JOURNAL OF COMPLEXITY 3, 201–229 (1987)

The Geometry of Ill-Conditioning

JAMES W. DEMMEL

Department of Computer Science, Courant Institute, New York, New York 10012

Numerous problems in numerical analysis, including matrix inversion, eigenvalue calculations, and polynomial zero finding, share the following property: the difficulty of solving a given problem is large when the distance from that problem to the nearest "ill-posed" one is small. For example, the closer a matrix is to the set of noninvertible matrices, the larger its condition number with respect to inversion. We show that the sets of ill-posed problems for matrix inversion, eigenproblems, and polynomial zero finding all have a common algebraic and geometric structure which lets us compute the probability distribution of the distance from a "random" problem to the set. From this probability distribution we derive, for example, the distribution of the condition number of a random matrix. We examine the relevance of this theory to the analysis and construction of numerical algorithms destined to be run in finite precision arithmetic. © 1987 Academic Press, Inc.

1. INTRODUCTION

To investigate the probability that a numerical analysis problem is difficult, we need to do three things:

(1) Choose a measure of difficulty,

(2) Choose a probability distribution on the set of problems,

(3) Compute the distribution of the measure of difficulty induced by the distribution on the set of problems.

The measure of difficulty we shall use in this paper is the *condition number*, which measures the sensitivity of the solution to small changes in the problem. For the problems we consider in this paper (matrix inversion, polynomial zero finding, and eigenvalue calculation), there are well-known condition numbers in the literature of which we shall use slightly modified versions to be discussed more fully later. The condition number is an appropriate measure of difficulty because it can be used to measure the expected loss of accuracy in the computed solution, or even the

0885-064X/87 $3.00

number of iterations required for an iterative algorithm to converge to a solution.

The probability distribution on the set of problems for which we will attain most of our results will be the "uniform distribution" which we define as follows. We will identify each problem as a point in either $\mathbf{R}^N$ (if it is real) or $\mathbf{C}^N$ (if it is complex). For example, a real n by n matrix A will be considered to be a point in $\mathbf{R}^{n^2}$, where each entry of A forms a coordinate in $\mathbf{R}^{n^2}$ in the natural way. Similarly, a complex nth degree polynomial can be identified with a point in $\mathbf{C}^{n+1}$ by using its coefficients as coordinates. On the space $\mathbf{R}^N$ (or $\mathbf{C}^N$) we will take any spherically symmetric distribution; i.e., the induced distribution of the normalized problem $x/\|x\|$ ($\|\cdot\|$ is the Euclidean norm) must be uniform on the unit sphere in $\mathbf{R}^N$. For example, we could take a uniform distribution on the interior of the unit ball in $\mathbf{R}^N$, or let each component be an independent Gaussian random variable with mean 0 and standard deviation 1. Our answers will hold for this entire class of distributions because our condition numbers are homogeneous (multiplying a problem by a nonzero scalar does not change its condition number).

The main justification for using a uniform distribution is that it appears to be fair: each problem is as likely as any other. However, it does not appear to apply in practice for a variety of reasons, including the fact that any set of problems which can be represented in a computer is necessarily discrete rather than continuous. We will discuss the validity of our choice of uniform distribution as well as alternatives at length in Section 6.

Finally, given this distribution, we must compute the induced probability distribution of the condition number. It turns out that all the problems we consider here have a common geometric structure which lets us compute the distributions of their condition numbers with a single analysis, which goes as follows:

(i) Certain problems of each kind are *ill-posed*, i.e., their condition number is infinite. These ill-posed problems form an algebraic variety within the space of all problems. For example, the singular matrices are ill-posed with respect to the problem of inversion, and they lie on the variety where the determinant, a polynomial in the matrix entries, is zero. Geometrically, varieties are possibly self-intersecting surfaces in the space of problems.

(ii) The condition number of a problem has a simple geometric interpretation: it is proportional to (or bounded by a multiple of) the reciprocal of the distance to the set of ill-posed problems. Thus, as a problem gets closer to the set of ill-posed ones, its condition number approaches infinity. In the case of matrix inversion, for example, the traditional condition number is exactly inversely proportional to the distance to the nearest singular matrix.

(iii) The last observation implies that the set of problems of condition number at least x is (approximately) the set of problems within distance c/x (c a constant) of the variety of ill-posed sets. Sets of this sort, called *tubular neighborhoods*, have been studied extensively by geometers. We will present upper bounds, lower bounds, and asymptotic values for the volumes of such sets. The asymptotic results, lower bounds, and some of the upper bounds are new. The formulas are very simple, depending only on x, the degree of N of the ambient space, the dimension of the variety, and the degree of the variety. These volume bounds in turn bound the volume of the set of problems with condition number at least x. Since we are assuming the problems are uniformly distributed, volume is proportional to probability.

Thus, for example, we will prove that a scaled version $\kappa(A) \equiv \|A\|_F \cdot \|A^{-1}\|$ of the usual condition number of a complex matrix with respect to inversion satisfies

$$\frac{(1 - x^{-1})^{2n^2-2}}{4n^4x^2} \leq \text{Prob}(\kappa(A) \geq x) \leq \frac{e^2n^5(1 + n^2/x)^{2n^2-2}}{x^2},$$

and that asymptotically

$$\text{Prob}(\kappa(A) \geq x) = \frac{n(n^2 - 1)}{x^2} + o\left(\frac{1}{x^2}\right).$$

In other words, the probability that the condition number exceeds x decreases as the square of the reciprocal of x. Even for moderate x the upper bound exceeds the asymptotic limit by a ratio of only about e^2n^2. If A is real we will show

$$\frac{C(1 - 1/x)^{n^2-1}}{x} \leq \text{Prob}(\kappa(A) \geq x) \leq \sum_{k=1}^{n^2} 2 \cdot \binom{n^2}{k} \cdot \left(\frac{2n}{x}\right)^k,$$

where C is a constant proportional to the $(n^2 - 1)$-dimensional volume of the set of singular matrices inside the unit ball. Thus, for real matrices the probability that the condition number exceeds x decreases as x^{-1}.

There are a number of open questions and conjectures concerning these volume bounds, in particular for how general a class of real varieties they apply (the case of complex varieties is simpler). We will discuss the history of this work and open problems in detail in Section 4.

It turns out that the reciprocal relationship between condition number and distance to the nearest ill-posed problem holds for a much wider class of problem than just matrix inversion, polynomial zero finding, and eigenvalue calculations: it is shared, at least asymptotically, by any problem

whose solution is an algebraic function. For simplicity we shall restrict ourselves to the three aforementioned problems, but our results do apply more widely, as discussed in Section 3 and Demmel (1986).

This work was inspired by earlier work in a number of fields. Demmel (1986), Gastinel (1966), Hough (1977), Kahan (1972), Ruhe (1970), Stewart (1973), Wilkinson (1972, 1984a,b) and others have analyzed the relationship between the condition number and the distance to the nearest ill-posed problem mentioned above in (ii). Gray (1982a, 1982b), Griffiths (1978), Hotelling (1939), Lelong (1968), Ocneau (1985), Renegar (1987), Santalo (1976), Smale (1981), and Weyl (1939) have worked on bounds of volumes of tubular neighborhoods. These volume bounds have been used by Smale (1981, 1986), Renegar (1987), and others to analyze the efficiency of Newton's method for finding zeros of polynomials. This latter work inspired the author (Demmel, 1983) to apply these bounds to conditioning. Ocneanu (to appear) and Kostlan (1985) have also analyzed the statistical properties of the condition number for matrix inversion.

The rest of this paper is organized as follows. Section 2 defines notation. Section 3 discusses the relationship between conditioning and the distance to the nearest ill-posed problem. Section 4 presents the bounds on the volumes of tubular neighborhoods we shall use and states some related open problems. Section 5 computes the distributions of the condition numbers of our three problems. Section 6 discusses the limitations of assuming a uniform distribution and suggests alternatives and open problems.

2. Notation

We introduce several ideas we will need from numerical analysis, algebra, and geometry. $\|x\|$ will denote the Euclidean norm of the vector x as well as the induced matrix norm

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

$\|A\|_F$ will denote the Frobenius norm

$$\|A\|_F \equiv \left(\sum_{ij} |A_{ij}|^2 \right)^{1/2}.$$

If P is a set and x is a point, we will let $\operatorname{dist}(x, P)$ denote the Euclidean distance from x to the nearest point in P.

A subset M of $\mathbf{R}^N$ is called an *n-dimensional manifold* if it is locally homeomorphic to $\mathbf{R}^N$. We also write $n = \dim(M)$. The *codimension* of M,

written codim(M), is $N - n$. In this paper dimension will always refer to the *real* dimension rather than the complex dimension, which is half the real dimension.

A *variety* is the set of solutions of a system of polynomial equations. A variety is *homogeneous* if it is cone-shaped; i.e., if x is in the variety so is every scalar multiple αx. A variety is not generally a manifold since it can have *singularities* in the neighborhood of which it is not homeomorphic to Euclidean space. However, points q with relatively open neighborhoods $U_q \subset P$ that are manifolds are dense in P (Kendig, 1977, Theorem 4.2.4) so that the following definition makes sense: the *dimension of P at p*, written $\dim_p(P)$, is

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

We in turn define the *dimension of the variety P* as the maximum over all $p \in P$ of $\dim_p(P)$. If $\dim_p(P)$ is constant for all p, we call P *pure dimensional*. A complex variety defined by a single nonconstant polynomial is called a *complex hypersurface*. Complex hypersurfaces are pure dimensional with codimension 2. A *real hypersurface* has codimension 1 everywhere. The real variety defined by the polynomials $f_1, \ldots, f_p$ is called a *complete intersection* if it is pure dimensional of codimension p.

Now we define the degree of a purely n-dimensional variety P in $\mathbf{R}^N$. Let L^{N-n} be a $(N - n)$-dimensional linear manifold (plane) in $\mathbf{R}^N$. Since $\dim(L^{N-n}) + \dim(P) = \dim(\mathbf{R}^N) = N$ we say that L^{N-n} and $\mathbf{R}^N$ are of *complementary dimension*. Generically, L^{N-n} and P will intersect in a surface of codimension equal to the sum of their codimensions, that is, N. In other words, their intersection will be of dimension 0 (a finite collection of points). If P is a complex homogeneous variety, then for almost all planes L^{N-n} this collection will contain the same number of points, and this common number is called the *degree* of P, and is written deg(P) (see Kendig, 1977, Theorem 4.6.2). Intuitively, deg(P) gives the number of "leaves" of the variety P that a typical plane L^{N-n} will intersect. In the case of a nonhomogeneous or real variety P, deg(P) is defined analogously as the maximum (finite) intersection number of a plane L^{N-n} and the n-dimensional variety P in $\mathbf{R}^N$, although the intersection number will not generally be constant for almost all L^{N-n}.

This concept of degree is a generalization of the degree of a polynomial. Indeed, if P is complex and defined as the solution set of a single irreducible polynomial, then the degree of the polynomial equals the degree of P as defined above (Kendig, 1977).

By *l-volume* of an n-dimensional manifold M ($l \geq n$) we mean the

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

3. Condition Numbers and the Distance to the Nearest Ill-Posed Problem

We claim that many classes of numerical analysis problems permit the following geometric characterization of their condition numbers:

(i) Certain problems of each class are *ill-posed*; i.e., their condition numbers are infinite. These problems form a variety within the space of all problems.

(ii) The condition number of a problem has a simple geometric interpretation: it is proportional to (or bounded by a multiple of) the reciprocal of the distance to the set of ill-posed problems. Thus, as a problem gets closer to the set of ill-posed ones, its condition number approaches infinity.

In this section we will cite results from the literature to prove these claims for the following three classes of problems: matrix inversion, polynomial zero finding, and eigenvalue calculation. Afterward we will outline why this characterization applies to many other problems as well (Demmel, 1986).

First we need to define condition number more precisely. If X is our space of problems equipped with norm $\|\cdot\|_X$, Y our space of solutions equipped with norm $\|\cdot\|_Y$, and $f: X \to Y$ is the solution map for our problem, the usual definition of the *relative condition number* is

$$\kappa_{\text{rel}}(f, x) \equiv \lim_{\delta x \to 0} \sup \frac{\|f(x + \delta x) - f(x)\|_Y / \|f(x)\|_Y}{\|\delta x\|_X / \|x\|_X} \tag{3.1}$$

$$= \frac{\|Df(x)\|_{XY} \cdot \|x\|_X}{\|f(x)\|_Y},$$

if the Jacobian Df exists ($\|\cdot\|_{XY}$ is the induced norm). Note that the essential information about the conditioning is contained in the $\|Df\|_{XY}$ factor. We may therefore use a multiple of $\|DF\|_{XY}$ instead of κ_{rel} without losing essential information.

All three of our problems are homogeneous: multiplying the problem by a scalar does not change the condition number. Therefore, the set of ill-posed problems will also be homogeneous, or cone-shaped. This permits us to normalize all our problems to have unit norm (lie on the unit sphere

in either $\mathbf{R}^N$ or $\mathbf{C}^N$), and implies that any results on the distribution of the condition number will hold for any distribution of problems inducing the same distribution of $x/\|x\|$ on the unit sphere. We will also see that for all our problems the set IP of ill-posed problems forms a hypersurface.

Matrix Inversion. The usual relative condition number as defined in (3.1) with the $\|\cdot\|$ norm on both the problem and solution spaces is (Golub and Van Loan, 1983)

$$\kappa_{\text{rel}}(A) \equiv \|A\| \cdot \|A^{-1}\|.$$

We shall use the nearly equivalent condition number

$$\kappa(A) \equiv \|A\|_{\text{F}} \cdot \|A^{-1}\|.$$

These condition numbers are both homogeneous, and infinite when A is singular, so the set of ill-posed problems is a variety defined by the single nth degree homogeneous irreducible polynomial $\det(A) = 0$, where $n = \dim(A)$ (Van der Waerden, 1953). Denote the set of ill-posed problems by IP. From the last section, we see that if A is complex, IP is a complex hypersurface. If A is real, it is easy to verify that IP is still a real hypersurface by using the explicit parameterization provided by Gaussian elimination (Demmel, 1983).

A theorem of Eckart and Young (1936) gives the distance from a nonsingular matrix to IP:

THEOREM 3.1 (Eckart and Young, 1936). $\text{dist}(A, IP) = \|A^{-1}\|^{-1}$. *Therefore, we see that in terms of* κ *we may write*

$$\textit{if } \|A\|_{\text{F}} = 1, \qquad \textit{then } \text{dist}(A, IP) = 1/\kappa(A), \tag{3.2}$$

i.e., that the distance from a normalized problem A to the nearest ill-posed problem is the reciprocal of its condition number.

Polynomial Zero Finding. In this case we are interested in the sensitivity of the zeros of a polynomial to small perturbations in the coefficients. If $p(x)$ is an nth-degree polynomial, let $\|p\|$ denote the Euclidean norm of the vector of its coefficients. If $p(z) = 0$ and δp is a small perturbation of p, it is easy to verify that to first order the perturbed polynomial $p + \delta p$ has a zero at $z + \delta z$, where

$$\delta z = \frac{-\delta p(z)}{p'(z)},$$

implying that the relative condition number is

$$\kappa_{\text{rel}}(p, z) = \frac{n(z) \cdot \|p\|}{|p'(z)|},$$

where

$$n(z) \equiv |z^{-1}| \left(\sum_{i=0}^{n} |z^{2i}| \right)^{1/2}.$$

Note that the condition number depends both on the polynomial p and on the choice of zero z. For simplicity we will use the similar condition number

$$\kappa(p, z) \equiv \frac{\|p\|}{|p'(z)|}.$$

Both condition numbers are infinite when $p'(z) = 0$, i.e., when z is a multiple zero. Thus we will take the set IP of ill-posed problems to be those polynomials with multiple zeros. A necessary and sufficient condition for a polynomial to have a multiple zero is that its discriminant, an irreducible homogeneous polynomial of degree $2n - 2$ in the coefficients of p (Van der Waerden, 1953), be zero. If p is complex, this implies the set of polynomials with zero discriminant is a hypersurface. If p is real, this set of polynomials is still a hypersurface, as may be verified using the parameterization provided by the leading coefficient p_n and the zeros. The discriminant may also be zero if the two leading coefficients of p equal zero (corresponding to a double eigenvalue at ∞), but this set is a subvariety of double the codimension of the hypersurface in which it lies, and so forms a set of measure zero we may neglect.

Now we need to estimate the distance from a given polynomial to one with a multiple zero. The estimate we shall use is due to Hough (1977) and says

THEOREM 3.2 (Hough, 1977; Demmel, 1986). *The distance* $\text{dist}(p, IP)$ *from the polynomial* p *of degree at least 2 to one with a multiple zero is bounded by*

$$\text{dist}(p, IP) \leq \sqrt{2} \cdot |p'(z)|,$$

where $p(z) = 0$.

In fact, this is quite a weak result gotten by estimating the smallest

change in p needed to make a double zero at z, which turns out to be a linear least-squares problem. Thus we may write

$$\text{if } \|p\| = 1, \qquad \text{then } \operatorname{dist}(p, IP) \leq \frac{\sqrt{2}}{\kappa(p, z)}, \tag{3.3}$$

i.e., that the distance from a normalized problem p to the nearest ill-posed problem is bounded by a multiple of the reciprocal of its condition number.

To see how much (3.3) may overestimate $\operatorname{dist}(p, IP)$, we present a lower bound. Note that by changing the argument of p from x to αx, α a scalar, we may make the leading coefficient p_n larger than the other coefficients.

THEOREM 3.3 (Demmel, 1986). *Assume that p is an nth-degree polynomial satisfying $\|p\| = 1$ and $|p_i| < |p_n|/n$ for $i \leq n$. Then*

$$\operatorname{dist}(p, IP) \geq \min_{z_i : p(z_i)=0} \left(\frac{1}{n^2}, \frac{0.0235}{n^2 \cdot \kappa^2(p, z_i)}\right). \tag{3.4}$$

Thus we see that the distance to the nearest ill-posed problem is bounded below essentially by a multiple of the square of the condition number. This is a general phenomenon among algebraic problems to which we shall return below.

Eigenvalue Calculations. We will be interested both in the sensitivity of eigenvalues and eigenvectors. More precisely, we will consider the sensitivity of the projection associated with an eigenvalue (Kato, 1966). If T is a matrix with simple eigenvalue λ, right eigenvector x and left eigenvector, y, the *projection* P associated with λ is the matrix $P = xy^T/y^Tx$. The *reduced resolvent* associated with λ is the matrix

$$S \equiv \lim_{z \to \lambda} (I - P) \cdot (T - z)^{-1}.$$

If T has n distinct eigenvalues λ_i with projections P_i one can write

$$S = \sum_{\lambda_i \neq \lambda} (\lambda_i - \lambda)^{-1} P_i.$$

If δT is a small perturbation of T, one can show that to first order λ changes to $\lambda + \delta\lambda$ and P changes to $P + \delta P$ (Kato, 1966), where

$$\delta\lambda = \operatorname{tr} P\delta T \qquad \text{and} \qquad \delta P = -S\delta TP - P\delta TS.$$

It is easy to verify that $|\delta\lambda|$ can be as large as $\|P\| \cdot \|\delta T\|$ and $\|\delta P\|$ can be at least as large as $\|S\| \cdot \|P\| \cdot \|\delta T\|$ (and no more than twice as large as this). Therefore we may take as condition numbers

$$\kappa(T, \lambda) \equiv \|P\|$$

and

$$\kappa(T, P) \equiv \|P\| \cdot \|S\| \cdot \|T\|_F$$

both of which are homogeneous.

Both condition numbers are infinite when λ is a multiple eigenvalue. Thus we will take the set IP of ill-posed problems to be those matrices with multiple eigenvalues. We may see that IP is a variety as follows. Let $p(T, \lambda)$ be the characteristic polynomial of T. T will have multiple eigenvalues if and only if p has multiple zeros, which happens if and only if the discriminant of p, a homogeneous polynomial of degree $n^2 - n$ in the entries of T, is zero (note that p is monic) (Kendig, 1977; Van der Waerden, 1953). It is not hard to show that this polynomial is irreducible (Demmel, 1983). Thus we see that if T is complex, IP is a hypersurface. If T is real, IP is still a hypersurface (Arnold, 1971).

We now need to relate the above condition numbers of T to the distance from T to IP. A slight restatement of a theorem due to Wilkinson states.

THEOREM 3.4 (Wilkinson, 1972). $\text{dist}(T, IP) \leq \sqrt{2} \cdot \|T\|_F / \|P$.
Therefore, in terms of κ *we may write*

$$\text{if } \|T\|_F = 1, \qquad \text{then } \text{dist}(T, IP) \leq \frac{\sqrt{2}}{\kappa(T, \lambda)}. \tag{3.5}$$

Wilkinson's theorem provides a somewhat weak upper bound on $\text{dist}(T, IP)$. The condition for P on the other hand provides a lower bound on $\text{dist}(T, IP)$:

THEOREM 3.5 (Demmel, 1986). $\text{dist}(T, IP) \geq \|T\|_F / (7 \cdot \kappa(T, P))$.
This result lets us write

$$\text{if } \|T\|_F = 1, \qquad \text{then } \text{dist}(T, IP) \geq \frac{1}{7 \cdot \kappa(T, P)}. \tag{3.6}$$

For somewhat stronger results and discussion, see Demmel (1986).

The phenomenon described above for matrix inversion, polynomial zero finding, and eigenvalue calculation is actually quite common in numerical analysis. It turns out all the above results can be derived from the

same underlying principle, that 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 gradient of κ. The lower bound on $\|D\kappa\|$ implies that the upper bound on $\text{dist}(T, IP)$,

$$\text{dist}(T, IP) \leq \frac{1}{m \cdot \kappa(T)},$$

holds, and the upper bound on $\|D\kappa\|$ implies that the lower bound

$$\text{dist}(T, IP) \geq \frac{1}{M \cdot \kappa(T)}$$

holds. This phenomenon also appears in pole placement in linear control theory, Newton's method, and elsewhere (Demmel, 1986). In the case of algebraic functions, one can show that at least for asymptotically large condition numbers, differential inequalities of the form

$$m \cdot \kappa^2 \leq \|D\kappa\| \leq M \cdot \kappa^3$$

hold, the new upper bound on $\|D\kappa\|$ yielding the lower bound on dist (T, IP)

$$\text{dist}(T, IP) \geq \frac{1}{2M \cdot \kappa^2(T)},$$

which is the source of inequality (3.4) above.

Note also that the set of ill-posed problems is a hypersurface in our three examples above. Other kinds of varieties are possible as well. For example, polynomials with at most m distinct zeros form a subvariety of the variety of polynomials with at least one multiple zero and have codimension $2(n - m)$ (if complex) or $n - m$ (if real) (Demmel, 1983). Since j-tuble zeros are more sensitive than $(j - 1)$-tuple zeros (Wilkinson, 1965) there is a natural hierarchy of sets of ever more ill-posed problems, each one forming a subvariety of the previous set. Similar comments apply to eigenvalue calculations (j-tuple eigenvalues are more sensitive than $(j - 1)$-tuple eigenvalues) and rank-deficient linear least-squares problems (problems with higher rank deficiency are more sensitive than ones with lower rank deficiency). Unfortunately, not all our results on volumes of

tubular neighborhoods of varieties apply as yet to these more general varieties. We will discuss these open problems in the next section.

4. On Volumes of Tubes

In this section we state our main volume estimates. Proofs appear elsewhere (Demmel, 1987).

First we consider complex polynomials. The upper bounds in the following theorem are obtained by generalizing an argument of Renegar (1987), who obtained the upper bound for hypersurfaces:

THEOREM 4.1. *Suppose M is a complex, purely $2d$-dimensional variety in $\mathbf{C}^N$. Let $f(\varepsilon)$ be the fraction of the volume of the unit ball in $\mathbf{C}^N$ which is within distance $\varepsilon \le 1$ of M. Then*

$$f(\varepsilon) \le \frac{\sqrt{\pi}\,\Gamma(N + 1/2)}{\Gamma(N - d + 1/2)\Gamma(d + 1/2)} e^2 N^2 (N-1)^{2N-2d-2} \cdot \deg(M) \cdot \varepsilon^{2(N-d)} \cdot (1 + N\varepsilon)^{2d}. \tag{4.1}$$

If M is a hypersurface ($d = N - 1$), then this upper bound may be improved to

$$f(\varepsilon) \le e^2 N^3 \cdot \deg(M) \cdot \varepsilon^2 \cdot (1 + N\varepsilon)^{2(N-1)}. \tag{4.2}$$

If M passes through the origin, it is also true that

$$(1 - \varepsilon)^{2d} \varepsilon^{2(N-d)} \cdot \frac{\Gamma(N - d + 1/2)\Gamma(d + 1/2)}{\deg(M)\sqrt{\pi}\,\Gamma(N + 1/2)} \le f(\varepsilon). \tag{4.3}$$

If M is a hypersurface passing through the origin this lower bound may be improved to

$$\frac{(1 - \varepsilon)^{2N-2}\varepsilon^2}{N \deg(M)} \le f(\varepsilon). \tag{4.4}$$

Now we specialize to the case of M homogeneous. In this case the upper bound (4.1) may be improved to

$$f(\varepsilon) \le e^2 N^2 (N-1)^{2N-2d-2} \cdot \deg(M) \cdot \varepsilon^{2(N-d)} \cdot (1 + N\varepsilon)^{2d}. \tag{4.5}$$

The lower bound (4.3) may be improved to

$$(1 - \varepsilon)^{2d}\varepsilon^{2(N-d)} \cdot \frac{\Gamma(N - d + 1/2)\Gamma(d + 1/2)}{\sqrt{\pi}\,\Gamma(N + 1/2)} \leq f(\varepsilon). \tag{4.6}$$

If M is also a hypersurface, the lower bound (4.4) may be further improved to

$$\frac{(1 - \varepsilon)^{2N-2}\varepsilon^2}{N} \leq f(\varepsilon). \tag{4.7}$$

Finally, we have the following asymptotic expression for small ε:

$$f(\varepsilon) = \binom{N}{d} \cdot \deg(M) \cdot \varepsilon^{2(N-d)} + o(\varepsilon^{2(N-d)}). \tag{4.8}$$

Thus, we have upper bounds, lower bounds, and asymptotic formulas all of which only depend on ε, N, d, and $\deg(M)$. All our expressions are proportional to $\varepsilon^{2(N-d)}$, and so differ at most only by factors depending on the parameters N, $\deg(M)$, and d. All these results are new, except for the upper bound (4.2) for $d = N - 1$ (Renegar, 1987).

These results can be used to give bounds for $\text{Prob}(\text{dist}(p, M) \leq \varepsilon)$ when M is homogeneous and p is uniformly distributed on the unit sphere in $\mathbf{C}^N$:

THEOREM 4.2. *Suppose M is a complex, homogeneous, purely $2d$-dimensional variety in $\mathbf{C}^N$. Let p be distributed uniformly on the unit sphere centered at the origin in $\mathbf{C}^N$. Then for $d < N - 1$*

$$\text{Prob}(\text{dist}(p, M) \leq \varepsilon) \leq e^2N^2(N - 1)^{2N-2d-2} \cdot \deg(M) \cdot \varepsilon^{2(N-d)} \cdot (1 + N\varepsilon)^{2d}, \tag{4.9}$$

$$(1 - \varepsilon)^{2d}\varepsilon^{2(N-d)} \cdot \frac{\Gamma(N - d + 1/2)\Gamma(d + 1/2)}{4N\sqrt{\pi}\,\Gamma(N + 1/2)} \leq \text{Prob}(\text{dist}(p, M) \leq \varepsilon), \tag{4.10}$$

and for asymptotically small ε

$$\text{Prob}(\text{dist}(p, M) \leq \varepsilon) = \binom{N - 1}{d - 1} \cdot \deg(M) \cdot \varepsilon^{2(N-d)} + o(\varepsilon^{2(N-d)}). \tag{4.11}$$

For hypersurfaces ($d = N - 1$)

$$(1 - \varepsilon)^{2N-2}\frac{\varepsilon^2}{4N^2} \leq \text{Prob}(\text{dist}(p, M) \leq \varepsilon) \leq e^2N^2 \cdot \deg(M) \cdot \varepsilon^2 \cdot (1 + N\varepsilon)^{2(N-1)} \tag{4.12}$$

and for asymptotically small ε

$$\text{Prob}(\text{dist}(p, M) \leq \varepsilon) = (N - 1)\deg(M)\varepsilon^2 + o(\varepsilon^2). \tag{4.13}$$

It is estimates (4.12) and (4.13) we shall apply to condition numbers in the next section.

Now we turn to real varieties. The bounds are necessarily looser, since a d-dimensional real variety can have an arbitrarily small volume; this is in strict contrast to complex varieties, where we can bound the volume above and below just in terms of the degree. The next theorem is due to Ocneanu:

THEOREM 4.3 (Ocneanu, 1985). *Suppose M is a real, purely d-dimensional variety in $\mathbf{R}^N$. Suppose further that M is the complete intersection of the polynomials $g_1, \ldots, g_{N-d}$. Let $D \equiv \max \deg(g_i)$, and $f(\varepsilon)$ be the fraction of the volume of the unit ball in $\mathbf{R}^N$ which lies within distance ε of M. Then*

$$f(\varepsilon) \leq 2(N - d) \sum_{k=N-d}^{N} \binom{N}{k} \cdot (2D\varepsilon)^k. \tag{4.14}$$

It appears that Ocneanu's proof may be able to be extended to give an asymptotic formula for $f(\varepsilon)$, which we state as a

Conjecture. Suppose M is as in Theorem 4.3. Then for asymptotically small ε

$$f(\varepsilon) = \text{vol}(M) \cdot \varepsilon^{N-d} \cdot \frac{N\Gamma(N/2)}{(N - d)\pi^{d/2}\Gamma((N - d)/2)} + o(\varepsilon^{N-d}), \tag{4.15}$$

where $\text{vol}(M)$ is the d-dimensional volume of M.

Without any assumptions about complete intersection, we can compute a lower bound for $f(\varepsilon)$:

THEOREM 4.4 *Suppose M is a real, purely d-dimensional variety in $\mathbf{R}^N$. Let $\text{vol}(M[r])$ be the d-dimensional volume of the subset of M within distance r of the origin. Then*

$$\frac{\text{vol}(M[1 - \varepsilon])}{\deg(M)} \varepsilon^{N-d} \cdot \frac{d\Gamma(d/2)\Gamma((d + 1)/2)\Gamma((N - d + 1)/2)}{2\pi^{(d+1)/2}\Gamma((N + 1)/2)} \leq f(\varepsilon). \tag{4.16}$$

If M is homogeneous, $\text{vol}(M[1 - \varepsilon])$ may be replaced by $(1 - \varepsilon)^d \text{vol}(M[1])$.

Note that the ratio between the conjectured asymptotic value in (4.15) and the lower bound in (4.16) depends only on N, d, and $\deg(M)$.

As before, we can translate the estimates in the last two theorems into estimates on $\text{Prob}(\text{dist}(p, M) \leq \varepsilon)$, where p is uniformly distributed on the unit sphere:

THEOREM 4.5. *Let M be a real, purely d-dimensional homogeneous variety in $\mathbf{R}^N$. Suppose p is uniformly distributed on the unit sphere in $\mathbf{R}^N$. Then*

$$\frac{\text{vol}(M[1])}{\deg(M)} \cdot (1 - \varepsilon)^d \varepsilon^{N-d} \cdot \frac{d\Gamma(d/2)\Gamma((d+1)/2)\Gamma((N-d+1)/2)}{8N\pi^{(d+1)/2}\Gamma((N+1)/2)} \leq \text{Prob}(\text{dist}(p, M) \leq \varepsilon). \tag{4.17}$$

If, in addition, M is the complete intersection of $N - d$ polynomials, each of degree at most D, then

$$\text{Prob}(\text{dist}(p, M) \leq \varepsilon) \leq 2(N - d) \sum_{k=N-d}^{N} \binom{N}{k} \cdot (2D\varepsilon)^k. \tag{4.18}$$

If the conjecture (4.15) *is true, then this would yield the following estimate for asymptotically small ε when M is a complete intersection:*

$$\text{Prob}(\text{dist}(p, M) \leq \varepsilon) = \text{vol}(M) \cdot \varepsilon^{N-d} \cdot \frac{d\Gamma(N/2)}{(N-d)\pi^{d/2}\Gamma((N-d)/2)} + o(\varepsilon^{N-d}). \tag{4.19}$$

Summarizing these results for the case of a real, homogeneous hypersurface defined by a single polynomial, we have

$$\frac{\text{vol}(M[1])}{\deg(M)} \cdot (1 - \varepsilon)^{N-1}\varepsilon \cdot \frac{\Gamma(N/2)}{4N\pi^{N/2}} \leq \text{Prob}(\text{dist}(p, M) \leq \varepsilon) \leq 2 \sum_{k=1}^{N} \binom{N}{k} (2 \deg(M)\varepsilon)^k \tag{4.20}$$

and, for asymptotically small ε (if the conjecture (4.15) *is true):*

$$\text{Prob}(\text{dist}(p, M) \leq \varepsilon) = \text{vol}(M) \cdot \varepsilon \frac{(N-1)\Gamma(N/2)}{\pi^{N/2}} + o(\varepsilon). \tag{4.21}$$

It is estimate (4.20) we will use to estimate the distribution of condition numbers of real problems.

We may explain these theorems intuitively as follows. If M is a d-dimensional surface in $\mathbf{R}^N$, the dominating term in the expression for the volume of the set of points within distance ε of M turns out to be (Weyl, 1939)

$$(d\text{-dimensional volume of } M) \cdot (N-d)\text{-dimensional volume of a unit ball in } \mathbf{R}^{N-d}) \cdot \varepsilon^{N-d}. \tag{4.22}$$

Suppose, for example, M is a straight line of length l in $\mathbf{R}^2$. Then $d = 1$, $N = 2$, and the estimate of (4.22) is $l \cdot 2 \cdot e$, the area of a rectangle of length l and width 2ε centered on M. It turns out that even if M is curved that as long as its radius of curvature everywhere exceeds ε, the area of the stripe of radius 2ε centered on M is exactly $2l\varepsilon$. If M is a straight line of length l in $\mathbf{R}^3$, (4.22) gives the volume $l \cdot \pi \cdot \varepsilon^2$ of the right circular cylinder of length l and radius ε centered on M. If M is curved, this formula is still asymptotically correct for small ε. If M is a square of side l in $\mathbf{R}^2$, (4.22) correctly gives the volume $l^2 \cdot 2 \cdot \varepsilon$ of the rectangular parallelepiped of thickness 2ε centered on M. Again, bending M does not change the asymptotic correctness of (4.22). In fact, if M is a smooth compact manifold, for sufficiently small ε the volume of the set of points within distance ε of M is a *polynomial* in ε with leading term given in (4.22) (Weyl, 1939).

It remains to estimate the d-dimensional volume of M needed in (4.22). Here we make use of the fact that M is a variety, for there are formulas from integral geometry for estimating the volume of a set M in $\mathbf{R}^N$ in terms of the number of points in $M \cap L$, where L is a plane of dimension $N - d$. For varieties, this number is bounded by $\deg(M)$. In fact, if M is a complex homogeneous purely $2d$-dimension variety in $\mathbf{C}^N$, the $2d$-volume of the part of M inside the unit ball is *exactly* $\deg(M)\pi^N/N!$ (Thie, 1967). No such statement can be made about real varieties, so a formula like (4.4) cannot hold for real varieties.

Open Problems. Ocneanu's proof of Theorem 4.3 depends on being able to express the real variety M as a complete intersection. Not all varieties permit such a representation. For example, the 3 by 3 real matrices of rank at most 1 forms a variety of codimension 4 but 9 polynomials (the determinants of all 2 by 2 submatrices) are needed for its definition. Is there a bound for real varieties that does not depend on the property of complete intersection? Also, Ocneanu's bound contains the factor D^{N-d}, which by Bezout's theorem (Van der Waerden, 1953) is a possibly pessimistic upper bound for $\deg(M)$. Is there a bound which depends only linearly on $\deg(M)$? More generally, is there an upper bound which depends linearly on $\mathrm{vol}(M[1])$?

All the asymptotic expressions above depend on the contribution to $f(\varepsilon)$ from small neighborhoods of the singular set of M going to zero. For complex varieties, the proof of the upper bounds yields this fact. Ocneanu's proof appears to yield it as well, leading us to make conjecture (4.15).

The proof of our lower bounds uses arguments from integral geometry (Santalo, 1976) which cannot rule out the possibility that the surface "folds over" onto itself $\deg(M)$ times, leading to the $\deg(M)$ factor in the denominator of (4.3) and (4.4). This factor seems unnecessary; can it be removed?

5. Computing the Distributions of Condition Numbers

In this section we apply our geometrical estimates of the last section to compute the distributions of the condition numbers discussed in Section 3.

Matrix Inversion. Applying estimates (4.12) and (4.13) to Eq. (3.2) yields the following theorem:

THEOREM 5.1. *Let A be a random complex n by n matrix distributed in such a way that $A/\|A\|_F$ is uniformly distributed on the unit sphere. Let $\kappa(A) = \|A\|_F \cdot \|A^{-1}\|$. Then*

$$\frac{(1 - 1/x)^{2n^2-2}}{4n^4x^2} \leq \text{Prob}(\kappa(A) \geq x) \leq \frac{e^2n^5(1 + n^2/x)^{2n^2-2}}{x^2} \tag{5.1}$$

and

$$\text{Prob}(\kappa(A) \geq x) = \frac{n(n^2 - 1)}{x^2} + O\left(\frac{1}{x^2}\right). \tag{5.2}$$

Remark. The upper bound in (5.1) exceeds the asymptotic value in (5.2) by a factor of only about $e^2n^4/(n^2 - 1)$ for sufficiently large x. However, even for $n = 10$, x must exceed about 5300 for the upper bound to drop below 1. For $n = 100$, x must exceed $2.2 \cdot 10^7$ for the upper bound to drop below 1.

Applying estimate (4.20) to Eq. (3.2) yields

THEOREM 5.2. *Let A be a random real n by n matrix distributed in such a way that $A/\|A\|_F$ is uniformly distributed on the unit sphere. Let $\kappa(A) = \|A\|_F \cdot \|A^{-1}\|$. Then*

$$\frac{C(1 - 1/x)^{n^2-1}}{x} \leq \text{Prob}(\kappa(A) \geq x) \leq \sum_{k=1}^{n^2} 2\binom{n^2}{k} \cdot \left(\frac{2n}{x}\right)^k, \qquad (5.3)$$

where $C > 0$ is a constant proportional to the volume of the variety of singular matrices inside the unit ball.

Remark. When $n = 10$, x must exceed 4900 for the upper bound in (5.3) to be less than 1. More generally, for large n, x must exceed about $4.93n^3$ for the upper bound to be less than 1. One can prove this by noting that the upper bound may also be written as $2[(1 + 2n/x)^{n^2} - 1]$.

Other sets of interest are matrices of rank at most $r < n - 1$. The volumes of these sets can also be estimated from above and below using Theorem 4.2, provided we can bound the degree of these varieties.

Polynomial Zero Finding. Applying estimate (4.12) to inequality (3.3) yields the following theorem:

THEOREM 5.3. *Let p be a random complex nth degree polynomial distributed in such a way that $p/\|p\|_F$ is uniformly distributed on the unit sphere. Let $\kappa(p) = \max_z \|p\|/|p'(z)|$, where the maximum is over all zeros of p. Then*

$$\text{Prob}(\kappa(A) \geq x) \leq \frac{4e^2(n + 1)^2(n - 1)(1 + \sqrt{2}(n + 1)/x)^{2n}}{x^2}. \qquad (5.4)$$

Applying estimate (4.20) *to inequality* (3.3) *yields*

THEOREM 5.4. *Let p be a random real nth-degree polynomial distributed in such a way that $p/\|p\|_F$ is uniformly distributed on the unit sphere. Let $\kappa(p)$ be as in Theorem* 5.2. *Then*

$$\text{Prob}(\kappa(A) \geq x) \leq 2\sum_{k=1}^{n+1} \binom{n+1}{k} \left(\frac{2^{5/2}(n - 1)}{x}\right)^k. \qquad (5.5)$$

Eigenvalue Calculations. Applying estimate (4.12) to inequality (3.5) yields

THEOREM 5.5 *Let A be a random complex n by n matrix distributed in such a way that $A/\|A\|_F$ is uniformly distributed on the unit sphere. Let $\kappa_\lambda(A) \equiv \max_{\lambda(A)} \|P_{\lambda(A)}\|$, where the max is over all eigenvalues $\lambda(A)$ of A and $P_{\lambda(A)}$ is the projection associated with $\lambda(A)$. Then*

$$\text{Prob}(\kappa_\lambda(A) \geq x) \leq \frac{2e^2n^5(n - 1)(1 + \sqrt{2}n^2/x)^{2n^2-2}}{x^2}. \qquad (5.6)$$

Applying estimate (4.20) to inequality (3.5) yields

THEOREM 5.6. *Let A be a random real n by n matrix distributed in such a way that $A/\|A\|_F$ is uniformly distributed on the unit sphere. Let $\kappa_\lambda(A)$ be as in Theorem 5.4. Then*

$$\text{Prob}(\kappa_\lambda(A) \geq x) \leq 2 \sum_{k=1}^{n^2} \binom{n^2}{k} \left(\frac{2^{3/2}(n^2 - n)}{x}\right)^k. \tag{5.7}$$

Applying estimate (4.9) to inequality (3.6) yields

THEOREM 5.7. *Let A be a random complex n by n matrix distributed in such a way that $A/\|A\|_F$ is uniformly distributed on the unit sphere. Let $\kappa_P(A) \equiv \max_{\lambda(A)} \|P_{\lambda(A)}\| \cdot \|S_{\lambda(A)}\| \cdot \|A\|_F$, where the* max *is over all eigenvalues $\lambda(A)$ of A, $P_{\lambda(A)}$ is the projection associated with $\lambda(A)$, and $S_{\lambda(A)}$ is the reduced resolvent associated with $\lambda(A)$. Then*

$$\frac{(1 - 1/x)^{2n^2-2}}{196n^4x^2} \leq \text{Prob}(\kappa_P(A) \geq x) \tag{5.8}$$

One can also prove a lower bound on $\text{Prob}(\kappa_P(A) \geq x)$ for real matrices of the form C/x, but C is proportional to the volume of the variety of real matrices with multiple eigenvalues and lying inside the unit ball, and seems difficult to estimate.

6. PRACTICAL APPLICATIONS AND LIMITATIONS

In this section we show how to estimate the distribution of the error in results computed by finite precision algorithms for the problems we analyzed above. The new tool required is backward error analysis (Wilkinson, 1963); using it we show that except in the improbable situation that the problem to be solved is close to the set *IP* of ill-posed problems, a backward stable algorithm will supply an accurate answer. We analyze Gaussian elimination this way in Section 6.1.

Such an analysis assumes problems are distributed uniformly as discussed in Section 1. This assumption breaks down in two important situations. First, some algorithms produce problems which tend to lie very close to the set *IP* of ill-posed problems, or which in fact converge to *IP*. For example, inverse iteration to compute eigenvalues and eigenvectors involves solving a sequence of linear equations with increasingly ill-conditioned coefficient matrices. Another example is the numerical solution of differential equations; the resulting matrices are approximations of unbounded operators and are necessarily close to singular.

Second, the set of problems representable in a computer (in finite precision arithmetic) is necessarily finite and so any distribution we put on this set will necessarily be discrete, not continuous as assumed in our previous analysis. As long as the discrete points are dense enough to model the continuum (this depends on the individual problem), the continuous model is relevant. It will turn out, however, that this discreteness ultimately leads to qualitatively different behavior of algorithms than is predicted by the continuous model. We discuss this situation further in Section 6.2.

Finally, in Section 6.3, we discuss how this theory might be extended to the finite precision case and what such an extension would tell us about the design both of numerical algorithms and computer arithmetic. In particular, we show how it would tell us how many finite precision problems we could solve as a function of the extra precision used in intermediate calculations. This information would be of use in algorithm and even computer hardware design. Accomplishing this extension is an open problem.

6.1. *A Paradigm for Analyzing the Accuracy of Finite Precision Algorithms*

The paradigm for applying the probabilistic model to the analysis of algorithms is as follows:

(1) Within the space of problems, identify the set *IP* of ill-posed ones, and show that the closer a problem is to *IP* the more sensitive the solution is to small changes in the problem.

(2) Show that the algorithm in question computes an accurate solution for a problem close to the one it received as input (this is known as "backward stability" (Wilkinson, 1963)). Combined with the result of (1), this will show that the algorithm will compute an accurate solution to a problem as long as the problem is far enough from *IP*.

(3) Compute the probability that a random problem is close to *IP*. Using this probability distribution in conjunction with the result of (2) we can compute the probability of the algorithm computing an accurate result.

This paradigm is best explained by applying it to matrix inversion:

(1) The set of matrices *IP* which are ill-posed with respect to inversion are the singular matrices. As discussed in Section 3, the condition number

$$\kappa(M) = \|M\|_F \cdot \|M^{-1}\| \tag{6.1}$$

measures how difficult the matrix M is to invert, and when $\|M\|_F = 1$ it is the reciprocal of the distance to the nearest singular matrix.

(2) Gaussian elimination with partial pivoting is a standard algorithm for matrix inversion and is well known to be a backward stable algorithm (Wilkinson, 1963). Backward stability means that when applying Gaussian elimination to compute the solution of the system of linear equations $Mx = b$, one gets an answer $\hat{x}$ which satisfies $(M + \delta M)\hat{x} = b$, where δM is small in norm compared to M. More precisely, let X_i be the ith column of the approximation to M^{-1} computed using Gaussian elimination, where the arithmetic operations performed (addition, subtraction, multiplication, and division) are all rounded off to b bits of precision. Then X_i is the value of the ith column of the inverse of a matrix $M + \delta M_i$ where δM_i is small,

$$\|\delta M_i\|_F \leq f(n) \cdot 2^{-b} \cdot \|M\|_F, \tag{6.2}$$

where $f(n)$ is a function only of n, the dimension of M (Wilkinson, 1963). This last expression can be used to bound the relative error in the computed solution (Wilkinson, 1963):

$$\frac{\|X - M^{-1}\|_F}{\|M^{-1}\|_F} \leq \frac{\sqrt{n}\,\kappa(M) \cdot f(n) \cdot 2^{-b}}{1 - \kappa(M) \cdot f(n) \cdot 2^{-b}}. \tag{6.3}$$

In other words, as long as the bound (6.2) on $\|\delta M_i\|_F$ is not so large that $M + \delta M_i$ could be singular, i.e., as long as

$$\operatorname{dist}(M, IP) > f(n) \cdot 2^{-b} \cdot \|M\|_F$$

or, substituting from (6.1)

$$\kappa(M) < 2^b/f(n), \tag{6.4}$$

then the relative error in the computed inverse X is bounded, and the smaller the value of $\kappa(M)$ the more accurate is the solution.

(3) Assuming M is complex we can apply Theorem 5.1 (which gives the probability distribution of the condition number) to estimate the probability that a random matrix can be inverted accurately:

$$\operatorname{Prob}\left(\frac{\|X - M^{-1}\|_F}{\|M^{-1}\|_F} \leq \varepsilon\right) \geq \operatorname{Prob}\left(\frac{\sqrt{n}\,\kappa(M) \cdot f(n) \cdot 2^{-b}}{1 - \kappa(M) \cdot f(n) \cdot 2^{-b}} \leq \varepsilon\right), \tag{6.5}$$

which, after some rearrangement (and assuming $\varepsilon < 1$),

$$= \text{Prob}\left(\kappa(M) \leq \frac{\varepsilon}{f(n) \cdot (\sqrt{n} + \varepsilon)2^{-b}}\right)$$

$$\geq 1 - (e^2 n^5 (1 + n^2 f(n)(\sqrt{n} + \varepsilon) \cdot (2^{-b}/\varepsilon))^{2n^2-2} f^2(n)(\sqrt{n} + \varepsilon)^2) \cdot \left(\frac{2^{-b}}{\varepsilon}\right)^2.$$

This inequality implies that as we compute with higher and higher precision (b increases), the probability of getting a computed answer with accuracy ε goes to 1 at least as fast as $1 - O(4^{-b})$. Note that the inequality only makes sense for $2^{-b}/\varepsilon$ small, that is, if the error 2^{-b} in the arithmetic is smaller than the error ε demanded of the answer. This restriction makes sense numerically, since we cannot expect more precision than we compute with. The restriction also implies that the finite precision numbers are sufficiently dense to approximate the continuum, since the radius r of the neighborhood around IP, $r = f(n)(\sqrt{n} + \varepsilon)2^{-b}/\varepsilon$, is much larger than the distance between adjacent finite precision points 2^{-b}. This situation is depicted in Fig. 1 and discussed in the next section.

We may use the same kind of paradigm as discussed so far to analyze the speed of convergence of an algorithm rather than its accuracy. In this case the paradigm is

(1′) Identify the ill-posed problems IP.

(2′) Show that the closer a problem is to IP, the more slowly the algorithm converges.

(3′) Compute the probability that a random problem is close to IP. Combined with (2′) this yields the probability distribution of the speed of convergence.

This approach has been used by Smale (1981) and Renegar (1987) in their average speed analyses of Newton's method for finding zeros of polynomials.

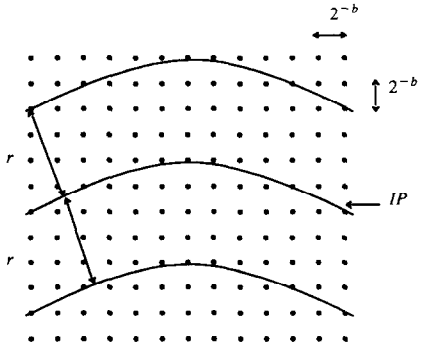

FIG. 1. An $r > 2^{-b}$ neighborhood of IP.

6.2. *Limitations of the Probabilistic Model*

In this section we discuss limitations to the applicability of our model. As mentioned before, the model does not apply in situations where the problems tend to be clustered about the ill-posed problems. One such example is inverse iteration for computing the eigenvalues and eigenvectors of a matrix:

$$x_{i+1} = (A - \lambda_i)^{-1} x_i$$
$$\lambda_{i+1} = (Ax_{i+1})^j / x^j_{i+1}, \qquad \text{where } |x^j_{i+1}| = \max_k |x^k_{i+1}|.$$

If λ_i is a good approximation to the simple eigenvalue λ, and x_i approximates the corresponding eigenvector x, then λ_{i+1} and x_{i+1} will be even better approximations to λ and x. As λ_i approaches λ, the matrices $A - \lambda_i$ become increasingly ill-conditioned. Thus, the set of matrices $\{A - \lambda_i\}$ being (conceptually) inverted (actually, one solves $(A - \lambda_i)x_{i+1} = x_i$ directly) converges to the set *IP* of ill-posed problems, and so is far from uniformly distributed. This invalidates the assumption of the model, even in exact arithmetic. In finite precision arithmetic, inverse iteration works very well, even though naive backward error analysis as in Section 6.1 might lead us to expect total loss of precision. This is because the rounding errors committed while solving $(A - \lambda_i)x_{i+1} = x_i$ provably conspire to produce an error lying almost certainly in the direction of the desired eigenvector (Golub and Van Loan, 1983).

The second way in which the model breaks down depends on the ultimate discreteness of the finite precision numbers which can be represented in a computer. The natural version of a "uniform distribution" in this case is simply counting measure. The continuous model is a good approximation to counting measure only as long as the finite precision numbers are dense enough to resemble the continuum. In Fig. 1, for example, the area of the set of points within distance r of the curve *IP* is a good approximation to the number of dots (finite precision points) within distance r of *IP* (scaled appropriately). This is true because the radius r of the neighborhood of *IP* is large compared to the spacing 2^{-b} between dots. When $r < 2^{-b}$ on the other hand as in Fig. 2 the area of the set of points within distance r of *IP* is not necessarily a good approximation of the number of dots within r of *IP*. For example, if *IP* were a straight line passing exactly halfway between two rows of dots, there would be no dots within distance 2^{-b-1} of *IP*. If on the other hand *IP* were a straight line running along a row of dots, there would be a constant nonzero number of dots within distance η of *IP* for all $\eta < 2^{-b}$. Thus, when the radius of the neighborhood of *IP* gets smaller than the interdot distance 2^{-b}, the model breaks down.

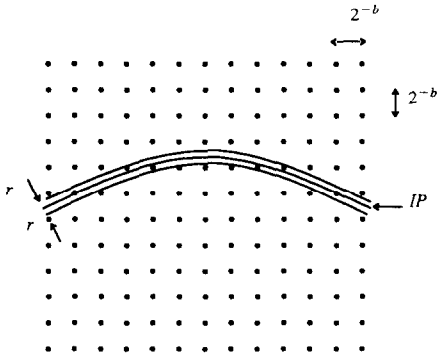

FIG. 2. An $r < 2^{-b}$ neighborhood of IP.

Specifically, let us consider matrix inversion. In the continuous model the exactly singular matrices form a set of measure zero, so the chance of a random problem being singular is zero. Also, there are nonsingular matrices arbitrarily close to the set of singular ones, and so of unbounded condition number. Consider now the finite (but large) set of matrices which can be represented in a computer using finite precision arithmetic. Some fraction of this finite set are exactly singular, so in choosing one member of this finite set at random (using counting measure) there is a nonzero probability of getting an exactly singular matrix. Furthermore, the remaining nonsingular matrices have condition numbers bounded by some finite value K. Thus, instead of $\text{Prob}(\kappa(A) \geq x)$ decreasing monotonically to 0 as x increases as in the continuous case, $\text{Prob}(\kappa(A) \geq x)$ becomes constant and nonzero for $x > K$. This is clearly significantly different behavior. It does not, however, invalidate the analysis of Gaussian elimination in the last section, because we assumed $2^{-b} < r$, i.e., the situation in Fig. 1.

In the next section we discuss what we could do if we could compute $\text{Prob}(\kappa(A) \geq x)$ in the discrete case for all x, in particular for x too large for the continuous approximation to apply.

6.3. *How to Use the Discrete Distribution of Points within Distance ε of a Variety*

Before proceeding, we need to say what probability measure we are going to put on the discrete set of finite precision points. The last section showed that no single distribution is good for all applications, but a uniform distribution remains a neutral and interesting choice. So far we have been implicitly using fixed point numbers, in which case assigning equal probability to each point (counting measure) gives a uniform distribution. For floating point numbers, however, this is no longer appropriate since

the floating point numbers are not evenly distributed on the number line. Since floating point numbers are much closer together near the origin than far away from it (the distance between adjacent numbers is approximately a constant times the number), counting measure would assign much more probability to equal length intervals near the origin than far away from it. A simple way to adjust for this nonuniform spacing is to assign to each point M a probability proportional to the volume of the small parallelepiped of points which round to M (i.e., the parallelepiped centered at M with sides equal in length to the distance between adjacent finite precision points). In the case of fixed point arithmetic, this just reproduces counting measure, whereas with floating point arithmetic points near 0 have smaller probability than larger points, and intervals of equal length have approximately equal probabilities. Actually, the question of the distribution of the digits of a floating point number has a large literature (Knuth, 1969; Hamming, 1970; and Bareiss, 1981) but the discussion in this section does not depend strongly on the actual distribution of digits chosen.

We claim that knowing the probability distribution of the distance of a random finite precision problem to the set IP of ill-posed problems will tell us how many finite precision problems we can solve as a function of the extra precision used in intermediate calculations. As mentioned before, programmers often resort to extra precision arithmetic to get more accurate solutions to problems which are given only to single precision. This extra precision has a cost (in speed and memory) dependent on the number of digits carried, so programmers usually avoid extra precision unless persuaded otherwise by bad experiences, an error analysis, or paranoia. Therefore an accurate estimate of how many problems can be solved as a function of the extra precision used would not only help programmers decide how much to use but possibly influence hardware designers when they decide how much precision to make available in their computer systems.

How does knowledge of this probability distribution tell us how much extra precision to use? The paradigm in Section 6.1 tells us how. Consider matrix inversion. Formula (6.3) tells us that using fixed point arithmetic of accuracy 2^{-b} permits us to compute inverses of matrices to within accuracy ε as long as their condition numbers are less than $\varepsilon/(f(n)(\sqrt{n} + \varepsilon)2^{-b})$. Suppose we choose our problems at random from the set of matrices with b_0-bit entries, and let $\text{Prob}_{b_0}(\kappa(M) \geq x)$ be the discrete distribution function of the condition number. Then

$$N_{b_0}(b) \equiv 1 - \text{Prob}_{b_0}\left(\kappa(M) \geq \frac{\varepsilon}{f(n)(\sqrt{n} + \varepsilon)2^{-b}}\right)$$

bounds from below the fraction of b_0-bit matrices we can invert with

accuracy ε as a function of the number of bits $b \geq b_0$ carried in the calculation. By examining $N_{b_0}(b)$ as a function of b, one can decide exactly how much improvement one gets for each additional bit of precision b. For example, we know from the previous discussion that there is a $\bar{b}$ such that when $b \geq \bar{b}$ $N_{b_0}(b)$ is constant and nonzero. Therefore, it clearly does not pay to increase b beyond $\bar{b}$

This discussion has assumed so far that the finite precision input is known exactly, i.e., that there is no error inherited from previous computations or from measurement errors. In general there will be such errors, and they will almost always be at least a few units in the last place of the input problem. In other words, there already is a ball of uncertainty around the input problem with a radius equal to a small multiple of the interpoint distance 2^{-b_0}. Therefore, it may make no sense to use higher precision to accurately solve problems lying very close to *IP* when the inherited input error is so large that the true answer is inherently very uncertain. In such situations programmers sometimes shrug and settle for the backward stability provided by the algorithm, even if the delivered solution is entirely wrong, because the act of solution has scarcely worsened the uncertainty inherited from the data, and the programmer declines to be held responsible for the uncertainty inherent in the data. Nevertheless, getting an accurate answer for as many inputs as possible is a worthwhile goal, so we will not concern ourselves with possible errors made in creating the input matrices.

We close with another application of the discrete distribution $\text{Prob}_{b_0}(\kappa(M) \geq x)$. Consider the rather simple problem of inverting real 2 by 2 matrices. This problem is small enough that we can exhaustively compute $\text{Prob}_{b_0}(\kappa(M) \geq x)$ for low precision arithmetic. We have done this for $b_0 =$ 3, 4, 5, 6, and 7 (all numbers lay between 0 and 1 in absolute value, and each fixed point matrix was assigned the same probability). Let $P(r) = \text{Prob}_{b_0}(\kappa(M) \geq 1/r)$. We recall that in the continuous case (Theorem 5.2) $P(r)$ would be approximately a linear function of r. For all values of b_0 tested, we observed approximately the behavior of $P(r)$ as shown in Fig. 3. Surprisingly, we observed linear dependence of $P(r)$ on r not only for r larger than 2^{-b_0} (corresponding to Fig. 1) but for r quite a bit smaller than 2^{-b_0} (Fig. 2). The fraction of problems within 2^{-b_0} of a singular matrix was about 2^{1-b_0}. This linear behavior of $P(r)$ continued until r reached approximately 2^{-2b_0}, and there the graph of the distribution became horizontal and remained so all the way to the origin, intersecting the vertical axis at about 2^{2-2b_0}. This means that all matrices closer to *IP* than approximately 2^{-2b_0} were exactly singular. The fraction of matrices which were exactly singular was 2^{2-2b_0}.

What does this tell us about the use of extra precision? Basically, as long as the distribution function $P(r)$ remains linear, it says that for every

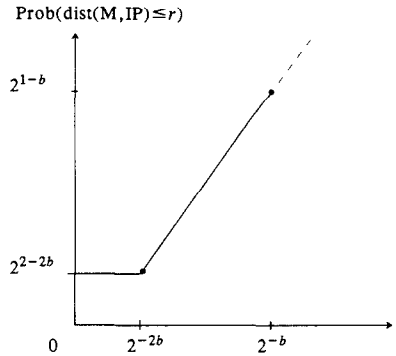

FIG. 3. Observed probability distribution of the distance r to the nearest singular matrix.

extra bit of intermediate precision, we can solve half the problems we couldn't solve before. This regime continues until we reach double precision, at which point the only problems we can't solve are exactly singular. Indeed, since

$$\begin{bmatrix} a & c \\ b & d \end{bmatrix}^{-1} = (ad - bc)^{-1} \begin{bmatrix} d & -c \\ -b & a \end{bmatrix}$$

we can clearly compute the inverse accurately if we can compute the determinant $ad - bc$ accurately. Since a, b, c, and d are given to single precision, double precision clearly suffices to compute $ad - bc$ exactly.

What if the discrete distribution function were similar for matrices of higher dimensions, that is, linear for a while and then suddenly horizontal when all worse conditioned matrices were exactly singular? It would again tell us that for a while, every extra bit of intermediate precision would let us solve half the problems we couldn't solve before. Eventually, after enough extra bits, (and for inverting fixed precision n by n matrices, this clearly occurs no later than reaching n-tuple precision), all finite precision matrices which are not exactly singular could be inverted, and more precision would contribute nothing. Thus a programmer (or hardware designer) could choose the number of bits b with which to compute in order to guarantee that the fraction of unsolvable problems is sufficiently close to its minimum. Of course, exhaustive evaluation of the distribution function is not reasonable for large problems, and estimating the distribution function becomes an interesting open question of number theory.

REFERENCES

ARNOLD, V. I. (1971), On matrices depending on parameters, *Russian Math. Surveys* **26,** 1–3.

BAREISS, E. H., AND BARLOW, J. L. (1981), "Probabilistic Error Analysis of Floating Point and CRD Arithmetics," Department of Electrical Engineering and Computer Science, Northwestern University, Report 81-02-NAM-1.

DEMMEL, J. (1983), "A Numerical Analysts's Jordan Canonical Form," Dissertation, Computer Science Division, University of California, Berkeley.

DEMMEL, J. (1986), On condition numbers and the distance to the nearest ill-posed problem, *Numer. Math.*, in press.

DEMMEL, J. (1987), "The Probability That a Numerical Analysis Problem is Difficult," Technical Report, Computer Science Department, Courant Institute, New York.

ECKART, C., AND YOUNG, G. (1936), The approximation of one matrix by another of lower rank, *Psychometrika,* **1,** 211–218.

KAHAN, W. (1966), Numerical linear algebra, *Canad. Math. Bull.* **9,** 757–801. (Gastinel's theorem appears here.)

GRAY, A. (1982a), An estimate for the volume of a tube about a complex hypersurface, *Tensor (N.S.)* **39,** 303–305.

GRAY, A. (1982b), Comparison theorems for the volumes of tubes as generalizations of the Weyl tube formula, *Topology* **21,** 2, 201–228.

GOLUB, G., AND VAN LOAN, C. (1983), "Matrix Computations," Johns Hopkins, Baltimore.

GRIFFITHS, PHILLIP A. (1978), Complex differential and integral geometry and curvature integrals associated to singularities of complex analytic varieties, *Duke Math. J.* **45,** 3, 427–512.

HAMMING, R. W. (1970), On the distribution of numbers, *Bell System Tech. J.* **49,** 8, 1609–1625.

HOTELLING, H. (1939), Tubes and spheres in n-spaces, and a class of statistical problems, *Amer. J. Math.* **61,** 440–460.

HOUGH, D. (1977), "Explaining and Ameliorating the Ill Condition of Zeros of Polynomials," Thesis, Mathematics Department, University of California, Berkeley.

KAHAN, W. (1972), "Conserving Confluence Curbs Ill-Condition," Technical Report 6, Computer Science Department, University of California, Berkeley, August 4.

KATO, T. (1966), "Perturbation Theory for Linear Operators," Springer-Verlag, Berlin.

KENDIG, KEITH (1977), "Elementary Algebraic Geometry," Springer-Verlag, New York.

KNUTH, D. (1969), "The Art of Computer Programming, " Vol. 2, Addison–Wesley, Reading, MA.

KOSTLAN, E. (1985), "Statistical Complexity of Numerical Linear Algebra," Dissertation, Mathematics Department, University of California, Berkeley.

LELONG, P. (1968), "Fonctions plurisousharmoniques et formes differentieles positiv," Gordon & Breach, Paris.

OCNEANU, A. (1985), "On the Volume of Tubes about a Real Variety," unpublished report, Mathematical Sciences Research Institute, Berkeley.

OCNEANU, A. On the loss of precision in solving large linear systems, to appear.

RENEGAR, J. (1987), On the efficiency of Newton's method in approximating all zeros of system of complex polynomials, *Math. Oper. Res.*, in press.

RUHE, A. (1970), Properties of a matrix with a very ill-conditioned eigenproblem, *Numer. Math.* **15,** 57–60.

SANTALÓ, LUIS A. (1976), "Integral Geometry and Geometric Probability, Encyclopedia of Mathematics and Its Applications," Vol. 1, Addison–Wesley, Reading, MA.

SMALE, S. (1981), The fundamental theorem of algebra and complexity theory, *Bull. Amer. Math. Soc. (N.S.)* **4,** No 1, 1–35.

SMALE, S. (1986), Algorithms for solving equations, presented at the International Congress of Mathematics, Berkeley.

STEWART, G. W. (1973), Error and perturbation bounds for subspaces associated with certain eigenvalue problems, *SIAM Rev.* **15,** No. 4, 752.

THIE, P. (1967), The Lelong number of points of a complex analytic set, *Math. Ann.* **172,** 269–312.

VAN DER WAERDEN, B. (1953), "Modern Algebra," Vol. 1, Ungar, New York.

WEYL, HERMANN (1939), On the volume of tubes, *Amer. J. Math.* **61,** 461–472.

WILKINSON, J. H. (1963), "Rounding Errors in Algebraic Processes," Prentice–Hall, Englewood Cliffs, NJ.

WILKINSON, J. H. (1965), "The Algebraic Eigenvalue Problem," Oxford Univ. Press, London/New York.

WILKINSON, J. H. (1972), Note on matrices with a very ill-conditioned eigenproblem, *Numer. Math.* **19,** 176–178.

WILKINSON, J. H. (1984a), On neighboring matrices with quadratic elementary divisors, *Numer. Math.* **44,** 1–21.

WILKINSON, J. H. (1984b), Sensitivity of eigenvalues, *Utilitas Math.* **25,** 5–76.