Numer. Math. 51, 251–289 (1987)

Numerische Mathematik

On Condition Numbers and the Distance to the Nearest Ill-posed Problem

James Weldon Demmel
Courant Institute of Mathematical Sciences, 251 Mercer Str. New York, NY 10012, USA

Summary. The condition number of a problem measures the sensitivity of the answer to small changes in the input. We call the problem ill-posed if its condition number is infinite. It turns out that for many problems of numerical analysis, there is a simple relationship between the condition number of a problem and the shortest distance from that problem to an ill-posed one: the shortest distance is proportional to the reciprocal of the condition number (or bounded by the reciprocal of the condition number). This is true for matrix inversion, computing eigenvalues and eigenvectors, finding zeros of polynomials, and pole assignment in linear control systems. In this paper we explain this phenomenon by showing that in all these cases, the condition number κ satisfies one or both of the differential inequalities $m \cdot \kappa^2 \leq \|D\kappa\| \leq M \cdot \kappa^2$, where $\|D\kappa\|$ is the norm of the gradient of κ. The lower bound on $\|D\kappa\|$ leads to an upper bound $1/(m\kappa(x))$ on the distance from x to the nearest ill-posed problem, and the upper bound on $\|D\kappa\|$ leads to a lower bound $1/(M\kappa(X))$ on the distance. The attraction of this approach is that it uses local information (the gradient of a condition number) to answer a global question: how far away is the nearest ill-posed problem? The above differential inequalities also have a simple interpretation: they imply that computing the condition number of a problem is approximately as hard as computing the solution of the problem itself. In addition to deriving many of the best known bounds for matrix inversion, eigendecompositions and polynomial zero finding, we derive new bounds on the distance to the nearest polynomial with multiple zeros and a new perturbation result on pole assignment.

Subject Classifications: AMS(MOS): 15A12, 15A60, 65F35; CR: F.2.1, G.1.0.

1. Introduction

The condition number of a problem measures the sensitivity of the answer to small changes in the input. We call the problem ill-posed if its condition number

number is infinite. The ill-posed problems typically form a lower dimensional surface in the space of problems. It turns out that for many problems of numerical analysis, there is a simple relationship between the condition number of a problem and the shortest distance from that problem to the surface of ill-posed ones: the shortest distance is proportional to the reciprocal of the condition number, or bounded by the reciprocal of the condition number. Sometimes, the distance is bounded below by the reciprocal of the condition number squared. This is true for matrix inversion, computing eigenvalues and eigenvectors, finding zeros of polynomials, and pole assignment in linear control systems.

For example, in the case of matrix inversion, if A is perturbed to $A+\delta A$, then to first order the solution A^{-1} becomes $A^{-1}+X$ where

$$\frac{\|X\|}{\|A^{-1}\|} \leq \|A^{-1}\| \cdot \|\delta A\|$$

($\|\cdot\|$ is any operator norm). Thus the condition number of this problem may be taken as $\|A^{-1}\|$. It is well known [8] that the shortest distance from A to the surface of singular (ill-posed) matrices is $1/\|A^{-1}\|$. Similar results for computing eigendecompositions and zeros of polynomials are due to Wilkinson [16] and Hough [7] and will be discussed further below.

In this paper we explain this phenomenon and unify the techniques used to obtain these results by showing that in all these cases, the condition number κ satisfies one or both of the following differential inequalities

$$m \cdot \kappa^2 \leq \|D\kappa\| \leq M \cdot \kappa^2 \tag{1.1}$$

almost everywhere, where $0<m \leq M$ and $\|D\kappa\|$ is the norm of the gradient of κ. From the lower bound on $\|D\kappa\|$ we will deduce that there is a curve $x(s)$ (the "steepest ascent" curve of κ), parametrized by arclength s and with $x(0)=x$, such that

$$\frac{d}{ds}\kappa(x(s)) \geq m \cdot \kappa^2(x(s)).$$

This last inequality can be integrated explicitly (see Lemma 1 below), yielding an upper bound on the distance $\operatorname{dist}(x, P)$ (which depends on $\|\cdot\|$) from x to the set P of ill-posed problems

$$\operatorname{dist}(x, P) \leq \frac{1}{m \cdot \kappa(x)}.$$

From the upper bound on $\|D\kappa\|$ we will deduce that if $x(s)$ is *any* smooth curve parametrized by arclength with $x(0)=x$, then

$$\frac{d}{ds}\kappa(x(s)) \leq M \cdot \kappa^2(x(s)).$$

This inequality can also be integrated explicitly (Lemma 2), yielding a lower bound

$$\frac{1}{M \cdot \kappa(x)} \leq \operatorname{dist}(x, P)$$

on the distance to the nearest problem with an infinite condition number.

For some problems we can prove a differential inequality of the form $\|D\kappa\| \leq M \cdot \kappa^3$ which yields a lower bound on the distance $1/(2M\kappa^2(x)) \leq \operatorname{dist}(x, P)$.

The attraction of this approach is that it uses purely local information (the norm of the gradient of the condition number) to answer a global question: how far away is the nearest point in a (generally quite complicated) set of ill-posed problems? This approach is quite similar in spirit to Wilkinson's "fast perturbation theory" for eigenvalues [17], with which we compare our method in Sect. 4 below. In fact, we argue in Sect. 4 that the idea behind using inequalities (1.1) to get distance estimates, doing steepest ascent on the condition number κ, generalizes Wilkinson's fast perturbation theory to a numerical approach applicable to a large class of problems.

The differential inequalities (1.1) have two simple and attractive interpretations. To state the first one we need to define the *relative condition number* of the mapping g at x as

$$\kappa_{\mathrm{rel}}(g, x) \equiv \limsup_{\delta x \to 0} \frac{\|g(x+\delta x) - g(x)\| / \|g(x)\|}{\|\delta x\| / \|x\|}$$

$$= \frac{\|D_g(x)\| \cdot \|x\|}{\|g(x)\|},$$

where the second definition is only true if the Fréchet derivative D_g of g exists. As is well known, κ_{rel} measures the maximum instantaneous relative change in g per relative change in x. It is easy to see that by multiplying (1.1) by $\|x\|/\kappa(x)$, we get

$$m \cdot \kappa(x) \leq \kappa_{\mathrm{rel}}(\kappa, x) \leq M \cdot \kappa(x) \quad \text{if } \|x\| = 1. \tag{1.2}$$

Inequalities (1.2) mean that solving the problem x, normalized so $\|x\| = 1$, is essentially just as hard (within factors m and M) as computing the condition number κ of the problem x. If we further assume that κ is *homogeneous*, i.e. $\kappa(\alpha x) = \alpha^k \kappa(x)$ for all real positive α, then inequalities (1.1) and (1.2) can be shown nearly equivalent (see Sect. 7). The near equivalence of inequalities (1.1) and (1.2) is very satisfying because it says that if the condition number κ has the utterly reasonable property of being just as hard to compute as the solution x itself, then it has the attractive geometric property of being the reciprocal of the distance to the nearest infinitely ill-conditioned problem. Indeed, the common formulas for relative condition numbers (e.g., $\|A\| \cdot \|A^{-1}\|$ for matrix inversion) lead one to believe that one must solve the problem (e.g., compute A^{-1}) to within reasonable accuracy to get a reasonably accurate condition number. This intuition is corroborated by the results in this paper.

The second interpretation of (1.1) is as a restatement of Newton's method. This interpretation applies only when the mapping g, which maps a problem to its solution, has as domain and range either the real or the complex numbers, and is smooth except for poles. For example, g may be a rational function of a single real or complex variable. As condition number we take the *absolute condition number*, which is just a multiple of the relative condition number defined above:

$$\kappa_{\mathrm{abs}}(g, x) \equiv \limsup_{\delta x \to 0} \frac{\|g(x+\delta x) - g(x)\| / \|g(x)\|}{\|\delta x\|} = \frac{\|Dg(x)\|}{\|g(x)\|}.$$

Since both domain and range are one dimensional, $\kappa_{\text{abs}}(g, x)$ can be written as $|g'(x)/g(x)|$. If g is smooth except at poles, the condition number can be infinite only at poles and zeros of g. The problem of finding the distance to the nearest ill-posed problem thus becomes the zero (or pole) finding problem. If we are close enough to a zero, we expect the absolute value of the Newton correction g/g' to be a good estimate of the distance to the nearest zero. But $|g/g'|$ is just the reciprocal of the condition number κ_{abs}. In Sect. 8 we show that (1.1) will asymptotically hold with $m = M = 1$ in a sufficiently small neighborhood of the set of ill-posed problems, thus yielding the Newton correction as the correct distance estimate. This also works in the neighborhood of multiple zeros and poles.

This connection with Newton's method becomes weaker when the domain and range of g are just linear spaces of the same dimension greater than one. Nonetheless, there is a connection which we also explore in Sect. 8.

For eigenvalue problems, polynomial zero finding, and pole placement this interpretation does not apply. The reasons are twofold: first, the problem space and solution space are of different dimensions, and ill-posed problems occur not at zeros and poles but at branch points. It is natural to ask if some general statement can be made about the condition number and distance to the nearest branch point. It turns out we can show that in a sufficiently small neighborhood of a branch point of *any* algebraic function, the distance is bounded below by a multiple of the reciprocal of the square of the condition number. We show this in Sect. 9. This result is reflected in gaps between the best known upper and lower bounds on the shortest distance for eigenvalue problems (Sect. 4)) and polynomial zero finding (Sect. 5).

The rest of the paper is organized as follows. In Sect. 2 we present our differential inequalities and solve them. Sects. 3 through 6 cover matrix inversion, eigendecompositions, polynomial zero finding, and pole assignment. Sects. 3 through 6 may be read independently. The results on matrix inversion and some of the results on eigendecompositions are known, but others are new. Much related work on the eigenvalue problem has been done by Wilkinson [17, 18] and we compare our approaches in Sect. 4. One of our upper bounds on the distance from a polynomial to one with a double zero is known but another is new. Our lower bound on this distance is new. Our results on pole assignment are also new. Sect. 7 discusses the equivalence of inequalities (1.1) and (1.2) when the condition number is homogeneous. In Sect. 8 we discuss the connection with Newton's method mentioned above. In Sect. 9 we show that when the solution of the problem is any algebraic function we expect a lower bound on the distance in terms of the reciprocal of the square of the condition number. Sect. 10 discusses extensions.

2. Differential Inequalities

The differential inequalities we need are given in the following lemmas. The first one will be used to derive an upper bound on the distance to the nearest ill-posed problem in terms of the condition number.

Lemma 1. *Suppose* $m>0$, $y_0>0$, $\alpha>1$ *and*

$$\frac{d}{ds} y(s) \geq m y^{\alpha}(s), \qquad y(0)=y_0.$$

Then $y(s)$ *becomes infinite for some satisfying*

$$0<s<\frac{1}{(\alpha-1)\cdot m\cdot y_0^{\alpha-1}}.$$

Proof. The differential inequality implies y is positive and strictly increasing, so by a standard result in the theory of ordinary differential equations [6, Thm. III. 4.1] it is bounded below by the solution of

$$\frac{d}{ds} z(s)=m z^{\alpha}(s), \qquad z(0)=y_0$$

which, as easily verified, is

$$z(s)=\frac{y_0}{(1-(\alpha-1)\, m y_0^{\alpha-1} s)^{(\alpha-1)^{-1}}}.$$

Since $z(s)$ has a pole at $1/((\alpha-1)\, m y_0^{\alpha-1})$, $y(s)$ must have a pole before that. q.e.d.

Now suppose κ were continuously differentiable wherever it was finite and that the norm $\|D\kappa(x)\|$ were an operator norm induced by some vector norm: $\|D\kappa(x)\| = \sup_{\|y\|=1} D\kappa(x)\cdot y$. Suppose further that $D\kappa(x)$ had a continuous dual vector field $y(x)$. In other words $y(x)$ should satisfy $\|y(x)\|=1$ and $D\kappa(x)\cdot y(x)=\|D\kappa(x)\|$. Then we could define a curve $x(s)$, parameterized by arclength ($\|\dot{x}(s)\|=1$), as the integral curve of the vector field $y(x)$ passing through $x(0)=x$. The curve $x(s)$ is simply the curve along which κ increases most rapidly at each point (the curve of "steepest ascent"). If $D\kappa(x)$ satisfied $\|D\kappa(x)\|\geq m\cdot\kappa^2(x)$, then by the chain rule $\kappa(x(s))$ would satisfy

$$\frac{d}{ds}\kappa(x(s))=D\kappa(x(s))\cdot\dot{x}(s)=\|D\kappa(x(s))\|\geq m\cdot\kappa^2(x(s))$$

so by Lemma 1 $1/(m\cdot\kappa(x))$ would be an upper bound on the distance in the $\|\cdot\|$ norm from x to the nearest ill-posed problem. Unfortunately, κ is not everywhere continuously differentiable for the problems we consider in this paper. Nonetheless, we will see that it is always smooth enough to construct a smooth curve $x(s)$ along which $\kappa(x(s))$ increases sufficiently fast to apply Lemma 1.

The next differential inequality will yield a lower bound on the distance to the nearest ill-posed problem in terms of the condition number. First define the *right derivate* of a continuous function f as

$$D^+ f(x)\equiv\limsup_{h\to 0+}\frac{f(x+h)-f(x)}{h}.$$

Lemma 2. *Suppose* $M>0$, $y(s)>0$ *for all* s, $y_0>0$, $\beta>1$ *and*

$$D^+ y(s)\leq M y^{\beta}(s), \qquad y(0)=y_0.$$

Then $y(s)$ *is finite for*

$$0 \leq s < \frac{1}{(\beta-1) \cdot M \cdot y_0^{\beta-1}}.$$

Proof. Since $y(s)$ is positive, the same standard result as used in the proof of Lemma 1 implies that it is bounded above by the solution of

$$\frac{d}{ds} z(s) = M z^{\beta}(s), \qquad z(0) = y_0$$

which, as in the last lemma, is

$$z(s) = \frac{y_0}{(1-(\beta-1)\, M y_0^{\beta-1} s)^{(\beta-1)^{-1}}}.$$

Since $z(s)$ is finite for all s less than $1/((\beta-1)\, M y_0^{\beta-1})$, so is $y(s)$. q.e.d.

Now suppose κ were continuously differentiable wherever it was finite and satisfied $\|D\kappa(x)\| \leq M \cdot \kappa^2(x)$, the norm on $D\kappa$ an operator norm as before. Then for every smooth curve $x(s)$ parameterized by arclength and passing through $x(0) = x$, $\kappa(x(s))$ would by the chain rule satisfy the conditions of Lemma 2, thus yielding a lower bound $1/(M \cdot \kappa(x))$ on the distance from x to the nearest ill-posed problem. As mentioned above, κ is not continuously differentiable for the problems we consider, but it is smooth enough to satisfy the constraints of Lemma 2 for smooth curves $x(s)$.

3. Matrix Inversion

In this section $\|\cdot\|$ will denote an arbitrary vector norm, $\|\cdot\|_D$ the dual norm:

$$\|y^T\|_D \equiv \sup_{x \neq 0} \frac{|y^T x|}{\|x\|}$$

and $\|A\|$ the induced matrix norm:

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}.$$

Let P be the set of singular matrices. Let $\operatorname{dist}(A, P)$ denote the minimum distance from the matrix A to the set P: $\operatorname{dist}(A, P) = \inf_{S \in P} \|A - S\|$.

As discussed in the introduction, $\|A^{-1}\|$ is a condition number for the problem of inverting the matrix A. This is true because to first order

$$(A + \delta A)^{-1} = A^{-1} - A^{-1} \delta A A^{-1} + O(\|\delta A\|^2) \equiv A^{-1} + X$$

so that for small δA

$$\frac{\|X\|}{\|A^{-1}\|} \leq \|A^{-1}\| \cdot \|\delta A\|.$$

The following result is originally due to Eckart and Young [4] when $\|\cdot\|$ is the Euclidean norm and to Gastinel [8] for arbitrary norm:

Theorem 1. [Gastinel] *Let P be the set of singular matrices. Then*

$$\operatorname{dist}(A,P)=\|A^{-1}\|^{-1},$$

i.e. *the reciprocal of the condition number $\|A^{-1}\|$ of the problem of inverting A.*

Proof. If $\|\delta A\|<\|A^{-1}\|^{-1}$ then $A+\delta A$ is invertible since

$$(A+\delta A)^{-1}=(I+A^{-1}\delta A)^{-1}A^{-1} \quad \text{and} \quad \|A^{-1}\delta A\|\leq\|A^{-1}\|\,\|\delta A\|<1.$$

Therefore $\operatorname{dist}(A,P)\geq\|A^{-1}\|^{-1}$. To show equality holds choose x and y such that $\|x\|=\|y^T\|_D=1$ and $y^T A^{-1}x=\|A^{-1}\|$ (the existence of x and y follows from the definitions of the norms). Let $\delta A=-\|A^{-1}\|^{-1}xy^T$. Clearly $\|\delta A\|=\|A^{-1}\|^{-1}$. To see $A+\delta A$ singular note that $(A+\delta A)(A^{-1}x)=0$. q.e.d.

We now prove this theorem using Lemmas 1 and 2. This alternate proof is no simpler than the above one, but illustrates the techniques we use later.

Theorem 2. *Let P be the set of singular matrices. Then*

$$\operatorname{dist}(A,P)=\|A^{-1}\|^{-1},$$

Proof. To show $\operatorname{dist}(A,P)\geq\|A^{-1}\|^{-1}$ let $A(s)$ be any smooth path from $A(0)=A$ to $A(s_0)\in P$ parameterized by arclength (i.e. $\|\dot{A}(s)\|=1$). Then

$$\begin{aligned}
D^+\|A^{-1}(s)\| &= \limsup_{h\to 0+}\frac{\|A^{-1}(s+h)\|-\|A^{-1}(s)\|}{h}\\
&= \limsup_{h\to 0+}\frac{\|(A(s)+h\dot{A}(s))^{-1}\|-\|A^{-1}(s)\|}{h}\\
&= \limsup_{h\to 0+}\frac{\|A^{-1}(s)-hA^{-1}(s)\,\dot{A}(s)\,A^{-1}(s)\|-\|A^{-1}(s)\|}{h}\\
&\leq \limsup_{h\to 0+}\frac{\|hA^{-1}(s)\,\dot{A}(s)\,A^{-1}(s)\|}{h}\leq\|A^{-1}(s)\|^2.
\end{aligned}$$

Applying Lemma 2 with $M=1$ and $\beta=2$ implies $\|A^{-1}(s)\|$ remains finite for $s<\|A^{-1}\|^{-1}$. Since the path $A(s)$ from A to P was arbitrary, we have $\operatorname{dist}(A,P)\geq\|A^{-1}\|^{-1}$.

To prove the opposite inequality we need to choose a path $A(s)$ along which $\|A^{-1}(s)\|$ increases as quickly as possible, i.e. we need an integrable vector field $X(A)$, $\|X(A)\|=1$, where $\|A^{-1}X(A)A^{-1}\|=\|A^{-1}\|^2$. Let $x(A)$ and $y(A)$ be defined as in Theorem 2 $\|x(A)\|=\|y^T(A)\|_D=1$ and $y^T(A)A^{-1}x(A)=\|A^{-1}\|$. Now let $X(A)=x(A)\,y^T(A)$. Assume for the moment that $X(A)$ is integrable, and let $A(s)$ be an integral curve parameterized by arclength such that $A(0)=A$ and $\|A^{-1}(s)\|$ is increasing. We will show that

$$\frac{d}{ds}\|A^{-1}(s)\|=\|A^{-1}(s)\|^2$$

so by Lemma 1 (with $m=1$ and $\alpha=2$) $\|A^{-1}(s)\|$ becomes infinite for $s=\|A^{-1}\|^{-1}$ as desired.

We show $X(A)$ is integrable by integrating it explicitly. Its integral curves are straight lines as they must be since they are geodesics in a normed linear space. To prove this it suffices to show that if $\|x\|=\|y^T\|_D=1$ and $y^T A^{-1} x = \|A^{-1}\|$, then $y^T(A-sxy^T)x=\|(A-sxy^T)^{-1}\|$ for s sufficiently small. This follows from the Sherman-Morrison formula [5]

$$(A-sxy^T)^{-1}=A^{-1}+\frac{sA^{-1}xy^T A^{-1}}{1-sy^T A^{-1} x}$$

so

$$\|(A-sxy^T)^{-1}\| \le \|A^{-1}\| + \frac{s\|A^{-1}\|^2}{1-s\|A^{-1}\|}=\frac{\|A^{-1}\|}{1-s\|A^{-1}\|}$$

and

$$y^T(A-sxy^T)^{-1}x=y^T A^{-1} x+\frac{s(y^T A^{-1} x)(y^T A^{-1} x)}{1-sy^T A^{-1} x}=\|A^{-1}\|+\frac{s\|A^{-1}\|^2}{1-s\|A^{-1}\|}$$
$$=\frac{\|A^{-1}\|}{1-s\|A^{-1}\|}$$

so

$$y^T(A-sxy^T)^{-1}x=\|(A-sxy^T)^{-1}\|=\frac{\|A^{-1}\|}{1-s\|A^{-1}\|}.$$

Finally, differentiating we see

$$\frac{d}{ds}\|(A-sxy^T)^{-1}\|\Big|_{s=0}=\frac{d}{ds}\frac{\|A^{-1}\|}{1-s\|A^{-1}\|}\Big|_{s=0}=\|A^{-1}\|^2$$

as desired. q.e.d.

In this proof, we explicitly integrated the vector field $X(A)$ in order to show integral curves existed. This is not generally possible or desirable, and in our later examples we only show that $X(A)$ is continuous, which is sufficient for integrability.

4. Eigenvalue and Eigenvector Computations

In this section we consider the problem of computing a simple eigenvalue or corresponding eigenvector of a general matrix T. In this section we let $\|\cdot\|$ denote the 2-norm $\|x\| \equiv \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}$,

$$\|T\| \equiv \sup_{x \neq 0} \frac{\|Tx\|}{\|x\|},$$

and $\|T\|_F \equiv (\sum_{ij} |T_{ij}|^2)^{1/2}$. Let λ be a simple eigenvalue of T, x its right eigenvector and y^T its left eigenvector, where we normalize so that $y^T x=1$. The projector

P belonging to λ is defined as xy^T and has norm $\|P\| = \|x\| \, \|y^T\|$. It is well known [15] that if we perturb T by δT, λ can change at most by $|\delta\lambda| \leq \|P\| \, \|\delta T\|$for small δT, and that this bound is attainable. Therefore, we call $\|P\|$ the condition number of the eigenvalue λ. It is also known [9, 16] that the distance from T to a matrix which has a double eigenvalue (at λ) is bounded by $\|T\|/(\|P\|^2 - 1)^{1/2}$, or approximately $\|T\|/\|P\|$ for large $\|P\|$. An n-tuple eigenvalue λ is infinitely ill-conditioned because a perturbation of size ε in the matrix can change λ by $\varepsilon^{1/n}$, whose derivative at $\varepsilon = 0$ is infinite. Therefore we may take the set of matrices with multiple eigenvalues as our surface of ill-posed problems. Thus the reciprocal of the condition number $\|P\|$ bounds the relative distance to the nearest infinitely ill-conditioned problem. We will obtain this result using Lemma 1.

Similar considerations show that the same surface is the set of ill-conditioned problems when computing eigenvectors, although the condition number for eigenvectors differs from the one for eigenvalues. It is known that the reciprocal of the eigenvector condition number is a lower bound on the distance to the nearest matrix with multiple eigenvalues [2, 14]. We will obtain this result using Lemma 2.

There can be quite a gap between these upper and lower bounds on the distance to the nearest matrix with a multiple eigenvalue. Several authors [2, 9, 13, 16, 17, 18] have explored the geometry of the set of matrices with multiple eigenvalues, and attempted to find simple ways to measure the distance from a given matrix to that set, but large gaps still remain between the best known upper and lower bounds. Wilkinson [17, 18] in particular has pointed out the inadequacy of the current bounds, and suggested a numerical approach to find the distance from any given matrix to the nearest matrix with multiple eigenvalues. His method also depends on using perturbations which cause the eigenvalues to rush together and increase their condition numbers nearly as quickly as possible (he calls this "fast perturbation theory"); a major point of our work is that this approach can be used on a wide variety of problems, not just the eigenproblem. Such a numerical method may find a much closer ill-posed problem then the bounds provided by our theorems can guarantee. We will point out the relation between our approach and his as we go, and at the end we will summarize and compare the various bounds in the literature. We will also propose an explanation of the gap between our upper and lower bounds as a feature of any algebraic function (see also Sect. 9).

First we need some notation. Since our matrix norm is invariant under orthogonal transformations, we may assume without loss of generality that our matrix T is in Schur canonical form [5]:

$$T = \begin{bmatrix} \lambda & x^T \\ 0 & B \end{bmatrix}.$$

Occasionally we will write $\lambda(T)$ to emphasize the (cointinuous) dependence of λ on T. Let $r \equiv (B - \lambda)^{-T} x$. It is easy to show that in this coordinate system the right and left eigenvectors of λ may be written $x = [1, 0, \ldots, 0]^T$ and $y^T = [1, -r^T]$, so that $(\|P\|^2 - 1)^{1/2} = \|r\|$. It is to this last quantity we shall apply Lemma 1. If we perturb T to $T + \delta T$, with

$$\delta T = \begin{bmatrix} \delta T_{11} & \delta T_{12} \\ \delta T_{21} & \delta T_{22} \end{bmatrix}$$

partitioned conformally with T, then to first order in δT P is perturbed to [10]

$$P + \delta P = P + S\delta T P + P \delta T S \tag{4.1}$$

where S is the *reduced resolvent*, or

$$S \equiv \lim_{z \to \lambda} (I - P) \cdot (T - z)^{-1} = \begin{bmatrix} 0 & r^T (B-\lambda)^{-1} \\ 0 & (B-\lambda)^{-1} \end{bmatrix}$$

in this coordinate system. Expanding (4.1) yields

$$\begin{aligned} P + \delta P = \begin{bmatrix} 1 & -r^T \\ 0 & 0 \end{bmatrix} + \Bigg[& \begin{matrix} r^T (B-\lambda)^{-1} \delta T_{21} \\ (B-\lambda)^{-1} \delta T_{21} \end{matrix} \\ & \delta T_{11} r^T (B-\lambda)^{-1} + \delta T_{12} (B-\lambda)^{-1} \\ & - r^T \delta T_{22} (B-\lambda)^{-1} - r^T (B-\lambda)^{-1} \delta T_{21} r^T - r^T \delta T_{21} r^T (B-\lambda)^{-1} \\ & - (B-\lambda)^{-1} \delta T_{21} r^T \Bigg]. \end{aligned} \tag{4.2}$$

We consider two perturbations, one where only δT_{22} is nonzero (this is the perturbation used in [16]) and one where only δT_{21} is nonzero. It is easy to find the smallest perturbation of δT_{22} that makes $T + \delta T$ have a double eigenvalue at λ:

Theorem 3. [Wilkinson] *If* $\|P\| > 1$ *then there exists a* δT *with only* $\delta T_{22} \neq 0$ *such that* $T + \delta T$ *has a double eigenvalue at* λ *and*

$$\|\delta T\| \leq \frac{\|T\|}{(\|P\|^2 - 1)^{1/2}}.$$

Proof. The δT_{22} we seek is the smallest perturbation such that $B - \lambda + \delta T_{22}$ is singular, which has norm $\|\delta T_{22}\| = \|(B-\lambda)^{-1}\|^{-1}$. But since $r = (B-\lambda)^{-T} x$, $\|r\| \leq \|x\| \, \|(B-\lambda)^{-1}\| \leq \|T\| \, \|(B-\lambda)^{-1}\|$ or

$$\|(B-\lambda)^{-1}\|^{-1} \leq \frac{\|T\|}{\|r\|} = \frac{\|T\|}{(\|P\|^2 - 1)^{1/2}}$$

as desired. q.e.d.

As pointed out by Wilkinson, this upper bound can be much weaker than the simpler upper bound $\|(B-\lambda)^{-1}\|^{-1}$. For example, if

$$\begin{bmatrix} \lambda & x^T \\ 0 & B \end{bmatrix} = \begin{bmatrix} 1+\varepsilon & 0 \\ 0 & 1 \end{bmatrix}$$

the upper bound of Theorem 3 is infinite and $\|(B-\lambda)^{-1}\|^{-1}$ is ε. Nonetheless, it does provide a one-sided inequality between the shortest distance and the reciprocal of the condition number.

Theorem 4. *If* $\|P\|>1$ *then there exist a* δT *with only* $\delta T_{22}\neq 0$ *such that* $T+\delta T$ *has a double eigenvalue at* λ *and*

$$\|\delta T\| \le \frac{\|T\|}{(\|P\|^2-1)^{1/2}}.$$

Proof. To apply Lemma 1 when only $\delta T_{22}\neq 0$ we need to compute

$$(\|P+\delta P\|^2-1)^{1/2}-(\|P\|^2-1)^{1/2}=\left(\left\|\begin{bmatrix}1 & -r^T-r^T\delta T_{22}(B-\lambda)^{-1}\\ 0 & 0\end{bmatrix}\right\|^2-1\right)^{1/2}$$

$$-\left(\left\|\begin{bmatrix}1 & -r^T\\ 0 & 0\end{bmatrix}\right\|^2-1\right)^{1/2}$$

$$=\|r^T+r^T\delta T_{22}(B-\lambda)^{-1}\|-\|r^T\|$$

when $\|\delta T\|$ approaches zero. Letting $r_u=r/\|r\|$, it is easy to see this last expression is

$$\operatorname{Re}\|r\|\, r_u^T\,\delta T_{22}(B-\lambda)^{-1}\bar{r}_u$$

to first order in δT_{22}. Now since $r=(B-\lambda)^{-T}x$,

$$\|r\|^2=x^T(B-\lambda)^{-1}(B-\lambda)^{-1*}\bar{x}=x^T(B-\lambda)^{-1}\bar{r}\le\|x\|\,\|(B-\lambda)^{-1}\bar{r}\|$$

so

$$\|(B-\lambda)^{-1}\bar{r}\|\ge\frac{\|r\|^2}{\|x\|}\ge\frac{\|r\|^2}{\|T\|}.$$

Therefore by choosing δT_{22} to be a small multiple of

$$\bar{r}r^T(B-\lambda)^{-1*} \tag{4.3}$$

we get

$$\operatorname{Re}\|r\|\, r_u^T\,\delta T_{22}(B-\lambda)^{-1}\bar{r}_u\ge\frac{\|\delta T_{22}\|\,\|r\|^2}{\|T\|}.$$

This implies that we may choose δT so that the rate of change of $(\|P\|^2-1)^{1/2}=\|r\|$ is at least $\|r\|^2/\|T\|$. The vector field given by (4.3) above is clearly smooth and integrable, so we may let $y(s)\equiv\|r(s)\|$ where $r(s)$ is computed along an integral curve of the vector field. Thus

$$\frac{d}{ds}y(s)\ge\frac{y^2(s)}{\|T\|}$$

and we may apply Lemma 1 to $y(s)$ with $m=\|T\|^{-1}$ and $\alpha=2$ to get the desired upper bound

$$\frac{\|T\|}{y(0)}=\frac{\|T\|}{(\|P\|^2-1)^{1/2}}$$

on the distance to the nearest matrix with λ as a double eigenvalue. q.e.d.

We may prove a similar thorem when we only perturb T_{21}. Intuitively we would expect such a perturbation to be at least as effective as one in the δT_{22} position since it can move both the eigenvalues of B and $\lambda(T+\delta T)$.

Theorem 5. *If* $\|P\|>1$ *then there exists a* δT *such that* $T+\delta T$ *has a double eigenvalue at* $\lambda(T+\delta T)$ *and*

$$\|\delta T\| \leq \frac{2\|T\|_F}{(\|P\|^2-1)^{1/2}}$$

Proof. Setting all but δT_{21} to 0 in (4.2) yields

$$P+\delta P=\begin{bmatrix} 1 & -r^T \\ 0 & 0 \end{bmatrix}$$
$$+\begin{bmatrix} r^T(B-\lambda)^{-1}\delta T_{21} & -r^T(B-\lambda)^{-1}\delta T_{21}\, r^T - r^T\delta T_{21}\, r^T(B-\lambda)^{-1} \\ (B-\lambda)^{-1}\delta T_{21} & -(B-\lambda)^{-1}\delta T_{21}\, r^T \end{bmatrix}.$$

When δT is small the square of the norm of this perturbed projector is at least

$$\begin{aligned} &1+2\operatorname{Re} r^T(B-\lambda)^{-1}\delta T_{21}+\|r\|^2 \\ &\qquad +2\|r\| \operatorname{Re}(r^T(B-\lambda)^{-1}\delta T_{21}\, r^T\bar{r}_u + r^T\delta T_{21}\, r^T(B-\lambda)^{-1}\,\bar{r}_u) \\ &\quad =1+\|r\|^2+2\operatorname{Re} r^T(B-\lambda)^{-1}[(\|r\|^2+1)+\|r\|^2\bar{r}_u r_u^T]\,\delta T_{21} \end{aligned}$$

where $r_u=r/\|r\|$. Now

$$\begin{aligned} \|2r^T(B-\lambda)^{-1}[(\|r\|^2+1)+\|r\|^2\bar{r}_u r_u^T]\| &\geq \frac{2\|r\|}{\|B-\lambda\|\,\|[(\|r\|^2+1)+\|r\|^2\bar{r}_u r_u^T]^{-1}\|} \\ &\geq \frac{\|r\|^3}{\|T\|_F} \end{aligned} \tag{4.4}$$

where $\|T\|_F$ is the Frobenius norm (the reason for this choice instead of the smaller $\|T\|$ will be clear in a moment). Thus by choosing δT_{21} a small multiple of $[2r^T(B-\lambda)^{-1}[(\|r\|^2+1)+\|r\|^2\bar{r}_u r_u^T]]^*$ we get

$$(\|P+\delta P\|^2-1)^{1/2}-(\|P\|^2-1)^{1/2} \geq \frac{(\|P\|^2-1)\,\|\delta T_{21}\|}{2\|T\|_F}.$$

Since all quantities defining δT_{21} are analytic [10] the vector field defined by δT_{21} is integrable. Furthermore, it is orthogonal to T (in the tr $T^*\delta T_{21}$ inner product) for all T. Therefore its integral curves lie on spheres of constant $\|T\|_F$. Thus, along an integral curve the function $y(s)=(\|P\|^2-1)^{1/2}=\|r(s)\|$ satisfies

$$\frac{d}{ds}y(s) \geq \frac{y^2(s)}{2\|T\|_F}$$

so we may apply Lemma 1 with $m=1/(2\|T\|_F)$ and $\alpha=2$. Note that $\|T\|_F$ is constant along the curve. Lemma 1 implies that there is a perturbation δT_{21} of 2-norm at most $2\|T\|_F/(\|P\|^2-1)^{1/2}$ that makes the eigenvalue $\lambda(T+\delta T)$ double. q.e.d.

Perturbations in T_{21} were also considered by Wilkinson [17, 18] under the name "fast perturbations," since for many nearly defective matrices they can make the eigenvalues rush together as fast as possible. In a series of examples

Wilkinson showed one could expect to find a smallest perturbation to make "close" eigenvalues coalesce of size approximately $\min_{\lambda' \in \sigma(B)} 1/2|\lambda-\lambda'|/(\|P_\lambda\|+\|P_{\lambda'}\|)$, where $\sigma(B)$ is the spectrum of B, instead of $\|T\|/\|P\|$. The intuition behind Wilkinson's estimate is this: the distance λ and λ' have to move to merge is $|\lambda-\lambda'|$ and their speeds are $\|P_\lambda\|$ and $\|P_{\lambda'}\|$. The factor 1/2 comes from their acceleration as they approach one another. At the end of this section we show that Wilkinson's estimate is nearly a lower bound on the norm of the smallest perturbation needed (although it may occasionally be a gross underestimate) and in fact belongs to a family of lower bounds including the one in Theorem 7 below.

We can motivate Wilkinson's estimate from the proof of Theorem 5. In making the estimate (4.4) we used the bound

$$\|r^T(B-\lambda)^{-1}\| \geq \frac{\|r\|}{\|B-\lambda\|}$$

which although attainable is often pessimistic. When λ is very close to an eigenvalue of B

$$\|r^T(B-\lambda)^{-1}\| \approx \frac{\|r\|}{\min_{\lambda' \in \sigma(B)} |\lambda-\lambda'|}$$

is a much better approximation, and leads to Wilkinson's estimate when $\|P_\lambda\| > \|P_{\lambda'}\|$.

We turn now to computing eigenvectors. Let $|||\cdot|||$ denote an arbitrary operator norm. From (4.1) we can see that if T is perturbed by δT, P can be perturbed to first order by at most $|||\delta P||| \leq 2|||S|||\cdot|||P|||\cdot|||\delta T|||$. A close examination of (4.2) shows this bound can be nearly attained, so $|||S|||\cdot|||P|||$ may be used as a condition number for P. The next theorem will use Lemma 2 to show that $1/(|||S|||\cdot|||P|||)$ is a lower bound on the distance to the surface of ill-posed problems. This result is essentially identical to other results in the literature [2, 14].

Theorem 6. *The distance in the $|||\cdot|||$ norm from T to the nearest matrix $T+\delta T$ where $\lambda(T+\delta T)$ is a multiple eigenvalue is at least*

$$\frac{1}{7\cdot|||S|||\cdot|||P|||}.$$

Proof. We need to compute the gradient of $|||S|||\cdot|||P|||$. Since we are only interested in an upper bound, it will suffice to use the first order bound

$$|||S+\delta S|||\cdot|||P+\delta P|||-|||S|||\cdot|||P||| \leq |||S|||\cdot|||\delta P|||+|||\delta S|||\cdot|||P|||.$$

Following Kato [10] we may compute to first order

$$\begin{aligned} S+\delta S &= (I-P-\delta P)(\lambda+\delta\lambda-(I-P-\delta P)(T+\delta T))^{-1}(I-P-\delta P) \\ &= S-P\delta TSS-SS\delta TP-2(\operatorname{tr} P\delta T)SS+S\delta TS \end{aligned}$$

so that

$$|||\delta S||| \leq 5|||P|||\cdot|||S|||^2\cdot|||\delta T|||.$$

Here we use the fact that P is of rank 1 to bound $|\operatorname{tr} \delta T P| \le |||\delta T||| \cdot |||P|||$. Similarly from (4.1)

$$|||\delta P||| \le 2\,|||S||| \cdot |||P||| \cdot |||\delta T|||.$$

Therefore

$$|||S+\delta S||| \cdot |||P+\delta P||| - |||S||| \cdot |||P||| \le 7(|||S||| \cdot |||P|||)^2 \cdot |||\delta T|||.$$

Now let $y(s) = |||S(s)||| \cdot |||P(s)|||$ where $T(s)$ is any smooth curve parameterized by arclength from $y(0) = T$ to $T(s_0)$ where $\lambda(T(s_0))$ is a double eigenvalue. Then

$$\frac{d}{ds} y(s) \le 7 y^2(s)$$

so we may apply Lemma 2 with $M=7$ and $\beta=2$ to get that s_0, the distance in the $|||\cdot|||$ norm to the matrix $T+\delta T$ with $\lambda(T+\delta T)$ a double eigenvalue, satisfies

$$s_0 \ge \frac{1}{7 \cdot |||S||| \cdot |||P|||}.$$

q.e.d.

By choosing a specific $|||\cdot|||$ we will see that this result recovers a number of lower bounds in the literature. For example, if we choose $|||X||| \equiv \|KXK^{-1}\|$ where K diagonalizes T, we get the lower bound of the Bauer-Fike Theorem. We can also recover one of the best current lower bounds in the literature to within a constant factor [2, 14]. Let

$$R = \begin{bmatrix} 1 & -r^T \\ 0 & \|P\|\, I \end{bmatrix}.$$

It is easy to see that $R\,TR^{-1} = \operatorname{diag}(\lambda, B)$, and in fact the condition number

$$\kappa(R) = \|R\|\,\|R^{-1}\| = \|P\| + (\|P\|^2 - 1)^{1/2}$$

of R is the minimum over all matrices which block diagonalize T [1]. Now choose

$$|||X||| \equiv \|RXR^{-1}\|.$$

Then it is easy to see that the lower bound of Theorem 6 becomes

$$\frac{1}{7 \cdot |||S||| \cdot |||P|||} = \frac{1}{7} \left\| \begin{bmatrix} 0 & 0 \\ 0 & (B-\lambda)^{-1} \end{bmatrix} \right\|^{-1} \cdot \left\| \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \right\|^{-1} = \frac{\sigma_{\min}(B-\lambda)}{7}.$$

Since $\|X\| \ge |||X|||/\kappa(R)$, we see that a lower bound in the $\|\cdot\|$ norm on the distance from T to the nearest matrix $T+\delta T$ where $\lambda(T+\delta T)$ is a double eigenvalue is

$$\frac{\sigma_{\min}(B-\lambda)}{7(\|P\| + (\|P\|^2 - 1)^{1/2})},$$

which is within a constant factor of the lower bound $\sigma_{\min}(B-\lambda)/(4\|P\|)$ on the distance in the $\|\cdot\|_F$ norm in the literature. (This lower bound may be improved by at most a factor of 4 to $\inf_{\lambda'} \max(|\lambda-\lambda'|, \sigma_{\min}(B-\lambda'))/(\|P\|+(\|P\|^2-1)^{1/2})$ [2]).

We can recover the lower bound $\sigma_{\min}(B-\lambda)/(4\|P\|)$ by using a quantity proportional to it rather than $|||S|||\cdot|||P|||$ as a condition number. In fact, we can prove a more general version by considering a group of eigenvalues of T rather than a single one λ, and using a condition number for the projector associated with the entire group. This condition number will be finite until one of the eigenvalues in the group merges with one outside the group; the eigenvalues within the group may or may not be multiple. Following Stewart [14], we assume T is initially in Schur canonical form

$$T=\begin{bmatrix} A & C \\ 0 & B \end{bmatrix}$$

whee the eigenvalues of A consist of the group in whose projector we are interested. We exhibit this projector by solving

$$\begin{bmatrix} I & -R \\ 0 & I \end{bmatrix}\cdot\begin{bmatrix} A & C \\ 0 & B \end{bmatrix}\cdot\begin{bmatrix} I & R \\ 0 & I \end{bmatrix}=\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix},$$

which upon manipulation yields the equation $RB-AR=C$, a nonsingular system of linear equations as long as the eigenvalue of A and B are disjoint, which we have assumed. Then the projector associated with the eigenvalues of A is

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

The equation $RB-AR=C$ can be rewritten using Kronecker products as $(B^T\otimes I - I\otimes A)\operatorname{col} R=\operatorname{col} C$, where $V\otimes W\equiv[V_{ij}W]$ and $\operatorname{col} C$ is a column matrix made from stacking the columns of C atop one another from left to right. If we define $\operatorname{sep}(A,B)\equiv\sigma_{\min}(B^T\otimes I-I\otimes A)$ then clearly $\|R\|_F\le\|C\|_F/\operatorname{sep}(A,B)$. The lower bound on the distance (measured with $\|\cdot\|_F$) from T to the nearest matrix where an eigenvalue of A merges with one from B is $\operatorname{sep}(A,B)/(4\cdot\|P\|)$ [14]. We prove this result by using $\|P\|/\operatorname{sep}(A,B)$ as a condition number and using Lemma 2:

Theorem 7. *Let T, A, B, and P be as above. Then the distance in the $\|\cdot\|_F$ norm from T to the nearest matrix where an eigenvalue of A merges with one in B is at least*

$$\frac{\operatorname{sep}(A,B)}{4\cdot\|P\|}.$$

Sketch of proof. This result requires some technical machinery found in [14]. Briefly, if T is perturbed to $T+\delta T$, then to first order $\|P\|$ and $\operatorname{sep}(A,B)$ are perturbed as follows:

$$|\delta\|P\||\le\frac{2\cdot\|P\|^2\cdot\|\delta T\|_F}{\operatorname{sep}(A,B)}$$

and

$$|\delta\operatorname{sep}(A,B)|\le 2\cdot\|P\|\cdot\|\delta T\|_F.$$

Combining these we get

$$\delta\left(\frac{\|P\|}{\operatorname{sep}(A,B)}\right)\leq 4\cdot\left(\frac{\|P\|}{\operatorname{sep}(A,B)}\right)^2\|\delta T\|_F.$$

Therefore along any smooth curve $T(s)$ parameterized by arclength and passing through $T(0)=T$, the quantity $\kappa(T(s))\equiv\|P(s)\|/\operatorname{sep}(A(s),B(s))$ satisfies $D^+\kappa(T(s))\leq 4\cdot\kappa^2(T(s))$ so by Lemma 2 the distance to the nearest $T(s)$ with $\kappa(T(s))$ infinite is at least $\operatorname{sep}(A,B)/(4\cdot\|P\|)$. (This lower bound may be improved to $\operatorname{sep}_\lambda(A,B)/(\|P\|+(\|P\|^2-1)^{1/2})$, where $\operatorname{sep}_\lambda(A,B)\equiv\inf_{\lambda'}\max(\sigma_{\min}(A-\lambda'),\sigma_{\min}(B-\lambda'))$. This improved lower bound is often close to $\operatorname{sep}(A,B)/(4\|P\|)$ but can be much larger [2].)

Now we compare and evaluate the various bounds on the distance from a matrix to the nearest one with a multiple eigenvalue. We assume for the moment that we have chosen an eigenvalue $\lambda(T)$ and wish to compute the distance $d_\lambda\equiv\|\delta T\|$ to the nearest matrix $T+\delta T$ where $\lambda(T+\delta T)$ is a multiple eigenvalue; later we will consider the choice of λ. Assume without loss of generality that the matrix T has norm $\|T\|_F=1$ and is in the form

$$T=\begin{bmatrix}\lambda & x_\lambda^T\\ 0 & B_\lambda\end{bmatrix}$$

Let P_λ denote the projector associated with λ. Then the upper and lower bounds on d_λ we have discussed so far may be summarized as:

$$\frac{1}{(\|P_\lambda\|^2-1)^{1/2}}\geq\sigma_{\min}(B_\lambda-\lambda)\geq d_\lambda\geq\frac{\sigma_{\min}(B_\lambda-\lambda)}{4\|P_\lambda\|}. \tag{4.5}$$

The lower bound on d_λ was improved by at most a small constant factor in [2] but is good enough for our purposes. Since $(\|P_\lambda\|^2-1)^{-1/2}\geq\sigma_{\min}(B_\lambda-\lambda)$, we see that the lower bound on d_λ can be no smaller than the square of the upper bound $\sigma_{\min}(B_\lambda-\lambda)$ (recall that $\|T\|_F=1$ so that these are all bounds on the relative distance and generally less than 1). This is the maximum gap between the upper and lower bounds. We explain in Sect. 9 why we expect such a gap for algebraic functions (such as eigenvalues) in general. Examples [2, 9, 17, 18] have shown that either bound may be a better approximation of d_λ; for example, the lower bound is nearly exact for 2 by 2 matrices.

By slightly modifying the proof of the Bauer-Fike theorem, we can show Wilkinson's estimate of d_λ is nearly a lower bound:

$$d_\lambda\geq\min_{\lambda\neq\lambda'}\frac{1}{n}\frac{|\lambda-\lambda'|}{\|P_\lambda\|+\|P_{\lambda'}\|}. \tag{4.6}$$

In fact, given any partitioning $\sigma(B)=\bigcup_{i=1}^{b=1}\sigma_i$ of the spectrum of B into disjoint pieces there is a lower bound of the form in (4.5) or (4.6); these correspond to the extreme partitions $b=1$, $\sigma_1=\sigma(B)$ in (4.5) and $b=n$, $\sigma_i=\{\lambda_i'\}$ in (4.6). These results are stated in the next theorem:

Theorem 8. *Suppose*

$$T=\begin{bmatrix} \lambda & * & * & \cdot & * \\ & B_1 & * & \cdot & * \\ & & B_2 & \cdot & * \\ & & & \cdot & \cdot \\ & & & & B_{b-1} \end{bmatrix}$$

is block upper triangular. Suppose B_i has spectrum σ_i, dimension n_i and associated projector P_i, with $B_0=[\lambda]$ for consistency. Let S_i be an n by n_i matrix of orthonormal columns spanning the right invariant subspace of T belonging to σ_i, and let $S=[S_0|\dots|S_{b-1}]$. Then S block diagonalizes T: $S^{-1}TS=D=\operatorname{diag}(D_0, \dots, D_{b-1})$, where D_i has spectrum σ_i. Then

$$d_\lambda \geq \frac{1}{2b} \min_{0<i<b} \frac{\sigma_{\min}(D_i-\lambda)}{\max(\|P_0\|, \|P_i\|)}.$$

Proof. We claim that if $\bar{\lambda}\equiv\lambda(T+\delta T)$ is an eigenvalue of the perturbed matrix $T+\delta T$, then δT satisfies

$$\|\delta T\| \geq \frac{1}{b} \min_i \frac{\sigma_{\min}(D_i-\bar{\lambda})}{\|P_i\|} \tag{4.7}$$

In other words, $\bar{\lambda}$ must lie in an inclusion region about one of the σ_i, but the size depends only on the sensitivity of the cluster $\|P_i\|$ rather than the maximum sensitivity $\max_i \|P_i\|$ which is the usual form of the Bauer-Fike theorem. We prove (4.7) as follows. It is easy to see that $\|S\| \leq \sqrt{b}$ [1, Thm 2] and that

$$S^{-1}=\begin{bmatrix} S_{(0)}^{-1} \\ \cdot \\ S_{(b-1)}^{-1} \end{bmatrix}$$

($S_{(i)}^{-1}$ is n_i by n) satisfies $\|S_{(i)}^{-1}\|=\|P_i\|$ [1, Thm 3]. Assuming that $\bar{\lambda}$ is not already an eigenvalue of T, we have

$$\begin{aligned} 0&=\det(T+\delta T-\bar{\lambda})=\det((T-\bar{\lambda})(I+(T-\bar{\lambda})^{-1}\delta T))=\det(I+(T-\bar{\lambda})^{-1}\delta T) \\ &=\det(I+(D-\lambda)^{-1}S^{-1}\delta TS) \end{aligned}$$

implying $1 \leq \|(D-\bar{\lambda})^{-1}S^{-1}\delta TS\|$.

So far this is the usual proof. What changes is that now we write

$$\begin{aligned} 1 \leq \|(D-\bar{\lambda})^{-1}S^{-1}\delta TS\| &\leq \sqrt{b}\cdot\max_i \|(D_i-\bar{\lambda})^{-1}S_{(i)}^{-1}\delta TS\| \\ &\leq b\cdot\|\delta T\| \max_i \|(D_i-\bar{\lambda})^{-1}\| \, \|P_i\| \end{aligned}$$

or

$$\|\delta T\| \geq \frac{1}{b} \min_i \frac{\sigma_{\min}(D_i - \bar{\lambda})}{\|P_i\|}$$

proving (4.7).

From (4.7) we proceed as follows. If $\bar{\lambda}$ is a multiple eigenvalue of $T+\delta T$, then it must lie within the inclusion regions

$$\|\delta T\| \geq \frac{1}{b} \frac{\sigma_{\min}(D_i - \bar{\lambda})}{\|P_i\|}$$

for both $i=0$ and at least one $i>0$. In other words

$$\begin{aligned}
\|\delta T\| &\geq \frac{1}{b} \min_{0<i<b} \min_{\bar{\lambda}} \max\left(\frac{|\lambda-\bar{\lambda}|}{\|P_0\|}, \frac{\sigma_{\min}(D_i-\bar{\lambda})}{\|P_i\|}\right) \\
&\geq \frac{1}{b} \min_{0<i<b} \frac{1}{\max(\|P_0\|, \|P_i\|)} \min_{\bar{\lambda}} \max(|\lambda-\bar{\lambda}|, \sigma_{\min}(D_i-\bar{\lambda})) \\
&\geq \frac{1}{2b} \min_{0<i<b} \frac{1}{\max(\|P_0\|, \|P_i\|)} \min_{\bar{\lambda}} \sigma_{\min}(D_i-\lambda)
\end{aligned}$$

(where we have used [2, Corollary 4.9]). This is the desired result. q.e.d.

When $b=1$, we can take $D_1 = B_1$, thus providing an alternate proof of the lower bound in (4.5). Similarly, when $b=n$, we can strengthen the result slightly to get (4.6). No particular partitioning of $\sigma(B)$ is always best; for example, sometimes the lower bound in (4.5) is stronger and sometimes (4.6) is stronger, and both may simultaneously be gross underestimates. These results have also been obtained by Wilkinson.

If we let $d \equiv \min_{\lambda} d_\lambda$ be the shortest distance from T to any matrix with a multiple eigenvalue, we see that any upper bound on any d_λ is an upper bound on d and that a lower bound on d is given by the minimum of all lower bounds of all d_λ. For example, (4.5) leads to the lower bound

$$d \geq \min_{\lambda} \frac{\sigma_{\min}(B_\lambda - \lambda)}{4\|P_\lambda\|} \tag{4.8}$$

and (4.6) leads to the lower bound

$$d \geq \min_{\lambda \neq \lambda'} \frac{1}{n} \frac{|\lambda-\lambda'|}{\|P_\lambda\| + \|P_{\lambda'}\|}. \tag{4.9}$$

Even though neither (4.5) nor (4.6) is uniformly better than the other, it is easy to show that the bound in (4.9) may never be much smaller than bound in (4.8), but it may be much larger. Unfortunately, (4.9) may occasionally also be pessimistically small.

In short, estimating d_λ and d explicitly seem to be quite difficult analytical problems. In practice, for a given matrix, the technique of following a condition number "uphill" until it reaches infinity is still a practical optimization method for getting upper bounds on d_λ and d even if it is hard to prove sharper explicit bounds in general.

A related problem to estimating d is computing the distance to the nearest matrix with an eigenvalue of higher multiplicity. Wilkinson has constructed examples where even though it is possible to coalesce any two eigenvalues with a perturbation of some very small size, it requires a far larger perturbation to make all the eigenvalues coalesce simultaneously [17, 18]. Much work remains to be done to understand the geometry of the set of matrices with multiple eigenvalues.

5. Polynomial Zero Finding

Let $p(z)=\sum_{i=0}^{n} p_i z^i$ be a complex polynomial with a simple zero at x: $p(x)=0$. Let $\|p\|$ denote the Euclidean norm of its vector of coefficients. If p is perturbed by adding a sufficiently small polynomial $e(z)=\sum_{i=0}^{n} e_i z^i$, then to first order $p+e$ will have a simple zero at $x+\delta x=x-e(x)/p'(x)$. This follows from simply solving the Taylor expansion

$$0=(p+e)(x+\delta x)=p(x)+\delta x\, p'(x)+e(x)+O(\|e\|^2)=\delta x\, p'(x)+O(\|e\|^2)$$

for δx. Thus, we may use $|1/p'(x)|$ as a condition number for the zero x. In this section we will find relationships between the reciprocal of the condition number $|p'(x)|$ and the distance from p (measuring using $\|\cdot\|$) to the nearest polynomial where x merges into a multiple zero. An n-tuple eigenvalue is infinitely ill-conditioned because a perturbation of p of norm ε can cause a perturbation of x of size $\varepsilon^{1/n}$ which has an infinite derivative at $\varepsilon=0$. Therefore we may take the set of polynomials with multiple zeros as our set of ill-posed problems.

In this section we will show that the distance to the nearest polynomial where x merges into a multiple zero is indeed bounded by a small multiple of the reciprocal of the condition number. As a lower bound we get a quantity proportional to the reciprocal of the condition number squared. The zero x is an algebraic function of the coefficients of p, and in Sect. 9 we explain why we expect a lower bound proportional to the reciprocal of the condition number squared for any algebraic function.

The most general previous result relating $|p'(x)|$ to the distance to the nearest polynomial where x becomes double is due to Hough [7]. We need some notation: if p is a polynomial $p(z)=\sum_{i=0}^{n} p_i z^i$ of degree at most n, let $\underline{p}\equiv[p_0, \ldots, p_n]^T$ denote the vector of its coefficients. For any complex number z, let $\underline{z}\equiv[1, z, z^2, \ldots, z^n]^T$. Therefore, $p(z)=\underline{p}^T\underline{z}$. Also, let $\underline{z}'\equiv[0, 1, 2z, 3z^2, \ldots, nz^{n-1}]$; thus $p'(z)=\underline{p}^T\underline{z}'$. Finally, let $\underline{z}''\equiv[0, 0, 2, 6z, \ldots, n(n-1)z^{n-2}]$; thus $p''(z)=\underline{p}^T\underline{z}''$.

Theorem 9. [Hough] *Suppose p is a polynomial of degree at least 2 and $p(x)=0$. Then the smallest polynomial e of degree no greater than p such that $p+e$ has a double zero at x has norm*

$$\|e\| = \frac{|p'(x)|}{\left\| \underset{\sim}{x}' - \frac{\underset{\sim}{x}^* \underset{\sim}{x}'}{\|\underset{\sim}{x}\|^2} \cdot \underset{\sim}{x} \right\|} \leq \sqrt{2}\,|p'(x)| \cdot \min(1, |x|^{2-n}).$$

Sketch of Proof. This is an underdetermined linear least squares problem

$$\begin{bmatrix} 1 & x & x^2 \ldots & x^i & \ldots & x^n \\ 0 & 1 & 2x \ldots & i x^{i-1} & \ldots & n x^{n-1} \end{bmatrix} \cdot \begin{bmatrix} e_0 \\ e_1 \\ \cdot \\ e_n \end{bmatrix} = \begin{bmatrix} 0 \\ -p'(x) \end{bmatrix}$$

which can be solved explicitly for a solution of minimum norm, giving the claimed solution.

Our approach proceeds by computing the gradient of $|1/p'(x)|$ under changes in p, subject to $p(x)=0$. We compute this gradient in the next lemma.

Lemma 3. *Let p be a polynomial of degree at least 1. Let D_e denote the directional derivative of $1/|p'(x)|$ (x a zero of p) in the direction of the polynomial e, where $\|e\|=1$. Then*

$$D_e = \frac{1}{|p'(x)|} \operatorname{Re} \underset{\sim}{e}^T \left[\frac{\underset{\sim}{x}\, p''(x)}{(p'(x))^2} - \frac{\underset{\sim}{x}'}{p'(x)} \right].$$

Furthermore,

$$|D_e| \leq \Delta \equiv \frac{\left\| \frac{\underset{\sim}{x}\, p''(x)}{p'(x)} - \underset{\sim}{x}' \right\|}{|p'(x)|^2}.$$

Proof. The directional derivative of $1/|p'(x)|$ in the direction of a unit vector $\underset{\sim}{e}$ is

$$D_e = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left[\frac{1}{|(p+\varepsilon e)'(x+\delta x)|} - \frac{1}{|p'(x)|} \right],$$

where $x+\delta x$ is a zero of $p+\varepsilon e$. From our earlier formula for δx we have

$$D_e = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left[\frac{1}{\left| p'(x) - \frac{\varepsilon (x)\, p''(x)}{p'(x)} + \varepsilon e'(x) + O(\varepsilon^2) \right|} - \frac{1}{|p'(x)|} \right]$$

$$= \lim_{\varepsilon \to 0} \frac{1}{\varepsilon |p'(x)|} \left[\frac{1}{\left| 1 - \frac{\varepsilon e(x)\, p''(x)}{(p'(x))^2} + \frac{\varepsilon e'(x)}{p'(x)} + O(\varepsilon^2) \right|} - 1 \right].$$

Noting that if η is a small complex number,

$$\frac{1}{|1+\eta|}=1-\operatorname{Re}\eta+O(|\eta|^2),$$

we see that

$$D_e=\frac{1}{|p'(x)|}\operatorname{Re}\left[\frac{e(x)\,p''(x)}{(p'(x))^2}-\frac{e'(x)}{p'(x)}\right]$$

$$=\frac{1}{|p'(x)|}\operatorname{Re}\underset{\sim}{e}^T\left[\frac{\underset{\sim}{x}\,p''(x)}{(p'(x))^2}-\frac{\underset{\sim}{x}'}{p'(x)}\right]$$

as claimed. We can clearly pick a unit vector $\underset{\sim}{e}$ to make D_e equal its upper bound

$$\Delta\equiv\frac{\left\|\dfrac{\underset{\sim}{x}\,p''(x)}{p'(x)}-\underset{\sim}{x}'\right\|}{|p'(x)|^2}.$$

q.e.d.

Applying this lemma for e perpendicular to $\underset{\sim}{x}$ (i.e. $e(x)=0$) yields the same result as in Theorem 9:

Theorem 10. *Suppose p is a polynomial of degree at least 2 and $p(x)=0$. Then the smallest polynomial e of degree no greater than p such that $p+e$ has a double zero at x has norm*

$$\|e\|=\frac{|p'(x)|}{\left\|\underset{\sim}{x}'-\dfrac{\underset{\sim}{x}^*\underset{\sim}{x}'}{\|\underset{\sim}{x}\|^2}\cdot\underset{\sim}{x}\right\|}.$$

Proof. By choosing e such that $e^T\underset{\sim}{x}=0$ in Lemma 3 we get a vector field in the space of polynomials along whose integral curves the zero x of p will not move. The unit vector e which satisfies $e^T\underset{\sim}{x}=0$ and maximizes D_e is clearly the one in the direction of the vector component of $\underset{\sim}{x}'$ orthogonal to $\underset{\sim}{x}$, or

$$e^T\underset{\sim}{x}'=\left\|\underset{\sim}{x}'-\frac{\underset{\sim}{x}^*\underset{\sim}{x}'}{\|\underset{\sim}{x}\|^2}\cdot\underset{\sim}{x}\right\|\equiv n(x).$$

This vector field is clearly continuous, so let p_s be an integral parameterized by arclength. Let $y(s)=|1/p_s'(x)|$. Then from Lemma 3 we have

$$\frac{d}{ds}y(s)=n(x)\,y^2(s)$$

so by Lemmas 1 and 2 $y(s)$ has a pole at $s_0=|p'(x)|/n(x)$, i.e. p_{s_0} has a double zero at x. To see that no closer polynomial to p has this property, note that under the constraint that $e^T\underset{\sim}{x}=0$, $|D_e|\le n(x)/|p'^2(x)|$, so by Lemma 2 $|p'(x)|/n(x)$ is the minimum distance to a polynomial with a double root at x as well. q.e.d.

Without much more effort, we get a similar theorem with a different constraint on e.

Theorem 11. *Let p be a polynomial of degree at least 3 and $p(x)=0$. Then there is a quadratic polynomial e of norm at most $2|p'(x)|$ such that $p+e$ has a double zero. This double zero corresponds to x in that as ε increases from 0 to 1, the polynomial $p+\varepsilon e$ has a simple zero which moves from x when $\varepsilon=0$ until it merges with another zero to form a double zero at $\varepsilon=1$.*

Proof. The first three components of

$$\frac{\underset{\sim}{x}\, p''(x)}{p'(x)} - \underset{\sim}{x}'$$

are

$$[p''(x)/p'(x),\, x\cdot p''(x)/p'(x)-1,\, x^2\cdot p''(x)/p'(x)-2x] \equiv [y,\, xy-1,\, x^2 y-2x] \tag{5.1}$$

For any x or y, one of the components of this last vector has to have absolute value at least 1/2. To show this, assume to the contrary that all there components are smaller than 1/2. Then $|xy-1|<1/2$ implies $|xy|>1/2$ or $|x|>1/(2|y|)>1$. Thus $|x^2 y-2x|=|x|\cdot|xy-2|>1-|xy-1|>1/2$, a contradiction. Therefore by choosing e a unit vector pointing the same direction as the vector in (5.1), we get

$$D_e \geq \frac{1}{2|p'(x)|^2}.$$

The vector field defined by this choice of e is smooth since the components in (5.1) are smooth, so let p_s be an integral curve with zero x_s, where s is arclength. Define $y(s)=|1/p_s(x_s)|$. Thus

$$\frac{d}{ds}\, y(s) \geq \frac{y^2(s)}{2}$$

so by Lemma 1 there is a polynomial p_{s_0} with a double zero with $s_0=\|p-p_{s_0}\| \leq 2|p'(x)|$. q.e.d.

Just as we used Lemma 1 to derive an upper bound on the distance to nearest polynomial with a multiple zero, we will use Lemma 2 to derive a lower bound.

Theorem 12. *Let p be a polynomial of degree $n\geq 2$. Let p_s be a continuous map from $s\in[0,s_0]$ to the space of polynomials of degree no greater than n, such that $p_0=p$, and such that s is the arclength parameter. Let x_s be a zero of p_s such that x_s is a continuous function of s and $x_0=x$. Then if x_{s_0} is a multiple zero of p_{s_0},*

$$\|p-p_{s_0}\| \geq \frac{|p'(x)|^2}{4n^3\cdot \max\limits_{0\leq s\leq s_0} (\|p_s\|,\, \|p_s\|\, |x_s|^{2n-2})}$$

Proof. From Lemma 3 we know that if $\|e\|=1$ then

$$\begin{aligned}|D_e| \le \Delta &= \frac{\|\underline{x}p''(x)-\underline{x}'p'(x)\|}{|p'(x)|^3}\\ &= \frac{\|\underline{x}p^T\underline{x}''-\underline{x}'p^T\underline{x}'\|}{|p'(x)|^3}\\ &= \frac{\|\underline{x}\underline{x}''^T p-\underline{x}'\underline{x}'^T p\|}{|p'(x)|^3}\\ &\le \frac{\|\underline{x}\underline{x}''^T-\underline{x}'\underline{x}'^T\|\cdot\|p\|}{|p'(x)|^3}\end{aligned}$$

where $\|\cdot\|$ is the 2-norm of a matrix. The matrix $\underline{x}\underline{x}''^T-\underline{x}'\underline{x}'^T$ is an $n+1$ by $n+1$ matrix whose norm we may bound simply by

$$\|\underline{x}\underline{x}''^T-\underline{x}'\underline{x}'^T\| \le \|\underline{x}\|\cdot\|\underline{x}''^T\|+\|\underline{x}'\|^2 \le 2n^3\max(1,|x|^{2n-2}).$$

This implies

$$|D_e| \le \frac{2n^3\max(1,|x|^{2n-2})\,\|p\|}{|p'(x)|^3}.$$

Now consider the function $y(s)\equiv 1/|p_s'(x_s)|$. We have just shown

$$\frac{d}{ds}y(s) \le 2n^3\max(1,|x_s|^{2n-2})\,\|p_s\|\cdot y^3(s)$$

so that by Lemma 2 $y(s)$ remains finite for

$$s < s_0 = \frac{|p'(x)|^2}{4n^3\cdot \max\limits_{0\le s\le s_0}(\|p_s\|,\|p_s\|\,|x_s|^{2n-2})}$$

as claimed. q.e.d.

At first glance it would seem hard to apply Theorem 12 since s_0 is defined in terms of itself. In practice, however, one would apply the theorem when it is possible to make x a multiple zero by only a small perturbation in p. Thus, x_s should not vary much from x nor should $\|p_s\|$ vary much from $\|p\|$. In such cases an approximate lower bound is thus provided by the expression

$$\frac{|p'(x)|^2}{4n^3\cdot\|p\|\cdot\max(1,|x|^{2n-2})}.$$

Alternatively, if we scale p appropriately, we can effectively estimate both $\|p_s\|$ and $|x_s|$. For example, if we assume that the coefficients p_i of p are no larger than the leading coefficient p_n divided by n: $|p_i|\le|p_n|/n$ for $i<n$ (which can be achieved by the simple change of variable $x\to\alpha x$ for appropriate α), we get the following bounds:

Corollary 1. *If p is a polynomial of degree n where the coefficients p_i satisfy $|p_i| \le |p_n|/n$ for $i<n$, and if z is a simple zero of p, then the distance from p to the nearest polynomial q where z coalesces into a multiple root (in the sense of Theorem 12) is bounded below by*

$$\|p-q\| \ge \min\left(\frac{\|p\|}{n^2}, \frac{|p'(z)|^2}{(5e^2+9(3/8)^{1/2})\, n^3 \|p\|}\right) \ge \min\left(\frac{\|p\|}{n^2}, \frac{0.0235 \cdot |p'(z)|^2}{n^3 \|p\|}\right)$$

Proof. Assume $\eta \equiv \|p-q\| \le \|p\|/n^2$; otherwise the theorem holds trivially. Since $|p_i| \le |p_n|/n$, it is easy to see any zero z of p satisfies $|z| \le 1$. Since $\|p-q\| = \eta$, if p_s is any polynomial on the line segment connecting p and q, $\|p-p_s\| \le \eta$, so the coefficients p_{si} of p_s satisfy $|p_{sn}| \ge |p_n| - \eta$ and $|p_{si}| \le |p_i| + \eta$. If z' is a zero of p_s, it is easy to see z'/α is a zero of the polynomial $\bar{p}_s(x) \equiv p_s(\alpha x)$ whose coefficients $\bar{p}_{si}$ satisfy $|\bar{p}_{si}| = |\alpha^i p_{si}$. If we choose $\alpha = (|p_n| + n\eta)/(|p_n| - \eta)$ then

$$\frac{|\bar{p}_{si}|}{|\bar{p}_{sn}|} \le \frac{1}{\alpha} \frac{|p_i| + \eta}{|p_n| - \eta} \le \frac{1}{\alpha} \frac{|p_n|/n + \eta}{|p_n| - \eta} = \frac{1}{n}$$

so $|z'/\alpha| \le 1$ or $|z'| \le \alpha$. Thus by Theorem 12 η must satisfy

$$\eta \ge \frac{|p'(z)|^2}{4n^3(\|p\| + \eta) \max(1, \alpha^{2n-2})} \ge \frac{|p'(z)|^2}{4n^3\left(1+\frac{1}{n^2}\right)\|p\| \left(\frac{1+\frac{n\eta}{|p_n|}}{1-\frac{n\eta}{|p_n|}}\right)^{2n-2}}$$

which is true only if

$$\frac{4n^3\left(1+\frac{1}{n^2}\right)\|p\|}{|p'(z)|^2} \eta \left(1+\frac{n\eta}{|p_n|}\right)^{2n-2} \ge \left(1-\frac{\eta}{|p_n|}\right)^{2n-2}$$

which, since $\|p\| \le |p_n|(1+1/n)^{1/2}$ and $n\eta/|p_n| \le \|p\|/(n^2|p_n|) \le (n^{-2}+n^{-3})^{1/2} \le (n-1)^{-1}$, is in turn true only if

$$\frac{4n^3\left(1+\frac{1}{n^2}\right)\|p\|}{|p'(z)|^2} \eta \left(1+\frac{1}{n-1}\right)^{2(n-1)} \ge 1 - \frac{(2n-2)\left(1+\frac{1}{n}\right)^{1/2} \eta}{\|p\|}$$

or

$$\eta \ge \frac{1}{\frac{4n^3(1+n^{-2})e^2\|p\|}{|p'(z)|^2} + \frac{(2n-2)(1+n^{-1})^{1/2}}{\|p\|}}.$$

Now

$$|p'(z)| \le n|p_n| + \sum_{j<n} j\frac{|p_n|}{n} \le |p_n| \frac{3n-1}{2} \le \|p\| \frac{3n-1}{2}$$

so η is in turn bounded below by

$$\eta \geq \frac{|p'(z)|^2}{\|p\| \left(4n^3(1+n^{-2})e^2 + \frac{(n-1)(3n-1)^2(1+n^{-1})^{1/2}}{2}\right)}$$

$$\geq \frac{|p'(z)|^2}{\|p\| \, n^3(5e^2+9(3/8)^{1/2})} \geq \frac{0.0235|p'(z)|^2}{n^3 \|p\|}$$

as claimed. q.e.d.

6. Pole Assignment

The pole assignment problem is defined as follows: given an n by n matrix A, an n by m matrix B, and a set $\{\lambda_i\}$ of n complex numbers, find an m by n *feedback matrix* F such that $A+BF$ has eigenvalues $\{\lambda_i\}$. The motivation for this problem is the following: given a *control system* $\dot{x}=Ax+Bu$, choose the *control input* u as a function Fx of x (feedback) to make the matrix of the controlled system $\dot{x}=(A+BF)x$ have a specified spectrum. It is well known that this problem has a solution for arbitrary $\{\lambda_i\}$ if and only if the pair (A, B) is *controllable*, i.e. $[B|AB|A^2B|\ldots|A^{n-1}B]$ has full rank n [19]. If the pair (A, B) is not controllable, then some eigenvalues (called the *uncontrollable modes*) of $A+BF$ will be independent of F (and be eigenvalues of A); the remaining eigenvalues can be set arbitrarily by choosing F.

The *robust pole assignment problem*, as defined in [11], is to find F subject to the additional condition that X, the eigenvector matrix of $A+BF=X\Lambda X^{-1}$, be as well conditioned as possible (here $\Lambda=\operatorname{diag}(\lambda_i)$). The condition number $\kappa(X)\equiv\|X\|\cdot\|X^{-1}\|$ (in this section $\|\cdot\|$ denotes the 2-norm and $\|\cdot\|_F$ the Frobenius norm) of the best conditioned X turns out to measure the size the sensitivity of both F and the time dependent solution of the control system $\dot{x}=(A+BF)x$. For example, if the λ_i are distinct, then $\|F\|$ will get larger as $\kappa(X)$ gets larger (see [11] for details). Therefore, we shall take $\kappa(X)$ as our condition number for the robust pole assignment problem.

We will also use a slightly different measure of distance than used before:

$$\operatorname{dist}((A_0, B_0), (A_1, B_1)) \equiv \frac{\|A_0-A_1\|_F}{\min(\|A_0\|_F, \|A_1\|_F)} + \frac{\|B_0-B_1\|_F}{\min(\|B_0\|_F, \|B_1\|_F)}. \tag{6.1}$$

This distance has the advantage of being insensitive to the scaling of A and B.

As in previous sections, we are interested in relating $\kappa(X)$ to the distance from (A, B) to the nearest ill-posed pair (for which no X exists). How do we characterize the set $\mathbf{S}$ of problems (A, B) (for fixed $\{\lambda_i\}$) which are ill-posed? From the above discussion, it is clear that it includes all uncontrollable (A, B) where $\{\lambda_i\}$ does not include the uncontrollable modes (e.g., if no λ_i is an eigenvalue of A).

Using Lemma 2, we will prove a theorem which gives a lower bound on the distance to the set $\mathbf{S}$ of ill-posed problems in terms of the reciprocal of the condition number $\kappa(X)$. Before doing so we state the following lemma from [11]:

Lemma 4. [Kautsky, Nichols, Van Dooren] *Let*

$$B=[U_0, U_1]\cdot\begin{bmatrix} Z \\ 0 \end{bmatrix}$$

where $U=[U_0, U_1]$ *is unitary and* Z *is of full* rank. *Let* $\mathbf{X}_i=\mathbf{N}(U_1^*(A-\lambda_i I))$, *where* $\mathbf{N}(\cdot)$ *denotes the null space. Then the i-th column* x_i *of* X *(the right eigenvector of* $A+BF$ *for the eigenvalue* λ_i*) satisfies* $x_i \in \mathbf{X}_i$.

Proof. Premultiply the equation $A+BF=X\Lambda X^{-1}$ by U^* and rearrange to get

$$\begin{bmatrix} Z \\ 0 \end{bmatrix} F=\begin{bmatrix} U_0^* \\ U_1^* \end{bmatrix}(AX-X\Lambda)\,X^{-1}$$

or, taking the last n-rank(B) rows

$$0=U_1^*(AX-X\Lambda).$$

q.e.d.

Let X_i be a matrix of orthonormal columns spanning $\mathbf{X}_i$, and let $S=[X_1|\ldots|X_n]$. If B is of full rank and λ_i is not an eigenvalue of A, X_i will be n by m. Lemma 4 says that the eigenvector matrix X can be written as

$$X=S\cdot\begin{bmatrix} u_1 & \ldots & 0 \\ \cdot & \cdot & \cdot \\ 0 & u_i & 0 \\ \cdot & \cdot & \cdot \\ 0 & \ldots & u_n \end{bmatrix} \tag{6.2}$$

where u_i is a column vector of the same dimension as $\mathbf{X}_i$. Since it is hard to characterize the condition number of X, we instead use the following lower bound based on S:

Lemma 5. *Let* X *and* S *be defined as above. Then*

$$\kappa(X)\geq\sigma_{\min}^{-1}(S)=\|(SS^*)^{-1}\|^{1/2}=\left\|\left(\sum_{i=1}^{n} X_i X_i^*\right)^{-1}\right\|^{1/2}.$$

Proof. Assume without loss of generality that $\|X\|=1$. This clearly implies that $\|u_i\|\leq 1$ in (6.2). Now let σ be the smallest singular value of S and v^* the corresponding left singular vector, i.e. $\|v^*\|=1$ and $\|v^*S\|=\sigma$. Then it follows simply that $\|v^*X\|\leq\sigma$. Thus

$$\kappa(X)\geq\sigma_{\min}^{-1}(S)=\|(SS^*)^{-1}\|^{1/2}=\left\|\left(\sum_{i=1}^{n} X_i X_i^*\right)^{-1}\right\|^{1/2}$$

as claimed. q.e.d.

It is to $\sigma_{\min}^{-2}(S)$ that we will apply Lemma 2 to get a lower bound on the distance to the nearest ill-posed problem.

Lemma 6. *Let $\sigma_{\min}^{-1}(S(A, B))$ be the value of $\sigma_{\min}^{-1}(S)$ for the S defined by A and B. Let $\mathbf{S}$ be the set of problems (A, B) for which $\sigma_{\min}(A, B))=0$. Let $\operatorname{dist}((A, B), \mathbf{S})$ be defined as*

$$\operatorname{dist}((A, B), \mathbf{S}) \equiv \inf_{(A_s, B_s)\in \mathbf{S}} \operatorname{dist}((A, B), (A_s, B_s)).$$

Assume B is of full rank and that none of the λ_i is an eigenvalue of A so that

$$\kappa_B \equiv \|B\|_F/\sigma_{\min}(B)$$

and

$$\kappa_A \equiv \max_i (\|A-\lambda_i\|_F \cdot \|(A-\lambda_i)^{-1}\|, \|A\|_F \cdot \|(A-\lambda_i)^{-1}\|)$$

are well defined. Then

$$\operatorname{dist}((A, B), \mathbf{S}) \geq \frac{0.187}{\sqrt{n}\cdot \kappa_B \cdot \kappa_A \cdot \sigma_{\min}^{-1}(S(A, B))}.$$

Proof. Consider perturbations $A+\delta A$ of A and $B+\delta B$ of B. We need to estimate for small δA and δB

$$\sigma_{\min}^{-2}(S) - \sigma_{\min}^{-2}(S_\delta) = \|(SS^*)^{-1}\| - \|(S_\delta S_\delta^*)^{-1}\| \equiv \delta \|(SS^*)^{-1}\|$$

where $S=S(A, B)$ and $S_\delta = S(A+\delta A, B+\delta B)$. Let X_i denote the matrices of orthonormal columns comprising $S=[X_1| \dots |X_n]$ and let σ be the smallest singular value of S. If the perturbations δA and δB yield perturbations δX_i in X_i, then S becomes $[X_1+\delta X_1| \dots |X_n+\delta X_n]$ and SS^* becomes (to first order)

$$SS^* + \sum_{i=1}^{n} \delta X_i X_i^* + X_i \delta X_i^* = SS^* + [\delta X_1| \dots |\delta X_n] S^* + S[\delta X_1| \dots |\delta X_n]^*$$

and $(SS^*)^{-1}$ becomes (again to first order)

$$(SS^*)^{-1} - (SS^*)^{-1}([\delta X_1| \dots |\delta X_n] S^* + S[\delta X_1| \dots |\delta X_n]^*)(SS^*)^{-1}.$$

What we need to estimate, then, is

$$\|(SS^*)^{-1}([\delta X_1| \dots |\delta X_n] S^* + S[\delta X_1| \dots |\delta X_n]^*)(SS^*)^{-1}\|.$$

Now it is easy to see that $\|S^*(SS^*)^{-1}\| = \sigma^{-1}$ so

$$\begin{aligned}
&\|(SS^*)^{-1}([\delta X_1| \dots |\delta X_n] S^* + S[\delta X_1| \dots |\delta X_n]^*)(SS^*)^{-1}\| \\
&\quad \leq 2\sigma^{-3} \cdot \|[\delta X_1| \dots |\delta X_n]\| \\
&\quad \leq 2\sqrt{n}\,\sigma^{-3} \cdot \max_{1\leq i\leq n} \|\delta X_i\|. \qquad (6.3)
\end{aligned}$$

To estimate $\|\delta X_i\|$ note that $\mathbf{X}_i = \mathbf{N}(U_1^*(A-\lambda_i)) = \mathbf{R}((A-\lambda_i)^* U_1)^\perp$, where $\mathbf{R}(\cdot)$ denotes the column space of $(\cdot)$. Now let Y be an n by $n-m$ matrix of orthonormal columns and $Y^\perp$ an n by m matrix of orthonormal columns orthogonal to the columns of Y. Then if Y is perturbed to $Y+\delta Y$ (δY in an arbitrary direction), it is easy to verify that $Y^\perp$ is perturbed to $Y^\perp - Y\delta Y^* Y^\perp$.

In our case we want Y to span $\mathbf{R}((A-\lambda_i)^* U_1)$ so we may take

$$Y = (A-\lambda_i)^* U_1 (U_1^*(A-\lambda_i)(A-\lambda_i)^* U_1)^{-1/2}.$$

X_i can be taken as $Y^\perp$ and so

$$\delta Y = (\delta A^* U_1 + (A-\lambda_i)^* \delta U_1)\cdot(U_1^*(A-\lambda_i)(A-\lambda_i)^* U_1)^{-1/2} + (A-\lambda_i)^* U_1 D$$

where $\|D\|$ is on the order of $\|\delta A\|$ and $\|\delta B\|$ and where the result of perturbing B to $B+\delta B$ is to perturb U_1 to $U_1+\delta U_1$. Thus

$$\delta X_i = -Y[(\delta A^* U_1 + (A-\lambda_i)^* \delta U_1)(U_1^*(A-\lambda_i)(A-\lambda_i)^* U_1)^{-1/2}] X_i$$

so

$$\begin{aligned}\|\delta X_i\| &\le (\|\delta A\| + \|A-\lambda_i\|\cdot\|\delta U_i\|)\,\|(U_1^*(A-\lambda_i)(A-\lambda_i)^* U_1)^{-1/2}\| \\ &\le (\|\delta A\| + \|A-\lambda_i\|\cdot\|\delta U_i\|)\,\|(A-\lambda_i)^{-1}\|.\end{aligned}$$

We estimate $\|\delta U_1\|$ similarly. Let $Y = B(B^*B)^{-1/2}$ be a set of orthonormal columns spanning the range of B. We may take U_0 as Y. Perturbing B to $B+\delta B$ makes an equivalent change of $\delta Y = \delta B(B^*B)^{-1/2} + BD$ in Y, where $\|D\|$ is on the order of $\|\delta B\|$, so

$$\|\delta U_1\| = \|-Y\delta Y^* Y^\perp\| \le \frac{\|\delta B\|}{\sigma_{\min}(B)}.$$

Putting these estimates together yields

$$\|\delta X_i\| \le \|\delta A\|\cdot\|(A-\lambda_i)^{-1}\| + \|\delta B\|\cdot\sigma_{\min}^{-1}(B)\cdot\kappa(A-\lambda_i)$$

and substituting into (6.3) we get the following bound on the change in $\|(SS^*)^{-1}\|$:

$$\begin{aligned}\delta\|(SS^*)^{-1}\| &\le 2\sqrt{n}\,\sigma^{-3}\cdot\max_i(\|\delta A\|\cdot\|(A-\lambda_i)^{-1}\| + \|\delta B\|\cdot\sigma_{\min}^{-1}(B)\cdot\kappa(A-\lambda_i)) \\ &\le 2\sqrt{n}\,\sigma^{-3}\cdot\left(\frac{\|\delta A\|_F}{\|A\|_F} + \frac{\|\delta B\|_F}{\|B\|_F}\right)\cdot\kappa_A\cdot\kappa_B. \qquad (6.4)\end{aligned}$$

We are now prepared to apply Lemma 2. Let $(A(s), B(s))$ be any smooth curve from $(A(0), B(0)) = (A, B)$ to $(A(s_0), B(s_0)) \in \mathbf{S}$ with the following properties:

(1) It is parameterized by arclength in the sense that

$$\frac{\left\|\frac{d}{ds}A(s)\right\|_F}{\|A(s)\|_F} + \frac{\left\|\frac{d}{ds}B(s)\right\|_F}{\|B(s)\|_F} = 1.$$

(2) $A(s)$ is the shortest smooth path from $A(0)$ to $A(s_0)$ such that $\|A(s)\|_F$ lies between $\min(\|A(0)\|_F, \|A(s_0)\|_F)$ and $\max(\|A(0)\|_F, \|A(s_0)\|_F)$ for all $0 \le s \le s_0$. It is easy to see that this assumption implies that the length of the curve $A(s)$ (measured using $\|\cdot\|_F$) is no more than $\pi/2$ times as long as the straight line distance $\|A(0)-A(s_0)\|_F$. We make the analogous assumption about $B(s)$.

Let $y(s)=\sigma_{\min}^{-2}(S(A(s), B(s)))$. Then from (6.4) we see

$$\frac{d}{ds}\, y(s) \le 2\sqrt{n}\, \kappa_{A(s)}\, \kappa_{B(s)}\, y^{3/2}(s)$$

so by applying Lemma 2 with $\beta=3/2$ the "distance" s_0 to set the S is at least

$$s_0 \ge \frac{1}{\sqrt{n} \max\limits_{0\le s\le s_0} (\kappa_{A(s)}\, \kappa_{B(s)})\, y^{1/2}(0)} = \frac{1}{\sqrt{n} \max\limits_{0\le s\le s_0} (\kappa_{A(s)}\, \kappa_{B(s)})\, \sigma^{-1}}.$$

We relate s_0 to the relative distance (6.1) as follows:

$$s_0 = \int_0^{s_0} \frac{\left\|\frac{d}{ds} A(s)\right\|_F}{\|A(s)\|_F} + \frac{\left\|\frac{d}{ds} B(s)\right\|_F}{\|B(s)\|_F}\, ds$$

$$\le \frac{\int_0^{s_0} \left\|\frac{d}{ds} A(s)\right\|_F ds}{\min\limits_{0\le s\le s_0} \|A(s)\|_F} + \frac{\int_0^{s_0} \left\|\frac{d}{ds} B(s)\right\|_F ds}{\min\limits_{0\le s\le s_0} \|B(s)\|_F}$$

$$\le \frac{\frac{\pi}{2}\|A(0)-A(s_0)\|_F}{\min(\|A(0)\|_F, \|A(s_0)\|_F} + \frac{\frac{\pi}{2}\|B(0)-B(s_0)\|_F}{\min(\|B(0)\|_F, \|B(s_0)\|_F}.$$

Therefore $2\cdot s_0/\pi$ is a lower bound on the distance measure (6.1).

It remains to estimate the maximum of $\kappa_{A(s)}\, \kappa_{B(s)}$ in the denominator.

Note that if

$$\frac{\|A(s)-A\|_F}{\min(\|A\|_F, \|A(s_0)\|_F)} + \frac{\|B(s)-B\|_F}{\min(\|B\|_F, \|B(s_0)\|_F)} \le \eta \tag{6.5}$$

then

$$\|\delta A\|_F \equiv \|A(s)-A\|_F \le \eta\, \|A\|_F \le \eta\, \kappa_A\, \|(A-\lambda_i)^{-1}\|^{-1}$$

and

$$\|\delta B\|_F \equiv \|B(s)-B\|_F \le \eta\, \|B\|_F = \eta\, \kappa_B\, \sigma_{\min}(B).$$

Therefore

$$\kappa_{A(s)} = \max_i (\|A(s)\|_F \cdot \|(A(s)-\lambda_i)^{-1}\|, \|A(s)-\lambda_i\|_F \cdot \|(A(s)-\lambda_i)^{-1}\|)$$

$$\leq \max_i \left(\frac{\kappa_A}{1-\|(A-\lambda_i)^{-1}\| \cdot \|\delta A\|_F} + \frac{\|\delta A\|_F \cdot \|(A-\lambda_i)^{-1}\|}{1-\|(A-\lambda_i)^{-1}\| \cdot \|\delta A\|_F} \right)$$

$$\leq \kappa_A \cdot \frac{1+\eta\kappa_A}{1-\eta\kappa_A}.$$

Similarly

$$\kappa_{B(s)} = \frac{\|B(s)\|_F}{\sigma_{\min}(B(s))} \leq \frac{\|B\|_F + \|\delta B\|_F}{\sigma_{\min}(B) - \|\delta B\|_F} \leq \kappa_B \cdot \frac{1+\eta\kappa_B}{1-\eta\kappa_B}.$$

Thus (6.5) implies

$$\kappa_{A(s)}\,\kappa_{B(s)} \leq \kappa_A\,\kappa_B \frac{(1+\eta\max(\kappa_A,\kappa_B))^2}{(1-\eta\max(\kappa_A,\kappa_B))^2}.$$

Thus, dist$((A, B), \mathbf{S})$ can be less than η only if

$$\eta \geq \frac{2}{\pi\sqrt{n}\,\kappa_A\,\kappa_B\,\sigma^{-1} \dfrac{(1+\eta\max(\kappa_A,\kappa_B))^2}{(1-\eta\max(\kappa_A,\kappa_B))^2}}$$

or, rearranging

$$\left(\frac{\pi}{2}\sqrt{n}\,\kappa_A\,\kappa_B\,\sigma^{-1}\eta\right) \cdot (1+\max(\kappa_A,\kappa_B)\,\eta)^2 \geq (1-\max(\kappa_A,\kappa_B)\,\eta)^2. \tag{6.6}$$

Since $\sigma \leq \|S\| \leq \sqrt{n}$, $\kappa_A \geq 1$ and $\kappa_B \geq 1$

$$\frac{\pi}{2}\sqrt{n}\,\kappa_A\,\kappa_B\,\sigma^{-1} \geq \max(\kappa_A,\kappa_B).$$

Thus (6.6) is true only if

$$\left(\frac{\pi}{2}\sqrt{n}\,\kappa_A\,\kappa_B\,\sigma^{-1}\eta\right) \cdot \left(1+\frac{\pi}{2}\sqrt{n}\,\kappa_A\,\kappa_B\,\sigma^{-1}\eta\right)^2 \geq \left(1-\frac{\pi}{2}\sqrt{n}\,\kappa_A\,\kappa_B\,\sigma^{-1}\eta\right)^2. \tag{6.7}$$

Letting $x = \frac{\pi}{2}\sqrt{n}\,\kappa_A\,\kappa_B\,\sigma^{-1}\eta$, we see (6.7) is equivalent to $x(1+x)^2 \geq (1-x)^2$, or $x^3+x^2+3x-1 \geq 0$, which is only true if $x \geq 0.295$. Thus

$$\operatorname{dist}((A, B), \mathbf{S}) \geq \frac{0.295}{\frac{\pi}{2}\sqrt{n}\,\kappa_A\,\kappa_B\,\sigma^{-1}} \geq \frac{0.187}{\sqrt{n}\,\kappa_A\,\kappa_B\,\sigma^{-1}}$$

as was to be proved. q.e.d.

Combining this with Lemma 5 yields

Theorem 13. *Let* **S** *be the set of* (A, B) *where no nonsingular matrix* X *of eigenvectors exists. Let* B *of full* rank, *no* λ_i *be an eigenvalue of* A, *and* κ_A *and* κ_B *be defined as in Lemma* 6. *Then*

$$\operatorname{dist}((A, B), \mathbf{S}) \geq \frac{0.187}{\sqrt{n}\,\kappa_A\,\kappa_B\,\kappa(X)}.$$

Let **U** *be the set of uncontrollable pairs* (A, B). *Then we also have*

$$\operatorname{dist}((A, B), \mathbf{U}) \geq \frac{0.187}{\sqrt{n}\,\kappa_A\,\kappa_B\,\kappa(X)}.$$

Proof. From Lemma 5 we have $\kappa(X) \geq \sigma_{\min}^{-1}(S)$, so the first claim follows immediately from Lemma 6. The second claim follows since the set of uncontrollable problems **U** is contained in **S**. q.e.d.

We can also write the second inequality of Theorem 13 as

$$\kappa(X) \geq \frac{0.187}{\sqrt{n}\,\kappa_A\,\kappa_B\,\operatorname{dist}((A, B), \mathbf{U})},$$

implying that the closer (A, B) is to being uncontrollable, the larger the condition number $\kappa(X)$ of the problem. Note that the factors κ_A and κ_B, both at least 1, tend to make the lower bound on $\kappa(X)$ smaller. The reason for this is as follows. If $\kappa(B)$ is large, a very small perturbation of B can change the space $\mathbf{R}(B)$ spanned by its columns greatly, in particular in such a way that the pole assignment problem becomes quite easy. Therefore we cannot guarantee that $\kappa(X)$ will be bad in this case. Similarly, if $\kappa(A)$ is large, some λ_i is nearly an eigenvalue of A. Thus, even if (A, B) is nearly uncontrollable, only a small perturbation may be needed to put a pole at λ_i. In the extreme case when $\{\lambda_i\}$ is the spectrum of A, $F=0$ solves the pole assignment problem even if (A, B) is exactly uncontrollable and so $\kappa(X)$ depends only on how hard it is to diagonalize A. A similar result to Theorem 13 was proven in [3] using more ad hoc techniques.

7. Interpretation of the Differential Inequalities

In this section we provide a numerical interpretation of the differential inequalities

$$m \cdot \kappa^2 \leq \|D\kappa\| \leq M \cdot \kappa^2 \tag{7.1}$$

stated in the introduction. We will use the relative condition number

$$\kappa_{\mathrm{rel}}(g, x) = \frac{\|Dg(x)\| \cdot \|x\|}{\|g(x)\|},$$

this formula holding only if g is differentiable.

As in the introduction, it is easy to see that by multiplying inequalities (7.1) by $\|x\|/\kappa(x)$, we get

$$m \cdot \kappa(x) \leq \kappa_{\text{rel}}(\kappa, x) \leq M \cdot \kappa(x) \quad \text{if } \|x\| = 1. \tag{7.2}$$

Inequalities (7.2) are equivalent to the statement: *solving the problem x, normalized so $\|x\|=1$, is essentially just as hard (within factors m and M) as computing the condition number κ of the problem x.*

If we further assume that the condition number κ of Eq. (1) is *homogeneous of degree k*, i.e. $\kappa(\alpha x) = \alpha^k \kappa(x)$ for any real positive α, then for $\kappa = 0$ or $\kappa = -1$ we will show that (7.2) is essentially equivalent to (7.1). All the condition numbers κ considered in this paper are homogeneous, either with $k = -1$ (for matrix inversion, eigenvectors, and polynomial zeros) or $k = 0$ (for eigenvalues and pole placement). κ is homogeneous in these cases because the problems themselves (i.e., the mapping from problem to solution) whose conditioning κ measures are homogeneous.

The main point of this paper has been that if (7.1), or equivalently (7.2), holds, then the distance d from the problem x to the nearest point in the set of ill-posed problems P is bounded by

$$\frac{1}{M\kappa(x)} \leq d \leq \frac{1}{m\kappa(x)} \tag{7.3}$$

(how d is measured depends on the norm $\|\cdot\|$ and whether the degree k of homogeneity is -1 or 0). Conversely, we will see that if we define $\kappa(x)$ to be $1/d$ then this $\kappa(x)$ satisfies the differential inequalities (7.1) and (7.2) with $m = M = 1$.

The near equivalence of (7.1), (7.2) and (7.3) is very satisfying, because it says that if the condition number κ has the utterly reasonable property of being just as hard to compute as the solution x itself, then it has the attractive geometric property of being 1 over the distance to the nearest infinitely ill-conditioned problem. Indeed, the common formulas for relative condition numbers (e.g., $\|A\| \cdot \|A^{-1}\|$ for matrix inversion) lead one to believe that one must solve the problem (e.g., compute A^{-1}) to within reasonable accuracy to get a reasonably accurate condition number. This intuition is corroborated by these theorems.

To state our results, we will need to measures of distance. If $\|\cdot\|_2$ is Euclidean distance, define

$$\operatorname{dist}_2(x, P) \equiv \inf_{y \in P} \|x - y\|_2,$$

where P is the set of ill-posed problems. Assume the set P is homogeneous, i.e. $x \in P$ implies $\alpha x \in P$ for all scalars α. This will be true if κ is homogeneous. Let $d_{GC}(a, b)$ denote shortest distance along a great circle between two points a and b on the unit sphere. Define the "great circle" distance between x and P

$$\operatorname{dist}_{GC}(x, P) \equiv \inf_{y \in P} d_{GC}\left(\frac{x}{\|x\|_2}, \frac{y}{\|y\|_2}\right).$$

For the first theorem we assume that $\kappa(x)$ is homogeneous of degree -1. This is the case for $\kappa(A) = \|A^{-1}\|$ (matrix inversion), $\kappa(A) = \|S\| \cdot \|P\|$, S a reduced resolvent and P a projector (eigenvectors), $\kappa(T)=\|P\|/\mathrm{sep}(A,B)$ (eigenvectors), and $\kappa(p)=1/|p'(x)|$, p a polynomial with zero x (polynomial zero finding). Then we have

Theorem 14. *Let a problem x have condition number $\kappa(x)>0$, where κ is homogeneous of degree -1 and has a continuous Fréchet derivative $D\kappa(x)$ almost everywhere it is finite. Let P denote the set of x where κ is infinite. Then* (7.4a) *and* (7.4b) *below are equivalent wherever $D\kappa$ exists:*

$$\exists 0<m\le M \quad \text{such that } m\cdot\kappa^2(x)\le\|D\kappa(x)\|\le M\cdot\kappa^2(x) \tag{7.4a}$$

$$\exists 0<m\le M \text{ such that } m\cdot\kappa(x)\le\kappa_{\mathrm{rel}}(\kappa,x)\le M\cdot\kappa(x) \quad \text{for all } \|x\|=1. \tag{7.4b}$$

In particular, when $D\kappa$ is continuously differentiable and $\|\cdot\|=\|\cdot\|_2$, the following three conditions are equivalent:

$$\kappa^2(x)=\|D\kappa(x)\|_2, \tag{7.5a}$$

$$\kappa(x)=\kappa_{\mathrm{rel}}(\kappa,x) \quad \text{for all } \|x\|_2=1, \tag{7.5b}$$

$$\frac{1}{\kappa(x)}=\mathrm{dist}_2(x,P). \tag{7.5c}$$

Proof. (7.4a) can be converted into (7.4b) by multiplying through by $\|x\|/\kappa(x)$ and taking $\|x\|=1$. Given (7.4b), (7.4a) can be derived by substituting $x/\|x\|$ for x, yielding inequalities true for all nonzero x, and using the fact that if κ is homogeneous of degree -1, $D\kappa(x)$ is homogeneous of degree -2. The equivalence of (7.5a) and (7.5b) follows from taking $m=M=1$ in (7.4a) and (7.4b). To show they imply (7.5c), use the arguments following Lemmas 1 and 2 in Sect. 2. To show (7.5c) implies (7.5a) and (7.5b), just differentiate. q.e.d.

For the second theorem assume that $\kappa(x)$ is homogeneous of degree 0. This is the case for $\kappa(A)=(\|P\|^2-1)^{1/2}$, P a projector (eigenvalues), and $\kappa(A,B,\Lambda)=\|X\|\cdot\|X^{-1}\|$, (A,B) a control system to be assigned the poles Λ via state feedback F: $A+BF=X\Lambda X^{-1}$.

Theorem 15. *Let a problem x have condition number $\kappa(x)>0$, where κ is homogeneous of degree 0 and has a continuous Fréchet derivative $D\kappa(x)$ almost everywhere it is finite. Let P denote the set of x where κ is infinite. Then* (7.6a) *and* (7.6b) *below are equivalent wherever $D\kappa$ exists:*

$$\exists 0<m\le M \quad \text{such that } m\cdot\kappa^2(x)\le\|D\kappa(x)\|\le M\cdot\kappa^2(x) \quad \text{for all } \|x\|=1, \tag{7.6a}$$

$$\exists 0<m\le M \quad \text{such that } m\cdot\kappa(x)\le\kappa_{\mathrm{rel}}(\kappa,x)\le M\cdot\kappa(x). \tag{7.6b}$$

In particular, if $D\kappa$ is continuously differentiable and $\|\cdot\|=\|\cdot\|_2$, the following three conditions are equivalent:

$$\kappa^2(x)=\|D\kappa(x)\|_2 \quad \text{for all } \|x\|_2=1, \tag{7.7a}$$

$$\kappa(x)=\kappa_{\mathrm{rel}}(\kappa,x), \tag{7.7b}$$

$$\frac{1}{\kappa(x)}=\mathrm{dist}_{GC}\left(\frac{x}{\|x\|_2},P\right). \tag{7.7c}$$

Proof. (7.6a) can be converted into (7.6b) by substituting $x/\|x\|$ for x, yielding inequalities true for all $x \neq 0$, and using the fact that if κ is homogeneous of degree 0, $\|D\kappa\|$ is homogeneous of degree -1. Given (7.6b), (7.6a) can be derived by multiplying through by $\kappa(x)$ and taking $\|x\|=1$. The equivalence of (7.7a) and (7.7b) follows from setting $m=M=1$ in (7.6a) and (7.6b). To derive (7.7c) from (7.7a, b) the argument is similar to that of the last theorem. The only difference is that since κ is homogeneous of degree 0, $D\kappa$ is orthogonal to x by Euler's theorem for homogeneous functions. Therefore integrating the vector field defined by $D\kappa$ yields a curve lying on a sphere of constant radius. This is why we deal with shortest paths along great circles between points of unit norm. To show (7.7c) implies (7.7a) and (7.7b), just differentiate. q.e.d.

8. Connections with Newton's Method

In this section we show that in case the function f which maps problems to solutions is scalar, the differential inequality (1.1) underlying our approach is nothing more than a restatement of Newton's method. Thus inequality (1.1), far from holding only coincidentally for the special problems considered here, actually holds locally (in a sufficiently small neighborhood of the set of ill-posed problems) for a quite general class of problems. When f maps between spaces of the same dimension greater than one, the relationship with Newton's method weakens but still holds to some extent.

In the scalar case, we let f be a smooth function from the real numbers to the real numbers (or the complex numbers to the complex numbers), which we take as the solution to some problem (such as "evaluate f"). As the condition number of the problem, we can take in principle any multiple of the derivative of f, but the one which will turn out to satisfy the differential inequalities (1.1) is the absolute condition number $\kappa_{\text{abs}}(x) \equiv |f'(x)/f(x)|$, the instantaneous relative change in the output per *absolute* change in the input. (We want to measure the absolute distance from the input to the nearst ill-posed input, so this condition number turns out to work instead of $\kappa_{\text{rel}}(x)=|f'(x)\,x/f(x)|$, the instantaneous relative change in the output per *relative* change in the input.) The set of ill-posed problems is the set of x where $\kappa(x)=f'(x)/f(x)$ is infinite. Since f is smooth, this is (in general) the set of zeros of f. Following the paradigm used so far, if $|\kappa'(x)|$ is close to $\kappa^2(x)$, then $1/|\kappa(x)|$ should be a good approximation to the distance to the nearest ill-posed problem, i.e. zero of f. Computing

$$\kappa'(x)=\frac{f''(x)\,f(x)-(f'(x))^2}{(f(x))^2} \tag{8.1}$$

we see that if x is close to a simple zero of f so that $f(x)$ is small, then

$$|\kappa'(x)| \approx \left|\frac{(f'(x))^2}{(f(x))^2}\right| = (\kappa(x))^2$$

as required by the paradigm. Thus, an even better approximation to a zero of f should be

$$x - \frac{1}{\kappa(x)} = x - \frac{f(x)}{f'(x)}$$

which is Newton's method.

If f is smooth except for poles, then these poles are also ill-posed problems. In this case, just consider $g = 1/f$, so the poles of f are zeros of g. The condition number $|g'/g|$ of g is identical to the condition number $|f'/f|$ of f, so again we get Newton's method, but now it converges to a pole instead of a zero.

Examining a little more closely the condition that $\kappa'(x) \approx (\kappa(x))^2$, we see that for this to be true $(f'(x))^2$ has to dominate $f(x) f''(x)$ in the numerator of (8.1) above, or at the very least $|f(x) f''(x)| < (f'(x))^2$. But this is just the condition that the Newton iteration contracts. For letting $g(x) = x - f(x)/f'(x)$ be the Newton iteration, we easily compute

$$g'(x) = \frac{f(x) f''(x)}{f'(x) f'(x)}.$$

What, if instead of applying Newton to find a zero of f, we apply Newton to find a zero of $1/\kappa(x)$? The formula is easily seen to be

$$x - \frac{1/\kappa(x)}{(1/\kappa(x))'} = x - \frac{f(x)}{f'(x)\left(1 - \dfrac{f(x) f''(x)}{(f'(x))^2}\right)}$$

which under the same conditions as above ($f(x)$ small so that $|f(x) f''(x)/(f'(x))^2| < 1$) is asymptotically the same as Newton applied to $f(x)$.

Thus, the property of condition number being the reciprocal of distance to the nearest ill-posed problem is quite universal, holding locally for all simple zeros and poles. For multiple zeros, it is easy to see you get a factor of the multiplicity of the zero/pole in the distance estimate from the paradigm, and this is just the usual modification to Newton for quadratic convergence to a multiple zero/pole.

The situation is not quite so simple when f maps between higher dimensional spaces of equal dimension. As we will see, the steepest ascent direction for κ is asymptotically $(Df)^T \cdot f$, whereas the Newton direction is $(Df)^{-1} f$. These two directions can be quite different, especially when Df is ill-conditioned. Nonetheless, we will show that asymptotically the paradigm supplies upper and lower bounds on the norm of the Newton correction.

In order to make the calculations easier, we use the 2-norm for vectors and Frobenius norm for matrices. Letting $\kappa(x) \equiv \|Df(x)\| / \|f(x)\|$, we compute

$$D\kappa = \frac{\left[\operatorname{tr}(Df)^T \cdot \dfrac{\partial}{\partial x_i} Df\right]}{\|f\| \cdot \|Df\|} - \frac{\|Df\|^2}{\|f\|^2} \frac{(Df)^T}{\|Df\|} \frac{f}{\|f\|}.$$

In the neighborhood of a zero of f, the second term dominates the first, yielding

$$\|D\kappa\| \approx \kappa^2 \frac{\|(Df)^T f\|}{\|Df\| \, \|f\|}.$$

It is easy to see that

$$\frac{\kappa^2}{\|Df\|\,\|Df^{-1}\|} \le \kappa^2 \frac{\|(Df)^T f\|}{\|Df\|\,\|f\|} \le \kappa^2$$

so that according to our paradigm, provided x is close enough to a zero of f,

$$\frac{1}{\kappa} = \frac{\|f\|}{\|Df\|} \le \operatorname{dist}(x, P) \le \|(Df)^{-1}\|\ \|(Df)^{-1}\|\ \|f\| = \frac{\|Df\|\ \|(Df)^{-1}\|}{\kappa}.$$

The norm of the Newton correction is bracketed by these bounds as expected:

$$\frac{\|f\|}{\|Df\|} \le \|(Df)^{-1} f\| \le (Df)^{-1}\|\ \|f\|.$$

Thus, the reciprocal of the condition number provides an asymptotic lower bound on the distance, and can underestimate the distance by a factor of at most $\|Df\|\ \|(Df)^{-1}\|$.

9. Algebraic Functions

In this section we show that in a neighborhood of a branch point of *any* algebraic function, we expect the distance to the branch point to be at least a multiple of the square of the reciprocal of the condition number. As in the last section, for utterly general functions the most we can prove is that this relationship holds in a neighborhood of a branch point, not globally. That such a relationship holds globally for eigenproblems and polynomial zero finding depends on exploiting the special structure of these problems.

We define an algebraic function as a root λ of the following equation:

$$0 = \sum_{i=0}^{n} p_i(x)\,\lambda^i \equiv P(p_i). \tag{8.1}$$

Here n must be at least two for there to be a branch. $p_i(x)$ is a scalar function of the vector variable x. By $P(z_i)$ we mean $\sum_{i=0}^{n} z_i \lambda^i$ where z_i is any subscripted quantity (scalar, vector, or matrix). Analogously, we let $P'(z_i)$ denote $\sum_{i=1}^{n} z_i\, i \lambda^{i-1}$. If λ is a particular simple root of (8.1), then we can compute the derivative $D\lambda$ of λ with respect to x by linearizing as follows:

$$\begin{aligned} 0 &= \sum_{i=0}^{n} (\lambda + \delta\lambda)^i\, p_i(x + \delta x) \\ &= P(p_i) + P'(p_i)\,\delta\lambda + P(Dp_i)\,\delta x + O(\|\delta x\|^2). \end{aligned}$$

Using the fact that $P(p_i) = 0$ we solve for $\delta\lambda$ and get

$$\delta\lambda = \frac{-P(Dp_i)}{P'(p_i)}\,\delta x$$

or

$$D\lambda = \frac{-P(Dp_i)}{P'(p_i)}$$

leading us to define

$$\kappa(x, \lambda) \equiv \|D\lambda\| = \frac{\|P(Dp_i)\|}{|P'(p_i)|}.$$

This condition number can be infinite only when $P'(p_i)=0$. When $p_i(x)=x_i$, i.e. we have a simple polynomial with coefficients x_i (the situation analyzed in Sect. 5), this condition $P'(p_i)=0$ is the usual condition for a double root. In x_i space, the set of points where $P'(p_i)$ vanishes forms a branch surface rather than branch point. The important thing to notice is that since the p_i are smooth, $\kappa(x, \lambda)$ is essentially a multiple of $1/|P'(p_i)|$ in a neighborhood of a branch surface (barring accidental cancellation in the numerator of $\kappa(x, \lambda)$.

To compute $D\kappa$ we proceed as above by linearing $\kappa(x+\delta x, \lambda+\delta\lambda)$. The result of this rather tedious calculation is

$$\begin{aligned}\|D\kappa\| = \Bigg\| & \frac{(P(Dp_i))^*}{\|P(Dp_i)\|}\,\frac{P(DDp_i)}{|P'(p_i)|} \\ & - \frac{(P(Dp_i))^*}{\|P(Dp_i)\|}\,\frac{(P'(Dp_i))^T P(Dp_i)}{|P'(p_i)|\,P'(p_i)} - \frac{\|P(Dp_i)\|}{|P'(p_i)|}\,\frac{P'(Dp_i)}{P'(p_i)} \\ & + \frac{\|P(Dp_i)\|}{|P'(p_i)|}\,\frac{P''(p_i)\,P(Dp_i)\,P(Dp_i)}{P'(p_i)\,P'(p_i)}\Bigg\|.\end{aligned}$$

As with κ itself, the only factors which contribute to $\|D\kappa\|$ going to infinity are the $P'(p_i)$ in the denominators. The first term has one, the second two terms have two, and the last term has three factors of $P'(p_i)$ in the denominator. Thus we expect $\|D\kappa\|$ to grow to infinity no faster than the third power of κ, and barring accidental cancellation in the numerator of the third term, it will grow this fast. Applying Lemma 2 in Sect. 2 with $\beta=3$, we see that a lower bound on the distance to the nearest ill-posed problem will be proportional to the reciprocal of the square of the condition number.

Looking back to the Sect. 4 and 5 on eigenvalue problems and polynomial zero finding we see this square dependence exhibited. In Sect. 5 it was explicit in Theorem 12 and its Corollary 1, which supplied lower bounds on the distance to the nearest polynomial with multiple roots as a multiple of the square of the condition number. In Sect. 4 one must look somewhat closer. The condition number for an eigenvalue λ was computed to be the norm of its associated projector $\|P\|$. A lower bound on the distance was computed to be proportional to sep$/\|P\|$ (sep and $\|P\|$ are defined in Sect. 4). At first glance, the lower bound does not appear to be the reciprocal of the square of the condition number, no matter how big $\|P\|$ is. In this case, the sufficiently small neighborhood of the set of ill-posed problems where this behavior occurs is the set where $\|P\|$ assumes its maximum value, which is proportional to 1/sep. In this neighborhood the condition number behaves like 1/sep and the lower bound like sep^2, as expected.

Finally, note that we did not need to assume that the p_i were actually polynomial functions of x, just that they were smooth. Thus the results of this section apply to a larger class of problems than just algebraic functions.

10. Extensions

In [9] Kahan relates the condition number and the distance to the nearest ill-posed problem for several problems of numerical analysis. In that paper he also observed that if one had an ill-posed problem, then by restricting it to lie within the set P of ill-posed problems it often became well-posed again in the sense that if x and $x+\delta x$ were both in P and δx were small, the difference between the solution at x and the solution at $x+\delta x$ would also be small. It would remain well-posed until it approached a subset P_1 of P, where it again become ill-posed. Restricted to lie within P_1, however, it again became well-posed until it approached a further subset P_2, and so on.

For example, computing the pseudo-inverse T^+ of a matrix T is equivalent to matrix inversion for square, nonsingular matrices, in which case the relative condition number of T for pseudo-inversion can be written as σ_1/σ_n, where $\sigma_1 \geq \ldots \geq \sigma_n$ are the singular values of T. The distance to the set P_1 of singular matrices in the $\|\cdot\|_F$ norm is σ_n, and as it approaches zero, the condition number σ_1/σ_n approaches infinity. If σ_n is exactly zero, the pseudo-inverse is well defined, and as long as T is restricted to have rank $n-1$ ($\sigma_n=0$, $\sigma_{n-1}\neq 0$), the condition number of its pseudo-inverse is σ_1/σ_{n-1}, where σ_{n-1} is the distance to the nearest matrix of rank $n-2$. So as T approaches the subset P_2 of rank $n-2$ matrices, its condition number again becomes infinite. If T is restricted to have rank $n-2$, its condition number becomes σ_1/σ_{n-3}, which remains finite until T approaches the set P_3 of matrices of rank $n-3$, and so on.

In the course of establishing the results of the last paragraph, Kahan [9] also showed that if $\kappa(T)=\|T^+\|$, then

$$\limsup_{\substack{\delta T\to 0\\ \operatorname{rank}(T)=\operatorname{rank}(T+\delta T)}} \frac{\kappa(T+\delta T)-\kappa(T)}{\|\delta T\|}=\kappa^2(T) \tag{10.1}$$

which is analogous to our differential inequalities (1.1). Using (10.1) as we did (1.1), one can find another proof that the distance within the set of matrices T of constant rank k to the nearest one of rank $k-1$ is $1/\|T^+\|=\sigma_k$.

The hierarchy of sets of ill-posed problems P_i plays an important role in numerical analysis because an ill-posed problem can often be regularized by restricting it to lie in a set P_i. For example, when solving rank deficient least squares problems one often regularizes by artificially forcing the smallest singular values to zero, thus solving a problem forced to lie in a set P_i [5]. Similarly, when computing eigenvalues and eigenvectors one often computes the invariant subspace belonging to a cluster of eigenvalues because it can be much better conditioned than the individual eigenvectors which span it. This is an important technique when computing functions of matrices [12].

In a future paper we hope to extend the techniques of this paper to estimating distances to hierarchical subsets P_i of other ill-posed problems. This should lend further geometrical and numerical insight into the solution of ill-posed problems.

Acknowledgements. Jim Wilkinson has provided many incisive comments on an earlier version of this paper. Gene Golub provided both financial support and a Leasant working environment for several weeks during summer 1985. The author has also received support from the National Science Foundation (grant number 8501708) and an IBM Faculty Development Award.

11. References

1. Demmel, J.: The condition number of equivalence transformations which block diagonalize matrix pencils. SIAM J. Numer. Anal. **20**, 599–610 (1983)
2. Demmel, J.: Computing stable eigendecompositions of matrices. Linear Algebra Appl. **79**: 163–193 (1986)
3. Demmel, J.: On the conditioning of pole assignment. Computer Science Dept. Report # 150, Courant Institute of Mathematical Sciences. New York, Jan. 1985
4. Eckart, C., Young, G.: A principal axis transformation for non-hermitian matrices. Bull. Am. Math. Soc., New Ser. **45**, 118–121 (1939)
5. Golub, G., Van Loan, C.: Matric computations. Baltimore: Johns Hopkins Press 1983
6. Hartman, P.: Ordinary differential equations. New York: John Wiley 1973
7. Hough, D.: Explaining and ameliorating the ill condition of zeros of polynomials. Thesis, Mathematics Department, University of California, Berkeley, CA 1977
8. Kahan, W.: Numerical linear algebra. Can. Math. Bull. **9**, 757–801 (1966)
9. Kahan, W.: Conserving confluence curbs ill-condition. Computer Science Dept. Report, University of California, Berkeley 1972
10. Kato, T.: Perturbation theorey for linear operators. Berlin, Heidelberg, New York: Springer 1966
11. Kautsky, J., Nichols, N., Van Dooren, P.: Robust pole assignment in linear state feedback. Int. J. Control **41**, 1129–1155 (1985)
12. Parlett, B.N., Ng, K.C.: Development of an accurate algorithm for $\exp(B\tau)$. Report PAM-294, Center for Pure and Applied Mathematics, University of California, Berkeley, August 1985
13. Ruhe, A.: Properties of a matrix with a very ill-conditioned eigenproblem. Numer. Math. **15**, 57–60 (1970)
14. Stewart, G.W.: Error bounds for approximate invariant subspaces of closed linear operators. SIAM J. Numer. Anal. **8**, 796–808 (1971)
15. Wilkinson, J.H.: The algebraic eigenvalue problem. Oxford: Clarendon Press 1965
16. Wilkinson, J.H.: Note on matrices with a very ill-conditioned eigenproblem. Numer. Math. **19**, 176–178 (1972)
17. Wilkinson, J.H.: On neighboring matrices with quadratic elementary divisors. Numer. Math. **44**, 1–21 (1984)
18. Wilkinson, J.H.: Sensitivity of eigenvalues. Util. Math., **25**, 5–76
19. Wonham, W.M.: Linear multivariable control.: A geometric approach. 2nd edition. Berlin, Heidelberg, New York: Springer 1979

Received November 11, 1986