvectors (respectively) of A, and that setting $\frac{d}{dt} f = 0$ gives

$$\dot{\lambda} u^H v = \langle uv^H, \dot{A} \rangle, \tag{5}$$

which recovers (2). The gradient f with λ fixed is the unit normal to M_λ, while the cosine of the eigenvectors weights the $\dot{\lambda}$ term. Figure 2 summarizes the distinction between the differential behaviors of e and f.

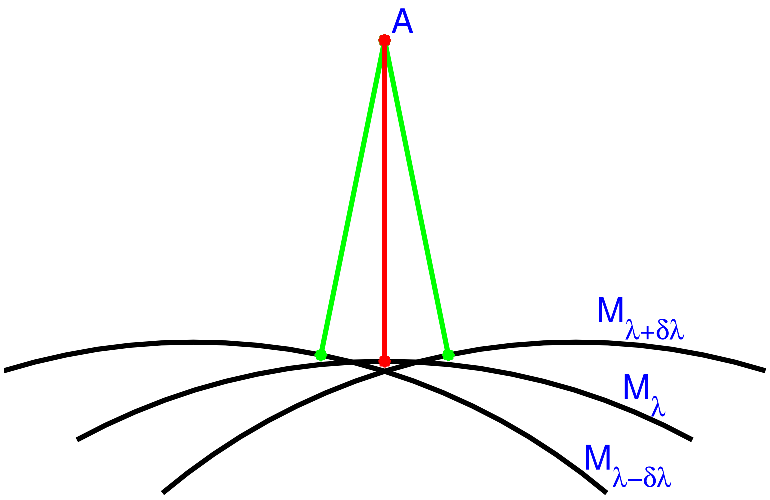

Figure 3: The envelope can be identified by the critical points of the distance from A to M_λ.

Another way to view $f(\lambda, A)$ is as the distance from A to M_λ. Since f is the distance function to M_λ, the magnitude of its gradient (with λ fixed) must be of unit length and perpendicular to M_λ. (One can derive 5 directly from this interpretation without differentiating f.)

2.2 The double eigenvalue variety

The double eigenvalue variety, $\mathcal{D}$, is the envelope of the family M_λ. Recall that the envelope of a family of surfaces S_t parameterized by t can be defined as a *tangent surface*, E, having the property that for $x \in E$, there exists a t_0 such that $x \in S_{t_0}$ and the tangent space of E at x coincides with the tangent space of S_{t_0} at x.

The characterization of the M_λ in terms of $f(\lambda, A) = \sigma_{\min}(A - \lambda I)$ allows us to analyze $\mathcal{D}$. It is well known ([21], pp. 9-10) that $A \in \mathcal{D}$ iff there exist λ, u, and v such that $Av = \lambda v$, $u^H A = \lambda u^H$, and $u^H v = 0$. Thus, for a differentiable curve $A(t) \in \mathcal{D}$ with repeated eigenvalue $\lambda(t)$, we have by (4)

$$f(\lambda(t), A(t)) = \frac{d}{dt} f(\lambda(t), A(t)) = 0 = u^H \dot{A} v = \langle uv^H, \dot{A} \rangle,$$

giving the unit normal $N = uv^H$ of $\mathcal{D}$ at $A(t)$, confirming that $\mathcal{D}$ is the envelope of the M_λ.

We can also recover an important property of envelopes since $f(\lambda, A)$ is the distance function from A to M_λ. It is the case for all differentiably parameterized families of surfaces, that the envelope is given by the the critical points of the distance function, i.e. $\hat{A} \in \mathcal{D}$ when $\frac{d}{dt} f(\lambda(t), A) = 0$ for some A where $\hat{A}$ is the nearest element of M_λ to A (see Figure 3).

In some sense, we have arrived at a geometric reasoning for the choice of Wilkinson's "locally fastest" perturbation to a matrix A near $\mathcal{D}$ which would move A to $\mathcal{D}$ nearly optimally. If one suspects that two eigenvalues, λ_1, λ_2, of A can be coalesced with a small perturbation

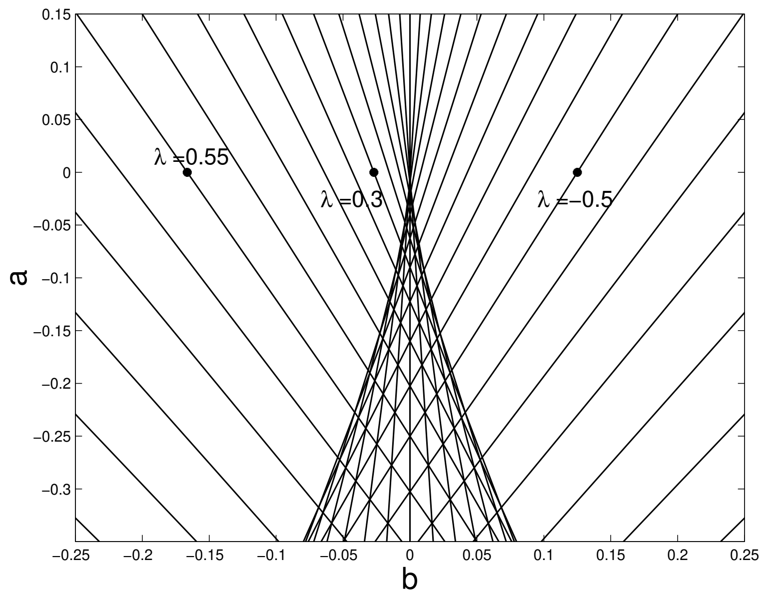

Figure 4: A graph in the a-b plane of the lines of polynomials of the form $p(x) = x^3 + ax + b$ satisfying $p(\lambda) = 0$.

E, Wilkinson suggested using a perturbation of the form $E = \sigma(y_1 + y_2)(x_1 + x_2)^H$ where the y_i and the x_i are the left and right eigenvectors of λ_i and σ is adjusted so that $A + E \in \mathcal{D}$ (Section 9 of [23]).

From the preceding results, we see that if $\hat{A} \in \mathcal{D}$ then the normal line $\hat{A} + \sigma uv^H$ (parameterized by σ) passes through $\mathcal{D}$ along a perpendicular, thus the perturbation $E = \sigma uv^H$ can be thought of as the perturbation which maximizes the change in the distance to $\mathcal{D}$ for small σ and is then the "locally fastest" perturbation to a nearby diagonalizable matrix A.

If one is close to $\mathcal{D}$ then one expects $(x_1 + x_2)/2$ will be close to v and $(y_1 + y_2)/2$ will be close to u. The "locally fastest" perturbation is then expected to be close to the actual normal of the variety at the actual minimizing point. We will elaborate on this heuristic in Section 3.1.

2.3 An illustration by characteristic polynomials

For purposes of "picturing" $\mathcal{D}$ and M_λ geometry, we consider monic n^{th} degree polynomials, which may be thought of as projections of the $n \times n$ matrices via the characteristic polynomial. Consider the cubic polynomials of the form $p(x) = x^3 + ax + b$ (corresponding to the traceless 3×3 matrices). Polynomials with a fixed root λ can be represented by a linear relation between a and b of the form

$$\lambda^3 + a\lambda + b = 0.$$

A collection of these fixed root lines is shown in Figure 4.

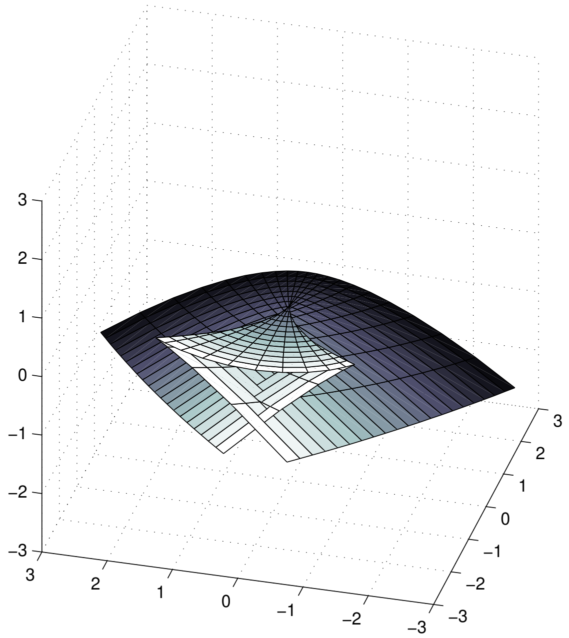

Figure 5: A graph in a-b-c space of the double root variety, or swallowtail.

The envelope of those lines is a cusp. By varying the constraint equation

$$3\lambda^2\dot{\lambda} + a\dot{\lambda} + \dot{a}\lambda + \dot{b} = 0,$$

we obtain the gradient of λ as a function of a and b,

$$\dot{\lambda} = \frac{-(\lambda, 1) \cdot (\dot{a}, \dot{b})}{3\lambda^2 + a}.$$

We may define a unit normal $N = \frac{(\lambda,1)}{\sqrt{\lambda^2+1}}$ and set the inverse magnitude of the gradient $s = \frac{3\lambda^2+a}{\sqrt{\lambda^2+1}}$.

The envelope of the lines of this single root family of polynomials is given by $s = 0$, or $a = -3\lambda^2$ (and thus $b = 2\lambda^3$ by the constraint). This gives a curve of polynomials with double root λ, seen as the cusp-shaped envelope in Figure 4. Observe that the cusp point is the polynomial $p(x) = x^3$, the triple root polynomial.

More complicated shapes arise for quartic polynomials of the form $x^4 + ax^2 + bx + c$. In this case, the double root variety is a collection of multiple cusps, the swallowtail.

In matrix space, the cusp may be formed from the 1-dimensional family of matrices

$$M(\lambda) = \begin{pmatrix} \lambda & \sqrt{1-2\lambda^2} & \lambda \\ 0 & -2\lambda & \sqrt{1-2\lambda^2} \\ 0 & 0 & \lambda \end{pmatrix}.$$

One then has $\det(xI - M(\lambda)) = x^3 - 3\lambda^2 x + 2\lambda^3$, and it is clear that $uv^H = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$.

One can verify that $\det(xI - M(\lambda) - \sigma uv^H) - \det(xI - M(\lambda)) = \sigma(\lambda x + 1)$, parallel to the normals of the double root polynomials. For the quartic case, we may use the two parameter surface

$$M(\lambda, \mu) = \begin{pmatrix} \lambda & 1 & -\lambda^3 + \frac{\lambda}{2} & \lambda^2 \\ 0 & -\lambda - \sqrt{\mu} & 1 + \lambda^4 - \lambda^2(1-\mu) & -\lambda^3 + \frac{\lambda}{2} \\ 0 & 0 & -\lambda + \sqrt{\mu} & 1 \\ 0 & 0 & 0 & \lambda \end{pmatrix},$$

for which $\det(xI - M(\lambda, \mu)) = (x^2 - \lambda^2)^2 - \mu(x - \lambda)^2$.

These matrix families embed the cusp and the swallowtail in matrix space as subsurfaces in (respectively) 9 and 16 dimensional spaces.

2.4 A pseudospectra "movie"

Since we wish to find the nearest matrix $\hat{A} \in \mathcal{D}$ to a matrix A, it is useful to study the normal line in matrix space, $\hat{A} + \sigma uv^H$ (parameterized by σ). We proceed by taking a journey along a normal to $\mathcal{D}$ and plotting the pseudospectra as we go. The spectral portrait will be our 2 dimensional window onto n^2 dimensional matrix space.

We have seen in the previous sections that along this line, the double eigenvalue separates into two distinct eigenvalues which move apart initially with great speed. Additionally, they will leave behind a stationary point whose location and character will remain invariant as we go up the normal line for some finite length.

We start out on the cubic polynomial example matrix

$$M(\lambda) = \begin{pmatrix} \lambda & \sqrt{1-2\lambda^2} & \lambda \\ 0 & -2\lambda & \sqrt{1-2\lambda^2} \\ 0 & 0 & \lambda \end{pmatrix}.$$

with $\lambda = 0.15$. The series of spectral portraits presented (Figures 6 through 10) are those of $M - \sigma uv^H$ for various interesting σ.

The reader studying these plots may wish to recall a certain "conservation law" which comes from planar bifurcation theory (see [16, 2]). The "law" is that if all the critical points of a surface are nondegenerate, they may be assigned charges, all extrema get a $+1$ charge and all saddles get a -1 charge, which will be locally conserved through all continuous deformations of the surface. Additionally, spectral portraits are continuous, and their contours approach concentric circles about the origin for large $|\lambda|$. Thus any spectral portrait can be continuously deformed into a bowl and therefore has a total charge of $+1$.

The contour lines in Figures 6 through 10 are not uniform, but have been chosen to be dense about the current value of σ and to highlight other important qualitative features.

We can summarize the sights we have seen on this journey in Table 1. The reader may wish to review this section as he/she reads subsequent sections to understand the mathematical significance of the observations listed here and to illustrate the mathematical phenomena described in subsequent sections.

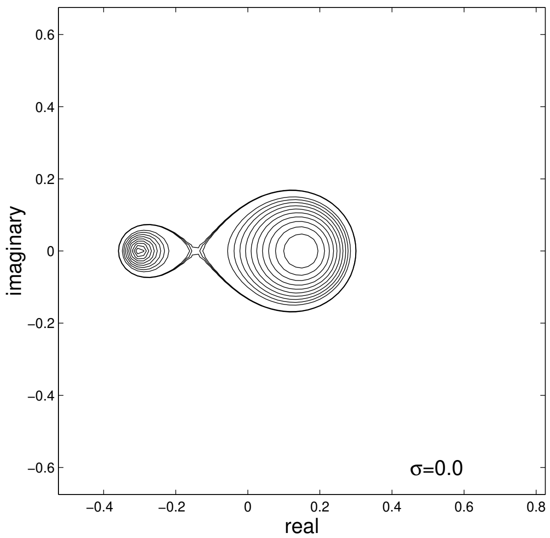

Figure 6: At $\sigma = 0$ we have a single eigenvalue at -0.3 and a parabolic valley for the double eigenvalue at 0.15. The conservation law indicates that there is also a saddle in the portrait (it sits between the valleys).

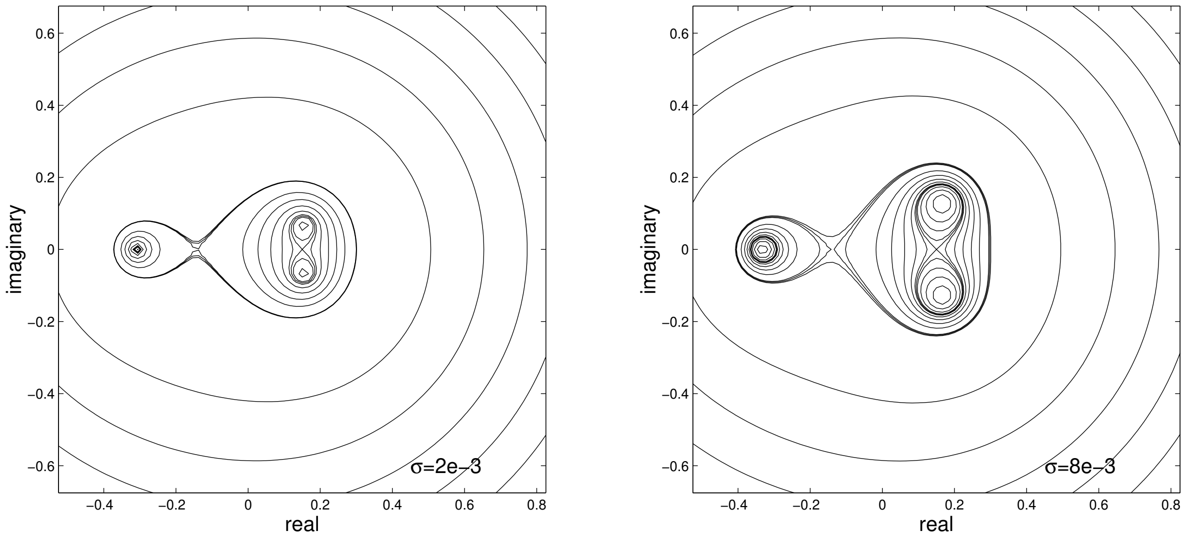

Figure 7: As we begin to move off $\mathcal{D}$, the double eigenvalue splits into two complex conjugates which move off from their origin very quickly. The valley of the double eigenvalue (charge $+1$) has split into two valleys ($+1$ each) with a saddle (-1) between them.

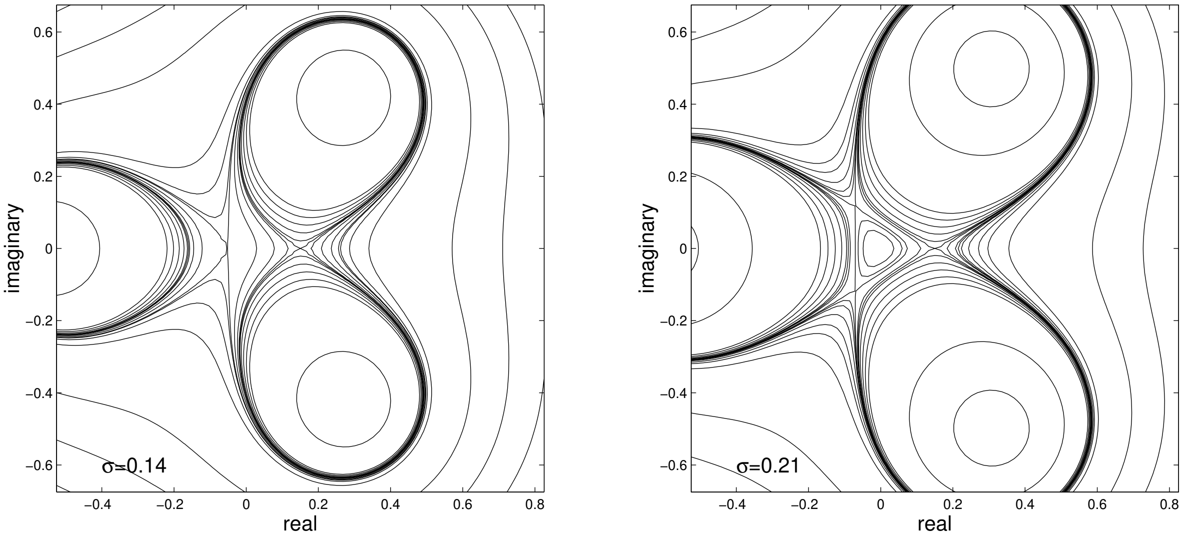

Figure 8: We then see a critical change to the left of $\lambda = 0.15$ as the saddle which initially sat between the two valleys at $\sigma = 0$ decays into two saddles and a maximum.

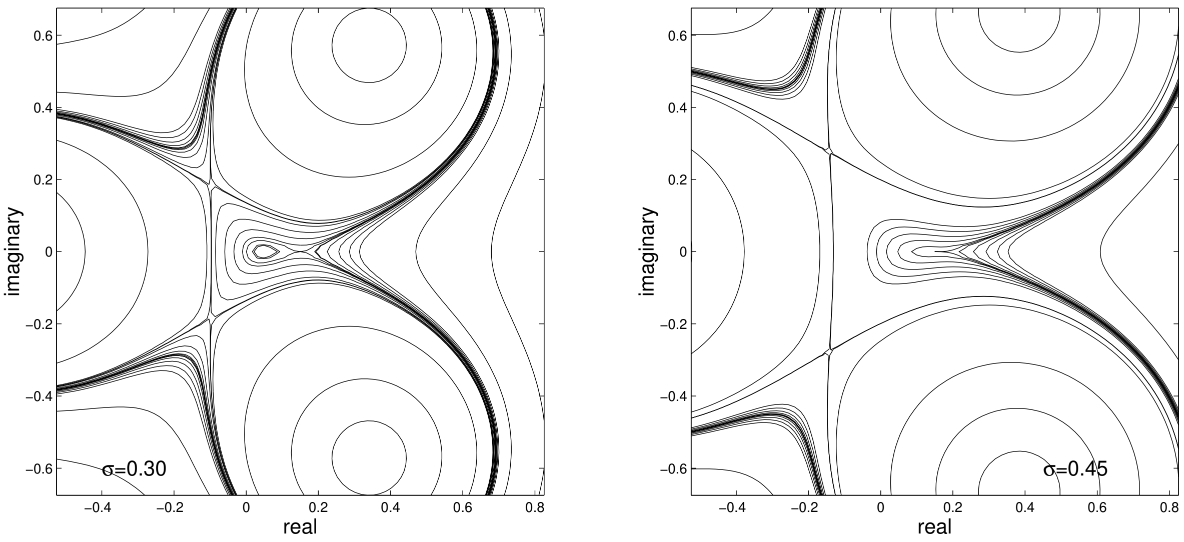

Figure 9: The maximum which was spun off from the decay collides with the saddle at $\lambda = 0.15$. At the focal distance $\sigma = 0.45$ the maximum and saddle have annihilated each other, creating a degenerate shoulder.

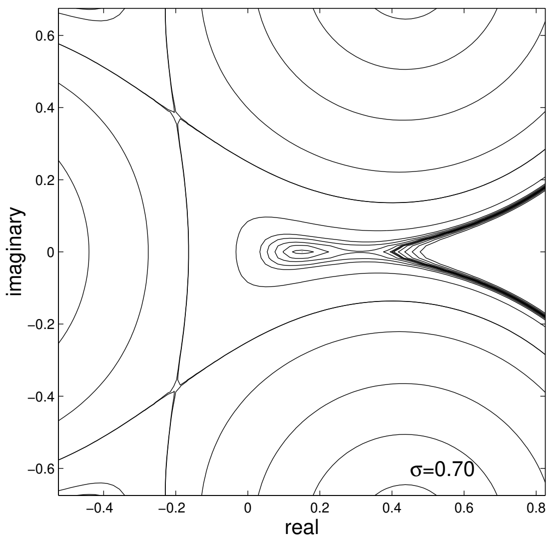

Figure 10: After the collision, the critical points have exchanged charges. The $\lambda = 0.15$ critical point is now a maximum while a saddle heads from it to the right.

σ	observation
$\sigma = 0.0$	double eigenvalue
$0.0 < \sigma < 0.45$	double eigenvalue splits leaving behind a saddle
$\sigma = 0.45$	the saddle becomes degenerate
$\sigma > 0.45$	the saddle has become a local maximum

Table 1: phenomena observed along the normal

3 Finding the nearest double eigenvalue

One way to find the nearest $\hat{A} \in \mathcal{D}$ to a given A is by directly finding the repeated eigenvalue λ of $\hat{A}$. If the repeated eigenvalue λ were known then one can take $\hat{A}$ to be the nearest matrix to A that has eigenvalue λ (using the singular value decomposition to find $\hat{A}$).

This section explores methods of calculating that λ. We will begin by exploring the common heuristic of averaging two nearby eigenvalues of A. A fixed point algorithm based on an improvement of this idea will be presented. We then show that λ can also be determined by an examination of pseudospectra of A, namely via the critical points of the spectral portrait.

3.1 A perturbative look at double eigenvalue averaging

A small perturbation in the normal direction of a generic point $\hat{A} \in \mathcal{D}$ results in a diagonalizable matrix $A = \hat{A} + \sigma u v^H$. The J_2 block that is part of the structure of $\hat{A}$ bifurcates into two J_1 blocks. In this section, we will examine the heuristic of approximating the double eigenvalue of $\hat{A}$ by the average of the two nearest eigenvalues of A by examining the Puiseux series (see [10]) in σ.

Puiseux series are nothing new in the description of the bifurcation of multiple eigenvalues, however, the additional contribution that we add is the derivation of the series along the geometric normal of $\mathcal{D}$. We will also derive the second order behavior of the series along the normal which can be used to measure the error of the averaging heuristic.

We quantify the non-surprising fact that the averaging heuristic tends to break sooner (i.e. for smaller σ) when $\hat{A}$ is close to a nongeneric point of $\mathcal{D}$, that is, is close to being more defective than J_2 or derogatory. Later sections will show that one cannot expect to find a well-conditioned double eigenvalue if the nearby $\hat{A}$ is close to being nongeneric, and thus the averaging heuristic must suffer from difficulties which will also plague the more sophisticated algorithms suggested in this paper.

We adopt the convention, in this section and throughout this paper (with the exception of Section 5.2), of using v and u for the right and left eigenvectors respectively, for $\hat{A} \in \mathcal{D}$ and of writing the singular value decomposition of $\hat{A}$ as

$$\hat{A} = \lambda I + \tilde{U}\tilde{\Sigma}\tilde{V}^H,$$

where $\tilde{U} = (\, u_1 \quad u_2 \quad \dots \quad u_{n-1}\,)$, $\tilde{V} = (\, v_1 \quad v_2 \quad \dots \quad v_{n-1}\,)$, and $\tilde{\Sigma} = \mathrm{diag}(\sigma_1, \dots, \sigma_{n-1})$. Additionally, we set $r = \tilde{U}^H v$ and $l = \tilde{V}^H u$, noting that $v = \tilde{U} r$ and $u = \tilde{V} l$ since $u^H v = 0$ and $(\, \tilde{U} \quad u\,)$ and $(\, \tilde{V} \quad v\,)$ are unitary matrices.

We now consider a Puiseux series expansion of the eigenvalues and right eigenvectors of $A(\sigma) = \hat{A} + \sigma u v^H$ given by

$$\lambda + \lambda(\sigma) = \lambda + \omega\sigma^{\frac{1}{2}}\lambda_1 + \omega^2\sigma\lambda_2 + \omega^3\sigma^{\frac{3}{2}}\lambda_3 + \cdots$$

and

$$v + v(\sigma) = v + \omega\sigma^{\frac{1}{2}}v_1 + \omega^2\sigma v_2 + \omega^3\sigma^{\frac{3}{2}}v_3 + \cdots$$

satisfying

$$(A(\sigma) - (\lambda + \lambda(\sigma))I)(v + v(\sigma)) = 0, \tag{6}$$

where $\omega^2 = 1$ and $v^H v_k = 0$. Since $v^H v_k = 0$ we will write $v_k = \tilde{V} r_k$ for some $n-1 \times 1$ vector r_k and let $r(\sigma) = \tilde{V}^H v(\sigma)$. The series for the average of the two eigenvalues is given by

$$\lambda + \frac{1}{2}(\lambda(\sigma)|_{\omega=1} + \lambda(\sigma)|_{\omega=-1}) = \lambda + \sigma\lambda_2 + \sigma^2\lambda_4 + \cdots.$$

Thus, averaging will be a good approximation to λ as long as the λ_2 term is relatively small.

Substituting $A(\sigma) - \lambda = \sigma u v^H + \tilde{U}\tilde{\Sigma}\tilde{V}^H$ into (6) we have

$$(\tilde{U}\tilde{\Sigma}\tilde{V}^H + \sigma u v^H - \lambda(\sigma))(v + v(\sigma)) = 0,$$

which can be simplified to

$$(\tilde{U}\tilde{\Sigma}\tilde{V}^H - \lambda(\sigma))v(\sigma) - \lambda(\sigma)v + \sigma u = 0. \tag{7}$$

We split (7) into two sets of equations, first by multiplying on the left by u^H,

$$\lambda(\sigma)\sigma = u^H v(\sigma), \tag{8}$$

and then by multiplying on the right by $\tilde{U}^H$,

$$\tilde{\Sigma} r(\sigma) - \lambda(\sigma)\tilde{U}^H \tilde{V} r(\sigma) = \lambda(\sigma) r. \tag{9}$$

We may invert Equation (9) to obtain

$$r(\sigma) = \lambda(\sigma)(\tilde{\Sigma} - \lambda(\sigma)\tilde{U}^H\tilde{V})^{-1} r.$$

To find $\lambda(\sigma)$ we substitute into (8) to get

$$\sigma = \lambda(\sigma) u^H v(\sigma) = \lambda(\sigma) l^H r(\sigma) = \lambda^2(\sigma) l^H (\tilde{\Sigma} - \lambda(\sigma)\tilde{U}^H\tilde{V})^{-1} r,$$

or

$$\sigma = \sum_{k \geq 1} u^H (\tilde{V}\tilde{\Sigma}^{-1}\tilde{U}^H)^k v \lambda^{k+1}(\sigma) = \sum_{k\geq 1} c_k \lambda^{k+1}(\sigma), \tag{10}$$

which must be inverted to obtain the series for $\lambda(\sigma)$ (the reader may note that $\tilde{V}\tilde{\Sigma}^{-1}\tilde{U}^H = (\hat{A} - \lambda I)^\dagger$, the Moore-Penrose inverse of $\hat{A} - \lambda I$, and that $c_1 = u(\hat{A} - \lambda I)^\dagger v = u^H w$ where w is the right generalized eigenvector of $\hat{A}$).

We seek only the first two coefficients of $\lambda(\sigma)$. They are given by

$$\lambda_1 = \frac{1}{\sqrt{u^H w}} \tag{11}$$

$$\lambda_2 = \frac{u^H(\tilde{V}\tilde{\Sigma}^{-1}\tilde{U})^2 v}{2(u^H w)^2}. \tag{12}$$

From (11) and (12) we see that the second order term in the expansion becomes significant when

$$|\sigma| \sim 4 \left| \frac{(u^H w)^3}{(u^H(\tilde{V}\tilde{\Sigma}^{-1}\tilde{U})^2 v)^2} \right|.$$

Thus we see that the averaging heuristic can be expected to work poorly if $u^H w$ $(= l^H \tilde{\Sigma}^{-1} r)$ is small or if $||\tilde{V}\tilde{\Sigma}^{-1}\tilde{U}^H||_2 = ||(\hat{A}-\lambda I)^\dagger||_2 = \frac{1}{\sigma_{n-1}}$ is large. There is an algebraic interpretation for both of these cases. If $u^H w = 0$ then it can be shown that $\hat{A}$ must have at least $J_3(\lambda)$ defectiveness. If $||(\hat{A}-\lambda I)^\dagger||_2 = \infty$, then $\hat{A} - \lambda I$ must have two vanishing singular values, and thus has a derogatory Jordan structure in λ.

In summary we see that the averaging heuristic will produce close approximations to λ for nearby $\hat{A}$ so long as the nearby $\hat{A}$ are not close to any less generic structure than a $J_2(\lambda)$ block. For all of the algorithms presented in this paper to calculate nearby double eigenvalues, analogous conditions will be found which govern their stability and/or accuracy.

3.2 An iterative fixed-point algorithm

Edelman and Elmroth [5] introduced an algorithm to find the nearest $\hat{A} \in \mathcal{D}$ to a given matrix A. The algorithm is the simultaneous solution of "dual" problems which can be stated as follows:

Problem 1:
Given a matrix A and an approximation to the perturbation bringing A to $\mathcal{D}$, find the repeated eigenvalue of $\hat{A}$.

Problem 2:
Given a matrix A and an approximation of the repeated eigenvalue of $\hat{A}$, find the smallest perturbation bringing A to $\mathcal{D}$.

For Problem 1, they use a clustering algorithm on the eigenvalues of $A - \sigma u v^H$, where $\sigma u v^H$ is the approximation to the normal bringing A to $\mathcal{D}$ (σ a scalar and uv^H an approximate normal). For Problem 2, they approximate u and v with the left and right singular vectors of $\sigma_{\min}(A - \lambda I)$ for the given approximate repeated eigenvalue λ.

This resulted in an extremely compact Matlab routine, which reads roughly as

```
lam = guess;
while (1)
    [u,s,v] = svd(A-lam*I);
    Ahat = A - u(:,n)*s(n,n)*v(:,n)';
    e = eig(Ahat);
    lam = cluster_routine(e);
end
```

where the cluster routine takes the mean of two nearest elements of `e`.

The resulting algorithm converges linearly much of the time, but it will sometimes go into cycles which are empirically not convergent. In Section 5.3, we present a detailed analysis of the convergence of the fixed point algorithm.

The Figures 11 and 12 show the fixed point algorithm both succeeding and failing for two different random matrices. We have plotted the new value of the variable `e` against its value in the previous iteration.

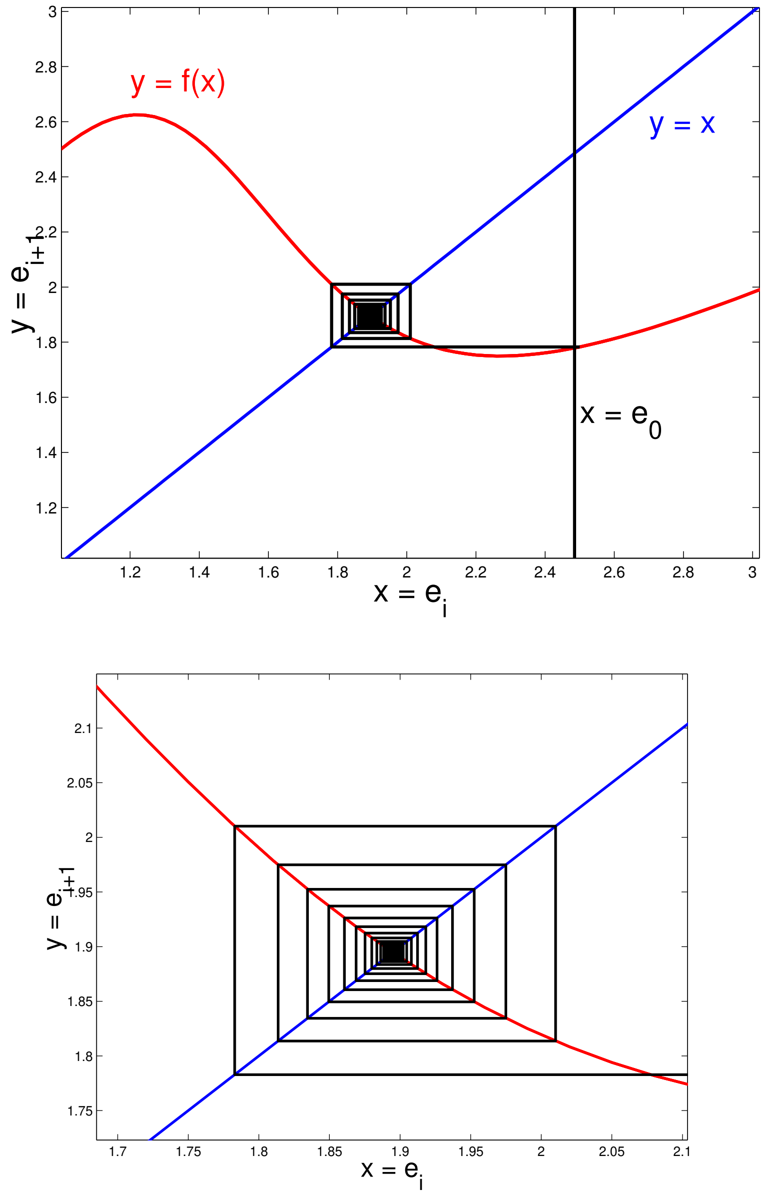

Figure 11: The fixed point algorithm converging.

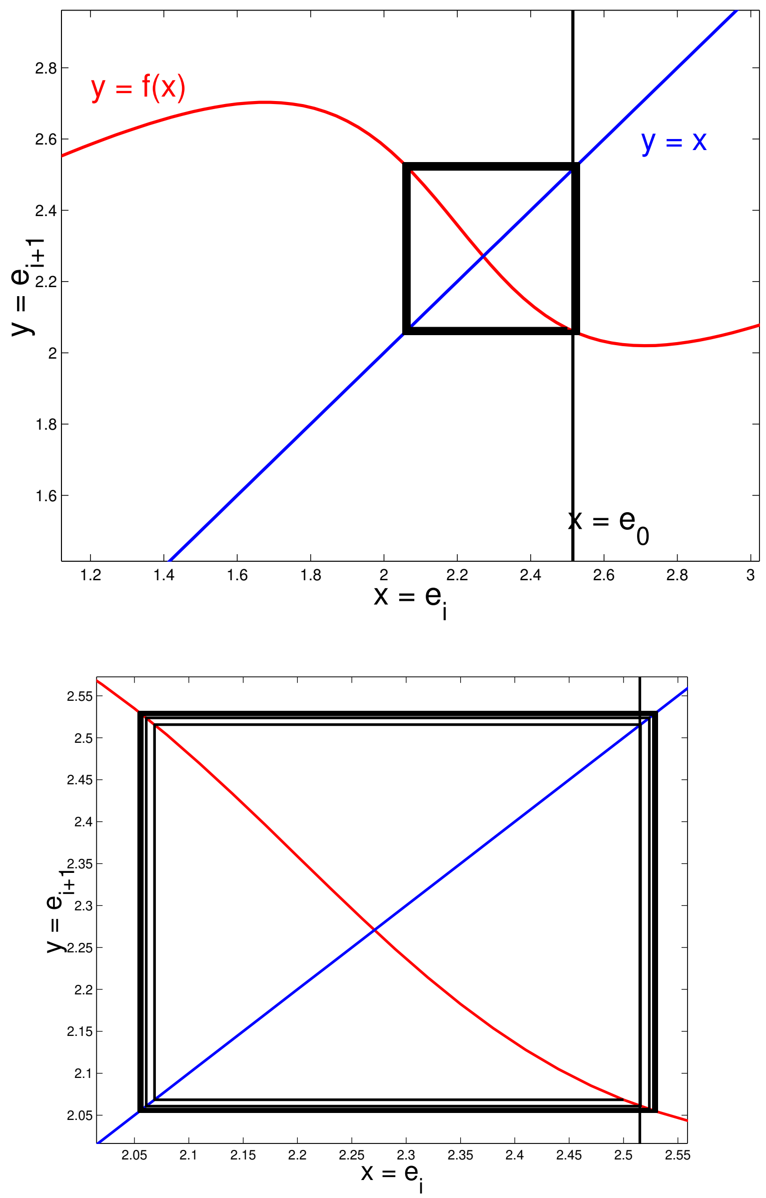

Figure 12: The fixed point algorithm approaching a limit-cycle.

3.3 The singularities of the spectral portrait

We have identified that the normal space of a matrix $\hat{A} \in \mathcal{D}$ is given by the line $\hat{A} + \sigma uv^H$, parameterized by σ, where u, v are left and right eigenvectors of $\hat{A}$. Thus if A lies in the normal space of its nearest matrix $\hat{A} \in \mathcal{D}$, then $A = \hat{A} + \sigma uv^H$. There is, however, no obvious way to compute u and v without knowing $\hat{A}$ in advance. In this section, we present a method for this based on the spectral portrait of A.

Let $\sigma(x, y) = \sigma_{\min}(A - (x + iy)I)$. In the case where $\sigma(x, y)$ is smooth, we have the following theorem about the derivatives of $\sigma(x, y)$ due to Sun (see [17]).

Theorem 3.1 *Let $U(x, y)\Sigma(x, y)V^H(x, y) = A - (x + iy)I$, where $\sigma_n(x_0, y_0)$ is simple, and let $\tilde{U} = (\, u_1 \quad u_2 \quad \ldots \quad u_{n-1}\,)$, $\tilde{V} = (\, v_1 \quad v_2 \quad \ldots \quad v_{n-1}\,)$, and $\tilde{\Sigma} = diag(\sigma_1, \ldots, \sigma_{n-1})$. We have then*

$$\begin{aligned} \frac{\partial \sigma}{\partial x}(x_0, y_0) &= -Real\{u_n^H v_n\}, \\ \frac{\partial \sigma}{\partial y}(x_0, y_0) &= Imag\{u_n^H v_n\}, \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 \sigma}{\partial x^2}(x_0, y_0) &= Real\{r^H \Phi r + l^H \Phi l + 2 l^H \Psi r\} + Imag\{u_n^H v_n\}^2/\sigma_n, \\ \frac{\partial^2 \sigma}{\partial x \partial y}(x_0, y_0) &= -Imag\{2 l^H \Psi r\} + Imag\{u_n^H v_n\} Real\{u_n^H v_n\}/\sigma_n, \\ \frac{\partial^2 \sigma}{\partial y^2}(x_0, y_0) &= Real\{r^H \Phi r + l^H \Phi l - 2 l^H \Psi r\} + Real\{u_n^H v_n\}^2/\sigma_n, \end{aligned}$$

where $\Phi = \sigma_n(\sigma_n^2 I - \tilde{\Sigma}^2)^{-1}$, $\Psi = \tilde{\Sigma}(\sigma_n^2 I - \tilde{\Sigma}^2)^{-1}$, $r = \tilde{U}^H v_n$, and $l = \tilde{V}^H u_n$.

Sun's theorem implies that $\sigma_x = \sigma_y = 0$ iff $u^H v = 0$. In this case, the matrix $\hat{A} = A - \sigma(x_0, y_0) uv^H \in M_{x_0 + iy_0}$ is a point on the envelope of the family of M_λ, i.e. $\hat{A} \in \mathcal{D}$. Thus we have

Corollary 3.2 *If $\sigma(x, y)$ is stationary for some $\lambda_0 = x_0 + iy_0$ then A lies on the normal line of the matrix $\hat{A} \in \mathcal{D}$ with repeated eigenvalue λ_0 and left and right eigenvectors u and v where $\hat{A} = A - \sigma(x_0, y_0) uv^H$ and u and v are the left and right singular vectors of $A - \lambda_0 I$ with singular value $\sigma(x_0, y_0)$.*

Since the stationary points of $\sigma(x, y)$ correspond to normal lines off of $\mathcal{D}$. An examination of the spectral portrait of A can be used to determine the distance from A to $\hat{A} \in \mathcal{D}$.

This would lead to an algorithm which is based on finding the repeated eigenvalue of $\hat{A}$ first, i.e.

```
(x0,y0) = find lowest critical points of the spectral portrait;
lam = x0 + iy0;
[u,s,v] = svd(A-lam*I);
Ahat = A - u(:,n)*s(n,n)*v(:,n)';
```

where one finds the critical point of lowest height by Newton's method or some other optimization routine over the complex plane.

Another result of Sun's formulae is that at stationary points

$$\det \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix} = |r^H \Phi r + l^H \Phi l|^2 - |2 l^H \Psi r|^2.$$

This formula will be useful in the later section on conditioning and stability (Section 5).

4 Nearest $\hat{A}$'s are basically J_2's

We have already seen that generically speaking, the matrix $A - \hat{A}$ is a rank one matrix and $\hat{A}$ has a J_2 structure. Of course, less generic Jordan structures may arise in special cases, but in this section we will show that even when $\hat{A}$ has a less generic Jordan structure, $A - \hat{A}$ still has the form of a generic J_2 normal.

In Section 3.3, we showed that, with perhaps few generic conditions on A, the nearest matrix, $\hat{A} \in \mathcal{D}$, to A is always given by $A - \sigma u v^H$ where $\sigma = \sigma_{\min}(A - \lambda I)$ is stationary in λ, u and v are the associated singular vectors, and λ is the repeated eigenvalue of $\hat{A}$ for which u and v are right and left eigenvectors.

If one were to be cavalier, one might assume generically that the nearest element of $\mathcal{D}$ was always some defective matrix with a J_2 structure. If this were the case, then $\mathcal{D}$ can be assumed to be locally differentiable about $\hat{A}$. Consequently, there would be a unique normal perturbation that would be of the form uv^H where $u^H v = 0$. Thus, the $\hat{A}$ would correspond to a critical point of the spectral portrait of $A - \lambda I$ since the spectral portrait is a graph of $\sigma_{\min}(A - \lambda I)$.

However, one might consider the ways in which this generic assumptions could be violated. If $\hat{A}$ were derogatory or had degeneracy less generic than J_2, then there could be multiple normal directions off of $\hat{A}$, and $A - \hat{A}$ would not be of the assumed rank 1 form. One particular example of where one might really be concerned is in the case where $\hat{A}$ actually has a J_3 structure and A might sit above all of the rank 1 normals of $\hat{A}$ as shown in Figure 13 (since a matrix with J_3 structure is a cusp point of $\mathcal{D}$, there is no reason to believe that it has a continuous local normal field).

Clearly, one needs to have some lemmae to understand the conditions under which such an event will or will not happen in order to make generalizations which are everywhere valid. The purpose of this section is to clarify these possibilities. We present in this section a series of lemmae which will progressively refine the possible form of $N = A - \hat{A}$ until we find that it has the form of a normal to a generic J_2. The reader who desires to skip the technicalities of this paper is advised to skip this section.

It will be useful to take a partial canonical decomposition of $\hat{A}$ in the repeated eigenvalue. That is, let

$$\hat{A} = \begin{pmatrix} X & \tilde{X} \end{pmatrix} \begin{pmatrix} J(\lambda) & 0 \\ 0 & M \end{pmatrix} \begin{pmatrix} Y & \tilde{Y} \end{pmatrix}^H,$$

where $\begin{pmatrix} Y & \tilde{Y} \end{pmatrix}^H \begin{pmatrix} X & \tilde{X} \end{pmatrix} = I$ and $J(\lambda)$ is the canonical matrix of the Jordan structure of $\hat{A}$ at λ.

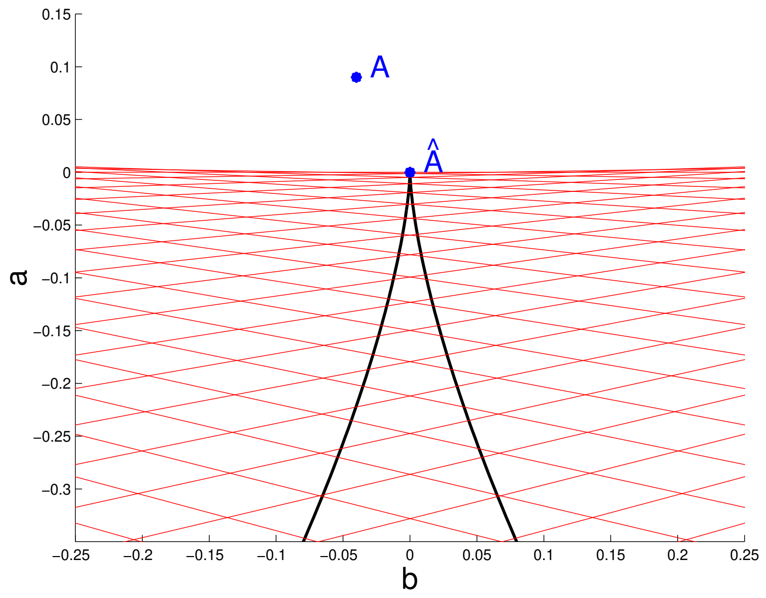

Figure 13: A possible problem finding a normal from $\hat{A}$ to A illustrated in characteristic polynomial space.

Lemma 4.1 *Let $\hat{A}$ be the nearest matrix in $\mathcal{D}$ to A. Then*

$$N = A - \hat{A} = YWX^H,$$

for some W.

Proof: We also decompose $N = A - \hat{A}$ according to the following (slightly twisted) decomposition,

$$N = \begin{pmatrix} Y & \tilde{X} \end{pmatrix} \begin{pmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{pmatrix} \begin{pmatrix} X & \tilde{Y} \end{pmatrix}^H,$$

one may check that $\begin{pmatrix} X & \tilde{Y} \end{pmatrix}^H \begin{pmatrix} Y & \tilde{X} \end{pmatrix} = I$, and that the displacements $YN_{11}X^H$, $\tilde{X}N_{12}X^H$, $YN_{21}\tilde{Y}^H$, and $\tilde{X}N_{22}\tilde{Y}^H$ are mutually orthogonal.

Since $\hat{A} + \tilde{X}N_{22}\tilde{Y}^H$ has the same canonical structure in λ as $\hat{A}$, it is clear that $N_{22} = 0$ for $\hat{A}$ to be the minimizer. Similarly, by observing that $(\hat{A} + YN_{21}\tilde{Y}^H)X = XJ$ and $Y^H(\hat{A} + \tilde{X}N_{12}X^H) = JY^H$, we see that N_{12} and N_{21} must vanish. $\square$

For the next lemma, we consider the case where J, the Jordan structure of $\hat{A}$ at λ, is derogatory and thus of the form $J = \begin{pmatrix} J^{(1)} & 0 \\ 0 & J^{(2)} \end{pmatrix}$ and the eigenvectors are correspondingly partitioned into $Y = \begin{pmatrix} Y_1 & Y_2 \end{pmatrix}$ and $X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}$. Decompose N as

$$N = \begin{pmatrix} Y_1 & Y_2 \end{pmatrix} \begin{pmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{pmatrix} \begin{pmatrix} X_1 & X_2 \end{pmatrix}^H.$$

Lemma 4.2 *Suppose that the Jordan structure of $\hat{A}$ is derogatory. Then N is of the form,*

$$N = Y_1 N_{11} X_1^H + Y_2 N_{22} X_2^H.$$

Proof: Observe that $\hat{A} + X_2 W Y_1^H$ has the Jordan structure of $\begin{pmatrix} J^{(1)} & W \\ 0 & J^{(2)} \end{pmatrix}$ which, while being possibly different from J, still has a repeated eigenvalue λ.

$\hat{A} + X_2 W Y_1$ is closer to A than $\hat{A}$ if $\langle X_2 W Y_1, N\rangle = \langle W, N_{12}\rangle$ is greater than zero for some arbitrarily small W. Thus, $N_{12} = 0$.

By a similar argument, $N_{21} = 0$. □

Finally, we can, in a sense, eliminate the derogatory $\hat{A}$ matrices by the following lemma.

Lemma 4.3 *If $\hat{A}$ has a derogatory canonical structure, J, then one of the irreducible blocks of J, say $J^{(1)}$ is defective, and*

$$N = Y_1 N_{11} X_1^H.$$

Proof: Suppose that all of the irreducible blocks of J were simple, i.e. $J = \lambda I$. Then, from the previous lemma, $N = y_1 n x_1^H$ (where x_1 and y_1 are vectors and n is a scalar).

$\hat{A} + \mu I$ has a repeated eigenvalue of $\lambda + \mu$.

$\hat{A} + \mu I$ will be closer to A if $\langle \mu I, N\rangle = \text{Real}\{\bar{\mu} n\}$ is greater than zero for some arbitrarily small μ. Thus $n = 0$ and $\hat{A} = A$ contradicting hypothesis.

Combining this with the result of the previous lemma gives the the conclusion. □

We summarize with the following corollary.

Lemma 4.4 *The nearest matrix $\hat{A} \in \mathcal{D}$ to A is reached by a perturbation of the form*

$$N = Y p(J(0)^T) X^H,$$

where $p(x)$ is a $k-1$-degree polynomial with vanishing constant term, and X, Y are the right and (corresponding dual) left generalized eigenvectors of a canonical $k \times k$ Jordan block, J, in the Jordan structure of $\hat{A}$, and $J(0)$ is the $k \times k$ Jordan block with eigenvalue 0.

Proof: Prior lemmae have shown that N is of the form $N = YWX^H$. The similarity transformation $e^{\sigma M} \hat{A} e^{-\sigma M} \in \mathcal{D}$ is closer to A for some arbitrarily small σ whenever the commutator, $M\hat{A} - \hat{A}M$ has positive inner product with N.

Thus $\langle [M, \hat{A}], N\rangle = \langle M, [N, \hat{A}^H]\rangle$ must vanish for all M. Giving $N\hat{A}^H - \hat{A}^H N = 0$ which implies that $WJ(0)^T = J(0)^T W$, and thus $W = p(J(0)^T)$. □

We now need only really consider the case where the Jordan structure of $\hat{A}$ is a single canonical Jordan block.

Lemma 4.5 *Let the Jordan block J of $\hat{A}$ be of size greater than 2. The polynomial $p(x)$ in*

$$N = Y p(J^T) X^H,$$

must be of the form $p(x) = ax^{k-1} + bx^{k-2}$.

Proof: Suppose that the coefficient, c, of x^j ($j < k-2$) is nonvanishing. Then we consider the matrix $\hat{A} + XMY^H$, where $M_{kj} = s$ ($M_{pq} = 0$ elsewhere).

$\hat{A} + XMY^H$, while not necessarily having the same canonical structure as J still has λ repeated at least twice, since

$$(\hat{A} + XMY^H)\begin{pmatrix} x_1 & x_2 \end{pmatrix} = \begin{pmatrix} x_1 & x_2 \end{pmatrix}\begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}.$$

However, $\hat{A} + XMY^H$ is closer to A so long as $\langle M, p(J(0)^T)\rangle = sc$ does not vanish for arbitrarily small s. Thus we must have $c = 0$. □

Thus, for all A not in $\mathcal{D}$, $N = A - \hat{A}$ is of the form that one generically expects from either a J_2 or a J_3 canonical structure. In what follows, we shall see that if we allow $\hat{A}$ to take values in $\mathcal{C}$ then the $A - \hat{A}$ will be rank 1.

Theorem 4.6 *Let $\hat{A} \in \mathcal{D}$ be the nearest matrix with a repeated eigenvalue to the given matrix A with distinct eigenvalues. Then*

$$N = A - \hat{A} = \sigma u v^H,$$

where u and v are the left and right eigenvectors of $\hat{A}$ and σ is a scalar.

Proof:

Assuming the canonical structure of $\hat{A}$ is $J = J_k(\lambda)$, we have from the previous lemmae, we have that

$$N = Y(a(J(0)^T)^{k-1} + b(J(0)^T)^{k-2})X^H,$$

$b \neq 0$.

Consider the matrix $\hat{A} + X(a'(J(0)^T)^{k-1} + b'(J(0)^T)^{k-2})Y^H$, where $b' = k\mu^{k-1}$ and $a' = -(k-1)\mu^k$. It is easy to show that $J + (a'(J(0)^T)^{k-1} + x(J(0)^T)^{k-2})$ has a repeated (J_2) eigenvalue $\lambda + \mu$.

if for some arbitrarily small μ $\langle X(a'(J(0)^T)^{k-1} + b'(J(0)^T)^{k-2})Y^H, N\rangle$ is greater than zero, $\hat{A} + X(a'(J(0)^T)^{k-1} + b'(J(0)^T)^{k-2})Y^H$ will be closer to A than $\hat{A}$.

$\langle X(a'(J(0)^T)^{k-1} + b'(J(0)^T)^{k-2})Y^H, N\rangle = \text{Real}\{\bar{b}'b + \bar{a}'a\}$. Substituting for a' and b' we have $\text{Real}\{\bar{b}'b + \bar{a}'a\} = \text{Real}\{\bar{\mu}^{k-1}(kb - (k-1)\bar{\mu}a\}$. If we take $|\mu| < \frac{b}{a}$ and let select the phase of μ such that $\mu^{k-1}b$ is positive real, then $\langle X(a'(J(0)^T)^{k-1} + b'(J(0)^T)^{k-2})Y^H, N\rangle > 0$.

Thus, we have $\hat{A} + X(a'(J(0)^T)^{k-1} + b'(J(0)^T)^{k-2})Y^H$ closer to A than $\hat{A}$.

We must conclude that $b = 0$ and therefore that N is rank 1 of the form asserted. □

The reader will note that it is not always possible to find such a' and b' if one is constrained to real matrices, but one can always do so over complex numbers. In fact, the illustration at the beginning of this section demonstrates this shortcoming of the reals. If this illustration had been drawn with complex a and b axes, then it would have been apparent that A was closer to a complex $\hat{A}$.

We have now established rigorously that any matrix A with distinct eigenvalues must lie along a rank 1 normal from the nearest $\hat{A} \in \mathcal{D}$. Thus $\hat{A}$ can be determined by examining the singular values of $A - \lambda I$.

What remains are some lemmae concerning the singular values themselves so that it can be determined what sort of behavior one must look for in the singular values in order to locate $\hat{A}$. The previous section shows that the critical points of $\sigma_{\min}$ are the points of interest in determining $\hat{A}$. However, one may also wonder what happens when $\sigma_{\min}$ is not simple (and hence not rigorously differentiable), and whether or not one must examine the critical points of the other singular values of $A - \lambda I$. We assure the reader that this will not be trouble for us. In Section 6, we will be able to show that the critical points of other singular values will not produce minima, and that, the minimum will occur on a critical point where $\sigma_{\min}$ is almost always simple.

5 Conditioning and stability

In this section, we wish to examine the conditioning of the problem of finding the $\hat{A} \in \mathcal{D}$ nearest to A. The conditioning of $\hat{A}$ is closely related to the behavior of the singular value decomposition of $\hat{A} - \lambda I$ under first order variations. We will find that the matrices with ill-conditioned $\hat{A}$ on $\mathcal{D}$ are sitting close to "focal points" of $\mathcal{D}$ analogous to focal points in geometric optics, explained in Section 5.2.

5.1 Sensitivity analysis of $A = \hat{A} + \sigma u v^H$

Consider the normal lines passing through a matrix $\hat{A} \in \mathcal{D}$. We can parameterize these lines by the real or complex line $A = \hat{A} + \sigma u v^H$, where σ is real and the phases on u and v are arbitrary (alternatively, we could fix the phase of uv^H and let σ be complex). For most A we expect this decomposition ($A = \hat{A} + \sigma u v^H$) to be locally invertible. We define a *focal point* as some matrix A for which the decomposition is not locally unique, i.e., where the Jacobian of this decomposition is singular.

Geometrically, we may think of the set of focal points as the evolute of $\mathcal{D}$, i.e., the envelope of normals of $\mathcal{D}$. A point which is in the evolute can be thought of as the intersection of two infinitesimally separated normals.

It will be useful to do some dimension counting in both complex and real versions of $\mathcal{D}$ so that we might better understand where the degrees of freedom are. (We shall always refer to one real parameter as one degree of freedom.)

The solution to the problem of finding a stationary point $\hat{A}$ to A can be expressed in terms of decompositions of A as follows:

$$A = \sigma u v^H + \hat{A}, \tag{13}$$
$$A = \sigma u v^H + \lambda I + \tilde{U}\tilde{\Sigma}\tilde{V}^H, \tag{14}$$
$$A = \lambda I + U \Sigma V^H \tag{15}$$

where $u^H v = 0$, $\tilde{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_{n-1})$, $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_{n-1}, \sigma) = \begin{pmatrix} \tilde{\Sigma} & 0 \\ 0 & \sigma \end{pmatrix}$, $U = (\,\tilde{U} \quad u\,)$, $V = (\,\tilde{V} \quad v\,)$, and U, V are unitary matrices. For the real case, Equation (13) decomposes the n^2 degrees of freedom of A into $n(n-1)/2$ for U, $n(n-1)/2$ for V, n for Σ, and 1 for λ, which leaves exactly 1 degree of freedom extra to satisfy the constraint $u^H v = 0$. For the complex case, A has $2n^2$ degrees of freedom, U has n^2, V has n^2, Σ has n, and λ has 2, leaving $n + 2$ surplus degrees of freedom. Since the complex decomposition is only

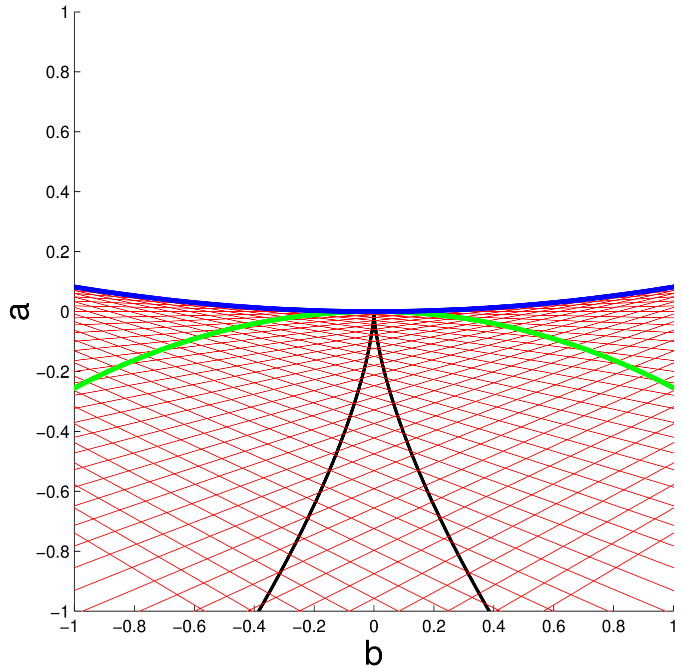

Figure 14: The normals of the cubic double root variety forming their parabola-like envelope. The concave down arc in the plot is another part of the envelope which occurs over cubics with complex coefficients.

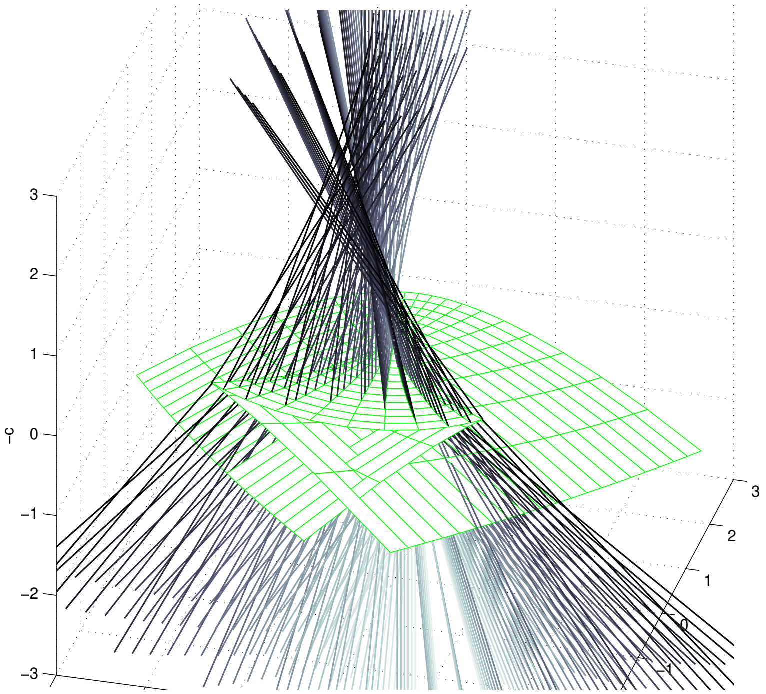

Figure 15: The normals of the swallowtail. A faint hood-like shape begins to emerge.

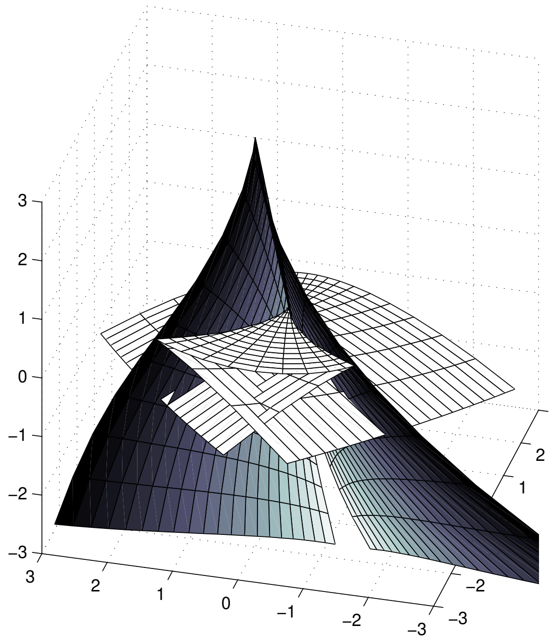

Figure 16: The swallowtail and its hood-like evolute of ill-conditioned problems.

unique up to transformation of the form $U, V \to U\Delta, V\Delta$ (Δ a diagonal unitary matrix), n of those surplus degrees of freedom are ineffective, leaving two to satisfy the constraints $\mathrm{Real}\{u^H v\} = 0, \mathrm{Imag}\{u^H v\} = 0$.

Thus, by counting dimensions we see that the decomposition which reveals the nearest $\hat{A} \in \mathcal{D}$ to A is really a singular value decomposition in which one dimension has been removed by the constraint $u^H v = 0$ and one dimension has been added by the presence of the λI term. Our strategy to vary $\hat{A}$ will be to vary the SVD (as in [17]), and then to use the variation in λ to satisfy the varied constraint ($u^H v = 0$) equation.

Note, if we were to vary this decomposition for an element of $A \in \mathcal{D}$, we set $\sigma = 0$ in Equation (13). In the real case, this eliminates one degree of freedom corresponding to $\mathcal{D}$'s real co-dimension of 1. In the complex case, this eliminates not only the degree of freedom in σ, but also the extra unitary freedom of u and v corresponding to $\hat{A}$'s complex co-dimension of 1.

Theorem 5.1 *Assuming that the singular values of $A - \lambda I$ are distinct, the first order behavior of the decomposition*

$$A = \hat{A} - \sigma u v^H,$$

is given by

$$\dot{A} = \dot{\hat{A}} + (u^H \dot{A} v) u v^H + \sigma \tilde{U}(\tilde{U}^H \dot{u}) v^H + \sigma u (\tilde{V}^H \dot{v})^H \tilde{V}^H,$$

with

$$\begin{pmatrix} \tilde{U}^H \dot{u} \\ \tilde{V}^H \dot{v} \end{pmatrix} = \begin{pmatrix} \Phi(\sigma) & \Psi(\sigma) \\ \Psi(\sigma) & \Phi(\sigma) \end{pmatrix} \begin{pmatrix} \tilde{U}^H (\dot{A} - \dot{\lambda} I) v \\ \tilde{V}^H (\dot{A}^H - \dot{\bar{\lambda}} I) u \end{pmatrix},$$

where $\Phi(\sigma) = \sigma(\sigma^2 I - \tilde{\Sigma}^2)^{-1}$, $\Psi(\sigma) = \tilde{\Sigma}(\sigma^2 I - \tilde{\Sigma}^2)^{-1}$, and $\dot{\lambda}$ is given by

$$\begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix} \begin{pmatrix} \dot{\lambda}_{real} \\ \dot{\lambda}_{imag} \end{pmatrix} = \begin{pmatrix} Real\{l^H \Psi p + l^H \Phi q + q^H \Psi r + p^H \Phi r\} \\ Imag\{l^H \Psi p + l^H \Phi q + q^H \Psi r + p^H \Phi r\} \end{pmatrix},$$

where σ_{xx}, σ_{xy}, and σ_{yy} are given by Sun's theorem.

Proof: Varying the decomposition (13) along a differentiable curve $A(t)$, we have

$$\dot{A} = \dot{\hat{A}} + \dot{\sigma} u v^H + \sigma \dot{u} v^H + \sigma u \dot{v}^H,$$

where $\dot{u}^H v + u^H \dot{v} = 0$. We will proceed by solving for an expression for $\dot{u}^H v + u^H \dot{v}$, finding values for σ and the phase of uv^H for which the variation of the constraint is satisfied (it is convenient for us to take σ and its variation to be real, adjusting the phases of uv^H and its variation implicitly).

Differentiating $(A - \lambda I)v = \sigma u$ and $u^H(A - \lambda I) = \sigma v^H$, we have that

$$(\dot{A} - \dot{\lambda} I) v - \dot{\sigma} u = \sigma \dot{u} - (A - \lambda I) \dot{v}$$

and

$$(\dot{A}^H - \dot{\bar{\lambda}} I) u - \dot{\sigma} v = \sigma \dot{v} - (A^H - \bar{\lambda} I) \dot{u}.$$

We may substitute $A-\lambda = U\Sigma V^H$, obtaining

$$U^H(\dot{A}-\dot{\lambda}I)v - \dot{\sigma}e_n = \sigma U^H\dot{u} - \Sigma V^H\dot{v}$$

and

$$V^H(\dot{A}^H-\dot{\bar{\lambda}}I)u - \dot{\sigma}e_n = \sigma V^H\dot{v} - \Sigma U^H\dot{u},$$

where $e_n = \begin{pmatrix} 0 & \cdots & 0 & 1 \end{pmatrix}^H$, which can be collected into $2n-2$ equations,

$$\begin{pmatrix} \sigma & -\tilde{\Sigma} \\ -\tilde{\Sigma} & \sigma \end{pmatrix} \begin{pmatrix} \tilde{U}^H\dot{u} \\ \tilde{V}^H\dot{v} \end{pmatrix} = \begin{pmatrix} \tilde{U}^H(\dot{A}-\dot{\lambda}I)v \\ \tilde{V}^H(\dot{A}^H-\dot{\bar{\lambda}}I)u \end{pmatrix} \tag{16}$$

and 2 identical equations,

$$u^H(\dot{A}-\dot{\lambda}I)v - \dot{\sigma} = \sigma(u^H\dot{u} - v^H\dot{v}). \tag{17}$$

If $\sigma = \sigma_n$ is distinct from the other σ_i then we can invert the matrix in (16) obtaining

$$\begin{pmatrix} \tilde{U}^H\dot{u} \\ \tilde{V}^H\dot{v} \end{pmatrix} = \begin{pmatrix} \Phi(\sigma) & \Psi(\sigma) \\ \Psi(\sigma) & \Phi(\sigma) \end{pmatrix} \begin{pmatrix} \tilde{U}^H(\dot{A}-\dot{\lambda}I)v \\ \tilde{V}^H(\dot{A}^H-\dot{\bar{\lambda}}I)u \end{pmatrix}, \tag{18}$$

where $\Phi(\sigma) = \sigma(\sigma^2 I - \tilde{\Sigma}^2)^{-1}$ and $\Psi(\sigma) = \tilde{\Sigma}(\sigma^2 I - \tilde{\Sigma}^2)^{-1}$. Since $U^H\dot{U}$ and $V^H\dot{V}$ must be skew-symmetric, (17) separates into real and imaginary parts

$$\begin{aligned} \dot{\sigma} &= \mathrm{Real}\{u^H\dot{A}v\} && (19) \\ -i(u^H\dot{u} - v^H\dot{v}) &= \mathrm{Imag}\{u^H\dot{A}v\}, && (20) \end{aligned}$$

where we have used $u^H v = 0$.

We now apply the differentiated constraint,

$$\dot{u}^H v + u^H\dot{v} = (U^H\dot{u})^H r + l^H(V^H\dot{v}) = 0,$$

where we have used $v = \tilde{U}r$ and $u = \tilde{V}l$ ($r = \tilde{U}^H v$ and $l = \tilde{V}^H u$). Substituting (18) into this, we have

$$\dot{\lambda}(2l^H\Psi r) + \dot{\bar{\lambda}}(l^H\Phi l + r^H\Phi r) = l^H\Psi p + l^H\Phi q + q^H\Psi r + p^H\Phi r, \tag{21}$$

where $p = \tilde{U}^H\dot{A}v$ and $q = \tilde{V}^H\dot{A}^H u$, which one can solve for $\dot{\lambda}$ solving

$$\begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix} \begin{pmatrix} \dot{\lambda}_{real} \\ \dot{\lambda}_{imag} \end{pmatrix} = \begin{pmatrix} \mathrm{Real}\{l^H\Psi p + l^H\Phi q + q^H\Psi r + p^H\Phi r\} \\ \mathrm{Imag}\{l^H\Psi p + l^H\Phi q + q^H\Psi r + p^H\Phi r\} \end{pmatrix}.$$

One then has

$$\begin{aligned} \dot{\hat{A}} &= \dot{A} - \mathrm{Real}\{u^H\dot{A}v\}uv^H - \sigma U(U^H\dot{u})v^H - \sigma u(V^H\dot{v})^H V^H && (22) \\ \dot{\hat{A}} &= \dot{A} - \mathrm{Real}\{u^H\dot{A}v\}uv^H - \sigma\tilde{U}(\tilde{U}^H\dot{u})v^H - \sigma u(\tilde{V}^H\dot{v})^H\tilde{V}^H - \sigma(u^H\dot{u} - \dot{v}^H v)uv^H && (23) \\ \dot{\hat{A}} &= \dot{A} - (u^H\dot{A}v)uv^H - \sigma\tilde{U}(\tilde{U}^H\dot{u})v^H - \sigma u(\tilde{V}^H\dot{v})^H\tilde{V}^H, && (24) \end{aligned}$$

where we have used Equation (20) to get (24) from (23). □

One can interpret the formula for $\dot{\hat{A}}$ as displacement resulting purely from the change in the phase and length of the normal (second term) plus a tangential variation of $\hat{A}$ (third and fourth terms) which results from a change in the direction of the normal from A to $\hat{A}$. The magnitude of the tangential part of the variation is $\sqrt{||\dot{u}||^2 + ||\dot{v}||^2}$.

The set of A with $\hat{A}$ having infinite condition number is given by finding those A for which $\dot{\hat{A}}$ can be nonzero when $\dot{A}$ vanishes. In this case, we set the p and q terms in (21) to 0 and solve

$$\det \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix} \begin{pmatrix} \dot{\lambda}_{real} \\ \dot{\lambda}_{imag} \end{pmatrix} = 0,$$

as an equation of σ equivalent to

$$|2l^H \Psi(\sigma) r| = |l^H \Phi(\sigma) l + r^H \Phi(\sigma) r|,$$

which gives the following corollary.

Corollary 5.2 *The nearby $\hat{A} \in \mathcal{D}$ to a generic matrix A is ill-conditioned whenever $|2l^H \Psi r| - |l^H \Phi l + r^H \Phi r|$ is small (i.e. $\begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{pmatrix}$ is nearly singular).*

5.2 Ill-posed problems and principal curvatures

In this section we explore the differential geometry of $\mathcal{D}$ itself, as a co-dimension 1 manifold embedded in $\mathcal{R}^{n^2}$ or $\mathcal{C}^{n^2}$. We will give an additional interpretation of the focal points of $\mathcal{D}$ in terms of principal curvatures.

The classical study of principal curvatures is concerned with surfaces in $\mathcal{R}^3$ (as in [4]). In this setting, one uses the unit normal field $\hat{n}$ of a surface M to construct a map from M to $\mathcal{S}^2$. Studying the local behavior of this map ultimately results in a quadratic form on the tangent space of M at a point. The reciprocal lengths of the principal axes of this quadratic form are called the principal curvatures, their sum the mean curvature and their product the Gaussian curvature.

One way to generalize this approach which is suitable for any manifold of co-dimension 1 is to use the normal to construct a natural quadratic form on the tangent space and then examine the eigenvalues of this form.

We first look at an n-sphere of radius l embedded in $\mathcal{R}^{n+1}$. Geodesics on $\mathcal{S}^n$ satisfy the equations

$$\begin{aligned} \frac{dp}{dt} &= v \\ \frac{dv}{dt} &= -\frac{v \cdot v}{l} \hat{n} \end{aligned}$$

where p is a point on the sphere and v is the velocity, satisfying $p^H v = 0$, and $\hat{n} = p/l$. One might recognize the derivative of v as the centripetal acceleration in classical mechanics.

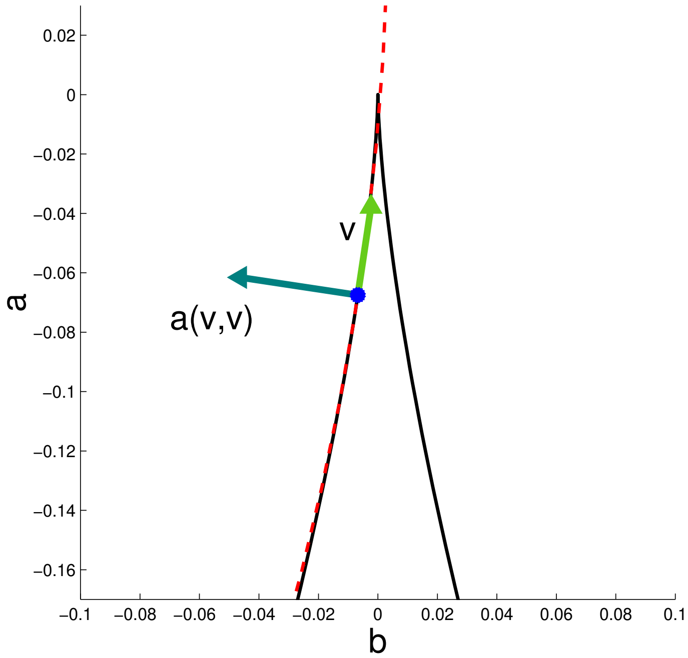

Figure 17: In the cubic polynomial space, a particle moves along the double root variety, with velocity v and centripetal acceleration $a(v,v)$.

For a more general manifold of co-dimension 1, one has (at least locally) a unit normal $\hat{n}$ and geodesic equations of the form

$$\frac{dp}{dt} = v \tag{25}$$

$$\frac{dv}{dt} = -\Gamma(v,v) \tag{26}$$

where $\Gamma(v,w)$ is the Levi-Civita connection. However, for manifolds of co-dimension 1, the Levi-Civita connection must be of the form $\Gamma(v,w) = a(v,w)\hat{n}$ where the aceleration, $a(v,w)$, is a symmetric bilinear (*not* complex sesquilinear) function defined on the tangent space at p and determined by $\frac{dv}{dt} \cdot \hat{n} + v \cdot \frac{d\hat{n}}{dt} = 0$.

If we consider $a(v,v)$ to be an instantaneous centripetal acceleration along a geodesic with velocity v through p (see Figure 17), then we may define the principal curvatures at p as the extrema of the function $|a(v,v)|$ over the unit ball in the tangent space at p. As with the sphere, the center of curvature is located at $p - \frac{1}{a(v,v)}\hat{n}$. We may refer to the reciprocals of the principal curvatures as the focal lengths along $\hat{n}$ from p (see Figure 18)

With this in mind, we now consider the differentiable subset of $\mathcal{D}$. For the remainder of this section any matrix A is a point in a differentiable neighborhood of $\mathcal{D}$ with decomposition

$$A = \tilde{U}\tilde{\Sigma}\tilde{V}^H + \lambda I$$

and normal $N = uv^H$ ($u^H v = 0$). Any element H of the tangent space of A must satisfy $u^H H v = 0$.

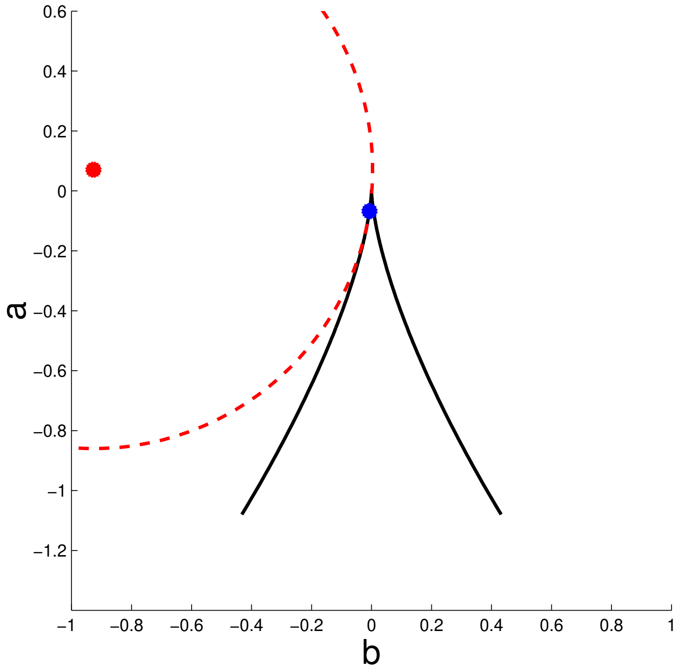

Figure 18: The centripetal acceleration of the particle moving along the double root variety can be associated with a circle centered at the focal point of the double root variety.

If $A(t)$ is a differentiable curve in $\mathcal{D}$ through A then we differentiate $u^H H v = 0$ to get

$$\begin{aligned} \frac{d}{dt}(u^H H v) = \dot{u}^H H v + u^H H \dot{v} - a(H,H) &= 0 \\ (\tilde{U}^H \dot{u})^H p + q^H \tilde{V}^H \dot{v} &= a(H,H), \end{aligned}$$

where $q = \tilde{V}^H H v$ and $p = \tilde{U}^H H v$ (note that since $u^H H v = 0$ we have $Hv = \tilde{U} q$ and $u^H H = p^H \tilde{V}^H$).

Since $A \in \mathcal{D}$ we may set $\sigma = 0$ and employ Equations (18, 21, 24) to see that

$$\begin{pmatrix} \tilde{U}^H \dot{u} \\ \tilde{V}^H \dot{v} \end{pmatrix} = \begin{pmatrix} 0 & \tilde{\Sigma}^{-1} \\ \tilde{\Sigma}^{-1} & 0 \end{pmatrix} \begin{pmatrix} \dot{\lambda} r - p \\ \dot{\lambda} l - q \end{pmatrix},$$

(where $r = \tilde{U}^H v, l = \tilde{V}^H u$) with $\dot{\lambda}$ given by

$$\dot{\lambda}(2 l^H \tilde{\Sigma}^{-1} r) = -l^H \tilde{\Sigma}^{-1} p - q^H \tilde{\Sigma}^{-1} r.$$

Substituting and simplifying, we have

$$a(H,H) = -2 q^H \tilde{\Sigma}^{-1} p - \frac{(l^H \tilde{\Sigma}^{-1} p + q^H \tilde{\Sigma}^{-1} r)^2}{2 l^H \tilde{\Sigma}^{-1} r}.$$

Since $\tilde{U} p v^H + u q^H \tilde{V}^H$ is a tangent vector for any p and q, we may consider $a(H,H)$ as a function of $2n-2$ independent p and q vectors.

We see that the quadratic form $a(H,H)$ can be written as $z^T W z$, where

$$z = \begin{pmatrix} \bar{p} \\ q \end{pmatrix} \tag{27}$$

and

$$W = \begin{pmatrix} 0 & -\tilde{\Sigma}^{-1} \\ -\tilde{\Sigma}^{-1} & 0 \end{pmatrix} - \frac{1}{2 \bar{l}^T \tilde{\Sigma}^{-1} r} \begin{pmatrix} \tilde{\Sigma}^{-1} r \\ \tilde{\Sigma}^{-1} \bar{l} \end{pmatrix} \begin{pmatrix} r^T \tilde{\Sigma}^{-1} & \bar{l}^T \tilde{\Sigma}^{-1} \end{pmatrix}. \tag{28}$$

By a simple orthogonal similarity transformation we have

$$Q W Q^T = Z - a a^T, \tag{29}$$

where $Z = \begin{pmatrix} -\tilde{\Sigma}^{-1} & 0 \\ 0 & \tilde{\Sigma}^{-1} \end{pmatrix}$ and $a = \frac{1}{2\sqrt{\bar{l}^T \tilde{\Sigma}^{-1} r}} \begin{pmatrix} \tilde{\Sigma}^{-1}(r + \bar{l}) \\ \tilde{\Sigma}^{-1}(r - \bar{l}) \end{pmatrix}$, a real diagonal matrix plus a rank-1 matrix.

We now seek the extrema of $z^T W z$ where $|H|^2 = |x|^2 + |y|^2 = 1$. Note that $W^T = W$. Varying $|z^T W z|$ we see that the condition for z to be an extremum of $|z^T W z|$ is

$$W z = \zeta \bar{z},$$

a somewhat unusual eigenvalue equation over $\mathcal{C}$.

Over $\mathcal{R}$ this reduces to the problem of finding the eigenvalues of a rank-1 update, $Z + a a^T$. An eigenvalue ζ of this matrix then satisfies

$$a^T (\zeta I - Z)^{-1} a - 1 = 0,$$

which, in our terms, is

$$\begin{pmatrix}(r+l)^T\tilde{\Sigma}^{-1} & (r-l)^T\tilde{\Sigma}^{-1}\end{pmatrix}\begin{pmatrix}(\zeta I+\tilde{\Sigma}^{-1})^{-1} & 0\\ 0 & (\zeta I-\tilde{\Sigma}^{-1})^{-1}\end{pmatrix}\begin{pmatrix}\tilde{\Sigma}^{-1}(r+l)\\ \tilde{\Sigma}^{-1}(r-l)\end{pmatrix}-4l^T\tilde{\Sigma}^{-1}r=0.$$

Multiplying this out and simplifying gives

$$r^T\frac{-1}{\zeta}(\frac{1}{\zeta^2}I-\tilde{\Sigma}^2)^{-1}r+l^T\frac{-1}{\zeta}(\frac{1}{\zeta^2}-\tilde{\Sigma}^2)^{-1}l+2l^T\tilde{\Sigma}(\frac{1}{\zeta^2}-\tilde{\Sigma}^2)^{-1}r=0.$$

$$r^T\Phi(\frac{-1}{\zeta})r+l^T\Phi(\frac{-1}{\zeta})l+2l^T\Psi(\frac{-1}{\zeta})r=0,$$

where $\Phi(\sigma)=\sigma(\sigma^2I-\tilde{\Sigma}^2)^{-1}$ and $\Psi(\sigma)=\tilde{\Sigma}(\sigma^2I-\tilde{\Sigma}^2)^{-1}$. The center of curvature is then $A+\sigma uv^H$ where $\sigma=\frac{-1}{\zeta}$. We may now interpret the vanishing of the second real-axis derivative at a stationary point of the spectral portrait as equivalent to being on a focal point of the real double eigenvalue variety.

Over $\mathcal{C}$, we solve $QWQ^H(Qz)=\zeta Q\bar{z}$ by first observing that given an eigen-pair z,ζ , the eigen-pair $ze^{i\frac{1}{2}\theta}$, $\zeta e^{-i\theta}$ is also a solution. Thus, while acknowledging that λ can have any phase. we may restrict ζ to the reals for purposes of computation, in which case we solve,

$$\begin{pmatrix}QWQ^H & -\zeta I\\ -\zeta I & Q\bar{W}Q^H\end{pmatrix}\begin{pmatrix}z\\ \bar{z}\end{pmatrix}.$$

Following a derivation of the usual rank-1 update formula, we rewrite this as

$$\begin{pmatrix}Z & -\zeta I\\ -\zeta I & Z\end{pmatrix}\begin{pmatrix}z\\ \bar{z}\end{pmatrix}-\begin{pmatrix}aa^T & 0\\ 0 & \bar{a}\bar{a}^T\end{pmatrix}\begin{pmatrix}z\\ \bar{z}\end{pmatrix}=0.$$

Letting $k=a^Tz$ we have

$$\begin{aligned}\begin{pmatrix}Z & -\zeta I\\ -\zeta I & Z\end{pmatrix}\begin{pmatrix}z\\ \bar{z}\end{pmatrix}-\begin{pmatrix}ak\\ \bar{a}\bar{k}\end{pmatrix} &= 0\\ \begin{pmatrix}z\\ \bar{z}\end{pmatrix}-\begin{pmatrix}Z(Z^2-\zeta^2I)^{-1}a & \zeta(Z^2-\zeta^2I)^{-1}\bar{a}\\ \zeta(Z^2-\zeta^2I)^{-1}a & Z(Z^2-\zeta^2I)^{-1}\bar{a}\end{pmatrix}\begin{pmatrix}k\\ \bar{k}\end{pmatrix} &= 0\\ \begin{pmatrix}k\\ \bar{k}\end{pmatrix}-\begin{pmatrix}a^TZ(Z^2-\zeta^2I)^{-1}a & a^T\zeta(Z^2-\zeta^2I)^{-1}\bar{a}\\ \bar{a}^T\zeta(Z^2-\zeta^2I)^{-1}a & \bar{a}^TZ(Z^2-\zeta^2I)^{-1}\bar{a}\end{pmatrix}\begin{pmatrix}k\\ \bar{k}\end{pmatrix} &= 0.\end{aligned}$$

Note that the off diagonals are real, while the diagonals are each others complex conjugates, and that $\begin{pmatrix}k\\ \bar{k}\end{pmatrix}$ is a fixed point of some matrix. A necessary and sufficient condition for such a k to exist is

$$\det\begin{pmatrix}a^TZ(Z^2-\zeta^2I)^{-1}a-1 & a^T\zeta(Z^2-\zeta^2I)^{-1}\bar{a}\\ \bar{a}^T\zeta(Z^2-\zeta^2I)^{-1}a & \bar{a}^TZ(Z^2-\zeta^2I)^{-1}\bar{a}-1\end{pmatrix}=0,$$

or equivalently,

$$|a^TZ(Z^2-\zeta^2I)^{-1}a+1|=|\bar{a}^T\zeta(Z^2-\zeta^2I)^{-1}a|.$$

Substituting and simplifying we have

$$|\bar{r}^T\Phi(\frac{-1}{\zeta})r+\bar{l}^T\Phi(\frac{-1}{\zeta})l|=|2\bar{l}^T\Psi(\frac{-1}{\zeta})r|.$$

Thus we have seen that an additional way to interpret these focal points is to imagine an osculating circle of radius $\frac{1}{\zeta}$, centered at the focal point and passing through $\mathcal{D}$ along a principal direction.

5.3 Stability of the fixed point algorithm

The fixed point algorithm can be viewed as a series of guesses at the repeated eigenvalue $\lambda = \lim_{i\to\infty} \mathtt{e}_i$, where $\mathtt{e}_{i+1} = f(\mathtt{e}_i)$. In iterated systems of this sort, convergence is to some fixed point $x_0 = f(x_0)$, and the linear multiplier governing convergence is $f'(x_0)$.

If $f'(x_0) > 1$ then the system diverges, either to another fixed point or to a limit cycle. If $f'(x_0) < 1$, but close to 1, then convergence can be very slow. We therefore need to examine $f'(x_0)$ to understand the stability of this algorithm. We linearize $f(x)$ about a fixed point λ.

$$f(\lambda + \delta\lambda) = \mathtt{cluster}(A - (\sigma + \delta\sigma)(u + \delta u)(v + \delta v)^H)$$

where $u + \delta u$, $v + \delta v$, and $\sigma + \delta\sigma$ are the singular vectors and value of $A - (\lambda + \delta\lambda)I$ and u, v, and σ are the singular vectors and value of $A - \lambda I$. This may be rewritten as

$$f(\lambda + \delta\lambda) = \mathtt{cluster}(\hat{A} - \delta\sigma uv^H - \sigma\delta uv^H - \sigma u\delta v^H).$$

Assuming that the `cluster` function simply averages the two nearest eigenvalues, we have a first variation giving

$$f(\lambda + \delta\lambda) = \lambda + \frac{1}{2}\mathrm{tr}((\, y_1 \quad y_2\,)^H(-\delta\sigma uv^H - \sigma\delta uv^H - \sigma u\delta v^H)\,(\, x_1 \quad x_2\,)),$$

where y_2 and x_1 are the left and right eigenvectors of $\hat{A}$, $y_2^H\hat{A} = y_1^H$, and $\hat{A}x_2 = x_1$.

Since v is a right eigenvector and u is a left eigenvector, we may take $x_1 = \frac{1}{l^H\tilde{\Sigma}^{-1}r}v$, $y_2 = u$, $x_2 = \frac{1}{l^H\tilde{\Sigma}^{-1}r}\tilde{V}\tilde{\Sigma}^{-1}r$, and $y_1 = \tilde{U}\tilde{\Sigma}^{-1}l$. Thus we see that

$$\delta f = -\frac{1}{2l^H\tilde{\Sigma}^{-1}r}(\sigma l^H\tilde{\Sigma}^{-1}\tilde{U}^H\delta u + \sigma\delta v\tilde{V}\tilde{\Sigma}^{-1}r).$$

Substituting from Theorem 5.1 (with A fixed), we have

$$\delta f = \frac{1}{2l^H\tilde{\Sigma}^{-1}r}(\delta\lambda 2l^H(\tilde{\Sigma}^{-1} + \Psi)r + \delta\bar{\lambda}(l^H\Phi l + r^H\Phi r)),$$

simplifying to

$$\delta f = \delta\lambda + \frac{1}{2l^H\tilde{\Sigma}^{-1}r}(\delta\lambda 2l^H\Psi r + \delta\bar{\lambda}(l^H\Phi l + r^H\Phi r)),$$

which is the linearization of the iterated map f on a small perturbation $\delta\lambda$.

The behavior of this map will determine the nature of the fixed point. We recognize elements of the second derivative of the spectral portrait in this formula. Using Sun's formulae, we have

$$\delta f = \delta\lambda + \frac{1}{2l^H\tilde{\Sigma}^{-1}r}(\delta\lambda(\frac{\sigma_{xx} - \sigma_{yy}}{2} - i\sigma_{xy}) + \delta\bar{\lambda}(\frac{\sigma_{xx} + \sigma_{yy}}{2})),$$

from which we may assert the following theorem.

Theorem 5.3 *The fixed point algorithm iteration linearizes to*

$$\begin{pmatrix} \delta\lambda_{real} \\ \delta\lambda_{imag} \end{pmatrix}_{i+1} = \left(I - \begin{pmatrix} l^H\tilde{\Sigma}^{-1}r_{real} & l^H\tilde{\Sigma}^{-1}r_{imag} \\ l^H\tilde{\Sigma}^{-1}r_{imag} & -l^H\tilde{\Sigma}^{-1}r_{real} \end{pmatrix}^{-1} \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix}\right) \begin{pmatrix} \delta\lambda_{real} \\ \delta\lambda_{imag} \end{pmatrix}_i. \tag{30}$$

We can make a number of qualitative statements about Equation (30). When the Hessian of $\sigma_{\min}$ is singular (i.e., near a focal point) this iteration can stagnate. Secondly, if $l^H\tilde{\Sigma}^{-1}r$ is small (corresponding to the Puiseux expansion in Section 3.1 having poor behavior), then the iteration is more likely to diverge. Lastly, since

$$\begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix} \to \begin{pmatrix} l^H\tilde{\Sigma}^{-1}r_{real} & l^H\tilde{\Sigma}^{-1}r_{imag} \\ l^H\tilde{\Sigma}^{-1}r_{imag} & -l^H\tilde{\Sigma}^{-1}r_{real} \end{pmatrix}^{-1}$$

as $\sigma \to 0$, we see that fast convergence is assured for small enough σ.

5.4 Sensitivity of the critical points of the spectral portrait

With results which will be derived in Section 6, the problem of finding the nearest $\hat{A} \in \mathcal{D}$ to A can be reduced to locating the critical points of the spectral portrait of A, $\sigma(x,y) = \sigma_{\min}(A-(x+iy)I))$ as suggested by Section 3.3. We now turn our attention to the sensitivity of this problem, investigating the sensitivity of the critical points of $\sigma(x,y)$ to small perturbations in A. To first order, a perturbation δA in A will cause a perturbation in the gradient of $\sigma(x,y) = (\sigma_x, \sigma_y)$ at the point (x_0, y_0), which we may call $(\delta\sigma_x, \delta\sigma_y)$. By inverting the Hessian (given by Sun's formulae) on this perturbation, one then has the first order correction to (x_0, y_0), which we write as $(\delta x_0, \delta y_0)$.

Theorem 5.4 *Let (x_0, y_0) be a stationary point of $\sigma(x,y)$ and Φ, Ψ, r, and l be defined as above. Then let $p = \tilde{U}^H\delta Av$ and $q = \tilde{V}^H\delta A^H u$. Then*

$$\begin{aligned} \delta\sigma_x(x_0,y_0) &= Real\{r^H\Phi p + r^H\Psi q + p^H\Psi l + q^H\Phi l\} \\ \delta\sigma_y(x_0,y_0) &= Imag\{r^H\Phi p + r^H\Psi q + p^H\Psi l + q^H\Phi l\}. \end{aligned}$$

Proof: From Sun's theorem, we have

$$\sigma_x + i\sigma_y = -v^H u$$

and thus

$$\delta\sigma_x + i\delta\sigma_y = -\delta v^H u - v^H\delta u.$$

Since $u^H v = 0$, we may substitute $v = \tilde{U}\tilde{U}^H v = \tilde{U}r$ and $u = \tilde{V}\tilde{V}^H u = \tilde{V}l$,

$$\delta\sigma_x + i\delta\sigma_y = -r^H\tilde{U}^H\delta u - (\tilde{V}^H\delta v)^H l.$$

From Theorem 5.1 we have a formula for $\tilde{U}^H\delta v$ and $\tilde{V}^H\delta u$, which we substitute to give

$$\sigma_x + i\sigma_y = r^H\Phi p + r^H\Psi q + p^H\Psi l + q^H\Phi l.$$

□

From this lemma and Sun's theorem, we see that the condition number on the critical points of the pseudospectra is given by largest singular value of the $2 \times (4n-4)$ real matrix

$$\begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{pmatrix}^{-1} \begin{pmatrix} \delta\sigma_x \\ \delta\sigma_y \end{pmatrix},$$

where $\delta\sigma_x$ and $\delta\sigma_y$ are the real linear functions of the real and imaginary parts of p and q defined by the lemma.

6 Nearest $\hat{A}$'s come from simple $\sigma_{\min}$

In Section 5.2 we derived expressions for the centripetal acceleration of a geodesic along $\mathcal{D}$. We will now use this machinery to give us the prove the claims made at the end of Section 4, namely, that the only critical points of singular values one need examine are those of $\sigma_{\min}$ and not any other σ_k. Once again, the reader who wishes to skip over technicalities is advised to skip this section.

First we will make some fairly general derivations over co-dimension 1 manifolds. These derivations will relate the extrema of centripetal acceleration through a point to the character of that point as a minimizer of euclidean distance. Then we will apply these results to the specific problem of minimizing $||A - \hat{A}||$.

We first consider a general manifold M of co-dimension 1 in $\mathcal{R}^n$ or $\mathcal{C}^n$. Suppose that $p \in M$ locally minimizes the function

$$D(p,q) = \frac{1}{2}\langle p-q, p-q\rangle = \frac{1}{2}d(p,q)^2 \tag{31}$$

for some point q in $\mathcal{R}^n$ or $\mathcal{C}^n$. Let $p(t) \in M$ be a geodesic passing through p at $t = 0$ satifying (25, 26) with unit velocity v ($\langle v, v\rangle = 1$). Differentiating (31) with respect to t, we have

$$\begin{aligned}
\frac{d}{dt}D(p(t),q) &= \mathrm{Real}\{\langle p-q, v\rangle\} \\
\frac{d^2}{dt^2}D(p(t),q) &= \mathrm{Real}\{\langle v,v\rangle + \langle p-q, -\Gamma(v,v)\rangle\} \\
\frac{d^2}{dt^2}D(p(t),q) &= 1 - \mathrm{Real}\{\langle p-q, \Gamma(v,v)\rangle\}.
\end{aligned}$$

Setting the first of these equations to zero at $t = 0$ for all v gives $p - q = \sigma\hat{n}$ where σ is a scalar ($|\sigma| = d(p,q)$) and $\hat{n}$ is the unit normal of M at p. This gives

$$\frac{d^2}{dt^2}D(p(t),q) = 1 - \mathrm{Real}\{\langle \sigma\hat{n}, \Gamma(v,v)\rangle\} \tag{32}$$

$$\frac{d^2}{dt^2}D(p(t),q) = 1 - \mathrm{Real}\{\bar{\sigma}a(v,v)\}, \tag{33}$$

where we have substituted the Levi-Civita connection, $\Gamma(v,v) = a(v,v)\hat{n}$.

In order for p to be a stationary point of $D(p,q)$ it is necessary that $q = p + \sigma\hat{n}$. However, in order for this stationary point to be a local minimum, the second derivative must be nonnegative for all v. Since $a(v,v)$ is bilinear in v we may select the phase of v so that the condition (33) becomes

$$1 - |\sigma||a(v,v)| \geq 0,$$

or

$$\frac{1}{\sigma} \geq a(v,v), \tag{34}$$

where we understand (34) as a comparison between magnitudes. Since (34) must be true for all unit v, and since the principal curvatures of M at p are given by the extrema of $|a(v,v)|$ over the set of all unit v, we have

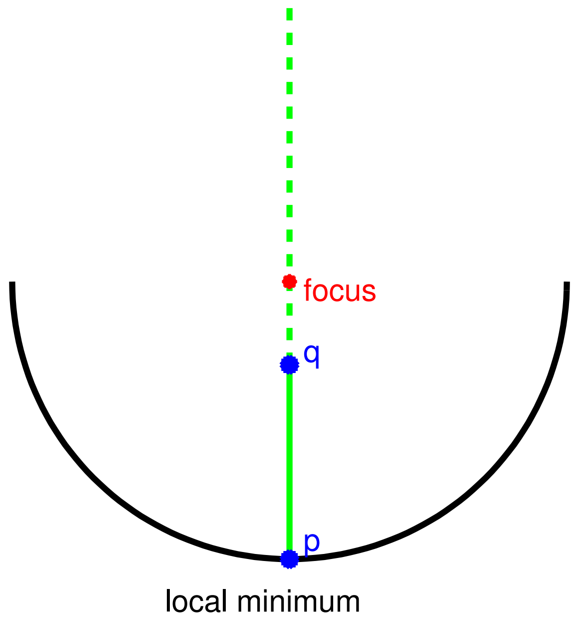

Figure 19: Below the focal point of a hemispherical surface, the distance minimizing p on the hemisphere to a given q is at the bottom of the surface.

Lemma 6.1 *If $d(p,q)$ is minimized by p then it is necessary that the focal lengths of M at p be greater than or equal to $d(p,q)$; i.e., if there is a focal point on the line segment $p+s\sigma\hat{n}$ $(0 \le s \le 1)$ connecting p to q, p cannot minimize $d(p,q)$.*

This relation of focal points to the characterization of the stationary points of $d(p,q)$ can be illustrated clearly in two dimensions. In Figures 19 and 20, we use a hemispherical surface to illustrate these lemmae.

We may now move to the specific case of minimizing $d(\hat{A},A)$ for $\hat{A} \in \mathcal{D}$.

Lemma 6.2 *Let*

$$\hat{A} = \tilde{U}\tilde{\Sigma}\tilde{V}^H + \lambda I$$

be an element of $\mathcal{D}$. Let $a(H,H)$ be the centripetal acceleration defined on the tangent vectors of $\hat{A}$ by Equation (27). Then, taking maximums over the unit tangent ball ($\langle H,H\rangle = 1$), $\frac{1}{\sigma_{n-1}} \le \max_H |a(H,H)|$, and, generically, $\frac{1}{\sigma_{n-1}} < \max_H |a(H,H)|$.

Proof: Using Equations (27, 28, 29) we write $a(H,H)$ in terms of a quadratic form

$$a(H,H) = z^T(Z - aa^T)z$$

where $Z = \begin{pmatrix} -\tilde{\Sigma}^{-1} & 0 \\ 0 & \tilde{\Sigma}^{-1} \end{pmatrix}$ and $a = \frac{1}{2\sqrt{\bar{l}^T\tilde{\Sigma}^{-1}r}} \begin{pmatrix} \tilde{\Sigma}^{-1}(r+\bar{l}) \\ \tilde{\Sigma}^{-1}(r-\bar{l}) \end{pmatrix}$.

We may bound the maximum magnitude of this form from below with the 2×2 quadratic form formed from a submatrix

$$|z^T(\begin{pmatrix} -\frac{1}{\sigma_{n-1}} & 0 \\ 0 & \frac{1}{\sigma_{n-1}} \end{pmatrix} - aa^T)z| \le \max_{z,||z||=1} |z^T(Z - aa^T)z|,$$

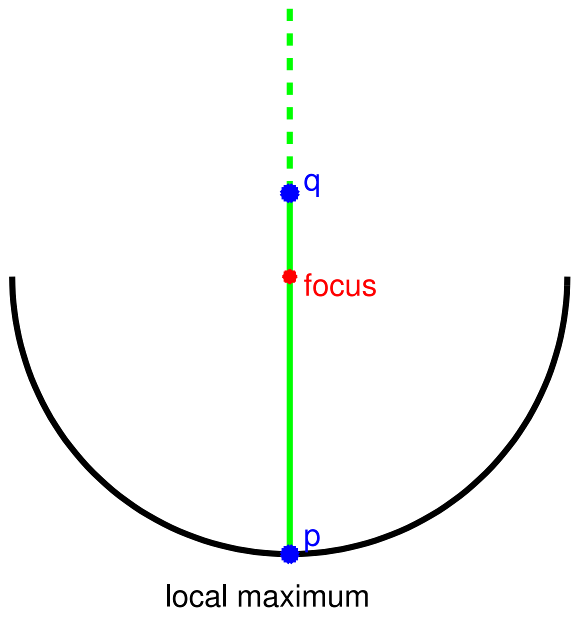

Figure 20: Above the focal point of the hemispherical surface, the bottom point of the hemisphere is now a maximum of distance (the nearest p on the hemisphere to q has become either of the end points of the hemisphere).

where $a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \frac{1}{2\sqrt{\bar{l}^T \tilde{\Sigma}^{-1} r}} \begin{pmatrix} \frac{1}{\sigma_{n-1}}(r_{n-1} + \bar{l}_{n-1}) \\ \frac{1}{\sigma_{n-1}}(r_{n-1} - \bar{l}_{n-1}) \end{pmatrix}$.

Let $z = \begin{pmatrix} i\sin(\theta) \\ \cos(\theta) \end{pmatrix}$ for some θ (we set $\theta = 0$ when working in $\mathcal{R}$). We have

$$z^T(\begin{pmatrix} -\frac{1}{\sigma_{n-1}} & 0 \\ 0 & \frac{1}{\sigma_{n-1}} \end{pmatrix} - aa^T)z = \frac{1}{\sigma_{n-1}} - (a_2\cos(\theta) + ia_1\sin(\theta))^2.$$

We see that if we select θ such that $\mathrm{Real}\{(a_2\cos(\theta) + ia_1\sin(\theta))^2\} \le 0$ then

$$\frac{1}{\sigma_{n-1}} \le |z^T(\begin{pmatrix} -\frac{1}{\sigma_{n-1}} & 0 \\ 0 & \frac{1}{\sigma_{n-1}} \end{pmatrix} - aa^T)z| \le \max_{z,||z||=1} |z^T(Z - aa^T)z|.$$

Generically, we can expect that we can select θ such that $\mathrm{Real}\{(a_2\cos(\theta)+ia_1\sin(\theta))^2\} < 0$. Thus,

$$\frac{1}{\sigma_{n-1}} < |z^T(\begin{pmatrix} -\frac{1}{\sigma_{n-1}} & 0 \\ 0 & \frac{1}{\sigma_{n-1}} \end{pmatrix} - aa^T)z| \le \max_{z,||z||=1} |z^T(Z - aa^T)z|,$$

and the inequality becomes strict.

$\square$

We are now in a position to prove

Lemma 6.3 *If λ_0 is a critical point of $\sigma_k(A-\lambda I)$ (with $\sigma_k(A-\lambda_0 I) > \sigma_n(A-\lambda_0 I)$) then (using singular vectors u and v) $\hat{A} = A - \sigma_k uv^H$ cannot locally minimize $||A-\hat{A}||$ over all $\hat{A} \in \mathcal{D}$.*

Proof: Since the smallest focal length (reciprocal principal curvature) of $\hat{A}$ is less than σ_n, it is impossible for $d(\hat{A}, A) = \sigma_k > \sigma_n$ to be locally minimized. □

We can also follow this sort of reasoning to establish that the points in the spectral portrait which are crossing points of singular values will usually not produce local minimizers of distance.

Corollary 6.4 *Generically, $\sigma_n(A-\lambda I)$ are simple in the neighborhood of λ_0.*

Proof: Generically, if $\sigma_{n-1} = \sigma_n$ at $\lambda = \lambda_0$, then $\hat{A}$ cannot be a local minimum of $d(\hat{A}, A)$. □

7 Triple eigenvalues

The variety of matrices with triple eigenvalues, $\mathcal{T}$, is a subvariety of $\mathcal{D}$. It is the variety of cusp points of $\mathcal{D}$, the points where $\mathcal{D}$ is not smooth. This might lead one to think that higher order derivatives of $\sigma_{\min}$ might shed light on the distance to $\mathcal{T}$. This is not quite the case, however, though some bounds are possible. This section will investigate relations which can be made between $\mathcal{T}$ and the focal points of $\mathcal{D}$.

It is easy to see that the intersection of $\mathcal{D}$ with its evolute is $\mathcal{T}$. For $\hat{A} \in \mathcal{D}$ with repeated eigenvalue λ, we may take $x_1 = v$, $x_2 = \tilde{U}\tilde{\Sigma}^{-1}r$ as the right eigenvector and generalized eigenvector and $y_2 = u$, $y_1 = \tilde{V}\tilde{\Sigma}^{-1}l$ as the left eigenvector and generalized eigenvector, with

$$\begin{pmatrix} y_1 & y_2 \end{pmatrix}^H \begin{pmatrix} x_1 & x_2 \end{pmatrix} = l^H\tilde{\Sigma}^{-1}r \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Note that if $l^H\tilde{\Sigma}^{-1}r = 0$ then the left and right eigenvectors cannot be duals and thus there must be additional right and left generalized eigenvectors of eigenvalue λ. Let us now suppose that $\hat{A}$ has a focal length of 0 (thus $\hat{A}$ is in the evolute of $\mathcal{D}$). Then, by the definitions of previous sections,

$$|r^H\Phi(0)r + l^H\Phi(0)l| = |2l^H\Psi(0)r|.$$

However, the LHS is zero. Thus we have $l^H\Psi(0)r = l^H\tilde{\Sigma}^{-1}r = 0$.

This is equivalent to the condition $\det \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix} = 0$. Thus $\mathcal{T}$ can be thought of as the locus of points satisfying

$$\begin{aligned} \sigma_{\min}(A-\lambda) &= 0 \\ \frac{d}{d\lambda}\sigma_{\min}(A-\lambda) &= 0 \end{aligned}$$

for some λ, with the additional constraint

$$\frac{d^2}{d\lambda^2}\sigma_{\min}(A-\lambda) = 0,$$

in contrast with the focal points which satisfy

$$\begin{aligned}\sigma_{\min}(A-\lambda) &= \sigma \\ \frac{d}{d\lambda}\sigma_{\min}(A-\lambda) &= 0 \\ \frac{d^2}{d\lambda^2}\sigma_{\min}(A-\lambda) &= 0.\end{aligned}$$

Unfortunately, while the first derivative of $\sigma_{\min}$ is independent of the value of $\sigma_{\min}$, the second derivative is not, preventing us from being able to use this set of equations as a method for finding the nearest element of $\mathcal{T}$.

We can bound the distance to the nearest $\tilde{A} \in \mathcal{T}$ from A if $\hat{A} \in \mathcal{D}$ is ill-conditioned. It is easy to show that one can always find a $\tilde{A}$ within $|\lambda - \lambda_i|(2l^H\tilde{\Sigma}^{-1}r)$ of $\hat{A}$, where λ_i is an eigenvalue of $\hat{A}$ distinct from and closest to the double eigenvalue λ.

References

[1] J. W. Bruce and P. J. Giblin. *Curves and Singularities (Second Edition).* Cambridge University Press, Cambridge, England, 1992.

[2] S.-N. Chow, C. Li, and D. Wang. *Normal Forms and Bifurcation of Planar Vector Fields.* Cambridge, University Press, 1994.

[3] J. W. Demmel. *A Numerical Analyst's Jordan Canonical Form.* PhD thesis, Univ. of California at Berkeley, Berkeley, CA, USA, 1983.

[4] M. P. do Carmo. *Differential geometry of curves and surfaces.* Prentice–Hall, 1976.

[5] A. Edelman and E. Elmroth. Personal communication, January 1997.

[6] A. Edelman and Y. Ma. Non-generic eigenvalue perturbations of jordan blocks. *Linear Algebra Appl.*, 273:45–63, 1998.

[7] A. Edelman and Y. Ma. Staircase failures explains by orthogonal versal forms. *Submitted to SIAM J. Matrix Ana. and Appl.*, 1998.

[8] T. F. Fairgrieve. The application of singularity theory to the computation of Jordan Canonical Form. *Master Thesis at Computer Science in Univ. of Toronto*, 1986.

[9] G. H. Golub and Charles F. Van Loan. *Matrix Computations.* Johns Hopkins University Press, Baltimore, Maryland, 2d edition, 1989.

[10] T. Kato. *Perturbation Theory for Linear Operators.* Springer-Verlag, New York, NY, USA, 1966.

[11] R. A. Lippert. *Nonlinear Eigenvalue Problems.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1998.

[12] R. A. Lippert and A. Edelman. Nonlinear eigenvalue problems with orthogonality constraints. In et al. Z. Bai, editor, *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*. SIAM (to appear), 1999(?).

[13] Ross Lippert and Alan Edelman. The computation and sensitivity of double eigenvalues. In Z. Chen, Y. Li, C. A. Micchelli, and Y. Xu, editors, *Advances in Computational Mathematics*, volume 202 of *Lecture Notes in Pure and Applied Mathematics*. Marcel Dekker, 1998.

[14] C. F. Van Loan. On estimating the condition of eigenvalues and eigenvectors. *Linear Algebra Appl.*, 88/89:715–732, 1987.

[15] A. N. Malyshev. On wilkinson's problem. Technical report, Report 140, Dept. of Informatics, University of Bergen, Norway, November 1997.

[16] S. H. Strogatz. *Nonlinear dynamics and Chaos: with applications to physics, biology, chemistry, and engineering*. Addison-Wesley, 1994.

[17] J.-G. Sun. A note on simple non-zero singular values. *Journal of Computational Mathematics*, 6:258–266, 1988.

[18] J.-G. Sun. Sensitivity analysis of multiple eigenvalues. I. *Journal of Computational Mathematics*, 6:28–38, 1988. Cited in [19].

[19] J.-G. Sun. A note on local behavior of multiple eigenvalues. *SIAM Journal on Matrix Analysis and Applications*, 10:533–541, 1989.

[20] L. N. Trefethen. Pseudospectra of matrices. In D. F. Griffiths and G. A. Watson, editors, *Numerical Analysis 1991*. Longman Scientific and Technical, Harlow, Essex, 1992.

[21] J. H. Wilkinson. *The Algebraic Eigenvalue Problem*. Clarendon Press, Oxford, England, 1965.

[22] J. H. Wilkinson. Sensitivity of eigenvalues. *Util. Math.*, 25:5–76, 1984.

[23] J. H. Wilkinson. Sensitivity of eigenvalues II. *Util. Math.*, 30:243–286, 1986.