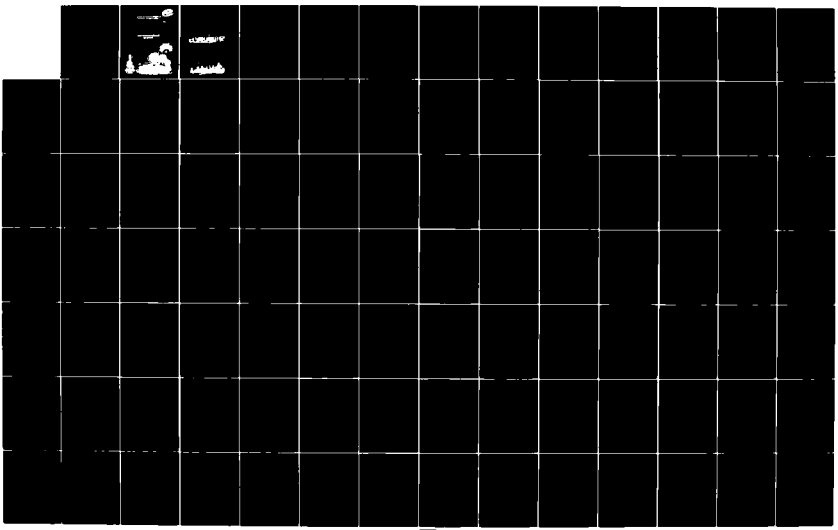
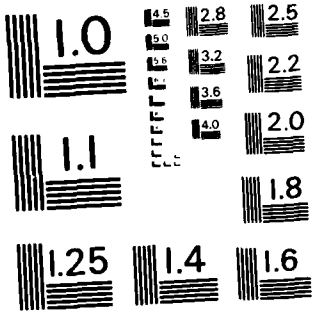
AD-A130 775 A NUMERICAL ANALYST'S JORDAN CANONICAL FORM(U) 1/2
CALIFORNIA UNIV BERKELEY CENTER FOR PURE AND APPLIED
MATHEMATICS J W DEMMEL MAY 83 PAM-156 N00014-76-C-0013

UNCLASSIFIED F/G 12/1 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS - 1963 - A

AD A130775

Security Classification

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Mathematics Department UC Berkeley	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP

3. REPORT TITLE

A Numerical Analyst's Jordan Canonical Form

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Scientific Final

5. AUTHOR(S) (First name, middle initial, last name)

James Weldon Demmel

6. REPORT DATE May 1983	7a. TOTAL NO. OF PAGES 142	7b. NO. OF REFS 47
8a. CONTRACT OR GRANT NO. N 00014-76-C-0013 b. PROJECT NO. c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)	

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Office of Naval Research

13. ABSTRACT

What does it mean to compute an eigendecomposition of an uncertain matrix? Because of measurement errors and roundoff errors, one must typically compute the eigenvalues and eigenvectors not of a single matrix but rather of a ball of matrices whose radius depends on the uncertainty in the data. We approach this problem by asking how to partition the eigenvalues of the matrices in the ball into nonoverlapping groups which cannot themselves be further partitioned. More specifically, we define the *dissociation* of two subsets σ_1 and $\sigma_2 \subset \sigma$ of the sets of eigenvalues σ of a matrix T as the smallest perturbation of T that will make some eigenvalue from σ_1 and some eigenvalue from σ_2 move together and become indistinguishable. If T is the center of the ball of matrices, and the dissociation of σ_1 and σ_2 is

(continued on back)

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6801

Security Classification

greater than the radius of the ball, then σ_1 and σ_2 are nonoverlapping groups of eigenvalues; otherwise the dissociation is less than or equal to the radius and σ_1 and σ_2 are not distinguishable groups. By computing the dissociation for various σ_1 and σ_2, we may compute our desired partition of σ.

The results of this thesis are of two kinds. First, we compute upper and lower bounds on the dissociation which improve bounds in the literature. Both upper and lower bounds are achievable or nearly so. The upper and lower bounds are often close together but occasionally far apart. Our second set of results quantifies this last statement by assuming a probability density on the set of matrices and computing the likelihood that the bounds are far apart. This approach leads to numerous other probabilistic results, such as the distribution of the condition number of a random matrix, and the distribution of the distance from a random matrix to one with a given Jordan form. We discuss the relevance of this probabilistic model to finite precision calculations.

A Numerical Analyst's Jordan Cononical Form

By

James Weldon Demmel

B.S. (California Institute of Technology) 1975

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Computer Science

in the

GRADUATE DIVISION

OF THE

UNIVERSITY OF CALIFORNIA, BERKELEY

Approved: W. Kahan ..... May 20, 1983
Chairman Date

Joseph Katz .....

B. N. Parlett .....

.....

Accession For
NTIS GRA&I ☒
DTIC TAB ☐
Unannounced ☐
Justification
By
Distribution/
Availability Codes
Dist | Avail and/or Special
A

DTIC COPY INSPECTED

DISTRIBUTION STATEMENT
Approved for public release;
Distribution Unlimited

A Numerical Analyst's Jordan Canonical Form

James Weldon Demmel

ABSTRACT

What does it mean to compute an eigendecomposition of an uncertain matrix? Because of measurement errors and roundoff errors, one must typically compute the eigenvalues and eigenvectors not of a single matrix but rather of a ball of matrices whose radius depends on the uncertainty in the data. We approach this problem by asking how to partition the eigenvalues of the matrices in the ball into nonoverlapping groups which cannot themselves be further partitioned. More specifically, we define the *dissociation* of two subsets σ_1 and $\sigma_2=\sigma \setminus \sigma_1$ of the set of eigenvalues σ of a matrix T as the smallest perturbation of T that will make some eigenvalue from σ_1 and some eigenvalue from σ_2 move together and become indistinguishable. If T is the center of the ball of matrices, and the dissociation of σ_1 and σ_2 is greater than the radius of the ball, then σ_1 and σ_2 are nonoverlapping groups of eigenvalues; otherwise the dissociation is less than or equal to the radius and σ_1 and σ_2 are not distinguishable groups. By computing the dissociation for various σ_1 and σ_2, we may compute our desired partition of σ.

The results of this thesis are of two kinds. First, we compute upper and lower bounds on the dissociation which improve bounds in the literature. Both upper and lower bounds are achievable or nearly so. The upper and lower bounds are often close together but occasionally far apart. Our second set of results quantifies this last statement by assuming a probability density on the set of matrices and computing the likelihood that the bounds are far apart. This approach leads to numerous other probabilistic results, such as the distribution of the condition number of a random matrix, and the distri-

bution of the distance from a random matrix to one with a given Jordan form. We discuss the relevance of this probabilistic model to finite precision calculations.

Acknowledgements

I am grateful to Professors W. Kahan, B. Parlett and S. Smale as well as numerous other friends and colleagues for many useful discussions. I also acknowledge the financial support of the U.S. Department of Energy, Contract DE-AM03-76SF00034, Project Agreement DE-AS03-79ER10358, the Office of Naval Research, Contract N00014-76-C-0013, IBM, Contract 82007PLP0446, and the IBM Graduate Fellowship Program.

Table of Contents

	Acknowledgements	i
	Table of Contents	ii
	List of Figures	iii
1	Introduction	1
2	Preliminary Definitions and Lemmas	12
3	Best Conditioned Diagonalizing Similarities	29
4	Lower Bounds on $\mathrm{diss}_g(\sigma_1, \sigma_2, \mathit{region})$	53
5	How Far Apart can the Upper and Lower Bounds on $\mathrm{diss}_g(\sigma_1, \sigma_2)$ Be?	65
6	A Probabilistic Model	76
7	Applications to Matrix Inversion, Eigenvalue Problems, and Polynomial Zero Finding	96
8	Probabilistic Estimates of $\mathrm{diss}_g(\sigma_1, \sigma_2)$	112
9	Relevance of the Probabilistic Model to Finite Precision Calculations	121
	References	138

List of Figures

6.1	Pinched Section in a Variety	95
6.2	Distances on the Circle	95
6.3	Counting the Area of the Doubly Shaded Region Twice	95
9.1	A 4ε Neighborhood of the Curve P	136
9.2	An $\varepsilon/4$ Neighborhood of the Curve P	136
9.3	Observed Probability Distribution of the Distance ε to the Nearest Singular Matrix	137

A Numerical Analyst's Jordan Canonical Form

James Weldon Demmel

Chapter 1: Introduction

Given a complex n by n matrix T_0 known only to within a tolerance $\varepsilon>0$, what does it mean to compute an eigendecomposition of T_0? By knowing T_0 only to a tolerance ε we mean that T_0 is indistinguishable from any matrix in the set

$$\mathbf{T}(\varepsilon) = \{T: \| T_0 - T \| < \varepsilon\} \ , \tag{1.1}$$

($\| T \|$ denotes some norm of the matrix T; we will specify the norm later.) We would like to produce an eigendecomposition which is valid in some way for all matrices in $\mathbf{T}(\varepsilon)$, and gives as much information as possible about all matrices in $\mathbf{T}(\varepsilon)$. This seemingly simple goal leads us along several interesting paths which will now be explored.

First some notation. An eigendecomposition of a matrix T will be written

$$T = S\Theta S^{-1} \tag{1.2}$$

where Θ is a block diagonal matrix $\Theta = \mathrm{diag}(\Theta_1, \ldots, \Theta_b)$. T's spectrum will be denoted by $\sigma(T)$ or merely σ if T is clear from context, and Θ_i's spectrum by σ_i for short. Thus $\sigma = \bigcup_1^b \sigma_i$. If σ_i contains m eigenvalues (counting multiplicities) we write $\#(\sigma_i)=m$.

The following sequence of examples will illustrate the difficulties encountered in computing an eigendecomposition of $\mathbf{T}(\varepsilon)$ for different values of ε. Consider the following matrix, which is essentially in Jordan canonical form:

$$T_0 = \begin{bmatrix} 1 & & & & \\ & 1.001 & & & \\ & & 0 & 100 & \\ & & & 0 & \\ & & & & 0 \\ & & & & & -1 \end{bmatrix} \tag{1.3}$$

(blanks and 0 both denote zero entries). This decomposition tells us several things: that T_0 has 4 distinct eigenvalues at 1.001, 1, 0, and -1, that each nonzero eigenvalue has a one dimensional invariant subspace (i.e. an eigenvector) associated with it, and that associated with 0 is one two dimensional and one one dimensional invariant subspace.

Does this information remains valid for all matrices in $\mathbf{T}(\varepsilon)$ as ε increases from 0? As soon as ε becomes nonzero, it is no longer true that all matrices in $\mathbf{T}(\varepsilon)$ have a triple eigenvalue at 0, nor two invariant subspaces associated with eigenvalues near 0. For example,

$$T_1 = \begin{bmatrix} 1 & & & & \\ & 1.001 & & & \\ & & 0 & 100 & \\ & & \varepsilon & 0 & \\ & & & & 0 \\ & & & & & -1 \end{bmatrix}$$

has 3 simple eigenvalues at 0, $10\sqrt{\varepsilon}$, and $-10\sqrt{\varepsilon}$ each with its own eigenvector, and

$$T_2 = \begin{bmatrix} 1 & & & & \\ & 1.001 & & & \\ & & 0 & 100 & \\ & & & 0 & \varepsilon \\ & & & & 0 \\ & & & & & -1 \end{bmatrix}$$

has one triple eigenvalue at 0 with just one three dimensional invariant subspace associated with it.

Thus, all matrices in $\mathbf{T}(\varepsilon)$ (for ε small enough) have three eigenvalues near to 0 which together have a three dimensional invariant subspace associ-

ated with them. We cannot, however, identify them individually because they could all simultaneously equal 0 (in the case of T_0); their only identities are as members of a cluster of three.

As ε increases to .0005, matrices occur in $\mathbf{T}(\varepsilon)$ which no longer have two simple eigenvalues around 1:

$$T_3 = \begin{bmatrix} 1.0005 & \eta & & & & \\ & 1.0005 & & & & \\ & & 0 & 100 & & \\ & & & 0 & & \\ & & & & 0 & \\ & & & & & -1 \end{bmatrix}.$$

T_3 has two eigenvalues at 1.0005 associated with a two dimensional invariant subspace and for $\eta \neq 0$ but arbitrarily small this subspace cannot be split into two one dimensional subspaces. Thus, when ε exceeds .0005 (but not .005), $\mathbf{T}(\varepsilon)$ has one three dimensional invariant subspace with three eigenvalues indistinguishable from 0, one two dimensional subspace with two eigenvalues indistinguishable from 1.0005, and one simple eigenvalue at -1.

In particular, one may draw three disjoint simple closed curves (Jordan curves) in the complex plane, one around 0, one around 1, and one around -1, such that any $T \in \mathbf{T}(.0005)$ will have three eigenvalues clustered strictly inside the first curve, two inside the second, and one inside the third. Furthermore, it is impossible to draw any larger number of such curves such that each one will strictly contain a fixed number of eigenvalues of each $T \in \mathbf{T}(.0005)$. This last statement is true because within $\mathbf{T}(.0005)$ there is a matrix (T_3) with a double eigenvalue at 1.0005 and a triple eigenvalue at 0.

For values of ε exceeding .005, say .01, the clustering of the eigenvalues changes again. The matrix

$$T_4 = \begin{bmatrix} 1 & & & & & \\ & 1.001 & & & & \\ & & 0 & 100 & & \\ & & .01 & 0 & & \\ & & & & 0 & \\ & & & & & -1 \end{bmatrix}.$$

has simple eigenvalues at 1.001 and 0, and double eigenvalues at 1 and -1. Looking at the eigenvalues as functions of the entry containing .01 (the 3,4 entry), that T_4 has a pair of eigenvalues at $\pm 10\sqrt{T_{4\,3,4}} = \pm 1$ when $T_{4\,3,4} = .01$. Thus, no Jordan curve can be drawn which separates the eigenvalues into disjoint regions as was done in the case of $\mathbf{T}(.0005)$ or for $\mathbf{T}(\varepsilon)$ with smaller ε. This is because the eigenvalues "near 0" can no longer be separated from the eigenvalues near -1 nor 1, and neither can the eigenvalue at 1.001 be separated from 1. Thus, one Jordan curve must be drawn containing all the eigenvalues.

In the case of $\mathbf{T}(.0005)$ we could find a matrix (T_3) with a single multiple eigenvalue within the region bounded by each Jordan curve. It seems natural to think of all matrices in $\mathbf{T}(.0005)$ as being small perturbations of one with a double eigenvalues at 1.0005, a triple eigenvalue at 0, and a simple eigenvalue at -1 (T_3). The existence of T_3 also provides a simple explanation for not being able to distinguish the three eigenvalues near 0 or the two near 1. Is it possible to find a matrix with a sextuple eigenvalue in $\mathbf{T}(.01)$? More generally, given a $\mathbf{T}(\varepsilon)$ and a clustering of eigenvalues which can not be refined by drawing more separating Jordan curves, is it possible to find T's in $\mathbf{T}(\varepsilon)$ which have single eigenvalues in place of each cluster? The answer is no. We and independently Wilkinson have produced examples such as [Wilkinson4]

$$T_0 = \begin{bmatrix} \eta & 1 & & \\ & 2\eta & 1 & \\ & & 3\eta & 1 \\ & & & 4\eta \end{bmatrix}$$

where for $\varepsilon > \eta^4$ ($\eta \ll 1$) one Jordan curve around the entire spectrum of $\mathbf{T}(\varepsilon)$ must be drawn, but where ε must exceed something of order $\eta^2 \gg \eta^4$ before a matrix with a single quadruple eigenvalue can be found in $\mathbf{T}(\varepsilon)$. We call the eigenproblem for $\mathbf{T}(\varepsilon')$ ($\eta^4 < \varepsilon' < \eta^2$) *ill posed*, because while no nonempty proper subset of the eigenvalues is distinguishable (by being separable by a Jordan curve from the remaining eigenvalues), matrices in $\mathbf{T}(\varepsilon)$ cannot be thought of as perturbations of some particular matrix in $\mathbf{T}(\varepsilon)$ with a single quadruple eigenvalue. The problem of locating the nearest matrix with just one eigenvalue is called the "nearest completely defective matrix problem."

The central problem in this sequence of examples has been how to cluster the eigenvalues into distinguishable groups, how to *name* the eigenvalues. There are at least two notions of clustering for eigenvalues. So far we have sought a collection of Jordan curves $\{J_i\}$ such that the region bounded by each J_i contains the same number of eigenvalues (counting multiplicities) of each $T \in \mathbf{T}(\varepsilon)$. This number of eigenvalues will be called the *content* of J_i. The easiest way to see how these curves depend on T_0 and ε is to consider the set $\sigma(\mathbf{T}(\varepsilon))$ of *all* eigenvalues of *all* $T \in \mathbf{T}(\varepsilon)$. $\sigma(\mathbf{T}(\varepsilon))$ is an open set and can be written as the disjoint union of its connected components $\sigma_i(\mathbf{T}(\varepsilon))$. Around each $\sigma_i(\mathbf{T}(\varepsilon))$ one can draw a Jordan curve J_i with $\sigma_i(\mathbf{T}(\varepsilon))$ strictly inside J_i and all other $\sigma_j(\mathbf{T}(\varepsilon))$ strictly outside. These Jordan curves cluster the eigenvalues of T_0 into regions in a way that also clusters the eigenvalues of each $T \in \mathbf{T}(\varepsilon)$. This notion of naming an eigenvalue by the component of $\sigma(\mathbf{T}(\varepsilon))$ in which it lies will be called *region clustering*.

There is another useful notion of clustering or naming. It will be described briefly here, with the formal definition left to the next chapter. Let λ_i be an eigenvalue of T_0, and let $T(x)$ be a continuous path starting at $T(0)=T_0$ and remaining in $\mathbf{T}(\varepsilon)$ for all $x>0$. Think of λ_i as a function of x, varying continuously as a function of x. As long as

$$\lambda_i(x)\neq\lambda_j(x) \text{ for all } x \quad \text{and} \quad j\neq i \tag{*}$$

$\lambda_i(x)$ can be unambiguously identified with λ_i. If (*) is true for *all* paths $T(x)$ in $\mathbf{T}(\varepsilon)$, then λ_i represents a cluster (of content 1) for all matrices in $\mathbf{T}(\varepsilon)$. This definition makes sense because as long as $\lambda_i(x)$ never equals $\lambda_j(x)$, it can be identified by naming it by the λ_i and the path $T(x)$ whence it came. If, on the other hand, there is a path $T(x)$ in $\mathbf{T}(\varepsilon)$ and an x_0 such that $\lambda_i(x_0)=\lambda_j(x_0)$, then we put λ_i and λ_j into the same cluster. In this way, a unique clustering of σ is constructed. This clustering method will be called ***path clustering***. It will be shown in the next chapter that given $\mathbf{T}(\varepsilon)$, this path clustering always produces at least as refined a clustering as does region clustering.

For numerical reasons to be discussed in a moment, one may add another constraint to the clustering of eigenvalues. Consider

$$T_0(x) = \begin{bmatrix} -x & 1 \\ & x \end{bmatrix} = \begin{bmatrix} 1 & 1/2x \\ & 1 \end{bmatrix} \cdot \begin{bmatrix} -x & \\ & x \end{bmatrix} \cdot \begin{bmatrix} 1 & -1/2x \\ & 1 \end{bmatrix} = S(x) \cdot \Theta(x) \cdot S(x)^{-1} .$$

As long as ε in $\mathbf{T}(\varepsilon)$ is less than $\varepsilon(x)=(\sqrt{4x^2+1}-1)/2\approx x^2$ for tiny x, the two eigenvalues must remain simple. However, as ε gets close to $\varepsilon(x)$, not only do the two eigenvalues get close, but the similarity transformation $S(x)$ which exhibits the decomposition in the last equation gets more and more ill-conditioned. That is, as $\varepsilon\to\varepsilon(x)$, $\|S(x)\| \cdot \|S(x)^{-1}\| \to\infty$. The ill-condition of $S(x)$ is numerically significant because it means computing $S(x)AS(x)^{-1}$ in floating point arithmetic is apt to lead to large errors (this phenomenon will

be discussed more in Chap. 3). Thus, one may add the constraint to a clustering that the matrix S which exhibits the decomposition must have a condition number less than some tolerance $\bar{\kappa}$:

$$\kappa(S) \equiv \|S\| \cdot \|S^{-1}\| < \bar{\kappa} \ .$$

At this point the reader might object to example (1.3) on the grounds that the 2 by 2 block

$$\begin{bmatrix} 0 & 100 \\ & 0 \end{bmatrix}$$

is "obviously" separate from the blocks containing 1, 1.001, and -1, because the off diagonal zeroes are "obviously" sacred. We can quantify this intuition by using only the condition number of the best conditioned diagonalizing similarity $\kappa(S)$ which displays the eigendecomposition as in (1.2): if $\kappa(S)<\bar{\kappa}$, then the decomposition is acceptable, otherwise it is not. In the case of (1.3), which is already diagonalized as much as possible, we may take $S=I$ so $\kappa(S)=1$, the smallest possible value of $\kappa(S)$ for any S. This criterion, which generally allows a finer clustering than the scheme in (1)-(3), can be used to decompose matrices in preparation for computing functions of them, such as the exponential. (This type of decomposition will be discussed further in Chapter 3.)

Let us review the discussion so far by describing a program to compute the eigendecomposition of an uncertain matrix.

(1) Given T_0 and ε, we must cluster the spectrum of the matrix into groups. As stated above, there are two possible criteria for performing the clustering. Whichever one is chosen, it will turn out that we need only consider clustering $\sigma(T_0)=\sigma_1 \cup \sigma_2$ into two disjoint pieces. Given such a clustering we must be able to compute the largest $\bar{\varepsilon}$ such that

Region Clustering: there is a Jordan curve or curves J dividing the complex plane into two regions such that the groups σ_1 and σ_2 remain on opposite sides of J for all $T \in \mathbf{T}_\varepsilon$, or

Path Clustering: for all paths $T(x)$ in $\mathbf{T}_\varepsilon$, $\lambda_1(x) \neq \lambda_2(x)$ for all $\lambda_i(0) \in \sigma_i$ and for all x.

This largest ε will be called the *(region or path) dissociation between* σ_1 *and* σ_2, and denoted by $\mathrm{diss}_E(\sigma_1, \sigma_2, T_0, \mathit{region})$ or $\mathrm{diss}_E(\sigma_1, \sigma_2, T_0, \mathit{path})$ (or $\mathrm{diss}_E(\sigma_1, \sigma_2)$ if both the choice between "region" or "path" and T_0 are clear from context). The subscript E indicates that perturbations are measured using the Euclidean norm. We will also consider $\mathrm{diss}_2(\sigma_1, \sigma_2)$, where perturbations will be measured using the 2-norm (these norms are defined in the next chapter). A better known synonym for dissociation is separation, but separation has already been used for related quantities [Stewart, Varah] which will be considered in the next chapter.

(2) Given T_0, ε and a clustering $\sigma = \bigcup_1^b \sigma_i$, how ill-conditioned must S_T be if it exhibits the *eigendecomposition*

$$T = S_T \mathrm{diag}(\Theta_1, \ldots, \Theta_b) S_T^{-1}$$

where $T \in \mathbf{T}(\varepsilon)$ and $\sigma(\Theta_i)$ is identified with σ_i? If it must be too ill-conditioned, then we need to combine the σ_i from step (1) into larger groups.

(3) Given a cluster σ_i which contains more than one distinct eigenvalue and cannot be split, is there a $T \in \mathbf{T}(\varepsilon)$ all of whose eigenvalues within this group are equal? If so, this matrix (or at least its existence) should be reported to the user as output; if such a $T \in \mathbf{T}(\varepsilon)$ does not exist, the user should be told that this part of his problem is ill-posed. This problem is addressed by Ruhe and Kågström [Ruhe2, Kågström1], and we will not pursue it in this thesis.

Now we describe the contributions of this thesis to the solution of this problem. We were not, alas, able to solve the problem completely, but we have made substantial progress and our results are applicable to other problems as well.

The results are of two kinds. Chapters 2 through 5 analyze the dissociation and compute both upper and lower bounds for path and region dissociation. Both upper and lower bounds are attainable or nearly so for various classes of matrices. The upper and lower bounds are usually close, but can be very far apart. Chapters 6 through 8 take a probabilistic approach to analyze how likely the bounds are to be close or far apart, and show, for example, how to compute the probability that all the matrices in $\mathbf{T}(\varepsilon)$ will be completely diagonalizable. Chapter 9 examines the applicability of the probabilistic model to finite precision calculations. More specifically, the results are as follows.

Chapter 2 further discusses the two notions of clustering (region and path) described above. In particular we show

$$\mathrm{diss}(\sigma_1, \sigma_2, \mathit{path}) \geq \mathrm{diss}(\sigma_1, \sigma_2, \mathit{region}) \ ,$$

and that we can choose a matrix norm that makes the two dissociation measures unequal. We also define the simple dissociation measures which will be combined to produce bounds on $\mathrm{diss}(\sigma_1, \sigma_2, \mathit{region})$ and $\mathrm{diss}(\sigma_1, \sigma_2, \mathit{path})$. We derive basic properties of these measures, in particular how they behave under similarity transforms of the matrix, and a "divide and conquer" property that makes them easier to compute when the matrix has a block diagonal structure. We also present an upper bound on $\mathrm{diss}_2(\sigma_1, \sigma_2, \mathit{path})$ and $\mathrm{diss}_F(\sigma_1, \sigma_2, \mathit{path})$ based on one of these measures.

Chapter 3 solves the problem in step (2) above by computing an S_T whose condition number is within a factor of $\sqrt{b}$ of best possible (b is the number of partitions), and by computing explicit upper and lower bounds on the best possible condition number which differ by at most a factor of b. We also discuss the possibility of partitioning Σ subject solely to the criterion that the condition number of the diagonalizing similarity be less than a threshold.

Chapter 4 presents a new lower bound on $\mathrm{diss}_2(\sigma_1, \sigma_2, \mathit{region})$ which is sharper than the previously best bound, compares it to other known bounds, and discusses when it is likely to be sharp. Combined with the upper bound of chapter 2 this yields the inclusion

$$\text{upper bound} \geq \mathrm{diss}(\sigma_1, \sigma_2, \mathit{path}) \geq \mathrm{diss}(\sigma_1, \sigma_2, \mathit{region}) \geq \text{lower bound}$$

Chapter 5 analyzes how far apart the upper bound of chapter 2 and the lower bound of chapter 4 can be, and presents worst case examples which show how far apart the bounds must be. We also compute $\mathrm{diss}_2(\sigma_1, \sigma_2)$ and $\mathrm{diss}_F(\sigma_1, \sigma_2)$ exactly for normal matrices, in which case all four notions of dissociation (path or region, 2-norm or Euclidean norm) are equal.

Chapter 6 presents a geometric/probabilistic model of the problem, by defining certain sets in matrix space which are the sets where the eigenproblem becomes difficult. We discuss the algebraic and geometric properties of these sets, which are algebraic varieties, and put a probability measure on matrix space which lets us analyze what fraction of matrix space consists of hard problems.

Chapter 7 uses the model of chapter 6 to compute probability distributions of the smallest distance

from a random matrix to one with a given rank (such as the nearest singular matrix),

from a random matrix to one with a given Jordan form (such as the nearest matrix with a double eigenvalue), and

from a random polynomial to one with a given zero structure (such as the nearest polynomial with a double zero).

Chapter 8 uses the model developed in chapter 6 and the results of chapter 7 to analyze when the bounds discussed in chapters 2 through 5 are likely to be accurate. We show, for example, that the ratio of the upper to lower bounds on $\mathrm{diss}_2(\sigma_1, \sigma_2)$ cannot exceed $K>1$ except on a set of matrices of probability proportional to K^{-2}.

Chapter 9 investigates the usefulness of the probabilistic model of chapters 6 and 7 for analyzing the performance (speed and accuracy) of algorithms for matrix inversion, eigendecompositions, and polynomial root finding. A paradigm for analyzing performance is presented, which, when applied to matrix inversion, yields a lower bound on the probability distribution of the relative error in Gaussian elimination. The model, because it ignores the effects of finite precision arithmetic, fails to provide any useful information at all about certain algorithms whose performance depends strongly on the effects of finite precision arithmetic. We show how extending the model to take finite precision arithmetic into account could be used to measure how many problems can be solved as a function of the amount of extra precision carried in intermediate computations.

Chapter 2: Preliminary Definitions and Lemmas

2.1 Introduction

In this chapter we introduce the notation and dissociation measures used in the rest of the thesis. These dissociation measures will be used later in the thesis to construct upper and lower bounds on $\mathrm{diss}(\sigma_1, \sigma_2)$. These upper and lower bounds can be far apart; just how far apart is the subject of chapters 5 and 8. However, it is unlikely that they are very far apart; chapters 6 and 7 will present a natural model for "picking a matrix at random" which we will use in chapter 8 to make this assertion precise and prove it.

The rest of this chapter is organized as follows. Section 2.2 discusses $\mathrm{diss}(\sigma_1, \sigma_2, \mathit{region})$ and $\mathrm{diss}(\sigma_1, \sigma_2, \mathit{path})$. $\mathrm{diss}(\sigma_1, \sigma_2, \mathit{path})$ must always be at least as large as $\mathrm{diss}(\sigma_1, \sigma_2, \mathit{region})$ although they may indeed differ in certain circumstances. We also show that they provide enough information to cluster the eigenvalues in the way discussed in chapter 1. Section 2.3 discusses the canonical form we use for matrices. Sections 2.4 and 2.5 define the dissociation measures $\|P\|$ (P a projector), $\mathrm{sep}(A,B)$, and $\mathrm{sep}_\lambda(A,B)$ and discuss their basic properties. In particular, $\mathrm{sep}(A,B)$, $\mathrm{sep}_\lambda(A,B)$ and $\|P\|$ share certain scaling and "divide and conquer" properties which we later exploit to compute relationships among them.

2.2 The Difference between $\mathrm{diss}(\sigma_1, \sigma_2, \mathit{region})$ and $\mathrm{diss}(\sigma_1, \sigma_2, \mathit{path})$

This section discusse the difference between $\mathrm{diss}(\sigma_1, \sigma_2, \mathit{region})$ and $\mathrm{diss}(\sigma_1, \sigma_2, \mathit{path})$, proves that $\mathrm{diss}(\sigma_1, \sigma_2, \mathit{path}) \geq \mathrm{diss}(\sigma_1, \sigma_2, \mathit{region})$, and shows why they provide sufficient information to compute the decomposition of chapter 1.

Let us restate the definition of diss(σ_1, σ_2, *region*). The definition depends, of course, on the matrix norm $|||\cdot|||$ which defines the shape of $\mathbf{T}(\varepsilon) = \{T: |||T-T_0||| < \varepsilon\}$. Consider the set $\sigma(\mathbf{T}(\varepsilon))$ of all eigenvalues of all matrices T in $\mathbf{T}(\varepsilon)$. $\sigma(\mathbf{T}(\varepsilon))$ is an open set and can be written as the disjoint union of its connected components. σ_1 and σ_2, being sets of eigenvalues of T_0, must lie in these connected components. Let $\sigma_1(\mathbf{T}(\varepsilon))$ be the union of those components containing points of σ_1 and $\sigma_2(\mathbf{T}(\varepsilon))$ be the union of the components containing σ_2. If $\sigma_1(\mathbf{T}(\varepsilon))$ and $\sigma_2(\mathbf{T}(\varepsilon))$ are disjoint, then we can draw a Jordan curve $J(\varepsilon)$ having $\sigma_1(\mathbf{T}(\varepsilon))$ strictly inside and $\sigma_2(\mathbf{T}(\varepsilon))$ strictly outside. As we increase ε, $\sigma_1(\mathbf{T}(\varepsilon))$ and $\sigma_2(\mathbf{T}(\varepsilon))$ will grow from tiny neighborhoods around the eigenvalues when ε is near 0 and eventually intersect for ε greater than some $\bar{\varepsilon}$, at which point the curve $J(\varepsilon)$ no longer exists. This $\bar{\varepsilon}$, the *supremum* of the set of ε for which separating curves $J(\varepsilon)$ do exist, is the definition of diss(σ_1, σ_2, *region*) (note the implicit dependence on the norm $|||\cdot|||$).

Now we define diss(σ_1, σ_2, *path*). Let $T(x)$ be a continuous path starting at $T(0)=T_0$ and remaining inside $\mathbf{T}(\varepsilon)$ for all $x \geq 0$. Let $\lambda_0=[\lambda_1, \ldots, \lambda_n]$ be some ordering of T_0's eigenvalues, possibly with repeated entries for multiple eigenvalues. We wish to define $\lambda(x)=[\lambda_1(x), \ldots, \lambda_n(x)]$ so that $\lambda(x)$ is a list of the eigenvalues of $T(x)$ and a continuous function of x. This is possible since the eigenvalues are continuous functions of the matrix. The only ambiguity arises when some some $T(x_0)$ has a multiple eigenvalue $\lambda_1(x_0)=\lambda_2(x_0)$ (say). In this case one may arbitrarily choose which eigenvalue to call $\lambda_1(x)$ and which to call $\lambda_2(x)$ for $x>x_0$ (this arbitrariness will not affect the definition of dissociation). Suppose that $\lambda_1(x_1)=\lambda_2(x_2)$ for some path $T(x)$ and possibly distinct x_1 and x_2. In the language of the last paragraph, this means that the connected components of $\sigma(\mathbf{T}(\varepsilon))$ containing λ_1 and λ_2 must coincide, since

both contain the point $\lambda_1(x_1)=\lambda_2(x_2)$. Said another way, if $\lambda_1\in\sigma_1$ and $\lambda_2\in\sigma_2$, then $\mathrm{diss}(\sigma_1, \sigma_2, \mathit{region})<\varepsilon$, because λ_1 and λ_2 belong to the same region cluster of $\mathbf{T}(\varepsilon)$. For path clustering we make a more stringent requirement on $\lambda(x)$, that $\lambda_1(x_0)=\lambda_2(x_0)$ for a particular value of x_0 and some path $T(x)$. In other words, $\lambda_1(x)$ and $\lambda_2(x)$ must be able to achieve the same value simultaneously. Now let $\bar{\varepsilon}$ be the supremum of the set of ε such that for all paths $T(x)$ in $\mathbf{T}(\varepsilon)$, $\lambda_1(x)$ never equals $\lambda_2(x)$ for any $\lambda_1(0)=\lambda_1\in\sigma_1$ and any $\lambda_2(0)=\lambda_2\in\sigma_2$. This $\bar{\varepsilon}$ is the definition of $\mathrm{diss}(\sigma_1, \sigma_2, \mathit{path})$ (note the implicit dependence on the norm $|||\cdot|||$).

Why have we bothered to draw this distinction between $\mathrm{diss}(\sigma_1, \sigma_2, \mathit{region})$ and $\mathrm{diss}(\sigma_1, \sigma_2, \mathit{path})$? The example in the next paragraph demonstrates that the two notions of dissociation can indeed differ, provided we are allowed to choose a matrix norm $|||\cdot|||$ other than $||\cdot||$ and $||\cdot||_F$. We do not know if $\mathrm{diss}(\sigma_1, \sigma_2, \mathit{region})$ and $\mathrm{diss}(\sigma_1, \sigma_2, \mathit{path})$ can differ if $|||\cdot|||$ is one of $||\cdot||$ or $||\cdot||_F$; this is an interesting open question.

For our example, we choose

$$T_0 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

and a norm $|||\cdot|||$ whose unit ball is a very narrow ellipsoid pointing in the

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

direction. For example, we may take

$$|||T|||^2 = B(|T_{12}|^2 + |T_{21}|^2 + |T_{11}-T_{22}|^2) + \frac{1}{4}|T_{11}+T_{22}|^2$$

where $B\gg 1$. The idea is that the unit ball in the $|||\cdot|||$ norm contains only matrices close to a multiple of the identity, so that points in $\mathbf{T}(\varepsilon)$ look like

$$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

plus an ε or smaller multiple of the identity. For ε near 1, we get essentially a pencil of matrices with eigenvalues $\{1+x\ ,\ -1+x\}$, $|x| \leq 1$. Thus, the region corresponding to $\lambda_1=1$ is a small neighborhood of the disk of radius 1 centered at 1, and the region corresponding to $\lambda_2=-1$ is a neighborhood of the disk of radius one centered at -1. These regions overlap at the origin and so $\mathrm{diss}(\sigma_1\ ,\ \sigma_2\ ,\ \textit{region})\leq 1$. However, λ_1 and λ_2 can not be made equal by perturbations of this size; indeed $\lambda_1-\lambda_2$ remains close to 2 until ε gets close to B. Therefore $\mathrm{diss}(\sigma_1\ ,\ \sigma_2\ ,\ \textit{path})$ can be arbitrarily larger than $\mathrm{diss}(\sigma_1\ ,\ \sigma_2\ ,\ \textit{region})$ if we are allowed to choose $|||\cdot|||$ other than $||\cdot||$ and $||\cdot||_E$.

It is true for any norm $|||\cdot|||$, however, that

$$\mathrm{diss}(\sigma_1\ ,\ \sigma_2\ ,\ \textit{path}) \geq \mathrm{diss}(\sigma_1\ ,\ \sigma_2\ ,\ \textit{region})\ ,$$

simply because if two distinct eigenvalues can be made equal by a perturbation of size $\mathrm{diss}(\sigma_1\ ,\ \sigma_2\ ,\ \textit{path})$, then no Jordan curve can be drawn separating the regions of the plane in which they lie.

It remains to show why being able to compute $\mathrm{diss}(\sigma_1\ ,\ \sigma_2\ ,\ \textit{path})$ (or $\mathrm{diss}(\sigma_1\ ,\ \sigma_2\ ,\ \textit{region})$) is sufficient to cluster the spectrum completely. By clustering the spectrum completely, we mean finding a partition $\Sigma = \{\sigma_1, \ldots, \sigma_b\}$ or T_0's spectrum such that

Region Clustering:

Σ is the finest partition for which we can find Jordan curves J_i surrounding disjoint regions of the complex plane containing $\sigma_i(\mathbf{T}(\varepsilon))$ ($\sigma_i(\mathbf{T}(\varepsilon))$ is the component of $\sigma(\mathbf{T}(\varepsilon))$ containing σ_i).

Path Clustering:

Σ is the finest partitioning for which no two distinct eigenvalues $\lambda_i \in \sigma_i$

and $\lambda_j \in \sigma_j$ can be continuously transformed to a common value $\bar{\lambda}$ along some path $T(x)$ in $\mathbf{T}(\varepsilon)$.

Σ is well defined (in both cases) because of the following property: if Σ_1 and Σ_2 are any two partitions satisfying the stated criterion (for paths or regions), then the partition $\Sigma_1 \cap \Sigma_2$ (the coarsest common refinement) also satisfies the criterion. Thus, Σ may be uniquely defined as the intersection of all partitions satisfying the criterion. (The set of all partitions satisfying the criterion is never empty, since it always contains the trivial partition $\Sigma=\{\sigma\}$.) Now note that diss(σ_1 , σ_2 , *region*) (or diss(σ_1 , σ_2 , *path*)) is sufficient to determine if $\Sigma=\{\sigma_1 , \sigma_2\}$ satisfies the criterion.

2.3 Schur Canonical Form

Throughout this thesis we will ask questions of the form:

"What matrix T possessing property P *minimizes* $\| T-T_0 \|$?"

where property P depends only on the Jordan canonical forms of T_0 and T. It will be useful to know what transformations we may perform on T_0 that either do not change this minimum distance, or change it in an easily measureable way. We will use two distance measures, the 2-norm $\| \cdot \|$ and the Euclidean (or Frobenius) norm $\| \cdot \|_E$, which we now define.

Let

$$\| x \| = (\sum_{i=1}^{n} | x_i |^2)^{1/2}$$

denote the Euclidean length of the n-vector x. Then $\| T \|$ is defined as

$$\| T \| = \sup_{x \neq 0} \frac{\| Tx \|}{\| x \|} .$$

$\| T \|_E$ is defined as

$$\| T \|_E = (\sum_{i,j} | T_{ij} |^2)^{1/2} .$$

Another definition of $\|\cdot\|_E$ we will need later is the following:

$$\|T\|_E \equiv (\mathrm{tr}(TT^*))^{1/2} \ . \tag{2.1}$$

This expresses $\|\cdot\|_E$ as the norm induced by the inner product

$$\langle A,B\rangle \equiv \mathrm{tr}(AB^*) \tag{2.2}$$

These norms have the following well known properties [Wilkinson2]:

$$\begin{aligned} &\|A\| \le \|A\|_E \le \sqrt{n}\ \|A\| \\ &\|AB\| \le \|A\|\ \|B\| \\ &\|AB\|_E \le \|A\|\ \|B\|_E \\ &\|AB\|_E \le \|A\|_E\ \|B\| \end{aligned} \ . \tag{2.3}$$

These last inequalities immediately imply

$$\mathrm{diss}_2(\sigma_1\,,\,\sigma_2) \le \mathrm{diss}_E(\sigma_1\,,\,\sigma_2) \le \sqrt{n}\ \mathrm{diss}_2(\sigma_1\,,\,\sigma_2) \ . \tag{2.4}$$

We also define the condition number of the nonsingular matrix S as

$$\kappa(S) \equiv \|S\|\ \|S^{-1}\| \ .$$

We may now ask how much our answer to our minimum distance question changes when we change T_0 to ST_0S^{-1}:

Lemma 2.1: Let $|||\cdot|||$ denote either $\|\cdot\|$ or $\|\cdot\|_E$. If $\delta = \inf_T |||T-T_0|||$, where the infimum is over matrices T possessing property P, then if S is nonsingular $\delta_S = \inf_T |||T-ST_0S^{-1}|||$ satisfies

$$\frac{\delta}{\kappa(S)} \le \delta_S \le \delta \cdot \kappa(S) \ . \tag{2.5}$$

Proof: Since property P depends only on the Jordan canonical form, T' has property P *if and only if* $ST'S^{-1}$ does as well. From (2.3) we see

$$\frac{|||T'-T_0|||}{\kappa(S)} \le |||ST'S^{-1}-ST_0S^{-1}||| \le |||T'-T_0|||\cdot\kappa(S)$$

whence follows (2.5). Q.E.D.

In other words, transforming T_0 to ST_0S^{-1} can only change the minimum distance by a factor of at most $\kappa(S)$. We will exploit this property

systematically throughout this thesis.

In particular, if $\kappa(S)=1$, the distance cannot change at all. It is well known that $\kappa(S)=1$ if and only if S is a scalar multiple of a unitary matrix. Schur's lemma, which we quote below, tells us that unitary transformations are enough to put any matrix into upper triangular form:

Lemma 2.2 (Schur's Lemma): Given any n by n complex matrix T there is a unitary matrix Q such that $QTQ^{-1}=U$ is upper triangular. Furthermore, Q may be chosen so that the eigenvalues appear on the diagonal of U in any prescribed order.

Proof: See [Isaacson].

These two lemmas tell us that we may assume without loss of generality that our original matrix T_0 is of the form

$$T_0 = \begin{bmatrix} A & C \\ & B \end{bmatrix}, \tag{2.6}$$

where $\sigma(A)=\sigma_1$ and $\sigma(B)=\sigma_2$. We will occasionally have need of a related unitary canonical form where A is upper triangular but B is lower triangular. This form is obtained from (2.6) by reversing the order of the last $\dim(B)$ rows and columns of T_0.

2.4 sep(A,B) and Projections

Projections and their norms have been used throughout the literature as measures of the sensitivity of an eigenvalue to perturbations [Schwarz, Kato2, Kahan1, Ruhe1, Wilkinson2, Wilkinson3], so it should be no surprise that we use them here, too.

There are several equivalent ways to define the projection P associated with σ_1. We will use the following two:

As described in the last section, we assume T_0 is in the form (2.6), where $\sigma(A)=\sigma_1$, $\sigma(B)=\sigma_2$, and $\sigma_1 \cap \sigma_2 = \phi$. We also assume B is lower triangular. Let $n_A=\dim(A)$ and $n_B=\dim(B)$. Now consider the system of linear equations

$$AR - RB = C \tag{2.7}$$

which we want to solve for R. It is easy to verify that if we renumber the entries of the n_A by n_B matrices R and C so that the first column is numbered 1 to n_A from top to bottom, the second column from n_A+1 to $2n_A$, and so on, then equation (2.7) can be rewritten as [Varah]

$$(A \otimes I - I \otimes B^T)R' \equiv \Psi_{A,B} R' = C' \ . \tag{2.8}$$

$\otimes$ denotes the Kronecker product, and R' and C' are the reordered versions of R and C (they are $n_A \cdot n_B$ dimensional column vectors). The matrix $\Psi_{A,B}$ is a square $n_A \cdot n_B$ dimensional upper triangular matrix with diagonal entries $\lambda_i(A)-\lambda_j(B)$. In other words, A and B have a common eigenvalue if and only if $\Psi_{A,B}$ is singular, a case we rule out by insisting $\sigma_1 \cap \sigma_2 = \phi$. For example, if $B=\{B_{ij}\}$ is 3 by 3, then

$$\Psi_{A,B} = \begin{bmatrix} A-B_{11}\cdot I & -B_{21}\cdot I & -B_{31}\cdot I \\ & A-B_{22}\cdot I & -B_{32}\cdot I \\ & & A-B_{33}\cdot I \end{bmatrix} . \tag{2.9}$$

Thus, we may solve (2.7) for R given any C. Now observe that

$$P \equiv \begin{bmatrix} 1 & R \\ 0 & 0 \end{bmatrix} = P^2 \tag{2.10}$$

so that P is a projection. Since

$$PT = TP = \begin{bmatrix} A & AR \\ 0 & 0 \end{bmatrix} \tag{2.11}$$

P projects onto the invariant subspace belonging to σ_1. Note that $\| P \|^2 = 1 + \| R \|^2 = \| I-P \|^2$.

Now define S as

$$S \equiv \begin{bmatrix} I & -R \\ & I \end{bmatrix} = [S_1 | S_2] \ , \tag{2.12}$$

where S_1 consists of the first n_A columns of S ($[I | 0]^T$) and S_2 consists of the remaining n_B columns ($[-R^T | I]^T$). Then it is easy to verify that

$$S^{-1} T_0 S = \begin{bmatrix} A & \\ & B \end{bmatrix} \ . \tag{2.13}$$

In fact, any S' which diagonalizes T_0 as in (2.13):

$$S'^{-1} T_0 S' = \begin{bmatrix} A' & \\ & B' \end{bmatrix} \tag{2.14}$$

with $\sigma(A')=\sigma_1$ and $\sigma(B')=\sigma_2$ is easily shown to be of the form

$$S' = S \cdot \begin{bmatrix} D_1 & \\ & D_2 \end{bmatrix} \tag{2.15}$$

where D_1 and D_2 are conforming nonsingular matrices. Conversely, any S' of the form (2.15) satisfies (2.14).

Given S as in (2.12), we express its inverse as

$$S^{-1} \equiv \begin{bmatrix} (S^{-1})^{(1)} \\ (S^{-1})^{(2)} \end{bmatrix} \ , \tag{2.16}$$

where $(S^{-1})^{(i)}$ contains as many rows as S_i contains columns. Then P may be written

$$P = S_1 \cdot (S^{-1})^{(1)} \ ; \tag{2.17}$$

This is equivalent to (2.10).

These facts will be important in chapter 3, where we exhibit D_1 and D_2 which minimize the condition number of S'.

A scaling property of $\| P \|$ analogous to the scaling property in lemma 2.1 follows from this last expression for P:

Lemma 2.3: If P is the projector of T_0 corresponding to σ_1, and P' is the projector of UT_0U^{-1} also corresponding to σ_1, then

$$\frac{\|P'\|}{\kappa(U)} \le \|P\| \le \kappa(U) \cdot \|P'\| \quad . \tag{2.18}$$

Proof: If $P = S_1 \cdot (S^{-1})^{(1)}$, then $P' = US_1 \cdot ((S^{-1})^{(1)} U^{-1}) = UPU^{-1}$. The result follows by taking norms. Q.E.D.

Now that we have defined the projection P, we turn our attention to sep(A,B). Following Stewart[Stewart], we define

Definition 2.4 The separation of two matrices A and B is denoted sep(A,B) and equals

$$\text{sep}(A,B) \equiv \inf_{R \neq 0} \frac{\|AR - RB\|_E}{\|R\|_E} \tag{2.19.a}$$

$$\equiv (\|\Psi_{A,B}^{-1}\|)^{-1} \tag{2.19.b}$$

$$\equiv \text{the smallest singular value of } \Psi_{A,B} \tag{2.19.c}$$

$$\equiv \text{the distance (measured either with } \|\cdot\| \text{ or } \|\cdot\|_E) \tag{2.19.d}$$

$$\text{from } \Psi_{A,B} \text{ to the nearest singular matrix} \quad .$$

If A and B are clear from context, we will abbreviate sep(A,B) by sep.

sep(A,B) has several important properties which we now enumerate. First of all sep(A,B)=sep(B,A); this is because $\Psi_{A,B}$ can be obtained from $-\Psi_{B,A}$ by simply reordering the rows and columns. More importantly we have

Lemma 2.5: sep(A,B), $\|P\|$ and $\|C\|_E$ satisfy the following inequalities:

$$\|P\|^2 \le 1 + \frac{\|C\|_E^2}{\text{sep}^2} \le 1 + \frac{\|T_0\|_E^2}{\text{sep}^2} \tag{2.20}$$

$$\|P\| \le 1 + \frac{\|C\|_E}{\text{sep}} \le 1 + \frac{\|T_0\|_E}{\text{sep}} \tag{2.21}$$

Proof: From (2.8) follows $\|R\|_E \le \|C\|_E / \text{sep}$. Since $\|P\|^2 = 1 + \|R\|^2$, (2.20) follows. (2.21) is simply a coarsening of (2.20). Q.E.D.

Thus, sep(A,B) (with $\| T_0 \|_E$) provides an upper bound on the norm of the projection associated with either A or B, and conversely, $\| P \|$ provides a lower bound for sep(A,B).

The next property shows that sep satisfies a property analogous to Lemma 2.1. We could hardly expect to be able to use sep to help measure distances (and dissociation) if it did not behave the same way under transformations as do distances.

Lemma 2.6 ([Stewart]):

$$\frac{\mathrm{sep}(A,B)}{\kappa(S_A)\cdot\kappa(S_B)} \le \mathrm{sep}(S_A A S_A^{-1}, S_B B S_B^{-1}) \le \mathrm{sep}(A,B)\cdot\kappa(S_A)\cdot\kappa(S_B) \quad . \tag{2.22}$$

Proof:

$$\begin{aligned}
\mathrm{sep}(S_A A S_A^{-1}, S_B B S_B^{-1}) &= \inf_{R\neq 0} \frac{\| S_A A S_A^{-1} R - R S_B B S_B^{-1} \|_E}{\| R \|_E} \\
&= \inf_{R\neq 0} \frac{\| S_A (A(S_A^{-1} R S_B) - (S_A^{-1} R S_B) B) S_B^{-1} \|_E}{\| R \|_E} \\
&\le \| S_A \| \cdot \| S_B^{-1} \| \cdot \inf_{R\neq 0} \frac{\| A(S_A^{-1} R S_B) - (S_A^{-1} R S_B) B \|_E}{\| R \|_E} \\
\le \| S_A \| \cdot \| S_A^{-1} \| \cdot \| S_B \| \cdot \| S_B^{-1} \| \cdot \inf_{R\neq 0} &\frac{\| A(S_A^{-1} R S_B) - (S_A^{-1} R S_B) B \|_E}{\| S_A^{-1} R S_B \|_E} \\
&= \kappa(S_A)\cdot\kappa(S_B)\cdot\mathrm{sep}(A,B) \quad .
\end{aligned}$$

This proves the second inequality of (2.22). The first inequality follows by symmetry. Q.E.D.

The next lemma shows that we may apply the "divide and conquer" paradigm to computing sep(A,B) when either A or B is block diagonal. We write $A = \bigoplus_i A_i$ when A is block diagonal with blocks A_i.

Lemma 2.7 [Stewart]:

$$\mathrm{sep}(\bigoplus_i A_i\,,\,\bigoplus_j B_j) = \min_{i,j} \mathrm{sep}(A_i,B_j) \tag{2.23}$$

Proof: It suffices to show $\mathrm{sep}(A,B_1\oplus B_2) = \min_i \mathrm{sep}(A,B_i)$. From (2.9) we see that if B is block diagonal, so is $\Psi_{A,B}$ with diagonal blocks Ψ_{A,B_i}. Since the singular values of any block diagonal matrix are the singular values of the blocks, the results follows directly from (2.19.c) in Definition 2.4. Q.E.D.

Lemmas 2.6 and 2.7 together tell us that if A and B can both be completely diagonalized by not too ill-conditioned similarities S_A and S_B (that is, neither $\kappa(S_A)$ nor $\kappa(S_B)$ is very large), then $\mathrm{sep}(A,B)$ differs from $\min_{i,j}|\lambda_i(A) - \lambda_j(B)|$ by the not too large factor $\kappa(S_A)\cdot\kappa(S_B)$. The expression $\min_{i,j}|\lambda_i(A) - \lambda_j(B)|$, the smallest difference between an eigenvalue of A and an eigenvalue of B, is the coarsest possible measure of the dissociation between $\sigma(A)$ and $\sigma(B)$. We record this fact as

Lemma 2.8: If $S_A^{-1}AS_A=\mathrm{diag}(\lambda_i(A))$ and $S_B^{-1}BS_B=\mathrm{diag}(\lambda_j(B))$, then

$$\min_{ij} |\lambda_i(A) - \lambda_j(B)| \ge \mathrm{sep}(A,B) \ge \frac{\min_{ij} |\lambda_i(A) - \lambda_j(B)|}{\kappa(S_A)\cdot\kappa(S_B)}$$

Proof: Combine Lemmas 2.0, 2.7 to obtain the lower bound, and Definition 2..d to obtain the upper bound. Q.E.D.

We will show in chapter 8 that the likely situation is that A and B are diagonalizable with reasonably well conditioned similarities so that $\min_{i,j}|\lambda_i(A) - \lambda_j(B)|$ is a reasonably accurate estimate of $\mathrm{sep}(A,B)$ and $\mathrm{diss}_2(\sigma_1\,,\,\sigma_2)$.

We present one more application of the divide and conquer paradigm used in Lemma 2.8: if A and B are block diagonal, the computation of R where $AR - RB = C$ can also be broken into smaller parts. We illustrate when both A and B have two diagonal blocks. The system of equations

$$AR - RB = \begin{bmatrix} A_1 & \\ & A_2 \end{bmatrix} \cdot \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} - \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \cdot \begin{bmatrix} B_1 & \\ & B_2 \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \equiv C$$

(where all blocks are of conforming sizes) breaks into four independent systems:

$$A_i R_{ij} - R_{ij} B_j = C_{ij} \quad . \tag{2.24}$$

If A or B has more than two blocks, equation (2.24) still applies.

2.5 $\mathrm{sep}_\lambda(A,B)$

Another measure of the separation of two matrices is $\mathrm{sep}_\lambda(A,B)$, the smallest perturbation to A and B which causes them to have a common eigenvalue. $\mathrm{sep}_\lambda(A,B)$ will be our upper bound on the dissociation between σ_1 and σ_2, and chapters 5 and 8 of this thesis will analyze how much sep and sep_λ may differ.

Modifying a definition of Varah [Varah] slightly, we define

Definition 2.9:

$$\mathrm{sep}_\lambda(A,B) \equiv \inf_\lambda \max(\|(A-\lambda I)^{-1}\|^{-1}, \|(B-\lambda I)^{-1}\|^{-1}) \tag{2.25.a}$$

$$= \inf_\lambda \max(\sigma_{\min}(A-\lambda I), \sigma_{\min}(B-\lambda I)) \tag{2.25.b}$$

where $\sigma_{\min}$ denotes the smallest singular value.

Varah defines sep_λ as the sum of the two singular values rather than the maximum, so his sep_λ cannot differ from ours by more than a factor of 2. We have modified his definition because it lets us state slightly sharper results later on.

$\mathrm{sep}_\lambda(A,B)$ is clearly an upper bound on $\mathrm{diss}_2(\sigma_1, \sigma_2, path)$. Of course, there might be a general perturbation of T_0 (not just one in A and B) of much smaller norm than $\mathrm{sep}_\lambda(A,B)$ that makes an eigenvalue of A coalesce with one of B, so in general $\mathrm{sep}_\lambda(A,B)$ only provides an upper bound on $\mathrm{diss}_2(\sigma_1, \sigma_2, path)$. Since chapter 8 contains the rather surprising result

that sep_λ provides a not too pessimistic overestimate of this dissociation, we record this fact in the following theorem.

Theorem 2.10:

$$\mathrm{sep}_\lambda(A,B) \geq \mathrm{diss}_2(\sigma_1\,,\,\sigma_2\,,\,path)$$

$$\sqrt{2}\,\mathrm{sep}_\lambda(A,B) \geq \mathrm{diss}_F(\sigma_1\,,\,\sigma_2\,,\,path)$$

The rest of this section proves several properties of sep_λ that we will need later. In particular, we will prove two lemmas analogous to lemmas 2.6 and 2.7. Since we are going to relate sep_λ and sep, it is important that they behave similarly under transformations (lemmas 2.6 and 2.11) and divide and conquer (lemmas 2.7 and 2.12).

First, however, we present a characterization of sep_λ analogous to Definition 2.4 of sep. The following lemma is essentially due to Varah [Varah]:

Lemma 2.11:

$$\mathrm{sep}_\lambda(A,B) \leq \inf_{\substack{R\neq 0\\ \mathrm{rank}(R)=1}} \frac{\|AR-RB\|_F}{\|R\|_F} \leq 2\,\mathrm{sep}_\lambda(A,B) \tag{2.26}$$

Proof: The infimum in definition 2.9 of sep_λ is clearly attained for some λ by compactness. Thus, there exist unit vectors u and v such that

$$\begin{aligned}\mathrm{sep}_\lambda(A,B) &= \max\left(\sigma_{\min}(A-\lambda)\,,\,\sigma_{\min}(B-\lambda)\right)\\ &= \max\left(\|Au-\lambda u\|\,,\,\|v^*B-v^*\lambda\|\right)\,.\end{aligned}$$

Let $R=uv^*$. Note that $\mathrm{rank}(R)=1$ and $\|R\| = \|R\|_F = 1$. Thus

$$\begin{aligned}\|AR-RB\|_F &= \|(A-\lambda)R - R(B-\lambda)\|_F\\ &= \|(Au-\lambda u)v^* - u(v^*B-v^*\lambda)\|_F\\ &\leq \|(Au-\lambda u)v^*\|_F + \|u(v^*B-v^*\lambda)\|_F\\ &\leq 2\,\mathrm{sep}_\lambda(A,B)\,.\end{aligned}$$

This proves the second inequality in (2.26).

To prove the other inequality, we again write R as uv^*, where u and v are unit vectors and uv^* attains the infimum in (2.26) (this is again possible by compactness). Then as before

$$\| AR-RB \|_E = \| (Au-\lambda u)v^* - u(v^*B-v^*\lambda) \|_E \ .$$

We now exploit the alternative definition of $\| \cdot \|_E$ given in (2.1) above:

$$\begin{aligned} \| AR-RB \|_E^2 &= \mathrm{tr}\{[(Au-\lambda u)v^* - u(v^*B-v^*\lambda)][(Au-\lambda u)v^* - u(v^*B-v^*\lambda)]^*\} \\ &= \mathrm{tr}[(A'-B')(A'-B')^*] \\ &= \mathrm{tr}(A'A'^*) + 2\cdot\mathrm{Re}\ \mathrm{tr}(A'B'^*) + \mathrm{tr}(B'B'^*) \\ &= \| A' \|_E^2 + 2\cdot\mathrm{Re}\ \mathrm{tr}(A'B'^*) + \| B' \|_E^2 \ . \end{aligned}$$

Now if we choose $\lambda = v^*Bv$ (or u^*Au), A' and B' are orthogonal:

$$\begin{aligned} \mathrm{tr}(A'B'^*) &= \mathrm{tr}(u(v^*B-v^*\lambda)v(Au-\lambda u)^*) \\ &= \mathrm{tr}(u(v^*Bv-\lambda)(Au-\lambda u)^*) \\ &= 0 \ . \end{aligned}$$

Thus with this choice of λ

$$\begin{aligned} \| AR-RB \|_E^2 &= \| (Au-\lambda u)v^* \|_E^2 + \| u(v^*B-v^*\lambda) \|_E^2 \\ &\geq \max\ (\| (A-\lambda)u \|_E \quad \| v^*(B-\lambda) \|_E)^2 \\ &\geq \cdot\ \mathrm{sep}_\lambda^2(A,B) \end{aligned}$$

as desired. Q.E.D.

An immediate consequence of the definition of $\mathrm{sep}(A,B)$ and this last lemma is

Lemma 2.12:

$$\mathrm{sep}(A,B) \leq 2\ \mathrm{sep}_\lambda(A,B)$$

If $\dim(A)=1$, then

$$\mathrm{sep}_\lambda(A,B) \leq \mathrm{sep}(A,B) = \sigma_{\min}(B-a_{11}\cdot I) \leq 2\ \mathrm{sep}_\lambda(A,B)$$

where $A=[a_{11}]$. An analogous inequality holds if $\dim(B)=1$.

Proof: The first inequality follows from lemma 2.11 and the first line of Definition 2.4 of sep. The second inequality holds because if $\dim(A)=1$ then the R in Definition 2.4 is either a row vector or column vector, and so necessarily of rank one. Q.E.D.

Just how much smaller sep can be than sep_λ is the subject of chapters 5 and 8.

The next lemma shows that sep_λ behaves similarly to sep under transformations.

Lemma 2.13: Let $\bar{\kappa} = \max(\kappa(S_A), \kappa(S_B))$. Then

$$\frac{\mathrm{sep}_\lambda(A,B)}{\bar{\kappa}} \le \mathrm{sep}_\lambda(S_A A S_A^{-1}, S_B B S_B^{-1}) \le \mathrm{sep}_\lambda(A,B)\cdot\bar{\kappa} \ . \tag{2.27}$$

Proof: For any λ

$$\frac{\|(A-\lambda)^{-1}\|^{-1}}{\bar{\kappa}} \le \frac{\|(A-\lambda)^{-1}\|^{-1}}{\kappa(S_A)} \tag{2.28}$$

$$\le \|(S_A A S_A^{-1} - \lambda)^{-1}\|$$

$$\le \|(A-\lambda)^{-1}\|^{-1}\cdot\kappa(S_A) \le \|(A-\lambda)^{-1}\|^{-1}\cdot\bar{\kappa} \ .$$

An analogous inequality holds for B and S_B. The lemma follows directly from these inequalities. Q.E.D.

There is an important difference between lemmas 2.6 and 2.13: while sep_λ may change by $\bar{\kappa}$ after a transformation, sep may change by as much as $\bar{\kappa}^2$. Although we do not exploit this difference further, it bolsters our conjecture of chapter 8 that $\mathrm{sep}/\|T_0\|_E$ is almost always bounded below by a constant times $(\mathrm{sep}_\lambda/\|T_0\|_E)^2$.

The next lemma shows that the same divide and conquer formula holds for sep_λ with block diagonal A and B as for sep.

Lemma 2.14:

$$\mathrm{sep}_\lambda(\bigoplus_i A_i\ ,\ \bigoplus_j B_j) = \min_{ij} \mathrm{sep}_\lambda(A_i\ ,\ B_j) \tag{2.29}$$

Proof: The singular values of a block diagonal matrix are the singular values of the blocks. Thus

$$\begin{aligned}
\mathrm{sep}_\lambda(\bigoplus_i A_i\ ,\ \bigoplus_j B_j) &= \inf_\lambda \max(\sigma_{\min}(\bigoplus_i (A_i-\lambda))\ ,\ \sigma_{\min}(\bigoplus_j (B_j-\lambda))) \\
&= \inf_\lambda \max(\min_i \sigma_{\min}(A_i-\lambda)\ ,\ \min_j \sigma_{\min}(B_j-\lambda)) \\
&= \min_{ij} \inf_\lambda \max(\sigma_{\min}(A_i-\lambda)\ ,\ \sigma_{\min}(B_j-\lambda)) \\
&= \min_{ij} \mathrm{sep}_\lambda(A_i\ ,\ B_j)
\end{aligned}$$

as desired. Q.E.D.

In analogy to Lemma 2.8 we have

Lemma 2.15: If $S_A^{-1}AS_A=\mathrm{diag}(\lambda_i(A))$ and $S_B^{-1}BS_B=\mathrm{diag}(\lambda_i(B))$, then

$$\frac{\min_{ij} |\lambda_i(A) - \lambda_j(B)|}{2} \geq \mathrm{sep}_\lambda(A,B) \geq \frac{\min_{ij} |\lambda_i(A) - \lambda_j(B)|}{2 \cdot \max(\kappa(S_A)\ ,\ \kappa(S_B))}\ .$$

Proof: The upper bound follows from the definition of sep_λ and the lower bound from the last two lemmas. Q.E.D.

Chapter 3: Best Conditioned Diagonalizing Similarities

3.1 Introduction

In this chapter we assume a partitioning $\Sigma=\{\sigma_1, \ldots, \sigma_b\}$ of σ has been chosen, and ask the following question: what is the best conditioned S such that

$$S^{-1}TS = \mathrm{diag}(\Theta_1, \ldots, \Theta_b) \tag{3.1}$$

and $\sigma(\Theta_i)=\sigma_i$? We need to use the answer as a tool in later chapters.

Actually, it is as easy to answer a more general question: how ill-conditioned must a matrix S be if its columns are constrained to span certain subspaces? We answer this question in order to find nearly best conditioned matrices S_R and S_L that block diagonalize a given matrix pencil $T = A+\lambda B$, i.e. $S_L^{-1}S_R = \Theta$ is block diagonal. We show that the best conditioned S_R has a condition number approximately equal to the cosecant of the smallest angle between right subspaces belonging to different diagonal blocks of Θ. Thus, the more nearly the right subspaces overlap the more ill-conditioned S_R must be. The same is true of S_L and the left subspaces.

For our original problem $T = A-\lambda I$, the standard eigenproblem, $S_L = S_R$ and the cosecant of the angle between subspaces turns out to be the norm $\|P\|$ of the projection associated with each subspace. More precisely, if P_i is the projection associated with σ_i, then

$$\max_i \|P_i\| \le \kappa(S_{OPTIMAL}) \le b \cdot \max_i \|P_i\| \tag{3.2}$$

where b is the number of blocks in (3.1). Furthermore, we can construct an S, denoted S_{ORTHO}, whose condition number $\kappa(S_{ORTHO})$ lies in the bounds given by (3.2): choose the rank(P_i) columns of S_{ORTHO} which span the invariant subspace belonging to σ_i to be orthonormal.

In particular, if $b=2$ so that we are dividing T into just two blocks, then we can compute an S (not S_{ORTHO}) such that

$$\kappa(S_{OPTIMAL}) = \kappa(S) = ||P|| + \sqrt{||P||^2 - 1}$$

and

$$S^{-1}TS = S^{-1}\begin{bmatrix} A & C \\ & B \end{bmatrix} S = \begin{bmatrix} A & \\ & B \end{bmatrix}.$$

We will need this construction for our lower bound on $\mathrm{diss}_2(\sigma_1, \sigma_2)$ in the next chapter.

The rest of this chapter is organized as follows. Section 3.2 defines the notions of invariant subspaces and angles between them more precisely, reviews some of the history of the problem, and summarizes the results of the chapter. Section 3.3 shows how to decompose T into $b=2$ diagonal blocks, and section 3.4 handles the case of $b \geq 3$ blocks. Section 3.5 contains the proof of a technical result needed in the proof of Theorem 3.1. Section 3.6 applies the main results to an error bound for computing a function of a matrix, such as $\exp(T)$. Section 3.7 discusses the possibility of partitioning σ using only projection norms as a criterion. Finally, section 3.8 presents some applications of the main results unrelated to eigenproblems.

Most of this chapter has been published already [Demmel], except for sections 3.7, 3.8 and part of 3.4.

3.2 Definitions and Summary of Results

Two measures of the ill-conditioning of the eigenvalues of a matrix have appeared frequently in the literature. One is the condition number of a matrix S which (block) diagonalizes T under similarity (i.e. $S^{-1}TS$ is block diagonal), and the other is the norm of the projection matrix P_i belonging to the spectrum of the i-th diagonal block of $S^{-1}TS$ (if the i-th block is 1 by 1,

the norm of P_i is usually denoted $1/|s_i|$ [Wilkinson2]). Many authors have shown that the larger the condition number of S, or the larger the norm of P_i, the more sensitive to perturbations are at least some of the eigenvalues of T. Bauer and Fike [Bauer1], Kato [Kato2], Kahan [Kahan1], Ruhe [Ruhe1], Wilkinson [Wilkinson2, Wilkinson3] and others have all contributed theorems stating this result in different ways. Recently Sun [Sun] has extended many of these results to regular matrix pencils.

Our goal in this paper is to show that these two measures of ill-conditioning are nearly equivalent. We state our result in terms of angles between subspaces because this makes sense for pencils $T=A+\lambda B$ as well as the standard eigenproblem $T=A-\lambda I$: the condition number of the best S which displays the block structure is within a small constant factor of the cosecant of the smallest angle between a subspace belonging to one diagonal block and the subspace spanned by all the other subspaces together. In the case of the standard eigenproblem this cosecant turns out equal to the largest of the norms of the projections P_i.

We exhibit a best S for decomposing T into two blocks and compute its condition number exactly in terms of the norm of a projection (see part 2 below). This result was obtained independently by Bart et. al. [Bart] and improves an earlier estimate of Kahan [Kahan1]. Wilkinson [Wilkinson2, p 89] and Bauer [Bauer4] relate the two measures when $S^{-1}TS$ is completely diagonal; we generalize their results to diagonal blocks of arbitrary sizes in theorems 3.3 and 3.3a below.

The angle between subspaces is defined as the smallest possible angle between a vector u in one subspace S^1 and a vector v in another subspace S^2:

$\vartheta(\mathbf{S}^1, \mathbf{S}^2) = \min\{\arccos |u^* v| \text{ when } u \in \mathbf{S}^1, v \in \mathbf{S}^2, \|u\| = \|v\| = 1\}$ (3.3)
(ϑ will be discussed more fully later).

If $\mathbf{S}^1, \ldots, \mathbf{S}^b$ is a collection of subspaces, the space spanned by their union is denoted $\text{span}\{\mathbf{S}^1, \ldots, \mathbf{S}^b\}$.

With this preparation, let us consider the subspaces associated with the block diagonal matrix $S_L^{-1} T S_R = \Theta = \text{diag}(\Theta_1, \ldots, \Theta_b)$, where Θ_i is r_i by c_i; r_i and c_i must be equal unless $T = A + \lambda B$ is a singular pencil [Gantmacher]. From $S_L^{-1} T S_R = \Theta$ follows $T S_R = S_L \Theta$ which implies that T maps the space $\mathbf{S}_R^i$ spanned by the first c_1 columns of S_R into a space $\mathbf{S}_L^i$ spanned by the first r_1 columns of S_L. Similarly, columns $c_1 + \cdots + c_{i-1} + 1$ to $c_1 + \cdots + c_i$ of S_R span a space $\mathbf{S}_R^i$ that T maps into a space $\mathbf{S}_L^i$ spanned by columns $r_1 + \cdots + r_{i-1} + 1$ to $r_1 + \cdots + r_i$ of S_L. Stewart [Stewart] calls the pairs $\mathbf{S}_R^i$, $\mathbf{S}_L^i$ deflating pairs since they deflate T to block diagonal form. For the standard eigenproblem $T = A - \lambda I$ we have $\mathbf{S}_R^i = \mathbf{S}_L^i$ [Gantmacher] in which case they are denoted by $\mathbf{S}^i$ and called invariant subspaces and then no generality is lost by assuming $S_R = S_L$. Henceforth we drop the subscripts R and L of S since they are unnecessary for the standard eigenvalue problem and since our results apply to each case separately for the general problem $T = A + \lambda B$.

Our problem is to choose the columns of S to minimize $\kappa(S)$ subject to the condition that the columns span the subspaces $\mathbf{S}^i$. (It is not important for the proofs of our results that the $\mathbf{S}^i$ be defined by an eigenvalue problem; we ask only that the $\mathbf{S}^i$ be linearly independent and together span all of euclidean space. Thus our results may be interpreted as results on one-sided block diagonal scaling of matrices.) Our first result will be that by choosing the columns spanning each subspace to be orthonormal, we will have an S whose condition number is within a factor $\sqrt{b}$ of optimal, where b is the number of diagonal blocks of Θ:

$$\kappa(S_{ORTHO}) \le \sqrt{b}\ \kappa(S_{OPTIMAL}) \ . \tag{3.4}$$

S_{ORTHO} denotes any matrix S whose columns are orthonormal in groups as described above, and $S_{OPTIMAL}$ denotes any matrix S whose condition number is as small as possible. This extends a result of Van der Sluis [vanderSluis] where all subspaces $\mathbf{S}^i$ are one-dimensional. Van Dooren and Dewilde [Van-Dooren] have also shown the choice of S_{ORTHO} is nearly best, and in fact optimal if the subspaces $\mathbf{S}^i$ are orthogonal.

Furthermore, we shall bound $\kappa(S_{ORTHO})$ above and below in terms of the angles between the subspaces $\mathbf{S}^i$ spanned by its columns. Let ϑ_i denote the smallest angle between $\mathbf{S}^i$ and the subspace spanned by all the other subspaces together:

$$\vartheta_i = \vartheta(\mathbf{S}^i\ ,\ \operatorname*{span}_{j \ne i}\{\mathbf{S}^j\}) \ . \tag{3.5}$$

We shall show

$$\max_i\,(\csc\vartheta_i + \sqrt{\csc^2\vartheta_i - 1}) \le \kappa(S_{OPTIMAL}) \tag{3.6}$$

$$\le \kappa(S_{ORTHO}) \le \sqrt{b}\ \sqrt{\sum_{i=1}^{b}\csc^2\vartheta_i}$$

When $b=2$ (i.e. we have only 2 diagonal blocks) S_{ORTHO} is in fact optimal, and

$$\kappa(S_{ORTHO}) = \kappa(S_{OPTIMAL}) = \csc\vartheta + \sqrt{\csc^2\vartheta - 1} = \cot\ \vartheta/2 \ . \tag{3.7}$$

$S_{OPTIMAL}$ is not unique, and we compute another S for the $b=2$ case which has the optimal condition number and which further satisfies

$$S^{-1}\begin{bmatrix} A & C \\ & B \end{bmatrix} S = \begin{bmatrix} A & \\ & B \end{bmatrix} \qquad ()$$

in the case of the standard eigenproblem where $\mathbf{S}^1$ is the invariant subspace belong to A, and $\mathbf{S}^2$ the invariant subspace for B.

For the standard eigenproblem we also have $\csc\vartheta_i = \|P_i\|$, where P_i is the projection associated with subspace i. It follows from (3.6) that the two

measures of ill-conditioning $\kappa(S_{OPTIMAL})$ and $\max_i \| P_i \|$ we wanted to show nearly equivalent can differ by no more than a constant factor:

$$\max_i \| P_i \| \le \kappa(S_{OPTIMAL}) \le b \cdot \max_i \| P_i \| \quad . \tag{3.8}$$

3.3 How to Decompose T into 2 blocks

In this section we show that the best conditioned S whose first c columns span a given subspace $\mathbf{S}^1$ and whose remaining $n-c$ columns span another given complementary subspace $\mathbf{S}^2$ has condition number

$$\kappa(S_{OPTIMAL}) = \csc \vartheta + \sqrt{\csc^2 \vartheta - 1} = \cot \vartheta/2 \tag{3.9}$$

where $\vartheta = \vartheta(\mathbf{S}^1, \mathbf{S}^2)$. Note that we assume $\mathbf{S}^1$ and $\mathbf{S}^2$ are linearly independent, for otherwise S would be singular.

To prove (3.9) we will need a technical result, Theorem 3.1, that bounds the norms of submatrices of a positive definite matrix in terms of its condition number. Theorem 3.1 is a slight generalization of an inequality of Wielandt [Bauer2] and the proof technique used here yields several other inequalities (Theorem 3.4) one of which (3.55) is an inequality of Bauer [Bauer3].

Let

$$H = \begin{bmatrix} A & B \\ B^* & C \end{bmatrix}$$

be a Hermitian positive definite matrix, partitioned so that A is n by n, B is n by m, and C is m by m. Let $\kappa = \| H \| \; \| H^{-1} \|$ be the condition number of H. Let $X^{-1/2}$ denote any matrix such that $X^{-1/2}(X^{-1/2})^* = X^{-1}$.

Theorem 3.1: If H and κ are defined as above, then

$$\| (A^{-1/2})^* B C^{-1/2} \| \le \frac{\kappa - 1}{\kappa + 1} \tag{3.10}$$

or, equivalently,

$$\kappa \geq \frac{1 + \|(A^{-1/2})^* BC^{-1/2}\|}{1 - \|(A^{-1/2})^* BC^{-1/2}\|} . \tag{3.11}$$

Furthermore, this bound is sharp. In fact, given any n by m matrix Z such that $\|Z\| < 1$, both sides of inequality (3.10) are equal for the matrix

$$H = \begin{bmatrix} I & Z \\ Z^* & I \end{bmatrix} .$$

This theorem will be proved in Part 3.5.

We also need another definition of the (smallest) angle ϑ between subspaces that is more useful than the one stated in the introduction. As stated there, ϑ is the smallest possible angle between a vector in one subspace and a vector in the other subspace (the largest possible angle may be much larger than the smallest if the subspaces are not one dimensional). If S_1 is an n by c matrix of orthonormal columns which form a basis of $\mathbf{S}^1$ and S_2 is an n by $n-c$ orthonormal basis of the second space $\mathbf{S}^2$, then ϑ may also be expressed as [Davis]

$$\vartheta(\mathbf{S}^1, \mathbf{S}^2) = \arccos \|S^*_1 S_2\| = \arccos \sup_{x,y} |y^* S^*_1 S_2 x| \tag{3.12}$$

$$= \inf_{u,v} \arccos |u^* v|$$

where the sup is over arbitrary unit vectors x and y, and where the inf is over unit vectors u in $\mathbf{S}^1$ and v in $\mathbf{S}^2$.

Now consider a candidate matrix S:

$$S_{ORTHO} = [S_1 \mid S_2] \tag{3.13}$$

where S_1 and S_2 are orthonormal bases of $\mathbf{S}^1$ and $\mathbf{S}^2$ respectively. We may describe every other possible S whose columns span $\mathbf{S}^1$ and $\mathbf{S}^2$ in terms of S_{ORTHO}:

$$S_D = S_{ORTHO}\, D = S_{ORTHO}\, \mathrm{diag}(D_1, D_2) = [S_1 D_1 \mid S_2 D_2] , \tag{3.14}$$

where D_1 is a nonsingular c by c matrix and D_2 is a nonsingular $n-c$ by $n-c$ matrix. (3.14) states simply that any basis of $\mathbf{S}^1$ can be expressed as a

nonsingular linear combination $S_i D_i$ of the columns of one basis S_i. We want to know which D minimizes $\kappa(S_D)$. We compute

$$\kappa^2(S_D) = \kappa(S_D{}^* S_D) \tag{3.15}$$
$$= \kappa \begin{bmatrix} D_1{}^* D_1 & D_1{}^* S_1{}^* S_2 D_2 \\ D_2{}^* S_2{}^* S_1 D_1 & D_2{}^* D_2 \end{bmatrix} .$$

We may now invoke Theorem 3.1 with $A^{-1/2} = D_1^{-1}$, $B = D_1{}^* S_1{}^* S_2 D_2$, and $C^{-1/2} = D_2^{-1}$ to find

$$\kappa^2(S_D) \geq \frac{1 + \| S_1{}^* S_2 \|}{1 - \| S_1{}^* S_2 \|}$$
$$= \frac{1 + \cos \vartheta}{1 - \cos \vartheta} \qquad (0 < \vartheta \leq \pi/2)$$
$$= \cot^2 (\vartheta/2)$$

or

$$\kappa(S_D) \geq \cot \vartheta/2 \ . \tag{3.16}$$

If D_1 and D_2 are unitary, it is easy to verify that we have equality in (3.16), proving (3.9) with $S_{OPTIMAL} = S_{ORTHO}$ as desired. Note, however, that $S_{OPTIMAL}$ is far from unique, since there are many orthonormal bases for a given space.

It remains to show $\csc \vartheta = \| P \|$ where P is the projection onto $\mathbf{S}^1$ parallel to $\mathbf{S}^2$. Recall from section (2.1) that if we assume (without loss of generality) that T is of the form (2.4?)

$$T = \begin{bmatrix} A & C \\ & B \end{bmatrix} = \begin{bmatrix} A & AR - RB \\ & B \end{bmatrix} ,$$

then any S which block diagonalizes T is of the form (2.10? and 2.12?)

$$S = \begin{bmatrix} I & -R \\ & I \end{bmatrix} \cdot \begin{bmatrix} D_1 & \\ & D_2 \end{bmatrix} .$$

Also, the P which projects onto $\mathbf{S}^1$ parallel to $\mathbf{S}^2$ is (2.8?)

$$P = \begin{bmatrix} I & R \\ 0 & 0 \end{bmatrix},$$

where $\|P\|^2 = 1 + \|R\|^2$.

By choosing

$$D_1 = I \quad \text{and} \quad D_2 = (I + R^*R)^{-1/2} \tag{3.17}$$

where D_2 can be any matrix such that $D_2D_2^* = (I + R^*R)^{-1}$ we obtain an

$$S = \begin{bmatrix} I & R \\ & I \end{bmatrix} \begin{bmatrix} D_1 & \\ & D_2 \end{bmatrix} = \begin{bmatrix} I & R(I + R^*R)^{-1/2} \\ & (I + R^*R)^{-1/2} \end{bmatrix}$$

$$= [S_1 \mid S_2]$$

where S_1 and S_2 contain orthonormal vectors.

Thus

$$\vartheta = \arccos \|S_1^*S_2\| \qquad (0 < \vartheta \le \pi/2) \tag{3.18}$$

$$= \arccos \|R(I + R^*R)^{-1/2}\|$$

so

$$\csc \vartheta = (1 - \cos^2 \vartheta)^{-1/2} \tag{3.19}$$

$$= (1 - \|R(I + R^*R)^{-1/2}\|^2)^{-1/2}$$

$$= (1 - \|(I + R^*R)^{-1/2}R^*R(I + R^*R)^{-1/2}\|)^{-1/2}$$

(since $\|H\|^2 = \|H^*H\|$ for any matrix H)

$$= (1 - \|R^*R(I + R^*R)^{-1}\|)^{-1/2}$$

$$= \left(1 - \frac{\|R^*R\|}{1 + \|R^*R\|}\right)^{-1/2}$$

$$= (1 + \|R^*R\|)^{1/2}$$

$$= \sqrt{1 + \|R\|^2}$$

$$= \|P\|$$

as desired.

It is possible to choose D_1 and D_2 which are multiples of the identity and also achieve the minimum condition number. We record this fact here because we will need it in Chapter 4. Choose

$$D_1 = I \quad \text{and} \quad D_2 = (1 + \|R\|^2)^{-1/2} \cdot I \ . \tag{3.20}$$

We will show that with this choice of D_1 and D_2 $\kappa(S)$ attains its minimum value. First pick unitary Q_1 and Q_2 so that $Q_1 R Q_2^T = \mathrm{diag}(r_i)$ is diagonal (i.e. the $\|R\| = r_1 \geq \cdots \geq r_n$ are the singular values.) Then

$$S' = \begin{bmatrix} Q_1 & \\ & Q_2 \end{bmatrix} \cdot S \cdot \begin{bmatrix} Q_1^T & \\ & Q_2^T \end{bmatrix} = \begin{bmatrix} Q_1 & \\ & Q_2 \end{bmatrix} \cdot \begin{bmatrix} I & -R \\ & I \end{bmatrix} \cdot \begin{bmatrix} D_1 & \\ & D_2 \end{bmatrix} \cdot \begin{bmatrix} Q_1^T & \\ & Q_2^T \end{bmatrix}$$

$$= \begin{bmatrix} I & \mathrm{diag}\left(\dfrac{r_i}{\sqrt{(1+r_i^2)}}\right) \\ & I/\sqrt{(1+r_i^2)} \end{bmatrix}$$

has the same condition number as S, and is the direct sum of 2 by 2 blocks and 1 by 1 blocks. It is a simple matter to compute the largest and smallest singular values of this new matrix, and to show in particular that

$$\|S\|^2 = \frac{\|R\| + \sqrt{\|R\|^2 + 1}}{\sqrt{\|R\|^2 + 1}}$$

and

$$\|S^{-1}\|^2 = \sqrt{\|R\|^2 + 1} \cdot (\|R\| + \sqrt{\|R\|^2 + 1}) \ .$$

The results follows from multiplying these two expressions to get $\kappa(S)$.

3.4 How to Decompose T into b Blocks when $b > 2$

In this section we first consider partitioned matrices

$$S = [S_1 \mid \cdots \mid S_b] \tag{3.21}$$

where each submatrix S_i must span a given subspace $\mathbf{S}^i$ and show that S is nearly best conditioned when each S_i's columns are orthonormal. Next we bound the condition number of the best such S above and below in terms of $\max_i \csc \vartheta_i$, where

$$\vartheta_i = \vartheta(\mathbf{S}^i \,,\, \underset{j \neq i}{\mathrm{span}}\{\mathbf{S}^j\}) \quad . \tag{3.22}$$

Finally we will discuss a different choice of S (also discussed in the literature [Smith,Wilkinson2]) which is harder to compute and has slightly different bounds on its condition number.

Theorem 3.2: Let S be

$$S = [S_1 \mid \cdots \mid S_b] \tag{3.23}$$

where S_i contains c_i columns.

If we choose the columns constituting S_i to be any orthonormal basis of the subspace $\mathbf{S}^i$, then S will have a condition number no larger than $\sqrt{b}$ times the smallest possible:

$$\kappa(S) \leq \sqrt{b} \cdot \kappa(S_{OPTIMAL}) \quad . \tag{3.24}$$

Said another way, choose S so that S^*S has identity matrices (of sizes c_i by c_i) as diagonal blocks.

Proof: This proof is a simple generalization of the proof that by diagonally scaling an n by n positive definite matrix to have unit diagonal, its condition number is within a factor of n of the lowest condition number achievable by diagonal scaling [vanderSluis]. We generalize diagonal scaling for unit diagonal to be block diagonal scaling for block unit diagonal, i.e. to have identity matrices (of various sizes) on the diagonal. We show that a block diagonal scaling with b blocks produces a matrix whose condition number is within a factor b of the lowest possible condition number.

Assume S_i forms an orthonormal basis of $\mathbf{S}^i$ and let D be a block diagonal nonsingular matrix whose blocks D_i are c_i by c_i. Then any S' whose columns S'_i span $\mathbf{S}^i$ can be written $S' = SD$ for some D. Now

$$\sqrt{b}\,\kappa(SD) = \sqrt{b}\,\frac{\max\limits_{w\neq 0}\dfrac{\|Sw\|}{\|D^{-1}w\|}}{\min\limits_{z\neq 0}\dfrac{\|Sz\|}{\|D^{-1}z\|}} \geq \frac{\|D^{-1}z_0\|}{\|D^{-1}w_0\|}\,\frac{\sqrt{b}\;\|Sw_0\|}{\sigma_{\min}(S)}\ , \tag{3.25}$$

where z_0 is chosen so that $\|z_0\| = 1$ and $\|Sz_0\| = \sigma_{\min}(S)$ = the smallest singular value of S, and w_0 is chosen so $\|w_0\| = 1$ and $\|D^{-1}w_0\| = \sigma_{\min}(D^{-1})$. With this choice of w_0 the factor $\|D^{-1}z_0\| / \|D^{-1}w_0\|$ is at least one. Since D is block diagonal, w_0 can be chosen to have nonzero components corresponding to only one block of D. Thus, $\|Sw_0\|^2 = \|w_0{}^*S^*Sw_0\| = \|w_0{}^*w_0\| = 1$. Since the largest singular value $\sigma_{\max}(S)$ satisfies

$$\sigma_{\max}(S) = \|S\| \leq \sqrt{\sum_{i=1}^{b}\|S_i\|^2} = \sqrt{\sum_{i=1}^{b}1} = \sqrt{b}\ ,$$

we get

$$\sqrt{b}\,\kappa(SD) \geq \frac{\sigma_{\max}(S)}{\sigma_{\min}(S)} = \kappa(S)\ . \tag{3.26}$$

Since (3.26) is true for any D, it is true in particular when $SD = S_{OPTIMAL}$. Q.E.D.

Van Dooren and Dewilde [VanDooren] have improved the factor $\sqrt{b}$ and shown, in particular, that if the subspaces are themselves orthogonal, then the above choice of S is in fact optimal.

In the case $b=2$ we expressed $\kappa(S_{OPTIMAL})$ in terms of csc ϑ, where ϑ was the smallest angle between $\mathbf{S}^1$ and $\mathbf{S}^2$. We can also bound $\kappa(S)$ here in terms of the csc ϑ_i, where ϑ_i is the angle between $\mathbf{S}^i$ and its complement $\operatorname{span}_{j\neq i}\{\mathbf{S}^j\}$:

Theorem 3.3: Let T, S and csc ϑ_i be defined as above. Then

$$\max_i\,(\csc\vartheta_i + \sqrt{\csc^2\vartheta_i - 1}) \leq \kappa(S) \leq \sqrt{b}\cdot\sqrt{\sum_{i=1}^{b}\csc^2\vartheta_i}\ , \tag{3.27}$$

or weakened slightly,

$$\max_i \csc \vartheta_i \le \kappa(S) \le b \cdot \max_i \csc \vartheta_i \ . \tag{3.28}$$

Proof: This proof is based on a similar result of Wilkinson's [Wilkinson2, p. 89] when all invariant subspaces are one dimensional. First we will prove the lower bound and then the upper bound.

From (3.16) we know that any S (not just the one defined above) which has one group of columns spanning $\mathbf{S}^i$ has a condition number bounded from below:

$$\kappa(S) \ge \cot \ \vartheta_i/2 \ = \csc \vartheta_i + \sqrt{\csc^2 \vartheta_i - 1} \ . \tag{3.29}$$

Since (3.29) is true for all i, the lower bound follows easily.

We compute the upper bound as follows:

$$\kappa(S) = \| S \| \ \| S^{-1} \| \le \sqrt{b} \ \| S^{-1} \| \tag{3.30}$$

since $\| S \| \le \sqrt{b}$ (as mentioned in the proof of Theorem 2). Using notation analogous to (2.14) define the matrix P_i

$$P_i = S_i \ (S^{-1})^{(i)} \tag{3.31}$$

(which would be the matrix projection onto $\mathbf{S}^i$ for the standard eigenproblem). Since S_i consists of orthonormal columns, (3.31) and then (3.19) yield

$$\| (S^{-1})^{(i)} \| \ = \| P_i \| \ = \csc \vartheta_i \tag{3.32}$$

Thus

$$\| S^{-1} \| \le \sqrt{\sum_{i=1}^{b} \| (S^{-1})^{(i)} \|^2} = \sqrt{\sum_{i=1}^{b} \csc^2 \vartheta_i} \tag{3.33}$$

and the upper bound follows. Q.E.D.

The lower bound in Theorem 3.3 has been proven by Bauer [Bauer4] in the case when all invariant subspaces are one-dimensional .

The other choice of S discussed in the literature is scaled so that the i-th diagonal block of S^*S is $\csc \vartheta_i$ times an identity matrix of size c_i by c_i. With this choice of S the i-th diagonal block of $(S^*S)^{-1}$ has the same norm as

the corresponding block of S^*S, namely csc ϑ_i. Smith [Smith] showed in the case when all invariant subspaces are one-dimensional that this choice of S is optimally scaled with respect to the condition number defined with the Euclidean norm:

$$\kappa_E(S) = \| S \|_E \, \| S^{-1} \|_E \ .$$

More generally, with this choice of S, Theorem 2 is weakened slightly to become:

Theorem 3.2a: With S chosen so that the i-th diagonal block of S^*S is csc ϑ_i times an identity matrix, we have

$$\kappa(S) \le b \cdot \kappa(S_{OPTIMAL}) \ . \tag{3.34}$$

Proof: Similar to Theorem 3.2.

Theorem 3.3, on the other hand, becomes slightly stronger:

Theorem 3.3a: With S chosen as in Theorem 3.2a, we can bound $\kappa(S)$ as follows:

$$\max_i \left(\csc \vartheta_i + \sqrt{\csc^2 \vartheta_i - 1}\right) \le \kappa(S) \le \sum_{i=1}^{b} \csc \vartheta_i \ . \tag{3.35}$$

Proof: Similar to Theorem 3.3.

The upper bound of Theorem 3.3a generalizes a result of Wilkinson [Wilkinson2, p 89] for one dimensional invariant subspaces. Note that the "spectral condition numbers" $1/|s_i|$ used by Wilkinson and others [Smith,Wilkinson2] are just csc ϑ_i (or $\| P_i \|$) when the invariant subspaces are one-dimensional. When $\sum_{i=1}^{b} \csc \vartheta_i$ is large the upper bound in (3.35) is comparable with the upper bound on $\kappa(S_{OPTIMAL})$ given by Bauer [Bauer4, Theorem VII] in the case of one-dimensional invariant subspaces.

This choice of S is more difficult to compute than the S of Theorems 3.2 and 3.3 because of the need to compute the csc ϑ_i, though not much more difficult if the subspaces are all one or two dimensional.

3.5 Proof of Theorem 3.1

This theorem was stated in section 3.3.

Unit vectors $x \in \mathbf{C}^m$ and $y \in \mathbf{C}^n$ satisfying

$$y^*(A^{-1/2})^*BC^{-1/2}x = \|(A^{-1/2})^*BC^{-1/2}\| \tag{3.36}$$

must exist. Use them to construct the unit vectors

$$z = A^{-1/2}y / \|A^{-1/2}y\| \quad , \quad w = C^{-1/2}x / \|C^{-1/2}x\| \quad , \tag{3.37}$$

and

$$s(\vartheta) = \begin{bmatrix} z \sin\vartheta \\ w \cos\vartheta \end{bmatrix} . \tag{3.38}$$

We want to consider H acting on the 2-dimensional subspace in which $s(\vartheta)$ lies. Now

$$s^*(\vartheta)Hs(\vartheta) \le \Lambda \tag{3.39}$$

implies

$$[z^* \sin\vartheta, w^* \cos\vartheta] \begin{bmatrix} A & B \\ B^* & C \end{bmatrix} \begin{bmatrix} z \sin\vartheta \\ w \cos\vartheta \end{bmatrix} \le \Lambda \ , \tag{3.40}$$

or

$$\sin^2\vartheta \cdot z^*Az + \cos^2\vartheta \cdot w^*Cw + \sin\vartheta\cos\vartheta\,(w^*B^*z + z^*Bw) \le \Lambda \ . \tag{3.41}$$

To simplify notation, let $a \equiv z^*Az$ and $c \equiv w^*Cw$.

From (3.36) and (3.37) we know that

$$\begin{aligned} z^*Bw &= \|(A^{-1/2})BC^{-1/2}\| / (\|A^{-1/2}y\| \cdot \|C^{-1/2}x\|) \\ &= \|(A^{-1/2})BC^{-1/2}\| \cdot \|A^{1/2}z\| \cdot \|C^{1/2}w\| \ . \end{aligned} \tag{3.42}$$

Since $(C^{1/2})^*C^{1/2} = C$, we get $c \equiv w^*Cw = \|w^*(C^{1/2})^*C^{1/2}w\| = \|C^{1/2}w\|^2$. Similarly, $a \equiv z^*Az = \|A^{1/2}z\|^2$, so (3.42) becomes

$$z^*Bw = \| (A^{-1/2})^* BC^{-1/2} \| \cdot \sqrt{ac} \quad . \tag{3.43}$$

Substituting (3.43) into (3.41) and rearranging, we obtain

$$\left(\frac{c+a}{2}\right) + \left(\frac{c-a}{2}\right) \cos 2\vartheta + \sqrt{ac} \; \| (A^{-1/2})^* BC^{-1/2} \| \; \sin 2\vartheta \le \Lambda \tag{3.44}$$

Since ϑ was arbitrary, we can maximize the L.H.S. of (3.44) over ϑ yielding

$$\left(\frac{c+a}{2}\right) + \sqrt{\left(\frac{c-a}{2}\right)^2 + ac \; \| (A^{-1/2})^* BC^{-1/2} \|^2} \le \Lambda \; , \tag{3.45}$$

or

$$\| (A^{-1/2})^* BC^{-1/2} \| \le \frac{\sqrt{(\Lambda - (c+a)/2)^2 - ((c-a)/2)^2}}{\sqrt{ac}} \tag{3.46}$$

$$= \frac{\sqrt{(\Lambda - a)(\Lambda - c)}}{\sqrt{ac}} \; .$$

Similarly, the inequality

$$\lambda \le s^*(\vartheta) Hs(\vartheta) \tag{3.47}$$

implies

$$\lambda \le \left(\frac{c+a}{2}\right) + \left(\frac{c-a}{2}\right) \cos 2\vartheta + \sqrt{ac} \; \| (A^{-1/2})^* BC^{-1/2} \| \; \sin 2\vartheta \; . \tag{3.48}$$

Minimizing the R.H.S. of (3.48) over ϑ we obtain

$$\lambda \le \left(\frac{c+a}{2}\right) - \sqrt{\left(\frac{c-a}{2}\right)^2 + ac \; \| (A^{-1/2})^* BC^{-1/2} \|^2} \tag{3.49}$$

or, rearranging,

$$\| (A^{-1/2})^* BC^{-1/2} \| \le \frac{\sqrt{(a-\lambda)(c-\lambda)}}{\sqrt{ac}} \; . \tag{3.50}$$

Combining (3.46) and (3.50) yields

$$\| (A^{-1/2})^* BC^{-1/2} \| \le \min\left[\sqrt{(a-\lambda)(c-\lambda)/(ac)} \, , \sqrt{(\Lambda - a)(\Lambda - c)/(ac)}\right] \; .$$

All we know about $z^*Az = a$ is that $\lambda \le a \le \Lambda$, and similarly $\lambda \le c = w^*Cw \le \Lambda$. Thus

$$\| (A^{-1/2})^* BC^{-1/2} \| \le \max_{\lambda \le \alpha, \gamma \le \Lambda} \min\left[\sqrt{(\alpha-\lambda)(\gamma-\lambda)/(\gamma\alpha)} \, , \sqrt{(\Lambda-\alpha)(\Lambda-\gamma)/(\gamma\alpha)}\right] \tag{3.51}$$

Since $(\alpha - \lambda)/\alpha$ is an increasing function of α and $(\Lambda - \alpha)/\alpha$ is a decreasing function of α in the range $\lambda \le \alpha \le \Lambda$, we see the max in the last inequality

occurs when the two arguments of the min are equal. This equality implies

$$(\alpha - \lambda)(\gamma - \lambda) = (\Lambda - \alpha)(\Lambda - \gamma) \tag{3.52}$$

or

$$\alpha + \gamma = \Lambda + \lambda \ . \tag{3.53}$$

Substituting (3.53) into (3.51) yields

$$\begin{aligned} \| (A^{-1/2})^* B C^{-1/2} \| &\le \max_{\lambda \le \gamma \le \Lambda} \sqrt{(\gamma - \lambda)(\Lambda - \gamma)} / \sqrt{\gamma(\Lambda + \lambda - \gamma)} \\ &= \frac{\Lambda - \lambda}{\Lambda + \lambda} \\ &= \frac{\kappa - 1}{\kappa + 1} \ . \end{aligned} \tag{3.54}$$

as desired.

Any 2 by 2 positive definite matrix whose diagonal entries are equal shows the the inequality of Theorem 3.1 is sharp.

We now show that given κ and $Z = (A^{-1/2})^* B C^{-1/2}$ such that $\| Z \| < 1$ and the inequality of the theorem is sharp, it is possible to construct an H with the given constraints. Simply choose

$$A = I \quad , \quad C = I \quad \text{and} \quad B = Z \tag{3.55}$$

corresponding to the (arbitrary) choice $\Lambda = 1 + \| Z \|$ and $\lambda = 1 - \| Z \|$. It is easy to verify that every inequality in the proof is sharp for this choice of A, B, and C. Q.E.D.

Theorem 3.4: Let H, Λ, λ, and κ be as above. Define $X^{-1/2}$ such that $X^{-1/2}(X^{-1/2})^* = X^{-1}$. Then the following inequalities are sharp:

$$\| B C^{-1} \| \le \frac{1}{2}(\sqrt{\kappa} - 1/\sqrt{\kappa}) \tag{3.56}$$

$$\| A^{-1} B \| \le \frac{1}{2}(\sqrt{\kappa} - 1/\sqrt{\kappa}) \tag{3.57}$$

$$\| B \| \le \frac{1}{2}(\Lambda - \lambda) \tag{3.58}$$

$$\| (A^{-1/2})^* B \| \le \sqrt{\Lambda} - \sqrt{\lambda} \tag{3.59}$$

$$\| BC^{-1/2} \| \le \sqrt{\Lambda} - \sqrt{\lambda} \tag{3.60}$$

Proof: All the proofs are analogous to the proof of Theorem 3.1. To prove (3.58), for example (also proved in [Bauer3]), choose z and y unit vectors such that

$$z^* BC^{-1} y = \| BC^{-1} \|$$

and let

$$x = C^{-1} y / \| C^{-1} y \| \quad .$$

Consider H restricted to the two dimensional subspace in which

$$s(\vartheta) = \begin{bmatrix} z \sin\vartheta \\ x \cos\vartheta \end{bmatrix}$$

lies. The rest of the proof follows similarly to that of Theorem 3.1.

We can also show that given κ and arbitrary $R = BC^{-1}$ such that (3.56) is sharp, it is possible to construct an H with the given constraints. Simply choose

$$C = I \quad , \quad A = \left(\frac{\kappa^2 + 1}{2\kappa}\right) I \quad \text{and} \quad B = R \tag{3.61}$$

corresponding to the (arbitrary) choice $\Lambda = (\kappa + 1)/2$ and $\lambda = (\kappa + 1)/2\kappa$. It is easy to verify that every inequality in the proof is sharp for this choice of A, B, and C.

Note that Theorems 3.1 and 3.4 are still true when A, B, and C are conforming submatrices extracted from a larger H (or $Q^* HQ$ with Q unitary) since the bounds are monotonic in κ (or Λ and λ). In particular, if A, B, and C are scalar Theorem 3.1 becomes an inequality of Wielandt [Bauer2].

3.6 Computing a Function of a Matrix

In this section we want to show why a well conditioned block diagonalizing matrix S is better than an ill-conditioned one for computing a function of

a matrix T. Assuming $f(T)$ is an analytic function of T, we compute $f(T)$ as follows:

$$f(T) = f(S\Theta S^{-1}) = Sf(\Theta)S^{-1} = S \begin{bmatrix} f(\Theta_1) & & \\ & \ddots & \\ & & f(\Theta_m) \end{bmatrix} S^{-1} \ . \tag{3.62}$$

The presumption is that it is easier to compute f of the small blocks Θ_i than of all of T. We will not ask about the error in computing $f(\Theta_i)$ but rather the error in computing $\Theta = S^{-1}TS$ and $f(T) = Sf(\Theta)S^{-1}$. In general, we are interested in the error in computing the similarity transformation $X = SYS^{-1}$.

We assume for this analysis that we compute with single precision floating point with relative precision ε. That is, when $*$ is one of the operations $+, -, *$ or $/$, the relative error in computing $fl(a*b)$ is bounded by ε:

$$fl(a*b) = (a*b)(1+e) \quad \text{where} \quad |e| \le \varepsilon \ . \tag{3.63}$$

Using (3.63) it is easy to show

Lemma 3.5: Let A and B be real n by n matrices, where $n\varepsilon < .1$. Let $|A|$ denote the matrix of absolute entries of A: $|A|_{ij} = |A_{ij}|$. Then to first order in ε the error in computing the matrix product AB is bounded as follows:

$$|fl(AB) - AB| \le n\varepsilon |A| \ |B| \ . \tag{3.64}$$

Proof: See [Wilkinson1].

Computing $X = SYS^{-1}$ requires two matrix products: $Z = fl(SY)$ and $X = fl(ZS^{-1})$, where we assume S and S^{-1} are known exactly. Applying Lemma 3.5 to these two products leads to

Lemma 3.6: To first order in ε

$$\| fl(SYS^{-1}) - SYS^{-1} \|_E \le 2n^2\varepsilon\kappa(S) \| Y \|_E \ . \tag{3.65}$$

Proof:

$$\| fl(SYS^{-1}) - SYS^{-1} \|_E$$
$$= \| \, | fl(SYS^{-1}) - SYS^{-1} | \, \|_E$$
$$= \| \, | fl(SYS^{-1}) - fl(SY)S^{-1} + fl(SY)S^{-1} - SYS^{-1} | \, \|_E$$
$$\leq \| \, | fl(SYS^{-1}) - fl(SY)S^{-1} | \, \|_E + \| \, | fl(SY)S^{-1} - SYS^{-1} | \, \|_E$$
$$\leq n\varepsilon \, \| \, |fl(SY)| \, |S^{-1}| \, \|_E + n\varepsilon \, \| \, |S| \, |Y| \, |S^{-1}| \, \|_E$$
$$\leq 2n\varepsilon \, \| S \|_E \, \| Y \|_E \, \| S^{-1} \|_E$$

(to first order in ε)

$$\leq 2n^2\varepsilon \, \kappa(S) \, \| Y \|_E$$

Q.E.D.

Assuming this bound is realistic, it is clear that picking S to keep $\kappa(S)$ small is advantageous. The error in computing similarity transformations of matrices is discussed in more detail in Wilkinson [Wilkinson2, chap 3].

3.7 On Projection Norms as a Partitioning Criterion

The analysis of the last section suggests that if the purpose of our eigendecomposition is to compute functions of matrices, then it may be sufficient to compute the partition $\Sigma=\{\sigma_1, \ldots, \sigma_b\}$ of σ subject only to the constraint that $\| P_i \|$ be less than some threshold $\bar{\kappa}$ for each i, rather than the more complicated requirement that the dissociation between σ_i and $\bigcup_{j \neq i} \sigma_j$ be larger than some threshold ε. This is because the error bounds depend only on projector norms. For example, it is trivial to compute the exponential of a diagonal matrix by exponentiating each diagonal entry, and any diagonal matrix is decomposable by the projector norm criterion (all projectors are of norm one). The more stringent dissociation criterion, however, would forbid any decomposition of the identity matrix, which is clearly a bad idea. Kågström [Kågström2] has used this partitioning criterion successfully for computing

matrix functions.

There is, however, a small problem with defining Σ in terms of bounded projector norms instead of the dissociation criterion: partitions defined by bounded projector norms do not satisfy the intersection property described in section 2.1. In other words, $\Sigma^1 = \{\sigma_1 \cup \sigma_2 , \sigma_3\}$ may be a legitimate partition since $\| P_1 + P_2 \| = \| P_3 \| < \bar{\kappa}$, and $\Sigma^2 = \{\sigma_1 \cup \sigma_3 , \sigma_2\}$ may be a legitimate partition since $\| P_1 + P_3 \| = \| P_2 \| < \bar{\kappa}$, but $\Sigma = \Sigma^1 \cap \Sigma^2 = \{\sigma_1 , \sigma_2 , \sigma_3\}$ may not be legitimate since $\| P_1 \|$ can be as large as $\| P_2 + P_3 \| \le \| P_2 \| + \| P_3 \| \le 2 * \bar{\kappa}$. Consider, for example,

$$T = \begin{bmatrix} 0 & z & z \\ & 1 & 0 \\ & & -1 \end{bmatrix}$$

with z^2 a little less than $\bar{\kappa}^2 - 1$, $\sigma_1 = \{0\}$, $\sigma_2 = \{1\}$, and $\sigma_3 = \{-1\}$. A factor of 2 is not bad, especially since it does not seem likely we can measure $\| P \|$ or the smallest $\| \delta T \|$ that accurately for a reasonable price. Nonetheless, it is unfortunate that we lose the intersection property which makes the best partition well defined.

3.8 Applications of a Variation of Theorem 3.1

It is more convenient here to use a slight variation on Theorem 1, stated as (3.56) in Theorem 3.4:

$$\| A^{-1} B \| \le \frac{1}{2} (\sqrt{\kappa} - 1/\sqrt{\kappa}) \ .$$

Application 1: Cholesky without square roots. The square root free Cholesky algorithm (triangular factorization) decomposes a positive definite Hermitian matrix H into the product of a unit lower triangular matrix L, a nonnegative diagonal matrix D, and L^*:

$$H = LDL^*$$

We wish to bound the entries of L. Consider the following partitioning of the decomposition:

$$H = \begin{bmatrix} A & B \\ B^* & C \end{bmatrix} = \begin{bmatrix} L_1 & \\ R & L_2 \end{bmatrix} \begin{bmatrix} D_1 & \\ & D_2 \end{bmatrix} \begin{bmatrix} L_1^* & R^* \\ & L_2^* \end{bmatrix} . \tag{3.66}$$

From (3.66) we see

$$L_1 D_1 R^* = B$$

or

$$\begin{aligned} R^* &= (L_1 D_1)^{-1} B \\ &= L^*_1 (L_1 D_1 L^*_1)^{-1} B \\ &= L^*_1 A^{-1} B \ . \end{aligned}$$

Since L^*_1 is unit upper triangular, the last row of R and the last row of $A^{-1}B$ are identical. But the last row of R^* is the conjugate transpose of a subdiagonal column of L. Thus

$$\begin{aligned} \| \text{subdiagonal column of } L \| &= \| \text{last column of corresponding } A^{-1}B \| \\ &\le \| \text{all of corresponding } A^{-1}B \| \ , \end{aligned}$$

and so Theorem 3.4 implies

$$\| \text{subdiagonal column of } L \| \le \frac{1}{2}(\sqrt{\kappa} - 1/\sqrt{\kappa}) \ .$$

A 2 by 2 example suggested by the proof of Theorem 3.4 (see (3.60)) shows this bound is achievable.

This bound is tighter than the simpler bound

$$|L_{ij}| \le \sqrt{(H_{ii} - D_{ii})/D_{jj}} \le \sqrt{(\Lambda - \lambda)/\lambda} = \sqrt{\kappa - 1} \ , \tag{3.67}$$

which is derived by considering the i,i-th entries of both sides of $H = LDL^*$:

$$L_{ij}^2 D_{jj} + D_{ii} + \text{positive terms} = H_{ii} \ .$$

This result can also be used to get a lower bound on $\kappa(H)$ given its Cholesky decomposition.

A similar application to Gauss-Jordan elimination appears in [Bauer3]

Application 2: Gram-Schmidt Orthogonalization Process. The Gram-Schmidt process takes a set of independent vectors $v_i \in \mathbf{C}^n$, $1 \le i \le m$, and produces a set of orthonormal vectors $q_i \in \mathbf{C}^n$, $1 \le i \le m$, where q_i is a linear combination of v_1 through v_i and orthogonal to v_1 through v_{i-1} for $i > 1$. We wish to bound the coefficients of q_1 to q_{i-1} (or v_1 to v_{i-1}) in the expression for q_i. We do this by showing Gram-Schmidt to be equivalent to square-root-free Cholesky performed on a certain matrix, and use Application 1.

The Gram-Schmidt process expresses q_i as a linear combination of v_i and q_1 through q_{i-1}. Let V be the n by m matrix whose columns are the vectors v_i and let Q be the n by m matrix with columns q_i. Then the Gram-Schmidt process may be expressed succinctly as

$$V = QD^{1/2}U \ , \tag{3.68}$$

where U is an n by n unit upper triangular matrix and D is an n by n nonnegative diagonal matrix. The entries of U are the coefficients we seek to bound. Multiplying both sides of (3.68) on the left by their transposes, we obtain

$$V^*V = U^*DU \ . \tag{3.69}$$

U is the factor of V^*V obtained by doing square root free Cholesky. Thus, from Application 2 we see

$$\| \text{superdiagonal column of } U \| \le \frac{1}{2}\left(\sqrt{\kappa(V^*V)} - 1/\sqrt{\kappa(V^*V)} \right) \ , \tag{3.70}$$

which is the desired bound.

If we wanted to express q_i as a linear combination of v_1 through v_i instead of v_i and q_1 through q_{i-1}, we would express the Gram-Schmidt process as

$$V\hat{U}\hat{D}^{-1/2} = Q \ . \tag{3.71}$$

What is a bound for the columns of $\hat{U}$? Multiply both sides of (3.71) on the the left by their transposes to obtain

$$\hat{D}^{-1/2}\hat{U}^{*}V^{*}V\hat{U}\hat{D}^{-1/2} = Q^{*}Q = I \ , \tag{3.72}$$

or

$$(V^{*}V)^{-1} = \hat{U}\hat{D}^{-1}\hat{U}^{*} \ . \tag{3.73}$$

$\hat{U}$ is the factor of $(V^{*}V)^{-1}$ obtained by doing square root free Cholesky starting at the lower right corner of $(V^{*}V)^{-1}$ instead of the upper left corner as is usual. Thus, from Application 2 we see

$$\begin{aligned} \|\text{ superdiagonal column of } \hat{U}\| &\leq \frac{1}{2}\left(\sqrt{\kappa((V^{*}V)^{-1})} - 1/\sqrt{\kappa((V^{*}V)^{-1})}\right) \\ &= \frac{1}{2}\left(\sqrt{\kappa(V^{*}V)} - 1/\sqrt{\kappa(V^{*}V)}\right) \ . \end{aligned} \tag{3.74}$$

since $\kappa(M) = \kappa(M^{-1})$ for all M. Thus, we get the same bound on the columns of $\hat{U}$ as on the columns of U.

Chapter 4: Lower Bounds on $\mathrm{diss}_2(\sigma_1, \sigma_2, region)$

4.1 Introduction

In this chapter we present a new lower bound on $\mathrm{diss}_2(\sigma_1, \sigma_2, region)$ which is sharper than previous lower bounds. In section 4.2 we present a history of previous lower bounds, in section 4.3 we present and prove our new bound, and in section 4.4 we present examples when the new lower bound is a good estimate of $\mathrm{diss}_2(\sigma_1, \sigma_2, region)$, and other examples showing how badly the lower bound can underestimate $\mathrm{diss}_2(\sigma_1, \sigma_2, region)$. These examples will be needed in chapter 8 on probabilistic bounds. Since we will only be using the "region" based dissociation notion, we will drop the word "region" from the arguments of the dist function for notational simplicity in this chapter.

4.2 Previous Lower Bounds

We discuss three previous lower bounds in this section. The first two, due to Dunford and Schwarz, and Bauer and Fike, are usually called inclusion theorems because they give upper bounds on the perturbations in eigenvalues given the norm of the perturbation. We will use the results of Chapter 3 to show that these results are essentially identical and derivable from Gerschgorin's Theorem [Isaacson]. The third result, due to Stewart, was until now the sharpest known bound. We will discuss lower bounds on $\mathrm{diss}_2(\sigma_1, \sigma_2)$, which are also lower bounds for $\mathrm{diss}_g(\sigma_1, \sigma_2)$ by inequality (2.4).

The first result is due to Dunford [Dunford, lemma 6] and Schwartz [Schwartz, lemma 3]. Taken together, these lemmas show:

Theorem 4.1 (Dunford and Schwartz): Let T be completely diagonalizable with eigenvalues λ_i and corresponding projections P_i. If λ' is an eigenvalue of $T+E$, then for some i

$$|\lambda' - \lambda_i| \le 4 \cdot \max_i \|P_i\| \; \|E\| \quad .$$

In other words, the eigenvalues λ' of $T+E$ lie in circles of radius $4 \max_i \|P_i\| \; \|E\|$ centered at the eigenvalues of T. From this result, it is easy to derive the following lower bound on $\mathrm{diss}_2(\sigma_1, \sigma_2)$:

Corollary 4.2:

$$\mathrm{diss}_2(\sigma_1, \sigma_2) \ge \frac{\min_{\lambda_1 \in \sigma_1} |\lambda_1 - \lambda_2|}{8 \cdot \max_i \|P_i\|}$$

Proof: This condition assures that no circle around any $\lambda_1 \in \sigma_1$ can intersect any circle around some $\lambda_2 \in \sigma_2$. Q.E.D.

The Bauer-Fike Theorem has similar assumptions and conclusions:

Theorem 4.3 (Bauer and Fike): Let T be completely diagonalizable with eigenvalues λ_i and diagonalizing similarity S. If λ' is an eigenvalue of $T+E$, then for some i

$$|\lambda' - \lambda_i| \le \inf_S \kappa(S) \; \|E\|$$

where the inf is over all diagonalizing similarities S.

This theorem yield a lower bound on $\mathrm{diss}_2(\sigma_1, \sigma_2)$ in the same way as Theorem 4.1:

Corollary 4.4:

$$\mathrm{diss}_2(\sigma_1, \sigma_2) \ge \frac{\min_{\lambda_1 \in \sigma_1} |\lambda_1 - \lambda_2|}{2 \cdot \inf_S \kappa(S)}$$

Proof: Analogous to Corollary 4.2.

We claim these two theorems are nearly equivalent because Chapter 3 showed that

$$\max_i \|P_i\| \le \inf_S \kappa(S) \le \dim(T) \cdot \max_i \|P_i\| \quad . \tag{3.8}$$

so that the expressions in the denominators of the corollaries cannot differ by more than a factor of 4 dim(T). Furthermore, the Bauer-Fike result is easily derivable by applying Gerschgorin's Theorem [Isaacson] to the matrix

$$STS^{-1} + SES^{-1} = \mathrm{diag}(\lambda_i) + SES^{-1} \ .$$

The drawbacks of these simple lower bounds on $\mathrm{diss}_2(\sigma_1, \sigma_2)$ are as follows. First, they assume that T is completely diagonalizable, and so do not apply to defective matrices. Second, even if T is diagonalizable, the $\max\|P_i\|$ term may be too large and so the lower bound too small because of nearly equal eigenvalues in some irrelevant part of the spectrum. For example, by making η small in

$$\begin{bmatrix} 100 & .01 & & \\ & 100+\eta & & \\ & & 1 & \\ & & & 0 \end{bmatrix} \tag{4.1}$$

(where $\sigma_1=\{0\}$ and $\sigma_2=\{100, 100+\eta, 1\}$), we can make the lower bound on $\mathrm{diss}_2(\sigma_1, \sigma_2)$ as small as desired, whereas $\mathrm{diss}_2(\sigma_1, \sigma_2)$ actually equals .5 (we will prove this later; the best E is nonzero in the lower right 2 by 2 corner only).

The heretofore best lower bound on $\mathrm{diss}_2(\sigma_1, \sigma_2)$ is due to Stewart [Stewart]:

Theorem 4.5 (Stewart): Assume without loss of generality that T is of the form

$$T = \begin{bmatrix} A & C \\ & B \end{bmatrix}$$

with $\sigma_1=\sigma(A)$ and $\sigma_2 = \sigma(B)$. Let P be the projection corresponding either to σ_1 or σ_2. Then

$$\mathrm{diss}_2(\sigma_1\,,\,\sigma_2) \geq \frac{\mathrm{sep}(A,B)}{4\cdot\|P\|}\,. \tag{4.2}$$

Proof: The proof seeks a unitary similarity of a special form which returns $T+E$ to block triangular form. The nontrivial part of this similarity satisfies a matrix Ricatti equation which can be solved by an iteration which is a contraction as long as $\|E\|$ is smaller than the expression on the right hand side of (4.2). The details of the proof are not needed in this thesis; see Stewart [Stewart] for more information.

Actually, Stewart's proof only appears to show that $\mathrm{sep}(A,B)/(4\|P\|)$ is a lower bound on $\mathrm{diss}_2(\sigma_1\,,\,\sigma_2\,,\,path)$. It will follow, however, from the comments following theorem 4.6 below that it really is a lower bound on $\mathrm{diss}_2(\sigma_1\,,\,\sigma_2\,,\,region)$.

Stewart's bound is generally much tighter than the bounds of Dunford-Schwarz or Bauer-Fike. We are pleased, therefore, to have found the improvement in the next section.

4.3 A New Lower Bound on $\mathrm{diss}_2(\sigma_1\,,\,\sigma_2)$

Our improvement of Theorem 4.5 is based on an approach used by Varga and Feingold [Varga] and more recently Meyer and Veselic [Meyer] to prove a block version of Gerschgorin's theorem.

Theorem 4.6: Let T, A, B, and C be as in Theorem 4.5. Then

$$\mathrm{diss}_2(\sigma_1\,,\,\sigma_2) \geq \frac{\mathrm{sep}_\lambda(A,B)}{\|P\| + \sqrt{\|P\|^2 - 1}} \tag{4.3}$$

Proof: If λ is an eigenvalue of $T+E$ but not of T then

$$0 = \det(\lambda - T - E) = \det(\lambda - T)\,\det(I - (\lambda - T)^{-1}E) = \det(I - (\lambda - T)^{-1}E)$$

implying that

$$1 \leq \|(\lambda - T)^{-1}E\| \leq \|(\lambda - T)^{-1}\|\;\|E\|$$

Now we choose a block diagonalizing similarity S as suggested in (3.20) so

that $S^{-1}TS = \mathrm{diag}(A,B)$. (Note that the other S suggested in section 3.3 would yield $\mathrm{diag}(A,B')$ where B' is similar to B but could be otherwise much different.) Thus

$$1 \le \| S^{-1} \mathrm{diag}((\lambda-A)^{-1}, (\lambda-B)^{-1}) S \| \cdot \| E \|$$
$$\le \kappa(S) \max(\| (\lambda-A)^{-1} \|, \| (\lambda-B)^{-1} \|) \| E \|$$

or

$$\kappa(S) \| E \| \ge \min(\| (\lambda-A)^{-1} \|^{-1}, \| (\lambda-B)^{-1} \|^{-1}) .$$

Just as the inequalities of Theorems 4.1 and 4.3 show that the eigenvalues of $T+E$ lie in circles centered at the eigenvalues of T, this last inequality shows that the eigenvalues of $T+E$ lie in certain regions around the eigenvalues of T. And just as in Corollaries 4.2 and 4.4, we can derive a bound on $\| E \|$ such that the regions belonging to $\sigma(A)$ and the regions belonging to $\sigma(B)$ remain disjoint.

Indeed, as long the region

$$\kappa(S) \| E \| = \| (\lambda-A)^{-1} \|^{-1}$$

remains disjoint from the analogous region for B, $\| E \| < \mathrm{diss}_2(\sigma_1, \sigma_2)$. Imagining these regions as functions of $\| E \|$, there is a smallest $\| E' \|$ for which these regions can intersect, which means there is a λ' such that

$$\kappa(S) \| E' \| = \| (\lambda'-A)^{-1} \|^{-1} = \| (\lambda'-B)^{-1} \|^{-1} .$$

From the definition of sep_λ, it is clear that $\mathrm{sep}_\lambda(A,B)$ is less than or equal to both $\| (\lambda'-A)^{-1} \|^{-1}$ and $\| (\lambda'-B)^{-1} \|^{-1}$. Thus,

$$\kappa(S) \| E' \| \ge \mathrm{sep}_\lambda(A,B)$$

or, substituting the value of $\kappa(S)$ and rearranging

$$\| E' \| \ge \frac{\mathrm{sep}_\lambda(A,B)}{\| P \| + \sqrt{\| P \|^2 - 1}} .$$

Since $\| E' \|$ is the smallest value for which the regions can possibly intersect, the proof is complete. Q.E.D.

Lemma 2.10, plus the inequality

$$\|P\| + \sqrt{\|P\|^2 - 1} \le 2\,\|P\|$$

show that the bound of Theorem 4.6 is always larger than the bound of Theorem 4.5. The two bounds can be significantly different because sep can be much smaller than sep_λ. Just how much smaller is the subject of the next chapter, but we present an example to illustrate typical behavior. If

$$A = \begin{bmatrix} \varepsilon & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \varepsilon \end{bmatrix} \quad \text{and} \quad B = -A$$

are both n by n matrices, then

$$\mathrm{sep}(A,B) = \Theta(\varepsilon^{2n-1}) \quad \text{and} \quad \mathrm{sep}_\lambda(A,B) = \Theta(\varepsilon^n)$$

(the notation $f = \Theta(g)$ means f is exactly of order g: $f = O(g)$ and $g = O(f)$). We discuss this example in detail in chapter 5. Experience in constructing examples like this led us to our conjecture in chapter 8 that even though sep and sep_λ are almost always close, when they are far apart they can differ by at most a square as in the example.

An interesting corollary of this theorem holds when T is block diagonal:

Corollary 4.7: Suppose T is block diagonal:

$$T = \begin{bmatrix} A & \\ & B \end{bmatrix}.$$

Then

$$\mathrm{diss}_2(\sigma_1, \sigma_2) = \mathrm{sep}_\lambda(A,B) \tag{4.4}$$

$$\sqrt{2}\,\mathrm{sep}_\lambda(A,B) \ge \mathrm{diss}_F(\sigma_1, \sigma_2) \ge \mathrm{sep}_\lambda(A,B) \tag{4.5}$$

In particular, if $A = \bigoplus_i A_i$ and $B = \bigoplus_i B_i$ then

$$\mathrm{diss}_2(\sigma_1, \sigma_2) = \min_{ij} \mathrm{sep}_\lambda(A_i, B_j) \tag{4.6}$$

Proof: $||P||$ is clearly 1. Thus, the lower bound in Theorem 4.5 equals the upper bound in Theorem 2.8, proving (4.4). (4.5) follows similarly. (4.6) follows from Lemma 2.12. Q.E.D.

Thus, $\mathrm{diss}_2(\sigma_1, \sigma_2)$ satisfies the same divide and conquer paradigm as do sep (Lemma 2.6) and sep_λ (Lemma 2.12). In particular, there is a smallest perturbation (measured with $||\cdot||$ and not $||\cdot||_E$) with the same block diagonal structure as T. This proves the claim made about the T in (4.1), since it is block diagonal.

4.4 When is the Lower Bound a Good Estimate?

In this section we discuss when the lower bound of Theorem 4.6 is likely to be sharp. We have already seen that it is sharp for block diagonal matrices. We will show that it is also sharp for 2 by 2 matrices, and nearly so when A is 1 by 1 and B is diagonal (we will need this special case in chapter 8 on probabilistic bounds). We then present two examples when the lower bound is much too low: in the first example A is 1 by 1 and B is a Jordan block, and in the second both A and B are 2 by 2 diagonal matrices. This second example leads to a "combinatorial" improvement on our new lower bound.

Lemma 4.8: If

$$T = \begin{bmatrix} a & c \\ & b \end{bmatrix}$$

is 2 by 2, then

$$\mathrm{diss}_E(\sigma_1, \sigma_2) = \mathrm{diss}_2(\sigma_1, \sigma_2) = \frac{\mathrm{sep}_\lambda(a,b)}{||P|| + \sqrt{||P||^2 - 1}}$$

$$= \frac{|b-a|^2}{2\left(|c| + \sqrt{|c|^2 + |b-a|^2}\right)} .$$

Furthermore, this last expressions also equals both $\mathrm{diss}_2(\sigma_1, \sigma_2, path)$ and

$\mathrm{diss}_F(\sigma_1, \sigma_2, path)$.

Proof: Choose ω on the unit circle so that $\omega c/(b-a)$ is real and nonnegative. Then choose ϑ so that

$$\cot \vartheta/2 = \frac{\omega c}{b-a}.$$

Let

$$p = \|P\| = \sqrt{1+|c|^2/|b-a|^2}.$$

It is easy to see that the matrix

$$Q = \begin{bmatrix} \cos\vartheta & \bar{\omega}\sin\vartheta \\ \sin\vartheta & \bar{\omega}\cos\vartheta \end{bmatrix}$$

is unitary, and that

$$T' = QTQ^* = \begin{bmatrix} \frac{a+b}{2} & \cdot \\ \frac{(b-a)}{2(p+\sqrt{p^2-1})} & \frac{a+b}{2} \end{bmatrix},$$

where $\cdot$ represents a complicated expression that is not important. Clearly, by changing the lower left entry of T' to 0 we will change T' to a matrix with a double eigenvalue at $(a+b)/2$. Both the two norm and Frobenius norm of this rank one perturbation are equal to the lower bound in Theorem 4.6. A little manipulation yields the expression in the statement of the lemma. This proves that the expression in the statement of the lemma is equal to both $\mathrm{diss}_2(\sigma_1, \sigma_2, path)$ and $\mathrm{diss}_F(\sigma_1, \sigma_2, path)$. Since $\mathrm{diss}(\sigma_1, \sigma_2, path) \geq \mathrm{diss}(\sigma_1, \sigma_2, region)$, the lemma follows. Q.E.D.

The next example of the lower bound being nearly sharp is the matrix T with a 1 by 1 block $A=[a]$, an n by n diagonal block $B=\mathrm{diag}(b_i)$, and a 1 by n block $C=[c_1, \ldots, c_n]$:

$$T = \begin{bmatrix} a & c_1 & \cdot & c_n \\ & b_1 & & \\ & & \cdot & \\ & & & b_n \end{bmatrix}.$$

We need to use this example in chapter 8 when we show that our new lower bound is likely to be sharp.

The geometry of this example is simple. $\mathrm{sep}_\lambda(A,B)=\min_i |a-b_i|/2$; say the minimum occurs at $i=i_S$. $\sqrt{\|P\|^2-1}=\max_i |c_i/(a-b_i)|$; say the max occurs at $i=i_P$. $|a-b_{i_S}|$ gives the distance from a to the closest eigenvalue of B, and $\|P\|$ gives the maximum instantaneous speed at which a can move under perturbations in T [Kato2]. Therefore it makes sense that the smallest perturbation needed to make a hit an eigenvalue of B be approximately distance/speed $= |a-b_{i_S}|/\|P\| = \mathrm{sep}_\lambda/(2\cdot\|P\|)$. The algebra is quite messy, but the proof is similar to that of lemma 4.8: pick a unitary matrix Q

$$Q = \begin{bmatrix} \cos\vartheta & \bar{\omega}\sin\vartheta \\ \sin\vartheta & \bar{\omega}\cos\vartheta \end{bmatrix}$$

(ϑ is a real angle and $|\omega|=1$) so that

$$Q\begin{bmatrix} a & c_{i_P} \\ & b_{i_P} \end{bmatrix}Q^* = \begin{bmatrix} b_{i_S} & \cdot \\ \delta & a+b_{i_P}-b_{i_S} \end{bmatrix}.$$

$\cdot$ represents a complicated expression that is not important, and δ is the perturbation which makes the eigenvalue a move to b_{i_S}. The exact formula for $|\delta|$ is rather complicated, but it is easy to see that when a is much closer to b_{i_S} than b_{i_P}, $|\delta|$ cannot exceed the lower bound in theorem 4.6 by more than a small factor.

From lemmas 2.1 and 2.3, we see that if B can be diagonalized by a similarity of condition number κ, then the lower bound can not be too low by more than a factor of κ^2. Thus, it should come as no surprise that our first

example of when the lower bound is not sharp has A 1 by 1 and B an n by n Jordan block, since a Jordan block is the "least diagonalizeable" matrix of all:

$$T = \begin{bmatrix} 0 & 1 & & & \\ & \varepsilon & 1 & & \\ & & \cdot & \cdot & \\ & & & \cdot & 1 \\ & & & & \varepsilon \end{bmatrix} . \tag{4.7}$$

We will see that the upper bound $\mathrm{sep}_\lambda(A,B)=\Theta(\varepsilon^n)$, the lower bound $=\Theta(\varepsilon^{2n})$, and $\mathrm{diss}_2(\sigma_1, \sigma_2)=\Theta(\varepsilon^{n+1})$; thus, neither the upper nor the lower bound is asymptotically correct, but the upper bound is a much better estimate than the lower bound for large n. The proof that $\mathrm{diss}_2(\sigma_1, \sigma_2)=O(\varepsilon^{n+1})$ will follow from considering perturbations only in the lower left corner, and the proof that $\mathrm{diss}_2(\sigma_1, \sigma_2)=\Omega(\varepsilon^{n+1})$ (i.e. $\mathrm{diss}_2(\sigma_1, \sigma_2)$ is decreases no more quickly than ε^{n+1}) will follow from analyzing the characteristic polynomial of $T+E$, where E is a general perturbation.

That $\mathrm{sep}_\lambda(A,B)=\Theta(\varepsilon^n)$ follows from Lemma 2.11. Similarly, it is easy to see that $\|P\|=\Theta(\varepsilon^{-n})$ so that our lower bound on $\mathrm{diss}_2(\sigma_1, \sigma_2)$ is $\Theta(\varepsilon^{2n})$.

To estimate $\mathrm{diss}_2(\sigma_1, \sigma_2)$, we consider perturbations in the lower left corner. A simple computation shows that if we change $T_{n+1,1}$ from 0 to

$$T_{n+1,1} = \frac{-1}{n+1}\left(\frac{n}{n+1}\right)^n \varepsilon^{n+1}$$

then the 0 eigenvalue and one of the eigenvalues at ε coalesce at $\varepsilon/(n+1)$. To show that $\mathrm{diss}_2(\sigma_1, \sigma_2)$ cannot be of order ε^x for $x>n+1$ we consider the characteristic polynomial of $T+E\cdot\varepsilon^x$, where $\|E\|=O(1)$. If $E=0$ the characteristic polynomial is

$$\det(\lambda I-T) = \sum_{j=0}^{n} \frac{n!}{(n-j)!j!}(-\varepsilon)^j \lambda^{n-j+1} = \sum_{j=0}^{n+1} a_j(\varepsilon)\lambda^j$$

where $a_j(\varepsilon)$ is a polynomial in ε with lowest order (dominating) term ε^{n+1-j} for $j\geq 1$ and $a_0=0$. It is easy to see that

$$\det(\lambda I - T - E) = \sum_{j=0}^{n+1} a'_j(\varepsilon)\lambda^j$$

where $a'_j(\varepsilon)$ has the same dominating term as $a_j(\varepsilon)$ for $j \geq 1$ and $a'_0(\varepsilon) = \Theta(\varepsilon^x)$. By changing variables to $u = \lambda/\varepsilon$ the characteristic polynomial becomes

$$\det(\lambda I - T - E) = \varepsilon^{n+1}(u(u-1)^n + \varepsilon p(u) + \varepsilon^{x-n-1})$$

where $p(0)=0$ and its remaining coefficients are $O(1)$. Clearly, if $x > n+1$ then the eigenvalue at $\lambda=0$ remains isolated from the eigenvalues at $\lambda = \varepsilon$ $(u=1)$ by the continuity of the roots of a polynomial as functions of the coefficients. If $x = n+1$ then this argument breaks down and indeed we have displayed a perturbation of that magnitude that makes eigenvalues coalesce.

The next kind of example that shows that the lower bound of theorem 4.6 can be low depends on $\dim(A)$ and $\dim(B)$ both being at least 2. The idea is that sep_λ will be small and $\|P\|$ large because of nonoverlapping parts of the spectra of A and B. In other words, sep_λ will be determined by $\lambda_1(A)$ and $\lambda_1(B)$, and $\|P\|$ by $\lambda_2(A)$ and $\lambda_2(B)$. It will turn out the $\mathrm{diss}_2(\sigma_1, \sigma_2)$ will be the smallest perturbation that either makes $\lambda_1(A)$ coalesce with $\lambda_1(B)$ or $\lambda_2(A)$ coalesce with $\lambda_2(B)$. This example will lead to a systematic improvement on theorem 4.6 obtained by considering all possible partitions of σ_1 and σ_2; it illustrates the combinatorial complexity of the problem.

Let

$$T = \begin{bmatrix} 1 & & 0 & \\ & -1 & & 1 \\ & & 1+\varepsilon & \\ & & & -1+\sqrt{\varepsilon} \end{bmatrix} = \begin{bmatrix} A & C \\ & B \end{bmatrix}.$$

Simple computations yield $\mathrm{sep}_\lambda(A,B) = \varepsilon/2$ and $\|P\| = \sqrt{(1+\varepsilon)/\varepsilon}$, providing a lower bound on $\mathrm{diss}_2(\sigma_1, \sigma_2)$ of $\Theta(\varepsilon^{3/2})$ and an upper bound of $\varepsilon/2$. We will show that the upper bound is a much better estimate of $\mathrm{diss}_2(\sigma_1, \sigma_2)$. Clearly, to make $\sigma_1 = \{-1, 1\}$ coalesce with $\sigma_2 = \{1+\varepsilon, -1+\sqrt{\varepsilon}\}$, either $\{1\}$ has

to coalesce with $\{-1, -1+\sqrt{\varepsilon}, 1+\varepsilon\}$, or $\{-1\}$ has to coalesce with $\{1, 1+\varepsilon, -1+\sqrt{\varepsilon}\}$. Thus

$$\mathrm{diss}_2(\sigma_1, \sigma_2) \geq \min\left(\mathrm{diss}_2(\{1\}, \{-1, -1+\sqrt{\varepsilon}, 1+\varepsilon\}), \mathrm{diss}_2(\{-1\}, \{1, 1+\varepsilon, -1+\sqrt{\varepsilon}\})\right)$$

$$= \min\left(\frac{\varepsilon}{2}, \frac{\varepsilon}{2(1+\sqrt{1+\varepsilon})}\right) = \frac{\varepsilon}{2(1+\sqrt{1+\varepsilon})}.$$

In other words, $\mathrm{diss}_2(\sigma_1, \sigma_2)$ is determined by the size of the smallest perturbation that makes -1 coalesce with $-1+\sqrt{\varepsilon}$; the rest of the spectrum is irrelevant.

In this example we used the fact that the lower bound of theorem 4.6 is exact for 2 by 2 matrices, but this was not necessary. In fact, if $\Sigma_1 = \{\sigma_{11}, \ldots, \sigma_{1j}\}$ is any partition of σ_1, the above argument shows that

$$\mathrm{diss}_2(\sigma_1, \sigma_2) \geq \min_i \mathrm{diss}_2(\sigma_{1i}, \sigma - \sigma_{1i})$$

since some eigenvalue from some σ_{1i} must coalesce with something in its complement $\sigma - \sigma_{1i}$. If we consider all possible partitions Σ_1 of σ_1 (including the trivial partition $\{\sigma_1\}$) and similarly all partitions Σ_2 of σ_2 we obtain the following equality:

Theorem 4.9: Let σ_i and Σ_i be as above for $i=1,2$. Then

$$\mathrm{diss}_2(\sigma_1, \sigma_2) = \max\left(\max_{\Sigma_1} \min_{\sigma_{1i} \in \Sigma_1} \mathrm{diss}_2(\sigma_{1i}, \sigma - \sigma_{1i}), \max_{\Sigma_2} \min_{\sigma_{2j} \in \Sigma_2} \mathrm{diss}_2(\sigma - \sigma_{2j}, \sigma_{2j})\right).$$

Proof: That the right hand side is no greater than the left hand side follows from the previous discussion. Equality must hold because $\mathrm{diss}_2(\sigma_1, \sigma_2)$ is one of the candidates of the maximum on the right. Q.E.D.

If we substitute the lower bound of theorem 4.6 for each diss_2 expression on the right hand side of the last equation, we obtain an improvement of the theorem, as illustrated by the last example.

Chapter 5: How Far Apart can the Upper and Lower Bounds on $\mathrm{diss}_2(\sigma_1, \sigma_2)$ Be?

5.1 Introduction

Let us summarize the upper and lower bounds already proven in theorems 2.10 and 4.6 (we assume T has the structure shown in (2.6), with $\sigma_1=\sigma(A)$ and $\sigma_2=\sigma(B)$):

Theorem 5.1:

$$\mathrm{sep}_\lambda(A,B) \geq \mathrm{diss}_2(\sigma_1, \sigma_2, path) \tag{5.1}$$

$$\geq \mathrm{diss}_2(\sigma_1, \sigma_2, region) \geq \frac{\mathrm{sep}_\lambda(A,B)}{\|P\| + \sqrt{\|P\|^2 - 1}}$$

and

$$\sqrt{2}\,\mathrm{sep}_\lambda(A,B) \geq \mathrm{diss}_E(\sigma_1, \sigma_2, path) \tag{5.2}$$

$$\leq \mathrm{diss}_E(\sigma_1, \sigma_2, region) \geq \frac{\mathrm{sep}_\lambda(A,B)}{\|P\| + \sqrt{\|P\|^2 - 1}}\,.$$

In this chapter we will analyze how far apart these bounds can be. We will present only global bounds, valid for all matrices and depending only on the dimensionality. Probabalistic bounds, which show when the upper and lower bounds are likely to be close together or far apart, are presented in chapter 8.

Our worst case analysis starts by substituting the upper bound for $\|P\|$ of lemma 2.5 in inequality (5.2) ((5.1) is so similar to (5.2) that we will not consider it further):

$$\sqrt{2}\,\mathrm{sep}_\lambda(A,B) \geq \mathrm{diss}_E(\sigma_1, \sigma_2) \geq \frac{\mathrm{sep}_\lambda(A,B)}{1+\dfrac{\|C\|_E}{\mathrm{sep}(A,B)}} \tag{5.3}$$

$$= \frac{\mathrm{sep}_\lambda(A,B)\cdot\mathrm{sep}(A,B)}{\mathrm{sep}(A,B) + 2\cdot\|C\|_E}$$

$$\geq \frac{\mathrm{sep}_\lambda(A,B)\cdot\mathrm{sep}(A,B)}{\mathrm{sep}(A,B) + 2\cdot\|T\|_E} .$$

We normalize by taking $\|T\|_E=1$ so that $\mathrm{diss}_E(\sigma_1, \sigma_2)$ actually measures the relative change in T. With this normalization sep≤ 2 and $\mathrm{sep}_\lambda \leq 1$, so the denominator of the last right hand side must lie in the interval [2,4] and is therefore not important:

Corollary 5.2:

$$\sqrt{2}\,\mathrm{sep}_\lambda(A,B) \geq \mathrm{diss}_E(\sigma_1, \sigma_2) \geq \mathrm{sep}_\lambda(A,B)\cdot\mathrm{sep}(A,B)/4 . \qquad (5.4)$$

For this coarsening of (5.2) to be realistic, $\|P\|$ must be near its largest possible value $1+\|T\|_E/\mathrm{sep}$. $\|P\|$ can fail to be near its bound because $\|C\|_E$ is much less than $\|T\|_E$; $\|C\|_E$ is a kind of generalized measure of nonnormality of T with respect to the partitioning $\{\sigma_1,\sigma_2\}$, and contributes to the bound in the simple way shown above. The way in which the upper and lower bound can differ greatly is for sep to be small. We know from lemma 2.12 that $2\cdot\mathrm{sep}_\lambda$ is an upper bound on sep; the question this chapter asks is how much smaller can sep be than sep_λ?

The results of this chapter are as follows (in this chapter we will make the convention that $n_A \equiv \dim(A) \leq \dim(B) \equiv n_B$). We show that sep can be no smaller than a constant multiple of $\mathrm{sep}_\lambda^{n_A}$. This means that the lower bound in (5.2) can be no smaller than a constant multiple of the $\dim(A)+1$-st power of the upper bound. We present examples where the lower bound is actually the cube of the upper bound, and other examples where it is the square and either the upper or lower bound may be the more accurate measure of $\mathrm{diss}_E(\sigma_1, \sigma_2, path)$. Since $\|P\|$ provides an upper bound on sep (from lemma 2.5), we may translate our results into upper bounds on $\mathrm{diss}_E(\sigma_1, \sigma_2, path)$ depending on $\|P\|$. These results, though necessarily

weaker than the results depending on sep and sep_λ, reproduce results found in the literature.

The remainder of this chapter is organized as follows. Section 5.2 surveys historical results. Section 5.3 handles the case $\dim(A)=1$ which works out especially simply. Section 5.4 discusses the general case $\dim(A)\geq 2$. Finally, in section 5.5, we compute $\mathrm{diss}_2(\sigma_1, \sigma_2)$ and $\mathrm{diss}_F(\sigma_1, \sigma_2)$ exactly for normal matrices, in which case the path and region dissociations coincide.

5.2 Survey of the Literature

As stated in the introduction, the literature has concentrated on using $\|P\|$ for upper bounds on $\mathrm{diss}_2(\sigma_1, \sigma_2, path)$ and $\mathrm{diss}_F(\sigma_1, \sigma_2, path)$. Here we mention three previous works, by Ruhe [Ruhe1], Wilkinson [Wilkinson3], and Kahan [Kahan1]. Since Kahan's result, stated below, essentially implies the other two results, we discuss it first.

Kahan proves the following theorem:

Theorem (Kahan): If $\|P\| > \sqrt{n_A+1}$ then

$$\frac{\mathrm{diss}_2(\sigma_1, \sigma_2, path)}{\|T\|} \leq \frac{1.22}{(\|P\|^2-1)^{1/(2n_A)}} .$$

We prove a stronger result in section 5.4, essentially replacing $\|P\|$ by an upper bound depending on sep. Kahan's proof, which is totally different than ours, could however be modified to use sep instead of $\|P\|$. This modified proof yields the insight that if the R attaining the infimum in definition 2.4 of sep has well separated singular values (i.e. some near $\|R\|$, the rest near zero), then the exponent $-1/n_A$ of $\|P\|$ appearing in the bound can be replaced by -1, but we do not pursue this approach further, since it would lead to a probabilistic analysis similar to that of chapters 6 and 7.

Wilkinson's result [Wilkinson3] only covers the case $n_A=1$ and is essentially the same as Kahan's result. Ruhe's bound is on the distance from a given diagonalizable matrix to the set **P** of matrices with at least one multiple eigenvalue, and, by an abuse of notation, may be written

$$\mathrm{dist}_E(T,\mathbf{P}) \le \frac{n}{4} \frac{\max_{ij} |\lambda_i - \lambda_j|}{\sqrt{(\max_i \|P_i\|^{2/(n-1)} - 1)}} ,$$

and is weaker than the previous two results in that it has a higher root of $\|P\|$ in the denominator. It does, however, explicitly depend on the difference of eigenvalues, which may be small even if $\|P\|$ is not large. This advantage is shared by $\sqrt{2}\mathrm{sep}_\lambda$ as an upper bound.

5.3 The Case $\dim(A)=1$

This case is particularly simple; $\mathrm{diss}_E(\sigma_1, \sigma_2)$ is the dissociation between a simple eigenvalue and the rest of the spectrum. From lemma 2.11 we know that sep and sep_λ cannot differ by more than a factor of 2:

$$\mathrm{sep}_\lambda(A,B) \le \mathrm{sep}(A,B) = \sigma_{\min}(B - \alpha \cdot I) \le 2\, \mathrm{sep}_\lambda(A,B)$$

so that inequality (5.4) becomes

$$\sqrt{2}\, \sigma_{\min}(B - \alpha \cdot I) \ge \mathrm{diss}_E(\sigma_1, \sigma_2) \ge \frac{\sqrt{2}}{16} \sigma_{\min}^2(B - \alpha \cdot I) .$$

Thus, the lower bound can behave at worst like the square of the upper bound (recall that $\|T\|_E=1$ so that all bounds are on the relative error). The distance between these bounds cannot be decreased, as two examples of the last chapter have shown. Lemma 4.8 shows that for a 2 by 2 matrix the lower bound in (5.2) is sharp. On the other hand, the example in equation (4.7), which also had a two point spectrum, shows that the upper bound in the last equation is more accurate for a case where the upper and lower bounds differ by a square. Thus, we cannot hope to improve the bounds in (5.2) much if we only use measures like sep_λ, $\|P\|$ and similar global measures. We will

see in chapter 8, though, that the lower bound of (5.2) will be accurate unless T falls into a set of small probability.

Since lemma 2.5 shows that $||P||$ provides a lower bound on sep and hence sep_λ, we can change the upper bounds in (5.1) and (5.2) to upper bounds in terms of $||P||$:

Lemma 5.3: Let $\dim(A)=1$. Then

$$\frac{\sqrt{2}}{\sqrt{||P||^2-1}} \geq \text{diss}_E(\sigma_1, \sigma_2, path)$$

$$\frac{1}{\sqrt{||P||^2-1}} \geq \text{diss}_2(\sigma_1, \sigma_2, path)$$

Proof: Follows immediately from lemma 2.5, theorem 2.9, and lemma 2.11. Q.E.D.

This yields the results of Kahan [Kahan1] and Wilkinson [Wilkinson3].

5.4 The Case $\dim(A)\geq 2$

The goal of this section is to prove

Theorem 5.4: *Let T have the structure of (2.6), and assume $||T||_E=1$. Then*

$$2 \cdot \text{sep}_\lambda(A,B) \geq \text{sep}(A,B) \geq \frac{\text{sep}_\lambda(A,B)}{n_A \cdot (1+\text{sep}_\lambda^{-1}(A,B))^{n_A-1}} \geq \frac{2}{n_A}\left[\frac{\text{sep}_\lambda(A,B)}{2}\right]^{n_A}.$$

In other words, sep can not be any smaller than some constant multiple of $\text{sep}_\lambda^{n_A}$. After proving this, we will give an example showing that sep can indeed be as small as sep_λ^2. We believe that this is worst case behavior, but have not been able to prove it.

Proof: The first inequality in the theorem is just lemma 2.11. The proof of the other inequalities are very simple given the expression $\text{sep}(A,B)=||\Psi_{B^{-1}A}^{-1}||^{-1}$ from definition 2.4. This expression means that an upper bound on $||\Psi_{B^{-1}A}^{-1}||$ provides a lower bound on $\text{sep}(A,B)$. To compute such an upper bound, con-

sider the following form of $\Psi_{B,A}$ for $n_A=3$, which is analogous to the expression in (2.9):

$$\Psi_{B,A}=\begin{bmatrix} B-a_{11}\cdot I & -a_{21}\cdot I & -a_{31}\cdot I \\ & B-a_{22}\cdot I & -a_{32}\cdot I \\ & & B-a_{33}\cdot I \end{bmatrix} .$$

Thus,

$$\Psi_{B,A}^{-1}=\begin{bmatrix} (B-a_{11}\cdot I)^{-1} & a_{21}(B-a_{11}\cdot I)^{-1}(B-a_{22}\cdot I)^{-1} & a_{21}a_{32}(B-a_{11}\cdot I)^{-1}(B-a_{22}\cdot I)^{-1}(B-a_{33}\cdot I)^{-1}+a_{31}(B-a_{11}\cdot I)^{-1}(B-a_{33}\cdot I)^{-1} \\ & (B-a_{22}\cdot I)^{-1} & a_{31}(B-a_{22}\cdot I)^{-1}(B-a_{33}\cdot I)^{-1} \\ & & (B-a_{33}\cdot I)^{-1} \end{bmatrix} .$$

The largest number of $(B-a_{ii}\cdot I)^{-1}$ terms that are multiplied together is three, and they appear in the upper right corner of $\Psi_{B,A}^{-1}$. Similarly, for any n_A, the largest product of $(B-a_{ii}\cdot I)^{-1}$ terms contains n_A of them and occurs in the upper right corner of $\Psi_{B,A}^{-1}$. Now since $\|(B-a_{ii}\cdot I)^{-1}\|^{-1}$ is clearly an upper bound for sep_λ for any i, $\|(B-a_{ii}\cdot I)^{-1}\|$ is a lower bound for $\mathrm{sep}_\lambda^{-1}$ and so the norms of the block entries of $\Psi_{B,A}^{-1}$ are bounded above by $\mathrm{sep}_\lambda^{-n_A}$ times a constant. Thus $\|\Psi_{B,A}^{-1}\|$ is itself bounded above by $\mathrm{sep}_\lambda^{-n_A}$ times some constant $c_{n_A}^{-1}$ depending only on n_A. Taking reciprocals, we see that

$$\mathrm{sep} = \|\Psi_{B,A}^{-1}\|^{-1} \geq c_{n_A}\cdot \mathrm{sep}_\lambda^{n_A}$$

as desired.

More precisely, we use the block matrix norm

$$\|M\|_b = \max_i \sum_j \|M_{ij}\|$$

where M_{ij} is a square subblock of M. If there are n_A blocks M_{ij} in any row or column of M, then it is straightforward to show that

$$\|M\| \leq n_A \cdot \|M\|_b .$$

For example, when $n_A = \dim(M)$, then $\| M \|_\infty$ is the usual infinity norm (maximum absolute row sum norm) of M.

Now we estimate $\| \Psi_{B,A}^{-1} \|$. Since 1 is a common upper bound for the magnitudes of the off diagonal elements of A, the sum of the upper bounds of the norms of the blocks in the first row of $\Psi_{B,A}^{-1}$ is

$$\begin{aligned}
\| \Psi_{B,A}^{-1} \| &\le n_A \cdot \| \Psi_{B,A}^{-1} \|_\infty \\
&\le n_A \left(\mathrm{sep}_\lambda^{-1} + \sum_{i=0}^{n_A-2} \mathrm{sep}_\lambda^{-2} (1 + \mathrm{sep}_\lambda^{-1})^i \right) \\
&= n_A \left(\mathrm{sep}_\lambda^{-1} (1 + \mathrm{sep}_\lambda^{-1})^{n_A-1} \right) \\
&\le n_A \, 2^{n_A-1} \, \mathrm{sep}_\lambda^{-n_A}
\end{aligned}$$

(since $\mathrm{sep}_\lambda \le 1$). This is clearly also an upper bound on the sum of norms of the blocks in the other rows too. Q.E.D.

Since we are intertested in the case when sep_λ is very small, bounding it by 1 as we did in the last equation is rather conservative. A more realistic bound, true asymptotically for small sep_λ, may be derived as follows: from the expression for $\Psi_{B,A}^{-1}$, we see that the coefficient for the $\prod_{i=1}^{n_A} (B - a_{ii} \cdot I)^{-1}$ $(\mathrm{sep}_\lambda^{-n_A})$ term is $\prod_{i=1}^{n_A-1} a_{i+1,i}$. Since the a_{ij} satisfy $\sum_{ij} |a_{ij}|^2 \le 1$ we see that the last product can be at most $(n_A - 1)^{-(n_A-1)/2}$, implying that for small sep_λ, sep is approximately bounded below by $(n_A - 1)^{(n_A-1)/2} \, \mathrm{sep}_\lambda^{n_A}$.

As in the $\dim(A)=1$ case, knowledge of $\| P \|$ provides a lower bound on sep which in turn provides an upper bound on $\mathrm{diss}_g(\sigma_1, \sigma_2, path)$.

$$\mathrm{diss}_g(\sigma_1, \sigma_2, path) \le \sqrt{2} \, \mathrm{sep}_\lambda \le \sqrt{2} \left(n_A \, 2^{n_A-1} \, \mathrm{sep} \right)^{1/n_A}$$

$$\leq \sqrt{2}\, 2 \left(\frac{5}{2}\right)^{\frac{1}{5}} \left(\frac{1}{\sqrt{\|P\|^2-1}}\right)^{1/n_A}$$

$$\approx 3.40 \left(\frac{1}{\sqrt{\|P\|^2-1}}\right)^{1/n_A}$$

Thus, we have an upper bound on $\mathrm{diss}_E(\sigma_1, \sigma_2, path)$ essentially proportional to the n_A-th root of $1/\|P\|$.

Theorem 5.4 improves results of Kahan [Kahan1], Ruhe [Ruhe1], and Wilkinson [Wilkinson3].

Now we present an example to show that sep can indeed be as small as sep_λ^2. Consider the n by n matrices

$$A = \begin{bmatrix} \varepsilon & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \varepsilon \end{bmatrix} \quad \text{and} \quad B = -A^T .$$

A simple computation shows that the upper right corner of $\Psi_{A,B}^{-1}\flat$ (the largest element) is

$$|(\Psi_{A,B}^{-1}\flat)_{1,n^2}| = \frac{(2n-2)!}{2^{2n-1}(n-1)!^2}\varepsilon^{1-2n}$$

$$\approx \frac{\varepsilon^{1-2n}}{2\sqrt{\pi(n-1)}} .$$

(by Stirling's formula) which means sep is bounded below by the reciprocal of this quantity and above by n^2 times the reciprocal. Symmetry considerations show that $\lambda=0$ is the value which attains the minimum in the definition of sep_λ. Since the largest entry in A^{-1} is

$$|(A^{-1})_{1,n}| = \varepsilon^{-n}$$

we see that sep_λ is bounded below by ε^n and above by $n\varepsilon^n$. Thus, for fixed n and ε approaching zero, we see that sep is bounded below by sep_λ^2.

5.5 $\mathrm{diss}_2(\sigma_1, \sigma_2)$ and $\mathrm{diss}_F(\sigma_1, \sigma_2)$ for Normal Matrices

For normal matrices it turns out that we can compute $\mathrm{diss}_2(\sigma_1, \sigma_2)$ and $\mathrm{diss}_F(\sigma_1, \sigma_2)$ exactly. This is because a matrix is normal if and only if if can be diagonalized by a unitary similarity. Thus, the upper triangular form of T that has been our starting point is actually diagonal, and all projectors are orthogonal and hence of norm 1. Thus, the upper bound and lower bound we have been comparing in this chapter are equal and we have

Theorem 5.5: If T is normal, then

$$\begin{aligned}
&\mathrm{diss}_2(\sigma_1, \sigma_2, \mathit{path}) = \mathrm{diss}_2(\sigma_1, \sigma_2, \mathit{region}) \\
&= \mathrm{diss}_F(\sigma_1, \sigma_2, \mathit{path}) = \mathrm{diss}_F(\sigma_1, \sigma_2, \mathit{region}) \\
&= \frac{\min_{\lambda_i \in \sigma_i} |\lambda_1 - \lambda_2|}{2}
\end{aligned}$$

Proof: The expressions for $\mathrm{diss}_2(\sigma_1, \sigma_2, \mathit{region})$ and $\mathrm{diss}_2(\sigma_1, \sigma_2, \mathit{path})$ follow from the discussion of the previous paragraph. Now let T' be the 2 by 2 diagonal submatrix of T containing λ'_1 and λ'_2 on its diagonal, where λ'_1 and λ'_2 attain the minimum in the statement of the theorem. The claims for $\mathrm{diss}_F(\sigma_1, \sigma_2, \mathit{region})$ and $\mathrm{diss}_F(\sigma_1, \sigma_2, \mathit{path})$ follow directly from applying lemma 4.8 to T'. Q.E.D.

The perturbation of lemma 4.8 has several interesting properties: from the construction of lemma 4.8, we see that the perturbed matrix T' is defective, and hence not normal. Furthermore, even if T is real the T' may not be, since by perturbing only two eigenvalues, the eigenvalues of T' may very well no longer occur in complex conjugate pairs, which means it could not be real. Furthermore, no other perturbation of minimal Euclidean norm can yield a normal T' because the Wielandt-Hoffman theorem for normal matrices [Wilkinson2] shows that $\| T - T' \|_F$ must be at least $\sqrt{2}\mathrm{sep}_\lambda$. We exhi-

bit such a perturbation in the next paragraph.

If, however, we measure perturbations with $\|\cdot\|$ instead of $\|\cdot\|_E$, we can find a minimum norm perturbation yielding a T' which is normal, although it may not be real if T is. We accomplish this by using a rank 2 perturbation instead of the rank 1 perturbation of lemma 4.8. Simply observe that δT in

$$T + \delta T = \begin{bmatrix} a & \\ & b \end{bmatrix} + \begin{bmatrix} \frac{b-a}{2} & \\ & \frac{a-b}{2} \end{bmatrix} = \begin{bmatrix} \frac{a+b}{2} & \\ & \frac{a+b}{2} \end{bmatrix} = T'$$

has 2-norm $|a-b|/2 = \mathrm{diss}_2(\sigma_1, \sigma_2)$, Euclidean norm $\sqrt{2}\mathrm{diss}_E(\sigma_1, \sigma_2)$, and rank 2. Applying this construction to the two closest eigenvalues λ'_1 and λ'_2 of σ_1 and σ_2 clearly produces a normal T'. This T' may not be real if T is, however. This may occur if λ'_1 and λ'_2 are complex but not complex conjugate pairs, because then the eigenvalues of T' will not occur in complex conjugate pairs, a necessary condition for being real. Sometimes we can still find a real T' if this happens: if both λ'_i have nonzero imaginary parts, then perturb their conjugates $\bar{\lambda}'_1$ and $\bar{\lambda}'_2$ to coalesce also. Since the eigenvector(s) belonging to any λ is(are) the complex conjugate(s) of the eigenvector(s) belonging to $\bar{\lambda}$, these two rank 2 perturbations are easily seen to be complex conjugates so their sum, the total perturbation, is real. If, however, one of the λ'_i is real, we may not be able to find a real T': consider

$$T = \begin{bmatrix} 0 & & \\ & 0 & 1 \\ & -1 & 0 \end{bmatrix}$$

with $\sigma_1=\{0\}$ and $\sigma_2=\{i,-i\}$. The only way for σ_1 and σ_2 to overlap and still have complex conjugate pairs is for all three eigenvalues to be real. By the Wielandt-Hoffman theorem [Kato2] this requires a perturbation of Euclidean norm at least $\sqrt{2}$, and hence a 2-norm of at least $\sqrt{(2/3)}$, whereas $\mathrm{diss}_2(\sigma_1, \sigma_2)=1/2$.

If T is Hermitian, we can say still more. Applying what we said about normal matrices, we see a minimum $\|\cdot\|_E$ perturbation yields a defective and hence nonhermitian T', but T' is clearly real if T is. The rank 2 perturbation above also produces a Hermitian T' which must be real if T is.

Chapter 6: A Probabilistic Model

6.1 Introduction

In the last chapter we presented upper and lower bounds on $\mathrm{diss}_E(\sigma_1, \sigma_2)$ and examples which showed that they could be equal or arbitrarily far apart. We did not provide any insight as to when they were likely to be close or distantly separated. We will provide this insight in this chapter and the next by filling in the details of the following description: There is a surface in the space of matrices such that our upper and lower bounds on $\mathrm{diss}_E(\sigma_1, \sigma_2)$ are far apart only for matrices within a small relative distance ε of the surface. (We say that a matrix M is within relative distance ε of a surface if there is a δM such that $M+\delta M$ is on the surface and $\|\delta M\|_E \le \varepsilon \cdot \|M\|_E$.) In fact, we will see that the closer to the surface, the farther apart the bounds. Furthermore, we can compute an asymptotic upper bound $n(n+1)(n-1)^2 \cdot \varepsilon^2$ on the fraction of the volume of the set of complex matrices that lie within relative distance ε of this surface (n is the dimension of the matrix). This upper bound shows that the volume of the set of points within ε of the surface goes to zero as ε goes to zero. If we interpret this fraction of the volume of a set of matrices as its probability then we may state our result as follows: the probability that the ratio of the upper bound to the lower bound is at most $K>1$ is at least $1-n(n+1)(n-1)^2 \cdot K^{-2} + o(K^{-2})$. In other words, the ratio of the bounds is large with a low probability. This same sort of description applies to our bounds for sep_λ in terms of sep, (the bound is accurate except within a small distance of a particular surface in matrix space, and the volume of points within ε of the surface goes to zero as a polynomial in ε), to the probability of being able to completely diagonalize a given matrix, and to other similar quantities which we will discuss in

chapter 7.

These results will follow from a more general theorem which we will prove in this chapter. The general problem is estimating the volume of points within distance ε of certain surfaces: homogeneous varieties. A homogeneous variety is the locus of solutions of a set of simultaneous polynomial equations which have the property that if $(x_1, \ldots, x_k)$ lies on the surface, so does $(\alpha x_1, \ldots, \alpha x_k)$ for any scalar α. Since we are interested in relative distance, it suffices to estimate the volume of points on the sphere of matrices of Frobenius norm 1 that lie within distance ε of a homogeneous variety. It is a remarkable fact that, for complex matrices, this volume can be computed to first order in terms of only two parameters of the variety: its dimension and its degree. The main result of this chapter is

Theorem 6.3: Let V be a complex purely $2n$ dimensional homogeneous variety of degree $\deg(V)$ in $\mathbf{C}^N$, with $n>0$. Then the fraction of unit sphere in $\mathbf{C}^N$ within Euclidean distance ε of V is

$$\binom{N-1}{n-1} \deg(V)\, \varepsilon^{2(N-n)} + o\left(\varepsilon^{2(N-n)}\right)$$

($\binom{N-1}{n-1}$ is the binomial coefficient $(N-1)!/\left((n-1)!\cdot(N-n)!\right)$).

Interpreting the fraction of area as a probability, we may restate this as

$$\mathrm{Prob}(\mathrm{dist}_E(M, V) \le \varepsilon) = \binom{N-1}{n-1} \deg(V)\, \varepsilon^{2(N-n)} + o\left(\varepsilon^{2(N-n)}\right) ,$$

where M is uniformly distributed on the surface of the unit sphere.

Thus, to first order the probability depends on only two parameters of the variety: the dimension and the degree. This simple result makes it easy to compute the probability in many interesting cases; we discuss singular matrices, defective matrices, and polynomials with multiple roots in the next chapter.

We may extend this result to real varieties, but we only obtain an upper bound on the probability:

Theorem 6.6: Let V be an n-dimensional homogeneous real variety of degree $\deg(V)$ in $\mathbf{R}^N$, where $n>0$. Then the fraction of the unit sphere in $\mathbf{R}^N$ within Euclidean distance ε of V, (or equivalently $\mathrm{Prob}(\mathrm{dist}_F(M\ ,\ V) \le \varepsilon)$ where M is uniformly distributed on the sphere), is less than or equal to

$$\frac{\Gamma(\frac{N+1}{2})\cdot\Gamma(\frac{N}{2})}{\pi^{\frac{N-n-1}{2}}\cdot\Gamma(\frac{N-n+1}{2})\cdot\Gamma(\frac{n+1}{2})\cdot\Gamma(\frac{n}{2})}\cdot\deg(V)\cdot\varepsilon^{N-n}+o(\varepsilon^{N-n}) \quad (6.1)$$

In Section 6.2 we will define the terminology just used, and state the theorems from geometry and algebra we need to prove our result. Specifically, we will use Weyl's theorem [Weyl, Griffiths] on volumes of spherical neighborhoods, and Lelong's theorem [Lelong, Thie, Griffiths] on the area of a homogeneous variety. In Section 6.3 we state and prove the main result, and show that the probabilistic interpretation holds for a large class of probability distributions on the set of matrices. In Section 6.4 we state Bézout's theorem [Kendig] on the degree of intersection of varieties, which we use to compute upper bounds on $\deg(V)$. In Section 6.5 we extend our results to real varieties by using Crofton's formula [Santaló, Griffiths].

Our approach was motivated by a similar analysis of varieties in the space of polynomials due to Smale [Smale].

6.2 Notation and Lemmas from Geometry and Algebra

To prove our main result we will need several theorems from geometry and algebra. The central result we need from geometry is Weyl's theorem [Weyl,Griffiths] which says that the volume of a spherical neighborhood (defined below) of a manifold of radius ε is well approximated by a polynomial

in ε for small ε. The dominating (lowest order) term of this polynomial contains, not surprisingly, the area of the manifold as a factor. Our central algebraic result is that the area of that part of a complex homogeneous variety within the unit ball is equal to the degree of the variety times a constant which depends only on the dimension of the variety [Lelong, Thie, Griffiths]. We will present several ways of computing the degree of a variety in Section 6.4.

Before discussing Weyl's theorem, we need several definitions from geometry. (See [Guillemin and Pollack] for more details). A subset M of Euclidean space $\mathbf{R}^N$ is called an *n-dimensional manifold* if it is locally homeomorphic to $\mathbf{R}^n$. $m=N-n$ is called the *codimension* of M and is denoted codim(M). M is called a *smooth manifold* if the homeomorphism and its inverse are infinitely differentiable and an *analytic manifold* if they are analytic.

By *l-volume* of an n-dimensional manifold M ($l \geq n$) we mean the l-dimensional Lebesgue measure of M, if it exists. Note that if $l > n$ this volume is zero. The notation vol(M) denotes the n-volume of the n-dimensional manifold M.

An *ε-tubular neighborhood of M*, denoted $\tau_\varepsilon(M)$, is, loosely speaking, the set of all points of $\mathbf{R}^N$ within Euclidean distance ε of M. More precisely, it is constructed as follows: for each $x \in M$, consider the space of all vectors starting at x and perpendicular to M (the *normal space* at x). The endpoints of all such vectors starting at x and of length at most ε form a closed disk of radius ε and center x perpendicular to M; the union of all these disks forms $\tau_\varepsilon(M)$. Moreover, each point in $\tau_\varepsilon(M)$ is required to lie in exactly one such disk. In other words, it must be possible to draw exactly one line segment

from any $y \in \tau_\varepsilon(M)$ to M that is perpendicular to M and of length at most ε. This constraint means that some manifolds only allow tubular neighborhoods for ε smaller than some bound; for example, a circle of radius r only has tubular neighborhoods for $\varepsilon < r$, because otherwise the center of the circle is not a distance $r \leq \varepsilon$ away from a *unique* point on the circle. In addition, some manifolds may not have an ε-tubular neighborhood at all (see figure 6.1).

If M lies entirely within the unit sphere $\mathbf{S}^{N-1}$ in $\mathbf{R}^N$, we may analogously define an *ε-spherical neighborhood* of M, denoted $\tau_\varepsilon^S(M)$, to be the set of points of $\mathbf{S}^{N-1}$ within spherical distance $\varepsilon < \pi$ of M. That is, the length of the great circle connecting any $y \in \tau_\varepsilon^S(M)$ to the nearest point $x \in M$ is at most ε. We construct $\tau_\varepsilon^S(M)$ as follows: construct the normal disks to M lying in $\mathbf{R}^N$ as above, but of radius $2 \sin \varepsilon/2$ rather than ε. Let $\tau_\varepsilon^S(M)$ be the intersection of the sphere $\mathbf{S}^{N-1}$ with the union of these disks, subject to the same constraints as before. Some simple trigonometry shows that these disks intersect the sphere in points at spherical distance at most ε from M. Note that ε must be less than π for an ε-spherical neighborhood to exist, since there are two geodesics of length π perpendicular to M connecting any $x \in M$ to its antipodal point.

Now we can state Weyl's theorem for the volume of ε-spherical neighborhoods. Let

$$\omega_m = \frac{2\pi^{(m+1)/2}}{\Gamma((m+1)/2)} \tag{6.2}$$

denote the surface area of the unit sphere $\mathbf{S}^m$ in $\mathbf{R}^{m+1}$ (i.e. the m-volume), and

$$\Omega_m = \frac{2\pi^{m/2}}{m\Gamma(m/2)} = \omega_{m-1}/m \tag{6.3}$$

denote the volume of the unit ball in $\mathbf{R}^m$ [Santaló].

Theorem 6.1 (Weyl): Let M be a smooth n-dimensional manifold in S^{N-1}, and $\tau_\varepsilon^S(M)$ an ε-spherical neighborhood of M. Let $m = N-1-n$ be the codimension of M (as a subset of S^{N-1}). Then

$$\mathrm{vol}(\tau_\varepsilon^S(M)) = \omega_{m-1} \cdot \sum_{\substack{0 \le e \le n \\ e \text{ even}}} k_e(M) \cdot J_e(\varepsilon) \ , \tag{6.4}$$

where the $k_e(M)$ are the integrals of certain differential forms over M, $k_0(M)$ is the volume of M, and

$$J_e(\varepsilon) \equiv \frac{\displaystyle\int_0^{\tan\varepsilon} \frac{r^{m+e-1}\,dr}{(1+r^2)^{N/2}}}{m\cdot(m+2)\cdots(m+e-2)} \tag{6.5}$$

(if $e=0$ the denominator in $J_e(\varepsilon)$ is 1).

First we will discuss the behavior of $J_e(\varepsilon)$ as a function of e and ε, and then we will discuss the geometric meaning of the theorem.

It is possible to evaluate the integral defining $J_e(\varepsilon)$ exactly, but since later approximations render the higher order terms in ε useless, we only consider its behavior for small ε. Thus, we need only examine the behavior of the integrand for small r as well, for which it obviously equals

$$\frac{r^{e+m-1}}{(1+r^2)^{N/2}} = r^{e+m-1} + O(r^{e+m+1})$$

so that

$$J_e(\varepsilon) = \frac{\displaystyle\int_0^{\tan\varepsilon} (r^{e+m-1} + O(r^{e+m+1}))\,dr}{m\cdot(m+2)\cdots(m+e-2)}$$

$$= \frac{\tan^{e+m}\varepsilon}{m\cdot(m+2)\cdots(m+e)} + O(\tan^{e+m+2}\varepsilon)$$

$$= \frac{\varepsilon^{e+m}}{m\cdot(m+2)\cdots(m+e)} + O(\varepsilon^{e+m+2})$$

(the last equality follows since $\tan\varepsilon = \varepsilon + O(\varepsilon^3)$).

Combining this last estimate with Weyl's theorem yields:

$$\mathrm{vol}(\tau_\varepsilon^S(M)) = \frac{\omega_{m-1}}{m}\mathrm{vol}(M)\,\varepsilon^m + O(\varepsilon^{m+2}) \tag{6.6}$$

$$= \Omega_m\ \mathrm{vol}(M)\,\varepsilon^m + O(\varepsilon^{m+2})\ .$$

What is the geometric meaning of this last expression? Take, for example, $N=3$ and $n=0$; this corresponds to M being a set of k distinct points on the unit sphere S^2 in R^3. $\mathrm{vol}(\tau_\varepsilon^S(M))$ is then just the area of k spherical caps centered at the points of M and each of radius ε. For small ε, this area is approximately $k\pi\varepsilon^2$, which is what (6.6) gives after plugging in $m=N-1-n=2$, $\mathrm{vol}(M)=k$, and $\Omega_m=\pi$. Similarly, we may take $N=3$ and $n=1$ corresponding to a one-dimensional curve M of length l, say, on S^2. $\mathrm{vol}(\tau_\varepsilon^S(M))$ is then just the area of a strip surrounding M of length l and width 2ε, which is clearly approximated by $2\cdot l\cdot\varepsilon$. Plugging $m=1$, $\Omega_m=2$, and $\mathrm{vol}(M)=l$ into (6.6) yields the same expression. In these simple examples Weyl's theorem tells us the intuitive fact that the volume of the spherical neighborhood is approximately equal to the volume of an m-ball of radius ε ($\Omega_m\ \varepsilon^m$) times the volume of M; Weyl's theorem extends the intuition of these small examples to higher dimensions.

There is also a version of Weyl's theorem for tubular neighborhoods [Weyl, Griffiths]. It says that the volume of $\tau_\varepsilon(M)$ is a polynomial in ε of degree at most N (where $M\subset R^N$), and that the lowest order (dominating) term in ε is $\Omega_{N-n}\ \mathrm{vol}(M)\,\varepsilon^{N-n}$ as expected. We will not use this version of Weyl's theorem.

Next we discuss Lelong's theorem on the volume of a homogeneous pure dimensional complex algebraic variety. First we need several definitions from algebraic geometry. (See [Kendig] for more details). A *variety* V is the zero

set of a collection $\{p_\alpha(z_1, \ldots, z_n)\}$: of polynomials:

$$V = \{(z_1, \ldots, z_n) \mid p_\alpha(z_1, \ldots, z_n)=0 \text{ for all } \alpha\} .$$

V is called is called real or complex according to whether the z_i are real or complex. Since V can in general have points of self intersection, it is generally not a manifold, since it is not homeomorphic to Euclidean space in the neighborhood of an intersection point. However, points q with relatively open neighborhoods $U_q \subset V$ that are analytic manifolds are dense in V [Theorem 4.2.4, Kendig] so that the following definition makes sense: the *dimension of V at p*, written $\dim_p(V)$, is

$$\dim_p(V) \equiv \limsup_{\substack{q \to p \\ q \in U_q \subset V \\ U_q \text{ a manifold}}} \dim(U_q) .$$

We define in turn the *dimension of V* as the maximum over all $p \in V$ of $\dim_p(V)$. V is called *pure dimensional* or *purely n-dimensional* if $\dim_p(V)=n$ for all $p \in V$. When we refer to the dimension of anything in this thesis, we will always mean the *real* dimension, i.e. the the dimension as a real manifold or variety, rather than the *complex* dimension often used for complex objects, which is exactly half the real dimension. To emphasize that we are dealing with a complex variety, we will write its real dimension as $2n$.

We call V *homogeneous* if it is a cone; that is if $(z_1, \ldots, z_n) \in V$ implies $(\alpha z_1, \ldots, \alpha z_n) \in V$ for all scalars α (real scalars if V is a real variety, complex if V is complex). In terms of the defining polynomials $\{p_\alpha\}$ this means that if

$$p_\alpha(z_1, \ldots, z_n) = \sum k_j z_1^{i_1^{(j)}} \times \cdots \times z_n^{i_n^{(j)}} ,$$

then

$$\sum_{k=1}^{n} i_k^{(j)} = d$$

where d does not depend on j. d is called the *order* of p_α, and written order(p). We say order instead of degree because we use degree to for the

more general concept in the next paragraph.

We define the *degree* of a purely $2n$-dimensional homogeneous complex variety V in $\mathbf{C}^N$ as follows. Let L^{2N-2n} be a $2N-2n$ dimensional linear manifold (plane) in $\mathbf{C}^N$. Since $\dim(L^{2N-2n}) + \dim(V) = \dim(\mathbf{C}^N) = 2N$, we say that L^{2N-2n} and V are of *complementary dimension*. Generically, L^{2N-2n} and V will intersect in a surface of codimension equal to the sum of their codimensions, that is $2N$. In other words, their intersection will be of dimension 0 (a finite collection of points) for almost all planes L^{2N-2n}. It turns out that for almost all L^{2N-2n}, this collection will contain the same number of points, and this common number is called the *degree of* V, and is written $\deg(V)$. (see [theorem 4.6.2, Kendig]). Intuitively, $\deg(V)$ gives the number of "leaves" of the variety V that a typical plane L^{2N-2n} will intersect.

Now we can state Lelong's theorem [Lelong, Thie, Griffiths] (or more precisely just the special case we need):

Theorem 6.2 (Lelong): Let V be a purely $2n$-dimensional homogeneous complex variety in $\mathbf{C}^N$, where $n>0$. Let $V[r]$ denote that part of V contained in $B_N(r)$ (the N-ball of radius r centered at the origin). Then the volume of $V[r]$ is given by

$$\mathrm{vol}(V[r]) = \Omega_{2n} \cdot \deg(V) \cdot r^{2n} \quad . \tag{6.7}$$

This remarkable theorem says the following: the volume of $V[r] = V \cap B_N(r)$ is identical to the volume of $B_N(r)$ intersected with the variety consisting simply of $\deg(V)$ *planes* of dimension $2n$ passing through the origin. This theorem makes the computation of the leading term in the expression for volume in Weyl's theorem simple, given the ability to compute $\deg(V)$. The preparation for proving the main result in the next section is now complete.

6.3 The Volume of a Spherical Neighborhood of a Complex Homogeneous Variety

The main result of this chapter is the following theorem. After the proof, which is quite easy given the preparation of the last section, we will discuss it.

Theorem 6.3: Let V be a complex purely $2n$ dimensional homogeneous variety of degree $\deg(V)$ in $\mathbf{C}^N$, where $n>0$. Then the fraction of unit sphere in $\mathbf{C}^N$ within Euclidean distance ε of V is

$$\binom{N-1}{n-1} \deg(V)\, \varepsilon^{2m} + o(\varepsilon^{2m}) \ , \tag{6.8.a}$$

where $2m=2N-2n$ is the codimension of V.

Interpreting the fraction of area as a probability, we may restate this as

$$\mathrm{Prob}(\mathrm{dist}_E(M,V) \le \varepsilon) = \binom{N-1}{n-1} \deg(V)\, \varepsilon^{2m} + o(\varepsilon^{2m}) \ , \tag{6.8.b}$$

where M is uniformly distributed on the surface of the unit sphere.

Proof: The surface to which we would like to be able to apply Weyl's theorem is $V' \equiv V \cap S^{2N-1}$, the intersection of V and the unit sphere in $\mathbf{C}^N$. From Figure 6.2 we see that a point $x \in S^{2N-1}$ is within Euclidean distance ε of V if and only if it is within spherical distance $\arcsin \varepsilon = \varepsilon + O(\varepsilon^3)$ of V'. Thus, the spherical neighborhood whose volume we would like to measure (and divide by $\mathrm{vol}(S^{2N-1})$ to get the fraction of S^{2N-1}) is $\tau^S_{\arcsin \varepsilon}(V')$. Lelong's theorem tells us $\mathrm{vol}(V[1])$ in terms of $\deg(V)$; it will turn out that $\mathrm{vol}(V') = 2n \cdot \mathrm{vol}(V[1])$. Thus, if $\tau^S_{\arcsin \varepsilon}(V')$ existed, we could use equations (6.1) through (6.4) to compute

$$\frac{\mathrm{vol}(\tau^S_{\arcsin \varepsilon}(V'))}{\mathrm{vol}(S^{2N-1})} = \frac{\Omega_{2m} \cdot \mathrm{vol}(V') \cdot (\arcsin \varepsilon)^{2m}}{\omega_{2N-1}} + O((\arcsin \varepsilon)^{2m+2})$$

$$= \Omega_{2m} \cdot 2n \cdot \Omega_{2n} \cdot \deg(V) \cdot \varepsilon^{2m} \,/\, \omega_{2N-1} + O(\varepsilon^{2m+2})$$

$$= \binom{N-1}{n-1} \cdot \deg(V) \cdot \varepsilon^{2m} + O(\varepsilon^{2m+2}) \tag{6.9}$$

and be done. Unfortunately, $\tau^{S}_{\text{arcsin}\,\varepsilon}(V)$ does not exist for reasons we will now discuss. Even so, we will show that the result of the above computation is valid provided we replace $O(\varepsilon^{2m+2})$ by $o(\varepsilon^{2m})$.

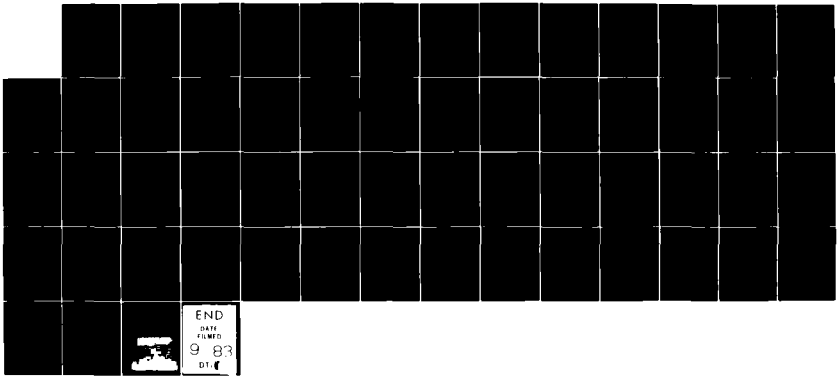
Since V is a variety, it will in general intersect itself and not be a manifold. If we remove the set V_s of these intersection points from V the remainder $V_{ns} \equiv V - V_s$ is a manifold [Griffiths]. Furthermore, $\dim(V_{ns}) = \dim(V)$, and V_{ns} is a connected, open and dense subset of V, so we lose nothing by considering V_{ns} instead of V [Kendig]. Since any homogeneous set like V_{ns} intersects the sphere S^{2N-1} transversally, the intersection $V_{ns}' = V_{ns} \cap S^{2N-1}$ is also a manifold of dimension one less than V_{ns}. Unfortunately, V_{ns}' also will not generally possess spherical neighborhoods because it can contain "pinched sections" as illustrated in Figure 6.1. However, if we remove the open set of all points within Euclidean distance η of V_s from V_{ns}', what remains will be a compact set $V_{ns}(\eta)$ which does have an ε neighborhood for ε less than a threshold $\bar{\varepsilon}(\eta)$ which may go to zero as η does. As η approaches zero, the volume of $V_{ns}(\eta)$ approaches the volume of V_{ns}, and the ratio of the volume of all points within ε of $V_{ns} - V_{ns}(\eta)$ to the volume of all points within ε of $V_{ns}(\eta)$ goes to zero. Thus, the estimate in (6.9) remains valid if we replace $O(\varepsilon^{2m+2})$ by $o(\varepsilon^{2m})$.

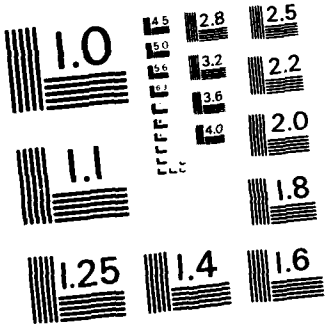
It remains to show that $\text{vol}(V) = 2n \cdot \text{vol}(V[1])$. We show more generally that if V is the union of d-dimensional cones, then $\text{vol}(V) = d \cdot \text{vol}(V[1])$. This follows from expressing $\text{vol}(V[1])$ as the integral in spherical coordinates of the volumes of concentric spherical sections of V:

$$\text{vol}(V[1]) = \int_0^1 \rho^{d-1}\, \text{vol}(V)\, d\rho$$

AD-A130 775 A NUMERICAL ANALYST'S JORDAN CANONICAL FORM(U) 2/2
CALIFORNIA UNIV BERKELEY CENTER FOR PURE AND APPLIED
MATHEMATICS J W DEMMEL MAY 83 PAM-156 N00014-76-C-0013

UNCLASSIFIED F/G 12/1 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS - 1963 - A

$$= \mathrm{vol}(V)/d \ .$$

This completes the proof. Q.E.D.

Our first remark concerns the likely size of the error term $o(\varepsilon^{2m})$. Since we are adding volumes of the form

$$\mathrm{const} \cdot \varepsilon^{2m} + O(\varepsilon^{2m+2}) \ ,$$

it is clear that if the first term in (6.9) is an underestimate, it is still an underestimate by at most $O(\varepsilon^{2m+2})$. Thus $o(\varepsilon^{2m})$ represents an error bounded above by $O(\varepsilon^{2m+2})$. More troublesome is the possibility that the first term of (6.9) is an overestimate. Consider the V of Figure 6.3, which is the union of parts of the plane curves $y=0$ and $y=x^{2n}$ (V is clearly not a homogeneous complex variety, but it illustrates our point). If we add the areas of the ε-tubular neighborhoods of these two parts, the sum overestimates the area we want to measure by the area of the shaded region, which is doubly covered by the two ε-neighborhoods. The area of this doubly covered region is approximately

$$\int_{-(2\varepsilon)^{1/(2n)}}^{(2\varepsilon)^{1/(2n)}} (2\varepsilon - x^{2n})\, dx = O(\varepsilon^{1+1/(2n)}) \ .$$

The area we want is clearly dominated by a linear term in ε, so we see that the overestimate depends on $2n$, the degree of V.

Our second remark concerns the probabilistic interpretation of (6.7.b). Instead of choosing a random point M uniformly distributed on S^{2N-1}, we consider choosing a random point M according to the density p and ask about the distribution $\mathrm{Prob}(\mathrm{dist}_E(M/\|M\|_E\ ,\ V) \le \varepsilon)$, where $\|M\|_E$ is the Euclidean norm of M (Frobenius norm if M is a matrix) so that $M/\|M\|_E$ must lie on S^{2N-1}. As long as the random variable $M/\|M\|_E$ is unformly distributed on S^{2N-1}, $\mathrm{Prob}(\mathrm{dist}_E(M/\|M\|_E\ ,\ V) \le \varepsilon)$ will still be given by the

expression in (6.7.b). Thus, we may apply our result to any density p for M which causes $M/\|M\|_E$ to be uniform on S^{2N-1}. A large class of such densities p is simply characterized by the symmetry condition $p(M) = f(\|M\|_E)$, i.e. that the density function p is really only a function of $\|M\|_E$. Two well known such densities are

$$p(M) = \begin{cases} \Omega_{2N}^{-1} & \text{if } \|M\|_E < 1 \\ 0 & \text{otherwise} \end{cases},$$

the uniform density on the interior of the unit ball, and

$$p(M) = (2\pi)^{1/N} \cdot e^{-\|M\|_E^2/2},$$

the normal density on $\mathbf{C}^N$ (i.e. each component of M is an independent Gaussian random variable with mean 0 and variance 1).

6.4 Estimating deg(V) of a Complex Homogeneous Variety

Next we turn to the problem of computing deg(V). There are two tools from algebra we will use. Both are standard results in algebraic geometry and can be found in [Chap 4, Kendig].

Theorem 6.4: If the complex homogeneous variety V is defined as the zero set of the single homogeneous polynomial p, then codim(V)=2, and V is called a ***hypersurface***. If in addition p is the product of distinct irreducible factors, then deg(V)=order(p).

Since several interesting varieties we encounter later are defined by a single irreducible polynomial whose order we know, this theorem supplies all data needed to compute the volumes of their spherical neighborhoods to first order.

Our second tool is a slightly nonstandard version of a well known theorem:

Theorem 6.5 (Bézout): Let V be a complex homogeneous variety given as the zero set of the finite collection of homogeneous polynomials $\{p_i\}_{i=1,m}$. Then we can bound $\deg(V)$ as follows:

$$1 \le \deg(V) \le \prod_{i=1}^{m} \operatorname{order}(p_i) \ . \tag{6.10}$$

The standard version of this theorem says that if varieties V_i, each defined by the single polynomial p_i, intersect *transversally*, that is, if

$$\operatorname{codim}(\bigcap_{i=1}^{m} V_i) = \sum_{i=1}^{m} \operatorname{codim}(V_i) \ ,$$

then $\deg(V)$ actually equals the product in (6.10). It is no surprise that this product provides an upper bound when the V_i do not intersect transversally. Unfortunately, it seems to give an atrocious upper bound on some occasions, but it is simple to compute.

6.5 Real Varieties

To extend the results of previous sections to real varieties, we need to estimate the volume of a real variety. The difficulties in doing this estimate are illustrated by the following example. Consider the variety $V(a,b,c)$ defined by the polynomial $p = ax^2 + by^2 + cz^2$, where neither a nor b nor c is zero. If x, y, and z were complex, theorem 6.4 would imply that V had codimension 2 and degree 2, and so by Lelong's theorem $\operatorname{vol}(V[1])$ would be $2\Omega_4 = \pi^2$, independent of a, b, and c (as long as they are nonzero). For x, y, and z real, we have the following possibilities, among others: If $a = b = 1$ and $c < 0$, then V is a circular cone with codimension 1 and area $2\pi\sqrt{c/(c-1)}$, which approaches 0 when c does, and approaches 2π as c goes to $-\infty$. If a, b and c all have the same sign, V degenerates to a single point at the origin.

Despite these problems, it turns out we can still derive an upper bound on the volume of a real variety V given only its dimension and degree, where

by degree we mean the largest finite number of intersection points of V and almost all planes L of complementary dimension. This bound uses Crofton's formula [Santaló]. We can use this bound in turn to extend Theorem 6.3 to homogeneous real varieties as follows:

Theorem 6.6: Let V be an n-dimensional homogeneous real variety of degree $\deg(V)$ in $\mathbf{R}^N$, where $n>0$. Then the fraction of the unit sphere in $\mathbf{R}^N$ within Euclidean distance ε of V is less than or equal to

$$\frac{\Gamma(\frac{N+1}{2})\cdot\Gamma(\frac{N}{2})}{\pi^{\frac{N-n-1}{2}}\cdot\Gamma(\frac{N-n+1}{2})\cdot\Gamma(\frac{n+1}{2})\cdot\Gamma(\frac{n}{2})}\cdot\deg(V)\cdot\varepsilon^m + o(\varepsilon^m) \quad (6.11)$$

where $m=N-n$ is the codimension of V.

Since this theorem includes pure dimensional homogeneous complex varieties as a special case, it provides an upper bound to the result in theorem 6.3:

Corollary 6.7: Let V be a purely $2n$-dimensional homogeneous complex variety of degree $\deg(V)$ in $\mathbf{C}^N$, where $n>0$. Then the fraction of the unit sphere in $\mathbf{C}^N$ within Euclidean distance ε of V is less than or equal to

$$\left[\frac{\binom{2N}{2n}}{\binom{N}{n}}\right]\cdot\binom{N-1}{n-1}\cdot\deg(V)\cdot\varepsilon^{2m} + o(\varepsilon^{2m}) = \binom{2N}{2n}\frac{n}{N}\cdot\deg(V)\,\varepsilon^{2m} + o(\varepsilon^{2m}) \ .$$

Proof: Convert the gamma functions in Theorem 6.6 to factorials.

Thus, this estimate is too big by the factor $\binom{2N}{2n}/\binom{N}{n}$ for complex homogeneous varieties. We will see why this factor appears later from Crofton's formula.

Proof of Theorem 6.6: Just as Lelong's theorem provides an estimate of $\mathrm{vol}(V[r]) = \mathrm{vol}(V\cap B_N(r))$ for a complex homogeneous variety V, Crofton's

formula lets us estimate $\mathrm{vol}(V[r])$ for real varieties. Given this estimate (Lemma 6.8), the rest of the proof is identical to that of Theorem 6.3.

Actually, Lemma 6.8 applies to much more general objects than varieties (note that the definition of $\deg(V)$ makes sense for V a union of manifolds):

Lemma 6.8: Let V be a countable union of manifolds in $\mathbf{R}^N$ of dimension n or lower such that $\deg(V)$ is finite. Let $V[r]=V\cap B_N(r)$ be that part of V within the ball of radius r. Then

$$\mathrm{vol}(V[r]) \le \left[\frac{\Gamma\left(\frac{N+1}{2}\right)\cdot\sqrt{\pi}}{\Gamma\left(\frac{N-n+1}{2}\right)\cdot\Gamma\left(\frac{n+1}{2}\right)}\right]\cdot\Omega_n\cdot\deg(V)\cdot r^n \quad . \tag{6.12}$$

Since this theorem includes complex pure dimensional homogeneous varieties as a special case, it provides an upper bound on the value of $\mathrm{vol}(V[r])$ given by Lelong's theorem:

Corollary 6.9: Let V be $2n$ dimensional in $\mathbf{C}^N$, and otherwise as described in Theorem 6.6. Then

$$\mathrm{vol}(V[r]) \le \left[\frac{\binom{2N}{2n}}{\binom{N}{n}}\right]\cdot\Omega_{2n}\cdot\deg(V)\cdot r^{2n} \quad . \tag{6.13}$$

Proof: Convert the gamma functions in Theorem 6.6 to factorials.

Thus, this estimate is too large by the factor $\binom{2N}{2n}/\binom{N}{n}$ for complex homogeneous pure dimensional varieties, which is the source of the overestimate in Corollary 6.7. Nevertheless, when $V=S^{N-1}$ in $\mathbf{R}^N$ (so $\deg(V)=2$) and $r=1$, the expression for $\mathrm{vol}(V[1])$ given by Theorem 6.8 is exact, so we see we have traded generality of the hypotheses (unions of manifolds instead of

homogeneous varieties) for tightness of the bound.

Proof of Lemma 6.8: The proof follows easily from several results of integral geometry, all of which can be found in [Santaló]. We assume without loss of generality that V is a manifold; the general result follows by applying the following analysis to each constituent manifold of V and adding the bounds.

Crofton's formula [equation 14.70, Santaló] expresses the volume of an n-manifold V in $\mathbf{R}^N$ in terms of an integral:

$$\mathrm{vol}(V) = \frac{\prod_{i=1,N-n} \omega_i}{\prod_{i=n+1,N} \omega_i} \int \#(L^{N-n} \cap V)\, dL^{N-n} \quad . \tag{6.14}$$

The integral is over all $N-n$ dimensional planes L^{N-n} where dL^{N-n} is the ***kinematic density*** for $N-n$ planes. This means that the measure of a set of planes is invariant under the group of rigid motions in $\mathbf{R}^N$. $\#(L^{N-n} \cap V)$ is the number of points in the intersection of L^{N-n} and V; by hypothesis this is a nonnegative integer bounded above by $\deg(V)$ for almost all L^{N-n}. Thus

$$\mathrm{vol}(V) \le \deg(V)\, \frac{\prod_{i=1,N-n} \omega_i}{\prod_{i=n+1,N} \omega_i} \int_{L^{N-n} \cap V \neq \phi} dL^{N-n} \quad . \tag{6.15}$$

Applying this to $V[r]$ instead of V, and noting that L^{N-n} can intersect $V[r]$ only if it intersects $B_N(r)$, we see that

$$\mathrm{vol}(V) \le \deg(V)\, \frac{\prod_{i=1,N-n} \omega_i}{\prod_{i=n+1,N} \omega_i} \int_{L^{N-n} \cap B_N(r) \neq \phi} dL^{N-n} \quad . \tag{6.16}$$

The integral in the last equation is known as a "cross-sectional integral", (quermassintegral: [Chap 13.6, Santaló]) because it gives the measure of the set of planes which slice $B_N(r)$. From equations (14.1) and (13.46) of Santaló, we find

$$\int_{L^{N-n}\cap B_N(r)\neq\phi} dL^{N-n} = \frac{r^n}{n}\cdot\frac{\prod_{i=n-1,N-1}\omega_i}{\prod_{i=0,N-n-1}\omega_i}\ . \tag{6.17}$$

Substituting this in the last inequality yields

$$\mathrm{vol}(V) \le \deg(V)\cdot r^n\cdot\frac{\omega_{N-n}\,\omega_n\,\omega_{n-1}}{n\,\omega_0\,\omega_N} \tag{6.18}$$

which after some manipulation using equation 6.1 yields the bound of the lemma. Q.E.D. of Lemma 6.8 and Theorem 6.6.

We turn now to estimating the dimension and degree of a homogeneous real variety $V_{\mathbb{R}}$. We assume $V_{\mathbb{R}}$ is given as the locus of zeros of a set $\{p_a\}$ of homogeneous polynomials in the real variables $\{x_i\}$. We may assume without loss of generality that each p_a has real coefficients. By allowing the x_i to be complex, $\{p_a\}$ naturally determines a complex homogeneous variety $V_{\mathbb{C}}$ and it is natural to ask about the relationship between the degree and dimension of $V_{\mathbb{C}}$ and the degree and dimension of $V_{\mathbb{R}}$.

Theorem 6.10: Let $V_{\mathbb{C}}$ and $V_{\mathbb{R}}$ both be determined by $\{p_a\}$ as described above. Then

$$\dim(V_{\mathbb{R}}) \le \frac{\dim(V_{\mathbb{C}})}{2} \tag{6.19}$$

and

$$\deg(V_{\mathbb{R}}) \le \deg(V_{\mathbb{C}})\ . \tag{6.20}$$

Proof: The relationship between dimensions follows from the implicit function theorem, which says that if a point $p\in V_{\mathbb{C}}$ has a neighborhood U which is a manifold of dimension $2n$, then there is an ordering of the coordinates $x_1,\ldots,x_N$ such that near p $V_{\mathbb{C}}$ can be parameterized as

$$(x_1,\ldots,x_n,\varphi_1,\ldots,\varphi_{N-n})$$

where

$$\varphi_i = \varphi_i(x_1, \ldots, x_n) \ .$$

By restricting the x_i to be real (but still within the region of definition), we see that $V_{\mathbb{R}}$ can have dimension at most n.

Conversely, such a local parameterization of a real manifold can define a complex manifold locally of twice the dimension if the functions φ_i are defined for complex arguments. In particular, a real plane $L_{\mathbb{R}}^{N-n}$ extends to a complex plane $L_{\mathbb{C}}^{2N-2n}$. If $L_{\mathbb{C}}^{2N-2n} \cap V_{\mathbb{C}}$ contains at most $\deg(V_{\mathbb{C}})$ points, then $L_{\mathbb{R}}^{N-n} \cap V_{\mathbb{R}} \subset L_{\mathbb{C}}^{2N-2n} \cap V_{\mathbb{C}}$ can also contain at most $\deg(V_{\mathbb{C}})$ points. Q.E.D.

All of the real and complex varieties we consider can be given as the the locus of zeros of $\{p_a\}$ where each p_a has real coefficients, so we can use this theorem to extend our knowledge about the degree and dimension of complex varieties (see section 6.4) to real varieties. Indeed, for all the varieties we study, the dimension of $V_{\mathbb{R}}$ will be exactly half the dimension of $V_{\mathbb{C}}$.

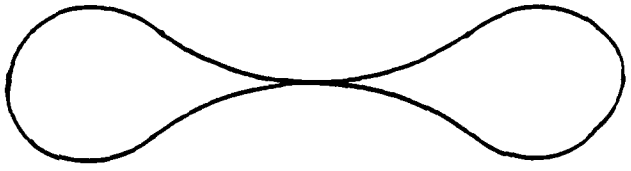


Figure 6.1 Pinched Section in a Variety

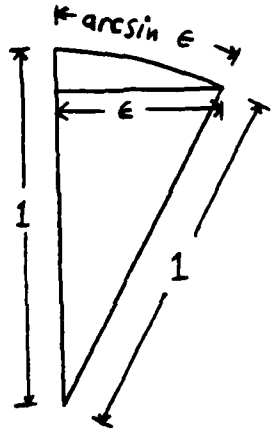


Figure 6.2 Distances on the Circle

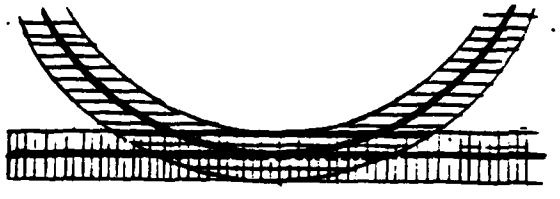


Figure 6.3 Counting the Area of the Doubly Shaded Region Twice

Chapter 7: Applications to Matrix Inversion, Eigenvalue Problems, and Polynomial Zero Finding

7.1 Introduction

In the last chapter we proved two theorems about the fraction of the unit sphere within distance ε of a homogeneous variety. In this chapter we apply this result to three problems of numerical analysis. In section 7.2 we compute the fraction of n by n matrices within ε of a matrix of rank at most r. When $r=n-1$ this result can be interpreted as the probability distribution of the condition number of a random matrix. In section 7.3 we compute the fraction of matrices within ε of a matrix with a given Jordan canonical form. In the simplest case, when the Jordan form contains (at least) one double eigenvalue, this result gives the probability distribution of the distance from a random matrix to the nearest defective matrix. In section 7.4 we compute the fraction of polynomials within ε of a polynomial with a given zero structure. For the zero structure containing (at least) one double zero, this result gives the probability distribution of the distance from a random polynomial to one with a double root. Section 7.5 contains the proofs of two algebraic lemmas needed earlier.

7.2 The Distribution of the Distance from a Random Matrix to a Matrix of Rank r

In this section we will apply the general results of the last chapter to the varieties of n by n matrices containing those of rank at most r. When $r=n-1$ we are talking about the variety of singular matrices. We adopt the probabilistic interpretation of the results of the last chapter and ask the following question: if a matrix M is chosen at random so that $M/\|M\|_g$ is uniformly distributed on the unit sphere, what is the probability distribution of the dis-

tance of $M/\|M\|_E$ to the set of matrices of rank r? Said another way, what is the distribution of the relative distance from M to the matrices of rank r? To describe our results, we let $V_{\mathbf{C}}^r$ denote the complex n by n matrices of rank at most r, and $V_{\mathbf{R}}^r$ denote the real matrices of rank at most r. When $r=n-1$, so that $V_{\mathbf{C}}^{n-1}$ ($V_{\mathbf{R}}^{n-1}$) is the set of singular matrices, we can state our result as follows:

Theorem 7.1: Let $M_{\mathbf{C}}$ be a complex n by n matrix chosen randomly as described above. Then

$$\text{Prob}(\text{dist}_E(M_{\mathbf{C}}/\|M_{\mathbf{C}}\|_E\,,\ V_{\mathbf{C}}^{n-1}) = n(n^2-1)\cdot\varepsilon^2 + o(\varepsilon^2)\ . \tag{7.1}$$

If $M_{\mathbf{R}}$ is a random real matrix, then

$$\text{Prob}(\text{dist}_E(M_{\mathbf{R}}/\|M_{\mathbf{R}}\|_E\,,\ V_{\mathbf{R}}^{n-1}) \le \frac{n(n^2-1)}{2}\cdot\varepsilon + o(\varepsilon)\ . \tag{7.2}$$

We can interpet this theorem in a fashion more common among numerical analysts. We define the *condition number* of a (real or complex) matrix as

$$\kappa(M) \equiv \|M\|_E\cdot\|M^{-1}\|\ .$$

As is well known [Eckart], $\|M^{-1}\|^{-1}$ is the Euclidean distance dist_E from M to the nearest singular matrix. The condition number is used by numerical analysts to measure the difficulty of inverting a matrix, because it gives the maximum relative perturbation that can be caused in M^{-1} by a unit relative perturbation in M. In this notation, Theorem 7.1 can be restated as

Corollary 7.2: If $M_{\mathbf{C}}$ is a random complex matrix, then

$$\text{Prob}(\kappa(M_{\mathbf{C}}) \ge K) = n(n^2-1)\cdot K^{-2} + o(K^{-2})\ . \tag{7.3}$$

If $M_{\mathbf{R}}$ is a random real matrix, then

$$\text{Prob}(\kappa(M_{\mathbf{R}}) \ge K) \le \frac{n(n^2-1)}{2}\cdot K^{-1} + o(K^{-1})\ . \tag{7.4}$$

Since the condition number is commonly used as a measure of how difficult a matrix is to invert accurately, Corollary 7.2 measures the likeli-

hood of a random matrix being hard to invert.

Note that this theorem provides no information at all unless $n(n^2-1)\cdot K^{-2}\le 1$, i.e. $K\ge\sqrt{n(n^2-1)}$ in the complex case, and unless $K\ge n(n^2-1)/2$ in the real case. For real 10 by 10 matrices, this requires K to be at least 495, already rather large by the standards of some numerical analysts. Until sharper versions of Theorem 6.3 and 6.6 are forthcoming, applying these asymptotic formulas in practical circumstances must be done carefully.

For matrices of rank $r<n-1$ our results are:

Theorem 7.3: If $M_{\mathbf{C}}$ is a random complex matrix, then

$$\mathrm{Prob}(\mathrm{dist}_F(M_{\mathbf{C}}/\|M_{\mathbf{C}}\|_F,\ V_{\mathbf{C}}^r)=\left\{\begin{matrix} n^2-1 \\ (n-r)^2\end{matrix}\right\}\cdot \deg(V_{\mathbf{C}}^r)\ \varepsilon^{2(n-r)^2}+o(\varepsilon^{2(n-r)^2}) \quad (7.5)$$

where

$$1\le \deg(V_{\mathbf{C}}^r)\le (r+1)^{\left\{\begin{matrix} n \\ r+1\end{matrix}\right\}^2}\ . \quad (7.6)$$

If $M_{\mathbf{R}}$ is a random real matrix, then

$$\mathrm{Prob}(\mathrm{dist}_F(M_{\mathbf{R}}/\|M_{\mathbf{R}}\|_F,\ V_{\mathbf{R}}^r)\le \quad (7.7)$$

$$\frac{\Gamma(\frac{n^2+1}{2})\cdot\Gamma(\frac{n^2}{2})}{\pi^{((n-r)^2-1)/2}\cdot\Gamma(\frac{(n-r)^2+1}{2})\cdot\Gamma(\frac{2nr+r^2+1}{2})\cdot\Gamma(\frac{2nr+r^2}{2})}\times$$

$$\deg(V_{\mathbf{R}}^r)\cdot\varepsilon^{(n-r)^2}+o(\varepsilon^{(n-r)^2})$$

where $1\le \deg(V_{\mathbf{R}}^r)\le \deg(V_{\mathbf{C}}^r)$.

The important aspect of these formulas is not the constant coefficient, but the exponent of ε, since this exponent describes the behavior of the probability as a function of ε. We see that matrices of lower rank become less common rather quickly: the exponent $2(n-r)^2$ (or $(n-r)^2$) going up as the

square of the rank deficiency $n-r$.

We can use Theorems 7.1 and 7.3 to compute various conditional probabilities. For example, one might ask about the probability of a random matrix being within relative distance ε of a matrix of rank r *given* that it is within ε of a matrix of rank $r+1$. We compute this simply by dividing the distribution of the distance to V^r by the distribution of the distance to V^{r+1} to get (in the complex case) a constant times $\varepsilon^{4(n-r)-2}$. This quantity tells us that surfaces of lower rank matrices become increasingly sparser within the surface of matrices of rank one higher. For example, the density of singular (rank $n-1$) matrices within all matrices behaves as ε^2, rank $n-2$ matrices within rank $n-1$ matrices as ε^6 and so on.

Proof of Theorem 7.1: The proof for $V_{\mathbb{C}}^{n-1}$ will follow immediately from Theorem 6.3 if we show that $V_{\mathbb{C}}^{n-1}$ is a purely $2n^2-2$-dimensional complex homogeneous variety of degree n. This in turn follows directly from Theorem 6.4 since $V_{\mathbb{C}}^{n-1}$ is the zero set of the single irreducible order n polynomial $\det(M)$. Similarly, the result for $V_{\mathbb{R}}^{1}$ will follow from Theorem 6.6 if we show that $V_{\mathbb{R}}^{1}$ is an n^2-1 dimensional real variety of degree at most n. The degree bound follows from Theorem 6.10 and the dimension from noting that $\det(M)=0$ is linear in each m^{ij} so that m^{ij} can easily be expressed as a rational function in the other n^2-1 real variables. Q.E.D.

Proof of Theorem 7.3: As with the last proof, the results follow from Theorems 6.3 and 6.6 given the dimension and bounds on the degrees of $V_{\mathbb{C}}^{r}$ and $V_{\mathbb{R}}^{r}$. To compute the dimension, we use Gaussian elimination to put the matrix in row echelon form: if M has rank r, then its rows and columns can be permuted so that the permuted M' can be factored as LU. L is lower triangular with ones on the diagonal and nonzeroes below the diagonal only in columns 1 through

r, and U is upper triangular with nonzeroes only in rows 1 through r. These nonzero entries serve as local coordinates for V^r, and there are $2rn-r^2$ of them, so that $\dim(V^r_{\mathbb{R}})=2rn-r^2$, and $\dim(V^r_{\mathbb{C}})=2\dim(V^r_{\mathbb{R}})$, since the nonzero entries are all complex in that case. To compute $\deg(V^r_{\mathbb{C}})$, we note that $V^r_{\mathbb{C}}$ is given by the collection $\{p_\alpha\}$ of all determinants of order $r+1$ minors of M, of which there are $\binom{n}{r+1}^2$. The bound on $\deg(V^r_{\mathbb{C}})$ comes from Bézout's Theorem, and the bound on $\deg(V^r_{\mathbb{R}})$ from Theorem 6.10. Q.E.D.

7.3 The Distribution of the Distance from a Random Matrix to a Matrix with a given Jordan Canonical Form

Let P^J denote the set of n by n matrices with Jordan canonical form given by the multiindex J ($P^J_{\mathbb{C}}$ denotes the set of complex matrices, and $P^J_{\mathbb{R}}$ the real matrices). J denotes the Jordan form with (generically) m distinct eigenvalues λ_k $(1\le k\le m)$, such that λ_k has b_k Jordan blocks of sizes $s^k_1\ge s^k_2\ge\cdots\ge s^k_{b_k}$. We say generic because within P^J lie lower dimensional surfaces where distinct eigenvalues λ_i and λ_j become equal, or where the number of Jordan blocks for a given eigenvalue increase. This is analogous to the situation in the last section, where the variety of matrices of rank at most r has the matrices of rank exactly r as a dense open subset whose complement (matrices of rank less than r) form a lower dimensional subvariety. (These statements about P^J require proof; we do not even know yet that P^J is a variety. These facts will be proven below.)

In this section we answer the following question: if the matrix M is chosen at random so that $M/\|M\|_E$ is uniformly distributed on the unit sphere, what is the probability distribution of the relative distance from M to P^J? The simplest case occurs when P^J is the variety of matrices with at least one double eigenvalue; in this case our result is:

Theorem 7.4: Let $M_{\mathbf{C}}$ be a complex n by n matrix chosen randomly as described above, and let $P_{\mathbf{C}}$ denote the set of complex matrices with at least one double eigenvalue. Then

$$\text{Prob}(\text{dist}_F(M_{\mathbf{C}} / \ \| M_{\mathbf{C}} \|_F \ , P_{\mathbf{C}}) = n(n-1)^2(n+1) \cdot \varepsilon^2 + o(\varepsilon^2) \ . \qquad (7.8)$$

If $M_{\mathbf{R}}$ is a random real matrix, and $P_{\mathbf{R}}$ the set of real matrices with at least one double eigenvalue, then

$$\text{Prob}(\text{dist}_F(M_{\mathbf{R}} / \ \| M_{\mathbf{R}} \|_F \ , P_{\mathbf{R}}) \le \frac{n(n-1)^2(n+1)}{2} \cdot \varepsilon + o(\varepsilon) \ . \qquad (7.9)$$

For matrices of a general Jordan form $\mathbf{J}$ we need to know the degree and dimension of $P^{\mathbf{J}}_{\mathbf{C}}$ and $P^{\mathbf{J}}_{\mathbf{R}}$. Given this data, the results will follow from Theorems 6.3 and 6.5 of the last chapter. We compute $\dim(P^{\mathbf{J}})$ explicitly below. We outline a procedure for computing a defining set of polynomials $\{p^{\mathbf{J}}_a\}$ for $P^{\mathbf{J}}$, thus proving that $P^{\mathbf{J}}$ is a variety and providing an upper bound on $\deg(P^{\mathbf{J}})$ by Bézout's theorem (Theorem 6.5). We will not display $\{p^{\mathbf{J}}_a\}$ or compute this upper bound, however, because they are very messy and do not illuminate the structure of $P^{\mathbf{J}}$ nearly as much as its dimension, which we do compute explicitly.

Theorem 7.5: Let $\mathbf{J}$ and $P^{\mathbf{J}}$ be defined as above. Then the codimension of $P^{\mathbf{J}}$ (and the exponent of ε in Theorems 6.3 and 6.5) is

$$\text{codim}(P^{\mathbf{J}}_{\mathbf{R}}) = \frac{\text{codim}(P^{\mathbf{J}}_{\mathbf{C}})}{2} = n - m + 2 \cdot \sum_{k=1}^{m} \sum_{i=2}^{b_k} (i-1) \cdot s_i^k \qquad (7.10)$$

where the sum from $i=2$ to b_k is zero if $b_k=1$. If all the $b_k=1$, so that there is one Jordan block per eigenvalue, this simplifies to

$$\text{codim}(P^{\mathbf{J}}_{\mathbf{R}}) = \frac{\text{codim}(P^{\mathbf{J}}_{\mathbf{C}})}{2} = n - m \ . \qquad (7.6)$$

Thus, if there is only one Jordan block per eigenvalue, the codimension depends only on the number of distinct eigenvalues (matrices with one Jordan block per eigenvalue are called *nonderogatory*).

Proof of Theorem 7.4: The proof for $P_{\mathbf{C}}$ will follow immediately from Theorem 6.3 if we show that $P_{\mathbf{C}}$ is a pure $2n^2-2$-dimensional complex homogeneous variety of degree $n(n-1)$. This will follow in turn from Theorem 6.4 if we show $P_{\mathbf{C}}$ is defined by a single irreducible homogeneous polynomial p of order $n(n-1)$. This polynomial is called the *resultant* of the characteristic polynomial of the matrix M and its first derivative, whose properties we record in the following lemma:

Lemma 7.7: Let M be a matrix of n^2 indeterminates m_{ij}. Let $p(\lambda, m_{ij}) = \det(M - \lambda \cdot I)$ be its characteristic polynomial. Then

$$r(m_{ij}) = \mathrm{res}(p(\lambda, m_{ij}), \frac{d}{d\lambda} p(\lambda, m_{ij}), \lambda)$$

i.e. the resultant of p and its derivative is a polynomial in the m_{ij} which is

1) *zero if and only if M has a multiple eigenvalue,*
2) homogeneous of degree $n(n-1)$, and
3) irreducible.

Proof: see Section 7.5.

The result for $P_{\mathbf{R}}$ follows from Theorems 6.6 and 6.10. This completes the proof of Theorem 7.4. Q.E.D. of Theorem 7.4.

Proof of Theorem 7.5: This theorem was originally proven in [Arnold] and discovered independently by us. Since the proof is short and provides and interesting application of Lie groups to numerical linear algebra, we sketch it here.

Let M be a matrix with Jordan form J. Frobenius's theorem [Gantmacher] characterizes all matrices which commute with M, and shows in particular that they form a linear manifold of dimension

$$z(J) = \sum_{k=1}^{m} \sum_{i=1}^{b_k} (2i-1) \cdot s_i^k$$

for real matrices, and $2 \cdot z(J)$ for complex matrices. Now consider the Lie

group $GL(n,\mathbf{C})$ of nonsingular n by n complex matrices, which has dimension $2n^2$. Let Z_M denote the centralizer of M in $GL(n,\mathbf{C})$, that is, the set of all nonsingular matrices which commute with M. We know $\dim(Z_M)$ by the result just stated. It is easy to see that Z_M is a closed subgroup of $Gl(n,\mathbf{C})$, so that the quotient space $GL(n,\mathbf{C})/Z_M$ is a manifold of dimension $2\cdot(n^2 - z(\mathbf{J}))$ [Sternberg]. This quotient space is naturally diffeomorphic to the set S_M of matrices which are similar to M, since $Z_1MZ_1^{-1} = Z_2MZ_2^{-1}$ if and only if Z_1 and Z_2 are in the same coset of Z_M. Now we claim $S_M\times\mathbf{C}^m$ (the Cartesian product), which has dimension $2\cdot(m + n^2 - z(\mathbf{J}))$, is locally diffeomorphic to $P^{\mathbf{J}}_{\mathbf{C}}$. The diffeomorphism simply takes the j-th component of the $\mathbf{C}^m$ factor and adds it to the m-th distinct eigenvalue of S_M. Subtracting $\dim(S_M\times\mathbf{C}^m)$ from $2n^2$ yields the value of $\operatorname{codim}(P^{\mathbf{J}}_{\mathbf{C}})$ claimed in the theorem.

The same proof works in the real case, yielding something of exactly half the codimension of $P^{\mathbf{J}}_{\mathbf{C}}$. We do need two additional facts: if two matrices over a field F are similar over an extension field K of F, then they are similar over F (because a matrix is similar to its rational canonical form over F), and two complex conjugate eigenvalues of a real n by n matrix are determined by two *real* parameters, so $S_M\times\mathbf{C}^m$ can be replaced by $S_M\times\mathbf{R}^m$ above. Q.E.D. of Theorem 7.5.

It remains to show how to construct a set of polynomials $\{p^{\mathbf{J}}_a\}$ which determine $P^{\mathbf{J}}$. The construction has two steps. First, we construct a set of polynomials in the matrix entries m^{ij} and the eigenvalues λ_i whose projection onto the m_{ij} coordinates is $P^{\mathbf{J}}$. Second, we show how to eliminate the λ_i variables. This elimination requires a generalization of the fundamental theorem on symmetric polynomials to symmetric varieties.

A polynomial $p(\lambda_1, \ldots, \lambda_n)$ is called *symmetric* if for all permutations σ of n objects we have

$$p_\sigma(\lambda_1, \ldots, \lambda_n) \equiv p(\lambda_{\sigma(1)}, \ldots, \lambda_{\sigma(n)}) = p(\lambda_1, \ldots, \lambda_n) \ .$$

The fundamental theorem on symmetric polynomials [VanderWaerden] says that $p(\lambda_1, \ldots, \lambda_n)$ is symmetric if and only if it can be expressed as a polynomial in the *elementary symmetric functions* Σ_i of the λ_i:

$$\Sigma_1 \equiv \sum_{1 \le i \le n} \lambda_i$$

$$\Sigma_2 \equiv \sum_{1 \le i < j \le n} \lambda_i \cdot \lambda_j$$

$$\Sigma_3 \equiv \sum_{1 \le i < j < k \le n} \lambda_i \cdot \lambda_j \cdot \lambda_k$$

$$\cdots$$

$$\Sigma_n \equiv \prod_{1 \le i \le n} \lambda_i$$

A variety V which is generated by $\{p_\alpha(\lambda_1, \ldots, \lambda_n)\}$ is called *symmetric* if for all permutations σ V is also generated by $\{p_{\alpha\sigma}\}$. Geometrically, this means V is invariant under the group generated by all reflections in planes $\lambda_i = \lambda_j$. Now we can state our generalization of the fundamental theorem on symmetric polynomials to symmetric varieties:

Lemma 7.7: Let the variety V be generated by $\{p_\alpha(\lambda_1, \ldots, \lambda_n)\}$. Then V is symmetric if and only if V is generated by a set of polynomials $\{q_\beta(\Sigma_1, \ldots, \Sigma_n)\}$ in the elementary symmetric functions of the λ_i.

Proof: See section 7.5.

For our construction of $\{p_\alpha^j\}$ we also need to know that the union and intersection of varieties are varieties. For if $\{p_i\}$ generates P and $\{q_j\}$ generates Q, then $\{p_i \ , \ q_j\}$ generates $P \cap Q$, and $\{p_i \cdot q_j\}$ generates $P \cup Q$.

Now we begin the construction. Let $m_k \equiv \sum_{i=1}^{b_k} s_i^k$ denote the algebraic

multiplicity of the k-th distinct eigenvalue. Now take the eigenvalues and constrain them with the following polynomials:

$$\lambda_1 = \lambda_2 \,,\, \lambda_1 = \lambda_3 \,,\, \ldots \,,\, \lambda_1 = \lambda_{m_1} \,,$$

$$\lambda_{m_1+1} = \lambda_{m_1+2} \,,\, \lambda_{m_1+1} = \lambda_{m_1+3} \,,\, \ldots \,,\, \lambda_{m_1+1} = \lambda_{m_1+m_2} \,,$$

$$\ldots$$

$$\lambda_{\sum_{i=1}^{j} m_i+1} = \lambda_{\sum_{i=1}^{j} m_i+2} \,,\, \lambda_{\sum_{i=1}^{j} m_i+1} = \lambda_{\sum_{i=1}^{j} m_i+3} \,,\, \lambda_{\sum_{i=1}^{j} m_i+1} \,,\, \ldots \,,\, \lambda_{\sum_{i=1}^{j+1} m_i} \,,\, \ldots$$

In other words, equate the first m_1 eigenvalues to λ_1, the next m_2 to λ_{m_1+1}, and so on. Next take the polynomials $\det(M-\lambda_j I)=0$, where j takes k values corresponding to distinct λ_j. This last set of equations guarantees that each λ_j is an eigenvalue of M. The constraints on the sizes of the Jordan blocks can be translated into constraints on the rank of powers of $M-\lambda_j \cdot I$, because $\mathrm{rank}(M-\lambda_j \cdot I)^{i-1} - \mathrm{rank}(M-\lambda_j \cdot I)^i$ equals the number of Jordan blocks of size at least i, which we denote B_i^j. Thus,

$$\mathrm{rank}(M-\lambda_j \cdot I)^i = n - \sum_{k=1}^{i} B_k^j \ .$$

Clearly, the s_k^j uniquely determine the B_k^j, and vice-versa. Also, any collapsing of λ_j's or breaking up of Jordan blocks (which occur on subvarieties) can only make $\mathrm{rank}(M-\lambda_j)^i$ drop. From Section 7.2 we know how to express the condition that $\mathrm{rank}(M-\lambda^j \cdot I)^i$ should be no more than some constant in terms of determinants of minors. All these polynomials taken together, for all m distinct eigenvalues λ_j and powers i, determine a variety in $\mathbf{C}^{n^2} \times \mathbf{C}^m$ space whose projection onto the $\mathbf{C}^{n^2}$ component is P^J.

However, there are many other varieties whose projection is P^J. If $\{p_a(\lambda_k \,,\, m_{kj})\}$ generates the variety of the last paragraph, and if σ is a permutation of the first n integers, then

$$V_\sigma = \{p_\alpha(\lambda_{\sigma(k)}, m_{ij})\} = \{p_{\alpha\sigma}(\lambda_k, m_{ij})\}$$

also has the same projection. Consider the variety $V = \bigcup_\sigma V_\sigma$, and let $\{q_\beta(\lambda_k, m_{ij})\}$ be a finite set of polynomials generating V (we know how to construct the q_β from the p_α and the rule for taking unions of varieties). It is clear from the construction of V that the ideal generated by $\{q_\beta(\lambda_k, m_{ij})\}$ is *symmetric*, that is $\{q_\beta(\lambda_{\sigma(k)}, m_{ij})\}$ generates the same ideal (and variety V) for any permutation σ. By Lemma 7.7, we see that there is a set of polynomials $\{r_\gamma\}$ which also generate V but which are functions of m_{ij} and the elementary symmetric functions Σ_i of the λ_i. But these Σ_i are nothing but the coefficients of the characteristic polynomial in M, which are polynomials in m_{ij}. Thus, the r_γ themselves are polynomials only in the m_{ij}. These r_γ are the desired polynomials which generate $P^{\mathbf{J}}$.

7.4 The Distribution of the Distance from a Random Polynomial to One With a Given Zero Structure

Let $Z^{\mathbf{K}}$ denote the set of n-th degree polynomials $p(z) = \sum_{i=1}^{n} p_i \cdot z^i$ with zero structure given by the multiindex $\mathbf{K}$ ($Z^{\mathbf{K}}_{\mathbb{C}}$ denotes the complex matrices, and $Z^{\mathbf{K}}_{\mathbb{R}}$ the real matrices). $\mathbf{K}$ denotes the zero structure with (generically) m distinct zeroes λ_k ($1 \le k \le m$), such that λ_k has multiplicity m_k. We say generic for the same reason as in the last section: within $Z^{\mathbf{K}}$ lie lower dimensional surfaces within which distinct roots coalesce. This will be proven below.

In this section we answer the question: if the n-th degree polynomial p is chosen at random so that $p \,/\, \|p\|_g$ is "uniformly distributed" on the unit sphere, what is the probability distribution of the relative distance from M to $Z^{\mathbf{K}}$? ($\|p\|_g$ is the norm $(\sum_{i=0}^{n} |p_i|^2)^{1/2}$.) The reason for the quotation marks around "uniformly distributed" is our insistence on choosing an n-th degree

polynomial at random, which means that we eliminate the hyperplane $p_n = 0$ from our sample space. This makes sense because if $p_n = 0$ we have a qualitatively different problem because we have a polynomial of different degree. Smale [Smale] considers the nonhomogeneous problem ($p_n = 1$), but by maintaining homogeneity we can still the results of chapter 6.

The simplest case occurs when $Z^{\mathbf{K}}$ is the variety of polynomials with at least one multiple root; in this case our result is:

Theorem 7.8: Let $p_{\mathbf{C}}$ be a complex degree n polynomial chosen randomly as described above, and let $Z_{\mathbf{C}}$ denote the set of complex matrices with at least one double zero. Then

$$\mathrm{Prob}(\mathrm{dist}_F(p_{\mathbf{C}} / \|p_{\mathbf{C}}\|_F, Z_{\mathbf{C}}) = n(n^2-n-1) \cdot \varepsilon^2 + o(\varepsilon^2) \ . \qquad (7.11)$$

If $p_{\mathbf{R}}$ is a random real polynomial, and $Z_{\mathbf{R}}$ the set of real polynomials with at least one double zero, then

$$\mathrm{Prob}(\mathrm{dist}_F(p_{\mathbf{R}} / \|p_{\mathbf{R}}\|_F, Z_{\mathbf{R}}) \le \frac{n(n^2-n-1)}{2} \cdot \varepsilon + o(\varepsilon) \ . \qquad ()$$

For polynomials of general zero structure $Z^{\mathbf{K}}$ we need to know the degree and dimension of $Z^{\mathbf{K}}_{\mathbf{C}}$ and $Z^{\mathbf{K}}_{\mathbf{R}}$ so we can apply Theorems 6.3 and 6.5 of the last chapter. We compute $\dim(Z^{\mathbf{K}})$ explicitly below, but just as in the last section we only outline a procedure for computing a defining set of polynomials $\{p^{\mathbf{K}}_a\}$ for $Z^{\mathbf{K}}$.

Theorem 7.9: Let $\mathbf{K}$ and $Z^{\mathbf{K}}$ be defined as above. Then the codimension of $Z^{\mathbf{K}}$ (and the exponent of ε in Theorems 6.3 and 6.5) is

$$\mathrm{codim}(Z^{\mathbf{K}}_{\mathbf{R}}) = \frac{\mathrm{codim}(Z^{\mathbf{K}}_{\mathbf{C}})}{2} = n - m$$

so that the codimension depends only on the number of distinct eigenvalues.

This theorem is analogous to Theorem 7.5 of the last section for nonderogatory matrices, which is no surprise since the rational canonical form

of a nonderogatory matrix is determined uniquely by the characteristic polynomial.

Proof of Theorem 7.9: As in the analogous Theorem 7.4, we use the resultant r of the polynomial p and its first derivative, which is a homogeneous polynomial of degree $n(n-1)$. Unfortunately, this polynomial does *not* have the property of being zero if and only if p has a multiple root, because it is also zero if $p_n=0$, the set of points we eliminated from our sample space above. If we divide r by p_n, we get a polynomial d called the *discriminant* of p, which is homogeneous of degree n^2-n-1 and irreducible (a proof of irreducibility follows from the proof of Lemma 7.7 below). d will be zero if and only if p is of degree n and has a multiple zero *or* $p_n=p_{n-1}=0$, but this last part is a lower dimensional subvariety of the forbidden part of our sample space, so does not contribute to vol(d). The result follows from applying Theorems 6.3, 6.4, 6.6 and 6.10 to d. Q.E.D. of Theorem 7.9.

Proof of Theorem 7.10: The $m+1$ parameters λ_i $(1\le i\le m)$ and p_n form a local coordinate system for the p_i as is easily seen by equating powers of z in the identity

$$\sum_{i=0}^{n} p_i \cdot z^i = p_n \cdot \prod_{i=1}^{m} (z-\lambda_i)^{m_i} .$$

Thus, $Z_{\mathbb{C}}^{\mathbb{K}}$ has dimension $2(m+1)$ and codimension $2(n-m)$ as claimed. The real case follows since complex zeroes occur in complex conjugate pairs, so all dimensions and codimensions are cut in half. Q.E.D. of Theorem 7.10.

It remains to show how to construct a set of polynomials $\{p_a^{\mathbb{K}}\}$ defining $Z^{\mathbb{K}}$. The construction is analogous to the construction in the last section for $P^{\mathbb{J}}$. First we construct a set of polynomials in the coefficients p_i and the zeroes λ_i which define a symmetric variety whose projection onto the first components is $Z^{\mathbb{K}}$. These polynomials simply equate different λ_i and express

the p_i as the usual symmetric functions of the λ_j. Second, we eliminate the λ_j using Lemma 7.8 just as in the last section.

7.5 Proofs of Lemmas 7.7 and 7.8:

Lemma 7.7: Let M be a matrix of n^2 indeterminates m_{ij}. Let $p(\lambda, m_{ij}) = \det(M - \lambda \cdot I)$ be its characteristic polynomial. Then

$$r(m_{ij}) \equiv \mathrm{res}(p(\lambda, m_{ij}),\ \frac{d}{d\lambda} p(\lambda, m_{ij}),\ \lambda)$$

i.e. the resultant of p and its derivative is a polynomial in the m_{ij} which is

1) zero if and only if M has a multiple eigenvalue,
2) homogeneous of degree $n(n-1)$, and
3) irreducible.

Proof: The *resultant* of the two polynomials $p(\lambda) = \sum_{i=0}^{m} p_i \lambda^i \in R[\lambda]$ and $q(\lambda) = \sum_{i=0}^{n} q_i \lambda^i \in R[\lambda]$ is denoted $\mathrm{res}(p, q, \lambda)$ (or $\mathrm{res}(p, q)$ if λ is clear from context) and defined as the determinant

$$\begin{vmatrix} p_0 & p_1 & \cdots & p_m & & & & \\ & p_0 & p_1 & \cdots & p_m & & & \\ & & & \cdots & & & & \\ & & & p_0 & p_1 & & \cdots & p_m \\ q_0 & q_1 & \cdots & & q_n & & & \\ & q_0 & q_1 & \cdots & & q_n & & \\ & & & \cdots & & & & \\ & & q_0 & q_1 & & \cdots & & q_n \end{vmatrix}$$

where there are m copies of the rows with q entries, and n copies of the rows with p entries. $\mathrm{res}(p, q)$ is clearly a polynomial in the p_i and q_j. If $p_m \neq 0 \neq q_n$ then $\mathrm{res}(p, q) = 0$ if and only if p and q have a common zero [VanderWaerden]. If we choose $q(\lambda) = p'(\lambda)$ to be the derivative of p and $p_m \neq 0$, then $\mathrm{res}(p, p')$ will be zero if and only if p has a multiple root [VanderWaerden]. Applying this to $p(\lambda) = \det(M - \lambda I)$ proves claim 1 above.

This choice of p is clearly homogeneous of degree n in the λ and m_{ij}, so p' is homogeneous of degree $n-1$. By Theorem 6.2 in [Kendig], their resultant is homogeneous of degree $n(n-1)$. This proves claim 2 above.

The proof of claim 3 takes two steps. First we show that if $r(\lambda)=\lambda^n + \sum_{i=0}^{n-1} r_i \lambda^i$, then $\mathrm{res}(r,r')$ is irreducible. Second we show that this implies the irreducibility of $\mathrm{res}(p,p')$.

To show $d = \mathrm{res}(r,r')$ is irreducible, we use another representation of it in terms of the zeros α_i of r: $d = \prod_{i<j} (\alpha_i - \alpha_j)^2$ [VanderWaerden]. Write d as the product $d=d_1 \cdot d_2$. Since the d_i are functions of the r_i, they are symmetric functions of the α_i. Since $\alpha_1-\alpha_2$ is a factor of d, it must divide either d_1 or d_2, say d_1. By symmetry, all the other factors $\alpha_i-\alpha_j$ must divide d_1, so d_1 is a constant multiple of d, and d is irreducible.

Now consider the *companion matrix* of r:

$$\begin{bmatrix} 0 & 1 & 0 & \cdot & 0 \\ 0 & 0 & 1 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & 1 \\ -r_{n-1} & -r_{n-2} & -r_{n-3} & \cdot & -r_0 \end{bmatrix}.$$

r is the characteristic polynomial of this matrix [Herstein]. Now if $\mathrm{res}(p,p')$ factored into $p_1 \cdot p_2$, this would induce a factorization of $d=\mathrm{res}(r,r')=d_1 \cdot d_2$. By the result of the last paragraph, p_1 (say), which corresponds to d_1, will be a constant multiple of d so that p_2 cannot depend on any entry m_{ni} of the last row of M, but only on entries $m_{i,i+1}$of the superdiagonal. Now take M and exchange rows i and n as well as columns i and n to obtain the similar matrix M'. M' has the same characteristic polynomial as M, so we conclude that p_2 cannot depend on entries from row i. Thus p_2 is constant and $\mathrm{res}(p,p')$ is irreducible as desired. Q.E.D. of Lemma 7.7.

Lemma 7.8: Let the variety V be generated by $\{p_\alpha(\lambda_1, \ldots, \lambda_n)\}$. Then V is symmetric if and only if V is generated by a set of polynomials $\{q_\beta(\Sigma_1, \ldots, \Sigma_n)\}$ in the elementary symmetric functions of the λ_i.

Proof: The if direction is trivial. If there is only one p_α, the symmetry condition implies $p_\alpha(\lambda_i) = p_\alpha(\lambda_{\sigma(i)})$ for all permutations σ, so the only if direction is equivalent to the fundamental theorem on symmetric polynomials. If there is more than one p_α we argue as follows. We let $p_{\alpha\sigma}$ denote the polynomial $p_\alpha(\lambda_{\sigma(i)})$, and V_σ denote the variety generated by $\{p_{\alpha\sigma}\}$. By assumption $V_\sigma = V$ for all σ. Then the variety $V_s = \bigcap_\sigma V_\sigma$ is generated by $\{p_{\alpha\sigma}$, all α and $\sigma\}$ which equals $\bigcap_\alpha V_\alpha$, where V_α is generated by $\{p_{\alpha\sigma}$, α fixed, all $\sigma\}$. If we can show the variety V_α is generated by a collection of polynomials over the Σ_i, we will be done. We claim this collection of polynomials is the set $\{\Sigma_i(p_\alpha)\}$ of all symmetric function of the $p_{\alpha\sigma}$ themselves. For all the $\Sigma_i(p_\alpha)$ can all be zero if and only if all the $p_{\alpha\sigma}$ are zero, so they generate the same variety. Furthermore, each $\Sigma_i(p_\alpha)$ is clearly a symmetric function of the λ_i, and so by the fundamental theorem on symmetric functions, is itself a function of the Σ_i. Q.E.D. of Lemma 7.8.

Chapter 8: Probabilistic Estimates of $\mathrm{diss}_g(\sigma_1, \sigma_2)$

8.1 Introduction

In this chapter we apply the probabilistic estimates of chapters 6 and 7 to measure the likelihood that the various bounds on $\mathrm{diss}_g(\sigma_1, \sigma_2, \mathit{path})$ and $\mathrm{diss}_g(\sigma_1, \sigma_2, \mathit{region})$ of chapter 5 are accurate.

In Section 8.2 we compute the probability that a randomly chosen matrix is completely diagonalizable, where our decomposability criterion is based on path clustering. We also compute upper bounds on the probability of being able to decompose a random matrix into blocks with no more than r eigenvalues per block, where $r>1$. In Section 8.3 we estimate the decomposition probabilities using a different decomposability criterion: σ is decomposable into $\bigcup_i \sigma_i$ if $\|P_i\| \leq K$ for all i (P_i is the projection belonging to σ_i). In Section 8.4 we ask how much larger the upper bound on $\mathrm{diss}_2(\sigma_1, \sigma_2, \mathit{path})$ is likely to be than the lower bound. We also consider how likely the lower bound on $\mathrm{diss}_g(\sigma_1, \sigma_2, \mathit{region})$ is to be accurate when σ_1 contains exactly one eigenvalue. Finally, in Section 8.5, we make probabilistic comparisons of sep and sep_λ, and compute the probability distribution of sep_λ. For ease of presentation we consider only complex matrices in this chapter; probabilistic statements for real matrices are in all cases similar and follow from analogous theorems for random real matrices in chapter 7.

8.2 The Probability of Being Able to Diagonalize a Matrix

We recall our path clustering criterion, introduced in chapter 1: we may decompose the spectrum σ of a matrix M into $\bigcup_i \sigma_i$ provided no perturbation of Euclidean norm ε or smaller can cause an eigenvalue in some σ_i to coalesce with an eigenvalue from some other σ_j. In this section we ask the

question: if the matrix M is chosen at random in the sense of chapters 6 and 7, what is the probability that M is decomposable into $\bigcup_i \{\lambda_i\}$? We may apply the result of section 7.3 to answer this question:

Theorem 8.1: Let M be a random complex matrix and $\varepsilon>0$ a constant. Then the probability that all matrices in the set M_ε of matrices within Euclidean distance ε of M are completely diagonalizable is

$$1 - n(n-1)^2(n+1) \cdot \varepsilon^2 + o(\varepsilon^2) \ .$$

Proof: A matrix $M' \in M_\varepsilon$ is not completely diagonalizable if and only if M is within ε of a matrix with a double eigenvalue. Now apply Theorem 7.4. Q.E.D.

This approach does *not* extend to computing the probability of being able to decompose σ of a random matrix M into $\bigcup_i \sigma_i$ where each σ contains at most r eigenvalues (rather than just 1). In other words, when $r>1$ it is not true that σ decomposes into $\bigcup_i \sigma_i$, $\#(\sigma_i) \le r$, if and only if M is not within ε of a matrix with an $r+1$-tuple eigenvalue. This is the point of Wilkinson's example

$$M = \begin{bmatrix} \eta & 1 & & \\ & 2\eta & 1 & \\ & & 3\eta & 1 \\ & & & 4\eta \end{bmatrix}$$

presented in chapter 1: when ε is on the order of η^5, $\sigma(M)$ cannot be decomposed at all even though M is not within η^2 of a matrix with a quadruple eigenvalue.

It is true, however, that the spectrum of a matrix within ε of one with an $r+1$-tuple eigenvalue is not decomposable $\sigma = \bigcup_i \sigma_i$ with $\#(\sigma_i) \le r$ for all i, so we still have an upper bound on the probability of such a decomposition:

Theorem 8.2: Let M be a random complex matrix and $\varepsilon>0$ a constant. Then the probability that $\sigma(M)$ is decomposable into σ_i of size at most $r<n$ (subject to the constraint imposed by ε) is at most

$$1 - f(n,r) \cdot \varepsilon^{2r} + o(\varepsilon^{2r}) \ ,$$

where $f(n,r)$ depends only on n and r.

Proof: If an n by n matrix has an $r+1$-tuple eigenvalue it can have at most $n-r$ distinct eigenvalues. Now apply Theorem 7.5. Q.E.D.

Whether the exponent $2r$ (or r in the real case) in this last theorem is best possible is an open question.

8.3 The Probability of a Random Matrix having all Projections of Small Norm

In this section we also are interested in the probability of being able to block diagonalize a random matrix, but now our decomposition criterion is different: we may decompose the σ of a matrix M into $\bigcup_i \sigma_i$ provided $\|P_i\| < K$ for all i, where P_i is the projection corresponding to σ_i. We consider the probability of completely diagonalizing M:

Theorem 8.3: Let M be a random complex matrix, with one dimensional projections P_i. Then

$$\text{Prob}(\|P_i\| < K \text{ for all } i) \geq 1 - 2n(n-1)^2(n+1) \cdot K^{-2} + o(K^{-2}) \ .$$

Proof: Since

$$\text{Prob}(\|P_i\| < K \text{ for all } i) = 1 - \text{Prob}(\text{ some } \|P_i\| \geq K) \ ,$$

it suffices to show that $2n(n-1)^2(n+1)\varepsilon^2 + o(\varepsilon^2)$ is an upper bound on Prob(some $\|P_i\| \geq K$). But by Lemma 5.1, $\|P_i\| \geq K$ implies that the relative distance from M to a matrix with a double eigenvalue is no more than $\sqrt{2/(K^2-1)}$, and the result follows from Theorem 7.4. Q.E.D.

We can combine this theorem with Theorem 3.3 to compute the approximate probability distribution of the condition number $\kappa(S) = \|S\| \cdot \|S^{-1}\|$ of either the best conditioned matrix S which diagonalizes a random matrix M: $S^{-1}MS = \mathrm{diag}(\lambda_i)$, or the nearly best conditioned S computed in chapter 3:

Theorem 8.4: Let M be a random complex matrix and let S be the nearly best conditioned diagonalizing similarity ($S^{-1}MS = \mathrm{diag}(\lambda_i)$) computed in chapter 3. Then

$$\mathrm{Prob}(\kappa(S) < K) \geq 1 - 2n^3(n-1)^2(n+1) \cdot K^{-2} + o(K^{-2}) \ .$$

This inequality is also true if S is the best conditioned similarity.

Proof: By Theorem 3.3, $\kappa(S) \leq n \cdot \max_i \|P_i\|$, so

$$\mathrm{Prob}(\kappa(S) < K) \geq \mathrm{Prob}(\max_i \|P_i\| < \frac{K}{n}) \ .$$

Now apply Theorem 8.3. Since $\kappa(S)$ is an upper bound for κ of a best conditioned $S_{OPTIMAL}$, these bounds also hold for $\kappa(S_{OPTIMAL})$. Q.E.D.

The probability of decomposing σ of M into $\bigcup_i \sigma_i$ where $\#(\sigma_i) \leq r$ and $\|P_i\| < K$, is clearly at least this large, but how much larger we do not know.

8.4 How Close are the Upper and Lower Bounds on $\mathrm{diss}_F(\sigma_1, \sigma_2)$**?**

In this section we consider the upper and lower bounds on $\mathrm{diss}_F(\sigma_1, \sigma_2)$

$$\sqrt{2}\ \mathrm{sep}_\lambda(A,B) \geq \mathrm{diss}_F(\sigma_1, \sigma_2) \geq \frac{\mathrm{sep}_\lambda(A,B)}{\|P\| + \sqrt{\|P\|^2 - 1}} \qquad (5.2)$$

for general σ_1 and also in the case where σ_1 contains just one simple eigenvalue. (The notation is from Chapter 5: we assume $\|M\|_F = 1$, and

$$M = \begin{bmatrix} A & C \\ & B \end{bmatrix} .)$$

We see immediately from Theorem 8.4 that with probability approaching 1 as $O(K^{-2})$ that the upper and lower bounds above will not differ by more than a factor of K *for all* σ_1.

To ask about the distance between the bounds for a *given* σ_1, however, is not a question that lends itself to a probabilistic interpretation, because there is no realistic way to probabilistically model the human choice of one σ_1 or another. Our approach still lets us measure the volume of the set of matrices for which one bound is more accurate than another, though; we only must not attribute a probabilistic interpretation to it.

We consider the special case where σ_1 contains a single eigenvalue of multiplicity one. In particular, we claim that for most matrices the lower bound in 5.1 above is more accurate than the upper bound. This claim is justified by considering the second example of section 4.4:

$$M = \begin{bmatrix} \alpha & c_1 & \cdot & c_n \\ & b_1 & & \\ & & \cdot & \\ & & & b_n \end{bmatrix} = \begin{bmatrix} A & C \\ & B \end{bmatrix},$$

for which we showed the lower bound in 5.1 is accurate to within a small factor. For a general matrix, B will not be diagonal as in the example but by Theorem 8.4 it will be diagonalizable by a similarity transformation whose condition number exceeds K on a set of matrices whose volume goes to zero as K goes to infinity. Thus, by Lemma 2.1, $\mathrm{diss}_F(\sigma_1, \sigma_2)$ will not exceed its lower bound by more than a factor of K where K is only large on a small set of matrices: those where B has a Jordan block of size at least 2 with eigenvalue α. In particular, as long as M is not close to having a *triple* eigenvalue at α (such matrices having codimension 4 in the complex case by Theorem 7.5), then the lower bound will be nearly accurate.

8.5 How much smaller than sep_λ can sep be?

We originally posed this question in Chapter 5, where we showed in Corollary 5.2 that the difference between upper and lower bounds on $\mathrm{diss}_E(\sigma_1, \sigma_2)$ depended on how much smaller sep could be than sep_λ. By Theorem 5.2

$$2\,\mathrm{sep}_\lambda \geq \mathrm{sep} \geq \frac{2}{n_A}\left(\frac{\mathrm{sep}_\lambda}{2}\right)^{n_A} .$$

The example in section 5.4

$$A = \begin{bmatrix} \varepsilon & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \varepsilon \end{bmatrix} \quad \text{and} \quad B = -A^T$$

showed that sep can indeed be as small as sep_λ^2 when $n_A>1$. When $n_A>2$ we have yet to find an example where sep can shrink as fast as $\mathrm{sep}_\lambda^{n_A}$, and experience in constructing examples suggests that none exist. In this section we show that it is unlikely that sep and sep_λ differ by much; indeed it is unlikely that they differ much from their trivial upper bounds (see Lemmas 2.8 and 2.15)

$$\mathrm{sep} \leq \min_{ij} |\lambda_i(A) - \lambda_j(B)| \tag{8.1.a}$$

and

$$\mathrm{sep}_\lambda \leq \frac{\min_{ij} |\lambda_i(A) - \lambda_j(B)|}{2} . \tag{8.1.b}$$

The experience in constructing examples mentioned above leads us to conjecture:

There is a constant c_n depending only on dimension such that

$$\mathrm{sep} \geq c_n \mathrm{sep}_\lambda^2 ,$$

but we will not pursue a proof of this claim here, except to say that it is possible to prove such an inequality for matrices bounded away from derogatory

ones.

To prove these claims, we need to introduce yet another nested family of pejorative varieties. Consider the Euclidean space $\mathbb{C}^{n_A^2+n_B^2}$, where the first n_A^2 coordinates represent A and the other n_B^2 coordinates represent B in the obvious way. Thus, we may refer to a point in $\mathbb{C}^{n_A^2+n_B^2}$ by its coordinates (A,B). We will need two norms on this space, the usual Euclidean one

$$\|(A,B)\|_E = \sqrt{\|A\|_E^2 + \|B\|_E^2}$$

and

$$\|(A,B)\| = \max(\|A\|_E, \|B\|_E) .$$

Clearly

$$\|(A,B)\| \le \|(A,B)\|_E \le \sqrt{2}\,\|(A,B)\| . \tag{8.2}$$

As usual we are primarily interested in what happens on the unit sphere $\{(A,B) : \|(A,B)\|_E = 1\}$. Theorem 5.2 is valid on (and inside) this set, and so implies that sep and sep_λ must approach zero simultaneously. This motivates investigating the set $P = \{(A,B) : \mathrm{sep}(A,B) = \mathrm{sep}_\lambda(A,B) = 0\}$. Not surprisingly, P is a homogeneous variety; in fact it is the zero set of the single irreducible order $n_A \cdot n_B$ polynomial $\det(\Psi_{A,B}) = \pm\det(\Psi_{B,A})$. What is the shortest distance from a given (A,B) to P? In analogy to the notation of chapters 6 and 7, we denote the minimum distance $\mathrm{dist}_E((A,B),P)$ if we use the $\|\cdot\|_E$ norm, and $\mathrm{dist}_2((A,B),P)$ if we use the $\|\cdot\|$ norm. It is immediate from the definition of $\mathrm{sep}_\lambda(A,B)$ that $\mathrm{sep}_\lambda(A,B)$ is precisely the distance from (A,B) to P in the $\|\cdot\|$ norm; we record this fact as

Lemma 8.5:

$$\mathrm{sep}_\lambda(A,B) = \mathrm{dist}_2((A,B), P)$$

This immediately suggests applying Theorems 6.3 and 6.4 to P to prove

Theorem 8.6: Let (A,B) be chosen at random so that $(A,B)/\|(A,B)\|_E$ is uniformly distributed on the unit sphere in $\mathbb{C}^{n_A^2+n_B^2}$. Then

$$(n_A^2+n_B^2-1)n_A n_B \cdot \varepsilon^2 + o(\varepsilon^2) \le \text{Prob}(\text{sep}_\lambda(\frac{A}{\|(A,B)\|_E}, \frac{B}{\|(A,B)\|_E}) \le \varepsilon)$$

$$\le 2(n_A^2+n_B^2-1)n_A n_B \cdot \varepsilon^2 + o(\varepsilon^2)$$

Proof: The proof is a by now routine application of Theorems 6.3 and 6.4, combined with Lemma 8.6 and inequality (8.2). Q.E.D.

Note that it is possible to randomly choose a pair (A,B) not only so that $(A,B)/\|(A,B)\|_E$ is uniformly distributed on the sphere as required by the theorem but also so that A and B are statistically independent (e.g. let each entry of A and B be an independent Gaussian random variable with mean 0 and variance 1). We do not know if this method of choosing A and B is a realistic model of the distributions induced by choosing the original T matrix at random, but we will use this method for the rest of this chapter anyway.

We begin by showing that neither sep nor sep_λ are likely to be significantly smaller than their trivial upper bounds in (8.2). From Lemma 2.8 we have

$$\min_{ij} |\lambda_i(A) - \lambda_j(B)| \ge \text{sep}(A,B) \ge \frac{\min_{ij} |\lambda_i(A) - \lambda_j(B)|}{\kappa(S_A) \cdot \kappa(S_B)}$$

where S_A is a (best conditioned) diagonalizing similarity for A and S_B similarly diagonalizes B. Since in our model A and B are chosen independently, we can use Theorem 8.4 to estimate the distribution of $\kappa(S_A) \cdot \kappa(S_B)$, the ratio of the upper to lower bounds in the last inequality. A little manipulation shows that the probability of this ratio exceeding K is $O(K^{-1})$.

The ratio of the upper to lower bounds on sep_λ in Lemma 2.15

$$\frac{\min_{i,j} |\lambda_i(A) - \lambda_j(B)|}{2} \geq \mathrm{sep}_\lambda(A,B) \geq \frac{\min_{i,j} |\lambda_i(A) - \lambda_j(B)|}{2 \cdot \max(\kappa(S_A), \kappa(S_B))}$$

is

$$\max(\kappa(S_A), \kappa(S_B)) .$$

An application of Theorem 8.4 shows that this last quantity exceeds K with probability $O(K^{-2})$.

The ratio of the upper bound on sep_λ to the lower bound on sep, a crude upper bound on $\mathrm{sep}_\lambda / \mathrm{sep}$, is

$$\frac{\kappa(S_A) \cdot \kappa(S_B)}{2} .$$

which from the above discussion exceeds K with probability $O(K^{-1})$.

A more detailed way to bound $\mathrm{sep}_\lambda / \mathrm{sep}$ is as follows: we identify a variety V with the property that for pairs (A,B) sufficiently far from V, $\mathrm{sep}_\lambda(A,B) / \mathrm{sep}(A,B)$ will be bounded above by a constant depending on the distance from V. From previous discussion we know V has to lie within the set P where $\mathrm{sep}=\mathrm{sep}_\lambda=0$. On the set P we know A and B have to have at least one eigenvalue λ in common. By lemmas 2.6 and 2.7 (for sep) and lemmas 2.13 and 2.14 (for sep_λ) we know it suffices to look at the parts of A and B with common eigenvalue λ. By lemma 2.12 we know that as long as λ is a simple eigenvalue of either A or B, then sep and sep_λ can not differ too much. Thus, V must consist of those pairs (A,B) where A and B have a common eigenvalue λ, and where both A and B have λ as a multiple eigenvalue. This is a total of 3 independent constraints, meaning V has codimension 3 in the real case and 6 in the complex. As long as the pair (A,B) does not fall into a small neighborhood of this set V of high codimension, $\mathrm{sep}_\lambda / \mathrm{sep}$ will not be too big.

Chapter 9: Relevance of the Probabilistic Model to Finite Precision Calculations

9.1 Introduction

What relevance does the probabilistic model of the last three chapters have to actual finite precision calculations? We will see that the model can predict certain behaviors of algorithms designed to solve the problems of chapter 7 (matrix inversion, eigendecompositions, polynomial zero finding). The tool required to analyze these algorithms is backwards error analysis; using it one can show that unless the problem to be solved is too close to some set P of ill-posed problems, a "backwards stable algorithm" will compute an accurate answer. For example, engineers have a rule of thumb that "to get an answer to a certain precision (say 3 decimal places) it suffices to do the intermediate calculations to about twice that precision (6 decimal places)" [Kahan2]. It will turn out that the model predicts this behavior by the measuring the relative rarity of matrices with triple eigenvalues (or polynomials with triple zeros) compared to matrices with double eigenvalues (or polynomials with double zeros).

The explanatory power of the model is limited by the underlying probability distribution of problems it assumes: $M / \| M \|_g$ should be uniformly distributed, where M is a random problem. Certain classes of problems simply do not generate this distribution. For example, we will see later that using Rayleigh quotient iteration to compute eigenvectors of a symmetric matrix requires solving a sequence of more and more nearly singular systems of linear equations. In fact, the more nearly singular the system, the better the resulting answer. It is clearly nonsense to model the set of matrices being inverted as coming from a uniform distribution. Other classes

of problems with predilections for producing nearly singular matrices are least squares problems (when solved using the normal equations) and finite difference schemes for differential equations. The most significant limitation of the model is that it uses a continuous distribution of problems; all points on the unit sphere are in a certain sense equally likely. In actual computations, however, the set of possible problems is discrete and finite. There are only a (huge) finite number of finite precision numbers representable in a computer, and hence only a finite number of finite precision matrices, polynomials, etc. It will turn out that this discreteness leads to qualitatively different behavior of algorithms than is predicted by the model. The continuous model makes sense only so long as the finite precision numbers are dense enough to resemble the continuum. In Figure 9.1, for example, the volume of the set of points within distance 4ε of the curve P is a good approximation to the number of dots (finite precision points) within distance 4ε of P. This is true because the radius of the neighborhood of P (4ε) is large compared to the spacing between finite precision points (ε). In Figure 9.2, on the other hand, the volume of points within distance $\varepsilon/4$ of P is not necessarily a good approximation of the number of dots within $\varepsilon/4$ of P. Thus, when the radius of the neighborhood of P get smaller than the interdot distance ε, the model breaks down. The breakdown of the model is critical if one is trying to analyze the behavior of real algorithms running in finite precision arithmetic. For example, we will see that one can measure the difficulty of inverting a matrix with the condition number $\kappa'(M)=\|M\|_E \cdot \|M^{-1}\|$. When using an iterative algorithm to compute the inverse, the number of iterations needed, if very large, is roughly proportional to $\kappa'(M)$ for many algorithms. Thus, we could ask what (according to the model) is the average number of iterations needed to invert a random matrix? This is roughly proportional to

the average condition number. In the case of real matrices, it will turn out that the model predicts an *infinite* average condition number. In fact, the model predicts that the average condition number of matrices whose condition numbers are restricted to be greater than K is infinite for any positive K. This is because the integral expressing the average condition number looks like

$$\int_K^\infty \frac{dx}{x} \tag{9.1}$$

which is infinite for all positive K. However, since there are only finitely many finite precision matrices, the average condition number of those that are not exactly singular must be finite. Thus, the model does not supply us any useful information in this case.

What if we could compute the actual probability distribution of the number of points within distance ε of the variety P of ill-posed problems for the finite precision case? It would tell us how many single precision problems we could solve as a function of the extra precision used in intermediate calculations. For example, in the case of inverting real matrices, if the actual probability distribution were roughly linear as in the continuous case, then each bit of extra precision used would allow us to solve half the problems we couldn't solve before. We present some simulations to substantiate this claim below. Clearly, such information would be of great use in the design of numerical algorithms or even computer arithmetic units, because it would tell the designer how to trade off the cost of arithmetic (which is an increasing function of the number of bits of precision) with the number of problems the system can solve.

The rest of this chapter is organized as follows. Section 9.2 shows how backwards error analysis makes the probability model relevant to finite

precision calculations. Section 9.3 discusses the limitations of the model mentioned above. Finally, section 9.4 demonstrates the usefulness of knowing the discrete distribution of problems for analyzing the use of extra precision arithmetic.

9.2 A Paradigm for Analyzing the Accuracy of Algorithms

The paradigm for applying the probabilistic model to the analysis of algorithms is as follows:

(1) Within the space of problems, identify the set P of ill-posed ones.

(2) Show that the closer a problem is to P, the more sensitive the solution is to small changes in the problem.

(3) Show that the algorithm in question computes an accurate solution for a problem close to the one it received as input (this is known as "backwards stability" [Wilkinson1]). Combined with the result of (2), this will show that the algorithm will compute an accurate solution to a problem so long as the problem is far enough from P.

(4) Compute the probability that a random problem is close to P. Using this probability distribution in conjunction with the result of (3) we can compute the probability of the algorithm computing an accurate result.

This paradigm is best explained by applying it to matrix inversion:

(1) The set of matrices P which are ill-posed with respect to inversion are precisely the singular matrices.

(2) As discussed in section 7.2, the condition number

$$\kappa'(M) = \| M \|_F \cdot \| M^{-1} \| \tag{9.2}$$

measures how difficult the matrix M is to invert. More precisely, it measures how much a relative perturbation in M can be magnified in

M^{-1} [Kahan1]:

$$\kappa'(M) = \limsup_{\delta M \to 0} \frac{\| (M+\delta M)^{-1} - M^{-1} \| \;/\; \| M^{-1} \|}{\| \delta M \|_E \;/\; \| M \|_E} . \tag{9.3}$$

As also discussed in section 7.2, the condition number can be expressed in terms of the distance from M to P:

$$\kappa'(M) = \| M \|_E \;/\; \mathrm{dist}_E(M\,,\,P) \tag{9.4}$$

as required by the paradigm.

(3) Gaussian elimination with partial pivoting is a standard algorithm for matrix inversion and is well known to be a backwards stable algorithm [Wilkinson1]. Backwards stability means that when applying Gaussian elimination to compute the solution of the system of linear equations $Ax=b$, one gets an answer $\hat{x}$ which satisfies $(A+\delta A)\hat{x}=b$, where δA is small in norm compared to A. More exactly, let X_i be the i-th column of the approximation to M^{-1} computed using Gaussian elimination, where the arithmetic operations performed (addition, subtraction, multiplication, and division) are all rounded off to b bits of precision. Then X_i is the *exact* value of the i-th column of the inverse of a matrix $(M(i))^{-1}$ where $M(i)$ is close to M (the subscript i means column i, and $M(i)$ is the i-th in a sequence of n by n matrices). In fact

$$\| M(i) - M \|_E \le f(n) \cdot 2^{-b} \cdot \| M \|_E \tag{9.5}$$

where $f(n)$ is a function only of n, the dimension of M [Wilkinson1]. This last expression can be used to bound the relative error in the solution X_i [Wilkinson1]:

$$\frac{\| X_i - (M^{-1})_i \|_E}{\| (M^{-1})_i \|_E} \le \frac{\kappa'(M) \cdot f(n) \cdot 2^{-b}}{1 - \kappa'(M) \cdot f(n) \cdot 2^{-b}} . \tag{9.6}$$

In other words, as long as the bound on the distance from $M(i)$ to M is not so large that $M(i)$ could be singular, i.e. as long as

$$\mathrm{dist}_E(M\ ,\ P) > f(n)\cdot 2^{-b}\cdot \|M\|_E \geq \|M(i)-M\|_E \tag{9.7}$$

or, substituting from equation (9.4)

$$\kappa'(M)\cdot f(n)\cdot 2^{-b} < 1\ , \tag{9.8}$$

then the relative error in the computed inverse X is bounded. So, as long the condition number $\kappa'(M)$ is smaller than $2^b/f(n)$, the solution will have some accuracy, and the smaller $\kappa'(M)$, the more accurate the solution.

(4) Now we can apply Corollary 7.2 which gives the probability distribution of the condition number to estimate the probability that a random matrix can be inverted accurately:

$$\mathrm{Prob}\left(\frac{\|X_i-(M^{-1})_i\|_E}{\|(M^{-1})_i\|_E}\leq \varepsilon\right) \geq \mathrm{Prob}\left(\frac{\kappa'(M)\cdot f(n)\cdot 2^{-b}}{1-\kappa'(M)\cdot f(n)\cdot 2^{-b}}\leq \varepsilon\right) \tag{9.9}$$

which, after some rearrangement (and assuming $\varepsilon<1$ of course) equals

$$= \mathrm{Prob}\left(\kappa'(M)\leq \frac{\varepsilon}{f(n)\cdot(1+\varepsilon)\,2^{-b}}\right)$$

$$= 1-n(n^2-1)\cdot f(n)^2\cdot\left(\frac{1+\varepsilon}{\varepsilon}\right)^2\cdot 2^{-2b} + o\left(f(n)^2\cdot\left(\frac{1+\varepsilon}{\varepsilon}\right)^2\cdot 2^{-2b}\right)$$

(assuming M is complex and applying Corollary 7.2). This last inequality only makes sense for

$$f(n)\cdot\frac{1+\varepsilon}{\varepsilon}\cdot 2^{-b}$$

small, that is if the precision 2^{-b} used in the computation is much smaller than the precision ε demanded of the answer. This restriction also makes sense numerically.

Similar analyses are possible of standard algorithms to compute eigenvalues and eigenvectors as well as zeros of polynomials [Wilkinson1, Wilkinson2]. In the case of eigenvalue problems, the ill-posed set P consists of those matrices with multiple eigenvalues, the higher the multiplicity the more ill-posed the problem. Why is this? It is well known that the eigenvalues

of a matrix are algebraic functions of the entries. In particular, if λ is a simple eigenvalue of the matrix A, and if B is another matrix, then the parameterized matrix $A+\varepsilon B$ will have an eigenvalue $\lambda(\varepsilon)$ such that $\lambda(0)=\lambda$ and for ε in a neighborhood of 0, $\lambda(\varepsilon)$ will be expressible as a power series in ε [Kato2]. Interpreting εB as a perturbation in A due to measurement or roundoff error, we see that the perturbation $\lambda(\varepsilon)-\lambda$ in A's eigenvalue λ depends at worst linearly on ε for small ε. Now what if λ is an m-tuple eigenvalue of A? Then it is well known [Wilkinson2] that $\lambda(\varepsilon)$ will generically be expressible as a power series in $\varepsilon^{1/m}$ for small ε, so that a perturbation εB of order ε in A results in a perturbation $\lambda(\varepsilon)-\lambda$ of order $\varepsilon^{1/m}$ in A's eigenvalue. Since for small ε and $m>1$, $\varepsilon^{1/m}$ is much larger than ε, this means that errors made in the computation of multiple *or nearly multiple* eigenvalues are large compared to the errors in a simple eigenvalue, and the higher the multiplicity of the eigenvalue, the worse the error.

We can relate the error made to the multiplicity of the eigenvalue being computed in a more precise way. Almost all algorithms used to compute eigenvalues are backwards stable in the sense that they compute the eigenvalues of a matrix near the one supplied as input. As is the case of matrix inversion, the distance from the input matrix to the nearby one depends on the precision 2^{-b} used in the calculations. Thus, the εB perturbation of the last paragraph is of order 2^{-b}. Therefore, the error in the computed value of an m-tuple eigenvalue will be of order $2^{-b/m}$ by the argument of the last paragraph. In other words, if we do our calculations using b bits of precision, we can only expect about b/m bits of precision in the computed value of an m-tuple eigenvalue. If $m=2$, for example (a double eigenvalue), we expect to lose half our precision. This analysis tells us how much precision is needed to compute eigenvalues accurately to the basic precision 2^{-b}. Since we lose half

the precision when computing double eigenvalues, double precision 2^{-2b} will get double eigenvalues accurate to single precision. Similarly, m-tuple precision 2^{-mb} is need to expect to compute m-tuple eigenvalues accurately.

How likely are multiple eigenvalues according to our probabilistic model? According to theorem 7.5, the codimension of the set of complex n by n matrices with at least one m-tuple eigenvalue is $2(m-1)$, and so by Theorem 6.3 the distribution of matrices within distance ε of one with an m-tuple eigenvalue is asymptotically proportional to $\varepsilon^{2(m-1)}$. As m grows, the exponent of ε increases, and $\varepsilon^{2(m-1)}$ decreases. Said another way, for small enough ε, matrices with multiple (double or more) eigenvalues are very rare in the set of all matrices, matrices with triple (or more) eigenvalues are very rare in the set of matrices with multiple eigenvalues, and so on.

Recall now the engineer's rule of thumb: "double precision in intermediate calculations is enough to get the answer to single precision." The model can be used to explain this empirical observation. Most eigenvalue problems involve simple eigenvalues, and for these single precision suffices to compute a satisfactory answer. Rarely, one has to compute a nearly double eigenvalue, and for these double precision suffices. Much more rarely, one needs yet higher precision, but the occurrence of these triple and higher multiple zeros is so rare that double precision is almost always enough. A similar analysis applies to computing multiple zeros of polynomials.

The discussion of the last few paragraphs has been far from rigorous, using asymptotic results of dubious validity to explain an empirical observation stated without evidence. Nonetheless, it demonstrates the power of the paradigm stated at the beginning of this section. In the next section we discuss when the results of the model are indeed inapplicable and misleading.

We may use the same kind of paradigm as discussed so far to analyze the speed of convergence of an algorithm rather than its accuracy. In this case the paradigm is

(1') Identify the ill-posed problems P'.

(2') Show that the closer a problem is to P', the slower the algorithm converges.

(3') Compute the probability that a random problem is close to P'. Combined with (2') this yields the probability distribution of the speed of convergence.

This approach has been used by Smale [Smale] in his average speed analysis of Newton's method for finding zeros of polynomials.

9.3 Limitations of the Probabilistic Model

In this section we discuss two examples illustrating the breakdown of the model. Both examples show behavior of widely used algorithms which disagrees with the predictions of the model because of the effects of finite precision arithmetic. In addition, the first example shows how the assumption of uniformity of $M / \| M \|_E$ breaks down even in exact arithmetic.

The first example is Rayleigh quotient iteration, which is used to compute the eigenvalues and eigenvectors of a symmetric matrix A. If x_0 is an initial guess at an eigenvector, the algorithm proceeds as follows:

$$\lambda_i = x_i^T A x_i \;/\; x_i^T x_i$$

$$x_{i+1} = (A - \lambda_i)^{-1} x_i \quad .$$

The idea is that if x_i is a good approximation to an eigenvector, then λ_{i+1} (the Rayleigh quotient) is a good approximation to an eigenvalue, and in turn x_{i+1} is an even better approximate eigenvector. In fact, the asymptotic convergence rate is cubic under some weak assumptions on the distribution of A's

eigenvalues (i.e. $|\lambda_{TRUE}-\lambda_{i+1}|$ is of the order of $|\lambda_{TRUE}-\lambda_i|^3$ when $|\lambda_{TRUE}-\lambda_i|$ is small enough [Parlett]). Note that as λ_i converges to an eigenvalue, the more nearly singular the matrix $A-\lambda_i$ becomes. In exact arithmetic, of course, if x_i is an exact eigenvector, $A-\lambda_i$ will be exactly singular.

The sequence of matrices to be inverted (actually, one solves the linear systems $(A-\lambda_i)x_{i+1}=x_i$ directly rather than compute $(A-\lambda_i)^{-1}$) becomes more and more nearly singular, so that the distribution of matrices to be (conceptually) inverted is far from uniformly distributed. This invalidates the assumption of the model, even in exact arithmetic. How does Rayleigh quotient iteration behave in finite precision arithmetic? The discussion of section 9.1 might lead us to doubt that it works at all, since we showed there that one can not expect an accurate solution to a nearly singular system of linear equations. In fact, Rayleigh quotient iteration works extremely well because the rounding errors committed in the course of computing x_{i+1} provably conspire to produce an error lying almost certainly in the direction of the desired eigenvector. In fact, when λ_i has almost converged to an eigenvalue, the rounding errors will swamp the computation so that x_{i+1} almost certainly becomes the desired eigenvector and further iterations serve only to make small, random changes in x without improving its accuracy. In other words, there is an effect due to finite precision arithmetic which makes the algorithm converge very quickly, so the asymptotically cubic convergence rate is rarely observed for long. Therefore, any average speed analysis of Rayleigh quotient iteration which ignores the effects of finite precision arithmetic may be misleading. For a further discussion of Rayleigh quotient iteration see [Parlett].

For the second example we return to matrix inversion. As discussed in section 9.1, the condition number is a measure of the anticipated accuracy in the computed inverse of a matrix. It can also be used to measure the speed of convergence of many iterative algorithms for computing the inverse to within a given precision [Wilkinson2]. Therefore, a reasonable question to ask is the following: what is the expected condition number of a random real matrix? Let us see what the model says. Even though we only have an asymptotic upper bound on the distribution of $\kappa'(M)$ of the form (Corollary 7.2):

$$\mathrm{Prob}(\kappa'(M) \geq K) \leq \mathrm{const} \cdot K^{-1} + o(K^{-1})$$

it is clear that for large enough K the probability distribution function will also be bounded below by a constant multiple of K^{-1}. This is because within the variety of singular matrices lies a manifold (perhaps small) of codimension 1 which does have an ε spherical neighborhood for sufficiently small ε (sufficiently large K) which by Weyl's theorem has a volume given asymptotically by a constant multiple of K^{-1}. Thus, the integral which expresses the expected condition number will be bounded below by

$$E(\kappa'(M)) \geq \int_{K_0}^{\infty} \mathrm{const} \cdot \frac{dK}{K}$$

and this integral diverges to infinity no matter how large K_0 is. However, since there are only a finite number of finite precision matrices, there is some K_0 such that no finite precision matrix that is not exactly singular has a condition number greater than K_0. Therefore the value of the integral is determined entirely by integrating over a range of condition numbers which do not correspond to any finite precision matrices. Clearly, this model is not telling us anything useful in this case.

In the case of complex matrices, the corresponding integral does converge, because for sufficiently large κ' it is dominated by

$$\int_{X_0}^{\infty} \text{const} \cdot \frac{dK}{K^2} ,$$

which converges. The results are however still not trustworthy because we are integrating over the region within distance 2^{-b} of the variety P of singular matrices, where 2^{-b} is the separation between adjacent finite precision numbers (see Figure 9.2), where the model breaks down. In the next section we discuss what we could do if we could extend the model to this region close to P.

9.4 How to Use the Discrete Distribution of Points Within Distance ε of a Variety

Before proceeding, we need to say what probability measure we are going to put on the discrete set of finite precision points. Section 9.3 showed that no single distribution is good for all applications, but a uniform distribution remains a neutral and interesting choice. Therefore for the sake of discussion the probability we assign to the point M will be proportional to the volume of the small parallelepiped of points which round to M (i.e. the parallelepiped centered at M with sides equal in length to the distance between adjacent finite precision points). In the case of fixed point arithmetic [Wilkinson1], this means that each point has equal probability, whereas with floating point arithmetic points near 0 have smaller probability than larger points, since points near 0 are closer together than points farther away. (Actually, the question of the distribution of the digits of a floating point number has a large literature [Hamming, Bareiss]. The discussion in this section does not depend on the actual distribution of digits chosen).

We claim that knowing the probability distribution of the distance of a random finite precision problem to the set P of ill-posed problems will tell us how many finite precision problems we can solve as a function of the extra

precision used in intermediate calculations. As mentioned before, programmers often resort to extra precision arithmetic to get more accurate solutions to problems which are given only to single precision. This extra precision has a cost (speed) dependent on the number of digits carried, so programmers usually avoid extra precision unless persuaded otherwise by bad experiences, an error analysis, or paranoia. Therefore an accurate estimate of how many problems can be solved as a function of the extra precision used would not only help programmers decide how much to use but possibly influence designers when they decide how much precision to make available in their computer systems.

How does knowledge of this probability distribution tell us how much extra precision to use? The paradigm in section 9.2 tells us how. A backwards stable algorithm using extra precision gets an accurate solution to a problem in a small ball around the input problem. The radius of this ball depends on the extra precision used. Therefore, we can expect to accurately solve problems lying within 2^{-b} of P, where 2^{-b} is the distance between adjacent finite precision numbers in the input data, since the small ball around the input problem will be bounded away from the set P. The probability distribution tells us as before how many problems lie within a given distance of P, and so it tells us how many problems we can solve that we couldn't solve before.

This discussion has assumed so far that the finite precision input is known exactly, i.e. that there is no error inherited from previous computations or from measurement errors. In general there will be such errors, and they will almost always be at least a few units in the last place of the input problem. In other words, there already is a ball of uncertainty around the input problem with a radius equal to a small multiple of the interpoint dis-

tance 2^{-b}. Therefore, it may make no sense to use higher precision to accurately solve problems lying very close to P when the inherited input error is so large that the true answer is inherently very uncertain. In such situations programmers usually shrug and settle for the backwards stability provided by the algorithm, even if the delivered solution is entirely wrong, because the act of solution has scarcely worsened the uncertainty inherited from the data, and the programmer declines to be held responsible for the uncertainty inherent in the data.

Nevertheless, we close with an example using the actual discrete distribution. Consider the rather simple problem of inverting 2 by 2 matrices. This problem is small enough that we can actually exhaustively compute the desired discrete probability distribution for low precision arithmetic. We did this for 3, 4, 5, 6 and 7 bit fixed point arithmetic (all numbers lay between 0 and 1 in absolute value), where each fixed point matrix was assigned the same probability. In all cases, we observed approximately linear behavior of the probability distribution (as predicted by the continuous model) both for distances ε to the nearest singular matrix larger than 2^{-b} ($3 \le b \le 7$), and for ε smaller than 2^{-b} (the fraction of problems within 2^{-b} of a singular matrix was about 2^{1-b}). This linear behavior continued until ε reached approximately 2^{-2b}, and there the graph of the distribution became horizontal and remained so all the way to the origin, intersecting the vertical axis at about 2^{2-2b}, meaning that all matrices closer to P than approximately 2^{-2b} were exactly singular, and that the fraction of matrices which were exactly singular was 2^{2-2b}. See Figure 9.3 for a rough sketch of this observed probability distribution. What does this tell us about the use of extra precision? Basically, as long as the distribution function remains linear, it says that for every extra bit of intermediate precision, we can solve half the problems we

couldn't solve before. This regime continues until we reach double precision, at which point the only problems we can't solve are exactly singular. Indeed, since

$$\begin{bmatrix} a & c \\ b & d \end{bmatrix}^{-1} = (ad-bc)^{-1} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

we can clearly compute the inverse accurately if we can compute the determinant $ad-bc$ accurately. Since a, b, c and d are given to single precision, double clearly suffices to compute $ad-bc$ exactly.

Of course, exhaustive evaluation of the distribution function is not reasonable for large problems, and evaluating the distribution function becomes an interesting question of Diophantine approximation.

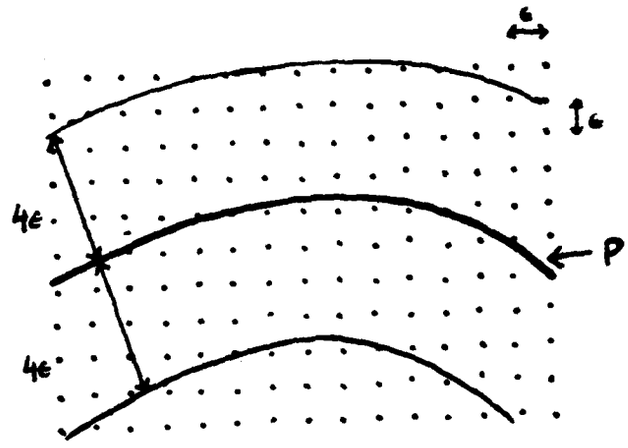


Figure 9.1 A 4ε neighborhood of the curve P

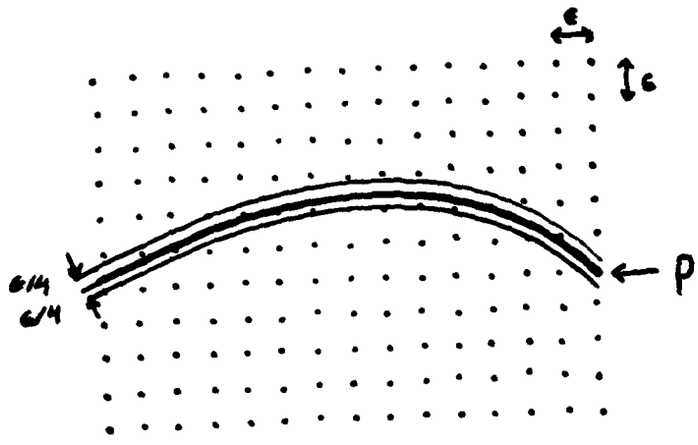


Figure 9.2 An ε/4 neighborhood of the curve P

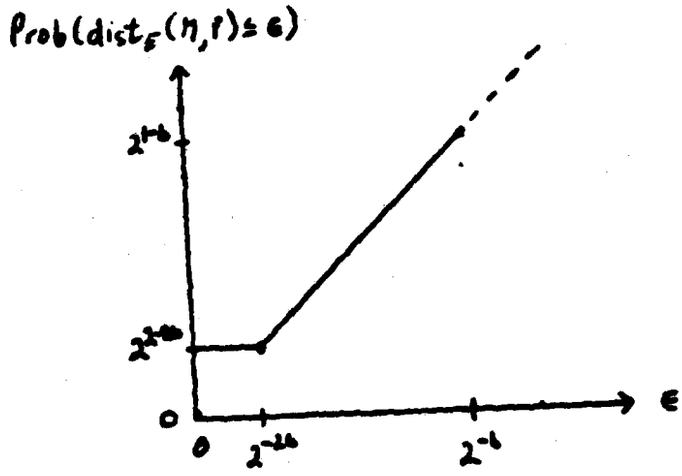


Figure 9.3 Observed Probability Distribution of the Distance ϵ to the Nearest Singular Matrix

References

[Arnold] V. I. Arnold, "On Matrices Depending on Parameters", Russian Math. Surveys, Jan-June 1971, vol 26:1-3

[Bareiss] E. H. Bareiss and J. L. Barlow, "Probabilistic Error Analysis of Floating Point and CRD Arithmetics", Dept. of Electrical Engineering and Computer Science, Northwestern University, Report 81-02-NAM-01, 1981

[Bart] H. Bart, I Gohberg, M. A. Kaashoek, P. van Dooren, "Factorizations of Transfer Functions", SIAM J. Control, vol 18, no 6, November 1980, pp 675-696

[Bauer1] F. L. Bauer and C. T. Fike, "Norms and Exclusion Theorems", Numer. Math., 2, pp 137-141, 1960

[Bauer2] F. L. Bauer, A. S. Householder, "Some inequalities involving the euclidean condition of a matrix", Numer. Math., 2, pp 308-311, 1960

[Bauer3] F. L. Bauer, "A further generalization of the Kantorovic inequality", Numer. Math., 3, pp 117-119, 1961

[Bauer4] F. L. Bauer, "Optimally scaled matrices", Numer. Math., 5, pp 73-87, 1963

[Davis] C. Davis and W. Kahan, "Some new bounds on perturbations of subspaces", Bull. A. M. S., 75, 1969, pp 863-8

[Demmel] J. Demmel, "The Condition Number of Equivalence Transformations that Block Diagonalize Matrix Pencils", to appear in SIAM J. Numer. Anal., also in *Lecture Notes in Mathematics #973: Matrix Pencils*, Springer-Verlag, Berlin, 1982

[Dunford] Nelson Dunford, "Spectral Operators", Pacific J. of Math. Vol 4, No 2, Sept 1954, pp321-354

[Eckart] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank", Psychometrika, 1 (1936), pp 211-218

[Gantmacher] F. R. Gantmacher, *The Theory of Matrices*, trans. K. A. Hirsch, Chelsea, 1959, vol. 1 and 2

[Griffiths] Phillip A. Griffiths, "Complex Differential and Integral Geometry and Curvature Integrals Associated to Singularities of Complex Analytic Varieties", Duke Math. J., vol 45, no 3, Sept 1978, pp427-512

[Guillemin] Victor Guillemin and Alan Pollack, *Differential Topology*, Prentice Hall, Englewood Cliffs, 1974

[Hamming] R. W. Hamming, "On the Distribution of Numbers", Bell System Technical Journal, vol. 49, no. 8, 1970, pp 1609-1625

[Herstein] I. N. Herstein, *Topics in Algebra*, Blaisdell, Waltham, Massachusetts, 1964

[Isaacson] E. Isaacson and H. B. Keller, *Analysis of Numerical Methods*, Wiley, New Jersey, 1966

[Kågström1] B. Kågström and A. Ruhe, "An Algorithm for Numerical Computation of the Jordan Normal Form of a Complex Matrix", Department of Information Processing, Report UMINF-51.74, Umeå University, Umeå, Sweden, 1974

[Kågström2] B. Kågström, "Numerical Computation of Matrix Functions", Department of Information Processing, Report UMINF-58.77, Umeå University, Umeå, Sweden, 1977

[Kahan1] W. Kahan, "Conserving Confluence Curbs Ill-Condition", Technical Report 6, Computer Science Dept., University of California, Berkeley, August 4, 1972

[Kahan2] W. Kahan, class notes for CS246, System Support for Scientific

Computation, University of Calfornia at Berkeley, 1980

[Kato1] T. Kato, "Estimation of Iterated Matrices, with Application to the von Neumann Condition", Numer. Math., 2, pp 22-29, 1960

[Kato2] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1966

[Kendig] Keith Kendig, *Elementary Algebraic Geometry*, Springer-Verlag, New York, 1977

[Lelong] P. Lelong, *Fonctions plurisousharmoniques et formes differentielles positiv*, Paris, Gordon Breach, 1968

[Meyer] D. Meyer and K. Veselic, "New Inclusion Theorems for Partitioned Matrices", Numer. Math. 34, pp431-437, 1980

[Parlett] B. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980

[Ruhe1] A. Ruhe, "Properties of a Matrix with a Very Ill-conditioned Eigenproblem", Numer. Math., 15, pp 57-60, 1970

[Ruhe2] A. Ruhe, "An Algorithm for Numerical Determination of the Structure of a General Matrix", BIT 10, pp 196-216, 1970

[Santalo] Luis A. Santaló, *Integral Geometry and Geometric Probability*, *Encyclopedia of Mathematics and Its Applications*, vol 1, Addison-Wesley, Reading, 1976

[Schwarz] J. Schwartz, "Perturbations of Spectral Operators, and Applications: I. Bounded Perturbations", Pacific J. Math, pp415-458

[Smale] S. Smale, "The Fundamental Theorem of Algebra and Complexity Theory", Bulletin (New Series) of the A.M.S., vol 4, no 1, 1981, pp1-35

[Smith] R. A. Smith, "The Condition Numbers of the Matrix Eigenvalue Problem", Numer. Math., 10, pp 232-240, 1967

[Sternberg] S. Sternberg, *Lectures on Differential Geometry*, Prentice-Hall, Englewood Cliffs, New Jersey, 1964

[Stewart] G. W. Stewart, "Error and Perturbation Bounds for Subspaces Associated with Certain Eigenvalue Problems", SIAM Review, vol. 15, no. 4, p 752, Oct 1973,

[Sun] J. G. Sun, "The Perturbation Bounds for Eigenspaces of Definite Matrix Pairs", to appear

[Thie] P. Thie, "The Lelong number of points of a complex analytic set", Math. Ann., vol 172 (1967) pp. 269-312

[vanderSluis] A. van der Sluis, "Condition Numbers and Equilibration of Matrices", Numer. Math., 14, pp 14-23, 1969

[VanderWaerden] B. Van der Waerden, *Modern Algebra*, vol 1, Ungar, New York, 1953

[VanDooren] P. Van Dooren and P. Dewilde, "Minimal Cascade Factorization of Real and Complex Rational Transfer Matrices", IEEE Trans. on Circuits and Systems, vol. CAS-28, no. 5, May 1981, p 395.

[Varah] J. M. Varah, "On the Separation of Two Matrices", SIAM J. Numer. Anal., vol 16, no 2, April 1979

[Varga] David G. Feingold and Richard S. Varga, "Block Diagonally Dominant Matrices and Generalizations of the Gerschgorin Circle Theorem", Pacific Journal of Math, vol 12, no 4, winter 1962, p1241-1250

[Weyl] Hermann Weyl, "On the Volume of Tubes", Amer. J. of Math. vol 61 (1939), pp 461-472

[Wilkinson1] J. H. Wilkinson, *Rounding Errors in Algebraic Processes*, Prentice Hall, Englewood Cliffs, New Jersey, 1963

[Wilkinson2] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford

University Press, 1965

[Wilkinson3] J. H. Wilkinson, "Note on Matrices with a Very Ill-Conditioned Eigenproblem", Numer. Math., 19, pp 176-178, 1972

[Wilkinson4] J. H. Wilkinson, private communication, 1982

ATE

LMED

8