MATHEMATICS OF COMPUTATION
VOLUME 50, NUMBER 182
APRIL 1988, PAGES 449-480

The Probability That A Numerical Analysis Problem Is Difficult

By James W. Demmel

Abstract. Numerous problems in numerical analysis, including matrix inversion, eigenvalue calculations and polynomial zerofinding, 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 zerofinding 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.

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 zerofinding 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 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

Received May 4, 1987; revised August 18, 1987.

1980 *Mathematics Subject Classification* (1985 *Revision*). Primary 15A12, 53C65, 60D05.

Parts of this paper appeared previously under the title "The Geometry of Ill-Conditioning," in *The Journal of Complexity*, Vol. III, pgs. 201-229, 1987. Copyright © 1987 Academic Press, Inc. Reprinted by permission of the publisher.

$\mathbf{R}^N$. 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 many practical cases for a variety of reasons, most fundamentally because real-world problems are not uniformly distributed. A lesser reason is that any set of problems which can be represented in a computer is necessarily discrete rather than continuous. We will discuss the limitations of our choice of uniform distribution as well as alternatives at length in Section 6 below.

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, sometimes 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 formulae are very simple, depending only on x, the degree 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

$$\frac{(1-x^{-1})^{2n^2-2}}{2n^4x^2} \le \operatorname{Prob}(\kappa(A) \ge x) \le \frac{e^2 n^5 (1+n^2/x)^{2n^2-2}}{x^2},$$

and that asymptotically

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

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

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

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 below.

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 zerofinding 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 below and in [4].

The work was inspired by earlier work in a number of fields. Demmel [4], Gastinel [14], Hough [13], Kahan [15], Ruhe [25], Stewart [29], Wilkinson [36], [37], [38] and others have analyzed the relationship between the condition number and the distance to the nearest ill-posed problem mentioned above in (ii). Gray [8], [9], Griffiths [10], Hotelling [12], Lelong [20], Ocneanu [21], Renegar [23], Santaló [26], Smale [27], and Weyl [33] have worked on bounds of volumes of tubular neighborhoods. These volume bounds have been used by Smale [27], [28], Renegar [23] and others to analyze the efficiency of Newton's method for finding zeros of polynomials. This latter work inspired the author [3] to apply these bounds to conditioning. Ocneanu [22] and Kostlan [19] 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. Section 7 contains the proofs of the theorems in Section 4.

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

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

$\kappa(p) = \max_z \|p\|/|p'(z)|$, *where the maximum is over all zeros of* p. *Then*

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

Applying estimate (4.20) to inequality (3.3) yields

THEOREM 5.4. *Let* p *be a random real* n*th 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* 5.3. *Then*

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

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 maximum is over all eigenvalues* $\lambda(A)$ *of* A *and* $P_{\lambda(A)}$ *is the projection associated with* $\lambda(A)$. *Then*

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

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* 5.5. *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 (4.12) 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 maximum 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*

$$\frac{(1-1/(7x))^{2n^2-2}}{98n^4x^2} \leq \text{Prob}(\kappa_P(A) \geq x). \tag{5.8}$$

One can also prove a lower bound on $\text{Prob}(\kappa_P(A) \geq x)$, for real matrices, which is 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 [34]; 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 Subsection 6.1.

Such an analysis assumes problems are distributed uniformly as discussed in Section 1 of this paper. This is an extremely strong assumption, which is not met in many practical situations. First, real problems often have a structure which produces 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 become more nearly singular the finer the approximation becomes.

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 Subsection 6.2. (This limitation does not invalidate our analysis of Gaussian elimination in finite-precision arithmetic in Subsection 6.1.)

Finally, in Subsection 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 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" [34]). 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.

The first two steps of this paradigm are quite standard [24], [35]; only the third is new. 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 pivoting is a standard algorithm for matrix inversion and is well known to be a backward stable algorithm [34]. Backward stability means that when applying Gaussian elimination to compute the solution of the system of linear equations $Mx = c$, one gets an answer $\hat{x}$ which satisfies $(M+\delta M)\hat{x} = c$. 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 [34]. The magnitude of $f(n)$ depends on the pivoting strategy, and can be as large as 2^n if partial pivoting is used. although this is very rare in practice [34]. $f(n)$ is much smaller if either complete pivoting is used or if we substitute the QR algorithm for Gaussian elimination [7]. For our analysis, however, we are not interested in how big $f(n)$ is, only that inequality (6.2) holds. Inequality (6.2) can be used to bound the relative error in the computed solution [34]:

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

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 $\kappa(M)$ is, 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 $\varepsilon < 1$) equals

$$\begin{aligned}
&\operatorname{Prob}\left(\kappa(M) \leq \frac{\varepsilon}{f(n) \cdot (\sqrt{n} + \varepsilon) 2^{-b}}\right) \\
&\quad \geq 1 - (e^2 n^5 (1 + n^2 f(n)(\sqrt{n} + \varepsilon) \cdot (2^{-b}/\varepsilon))^{2n^2-2} f^2(n)(\sqrt{n} + \varepsilon)^2) \cdot \left(\frac{2^{-b}}{\varepsilon}\right)^2 \\
&\quad \equiv 1 - g(n, \varepsilon, b) \cdot (2^{-b}/\varepsilon)^2.
\end{aligned}$$

The $g(n, \varepsilon, b)$ factor depends only weakly on ε and b; the interesting factor is $(2^{-b}/\varepsilon)^2$. 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 Figure 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 [27] and Renegar [23] in their average speed analyses of Newton's method for finding zeros of polynomials.

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 might even assert that most problems have this character, or at least constitute the majority of interesting problems numerical analysts encounter. One such example is iteration for computing the eigenvalues and eigenvectors of a matrix:

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

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 Subsection 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 [7].

Another area where matrices tend to cluster around the singular ones is the solution of differential equations. In this case the matrices encountered are generally approximations to unbounded operators, and so become more nearly singular the finer the approximation. For example, the usual centered difference approximation to the second derivative on the interval $[0, 1]$ with mesh spacing h yields a matrix with condition number on the order of h^{-2}.

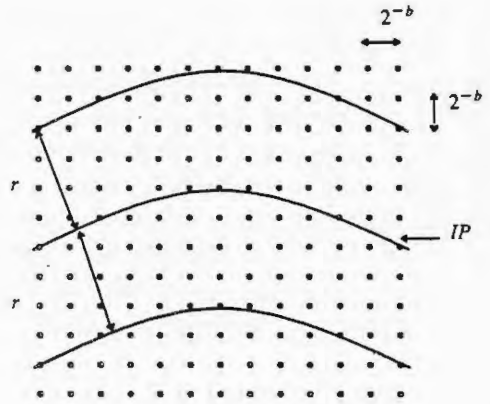

FIGURE 1

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

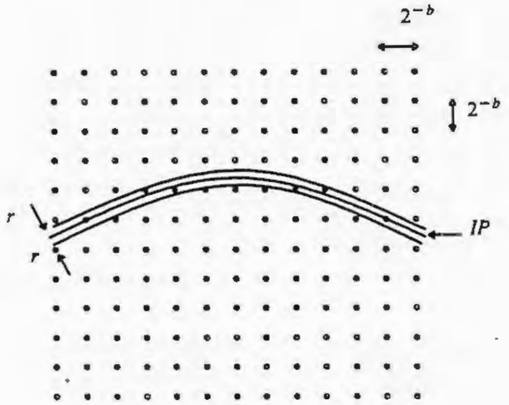

FIGURE 2

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

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 Figure 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 Figure 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 half way 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.

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 Figure 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 v hat 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 [2], [11], [18], but the discussion in this section does not depend strongly on the actual distribution of digits chosen.

This discussion does assume 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} (b_0 is the number of bits to which the input is stored). 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 claim that knowing the probability distribution of the distance of a random finite-precision problem to the set IP of ill-posed problems would tell us how many finite-precision problems we could 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 would help programmers decide how much to use.

How does knowledge of this probability distribution tell us how much extra precision to use? The paradigm in Subsection 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 $\mathrm{Prob}_{b_0}(\kappa(M) \geq x)$ be the discrete distribution function of the condition number. Then

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

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 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 $\bar{b}$.

We close with an example of the discrete distribution $\mathrm{Prob}_{b_0}(\kappa(M) \geq x)$. Consider the rather simple problem of inverting real 2 by 2 matrices. This problem is small enough that we can exhaustively compute $\mathrm{Prob}_{b_0}(\kappa(M) \geq x)$ 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 $P(r) = \mathrm{Prob}_{b_0}(\kappa(M) \geq 1/r)$. 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 $P(r)$ shown in Figure 3. Surprisingly, we observed linear dependence of $P(r)$ on r not only for r larger than 2^{-b_0} (corresponding to Figure 1), but for r quite a bit smaller than 2^{-b_0} (Figure 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 extra bit of intermediate precision, we can solve half the problems we could not solve before. This regime continues until we reach double precision, at which point the only problems we cannot 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.

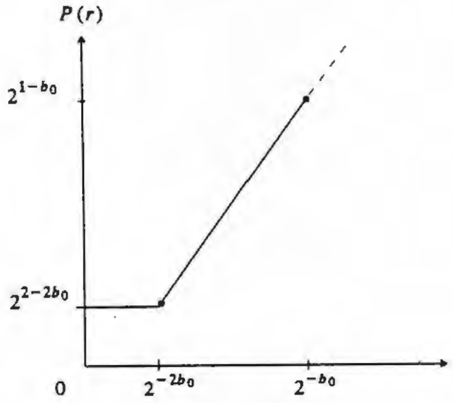

FIGURE 3

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

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 could not 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 system 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.

7. Proofs of Volume Estimates. In this section we present the proofs of the volume estimates of Section 4. In addition to the notation of Section 2, we will use

$$O_n = \frac{2\pi^{(n+1)/2}}{\Gamma((n+1)/2)} \quad \text{and} \quad \theta_n = \frac{O_{n-1}}{n} = \frac{2\pi^{n/2}}{n\Gamma(n/2)}.$$

O_n is the surface area of the n-dimensional unit sphere and θ_n is the volume of the n-dimensional unit ball [26]. $B(p,r)$ is the open ball of radius r centered at p, and $B(r)$ is centered at the origin. If M is any set, $M[r] = M \cap \overline{B}(r)$. If M is a variety, $NS(M)$ will denote the nonsingular part of M (a manifold) and $S(M)$ the remaining singular part (a lower-dimensional subvariety). $\mathrm{vol}(M)$ or $\mathrm{vol}_{\dim(M)}(M)$ will denote the $\dim(M)$-dimensional measure of $NS(M)$. $T(M,\varepsilon)$ is the set of points inside $\overline{B}(1)$ within distance ε of M:

$$T(M,\varepsilon) = \{z \colon \|z\| \le 1, \mathrm{dist}(z,M) \le \varepsilon\}.$$

If $M \subseteq \mathbf{R}^N$, $f(M,\varepsilon) = \mathrm{vol}_N(T(M,\varepsilon))/\theta_N$, the fraction of the unit ball within ε of M. $\#(S)$ will denote the cardinality of the discrete set S.

We will need the following estimates on the volumes of varieties inside balls.

LEMMA 7.1 ([31]). *Let M be a purely $2d$-dimensional homogeneous complex variety in $\mathbf{C}^N$. Then* $\mathrm{vol}_{2d}(M[r]) = \deg(M) \cdot \theta_{2d} \cdot r^{2d}$.

LEMMA 7.2 ([30, Theorem B]). *Let M be a purely $2d$-dimensional complex variety containing the origin. Then* $\mathrm{vol}_{2d}(M[r]) \ge \theta_{2d} \cdot r^{2d}$.

LEMMA 7.3 ([23, Proposition 5.3]). *Let M be a complex hypersurface in $\mathbf{C}^N$. Then* $\mathrm{vol}_{2N-2}(M[r]) \le \deg(M) \cdot N \cdot \theta_{2N-2} \cdot r^{2N-2}$.

LEMMA 7.4. *Let M be a purely d-dimensional real variety in $\mathbf{R}^N$. Then*

$$\mathrm{vol}_d(M[r]) \le \deg(M) \cdot \frac{O_{N-d} \cdot O_d}{2 \cdot O_N} \cdot \theta_d \cdot r^d.$$

Proof. That $\mathrm{vol}_d(M[r])$ is finite follows from [6, Section 3.4.10]. Let L_{N-d} denote an $(N-d)$-dimensional plane in $\mathbf{R}^N$. dL_{N-d} will denote the *kinematic density* on this set of planes [26, Chapter 12]. From [26, Eq. 12.38] we may write $dL_{N-d} = d\sigma_d \wedge dL_{d[0]}$, where $dL_{d[0]}$ is the kinematic density on d-planes through the origin and $d\sigma_d$ is the volume element on $L_{d[0]}$. This corresponds to parametrizing a plane L_{N-d} by the perpendicular plane through the origin $L_{d[0]}$ which it intersects, and where it intersects it. From [26, Eq. 14.70] we may write

$$\int_{M[r] \cap L_{N-d} \neq \varnothing} \#(M[r] \cap L_{N-d})\, dL_{N-d} = \frac{O_N \cdots O_{d+1}}{O_{N-d} \cdots O_1} \mathrm{vol}_d(M[r])$$

Computer Science Department
Courant Institute of Mathematical Sciences
New York University
251 Mercer Street
New York, New York 10012

1. V. I. ARNOL'D, "On matrices depending on parameters", *Russian Math. Surveys*, v. 26, 1971, pp. 29–43.

2. E. H. BAREISS & J. L. BARLOW, *Probabilistic Error Analysis of Floating Point and CRD Arithmetics*, Dept. of Electrical Engineering and Computer Science, Northwestern University, Report 81-02-NAM-01, 1981.

3. J. W. DEMMEL, *A Numerical Analyst's Jordan Canonical Form*, Dissertation, Computer Science Division, University of California, Berkeley, 1983.

4. J. W. DEMMEL, "On condition numbers and the distance to the nearest ill-posed problem," *Numer. Math.*, v. 51, 1987, pp. 251–289.

5. C. ECKART & G. YOUNG, "The approximation of one matrix by another of lower rank", *Psychometrika*, v. 1, 1936, pp. 211–218.

6. H. FEDERER, *Geometric Measure Theory*, Springer-Verlag, Berlin and New York, 1969.

7. G. H. GOLUB & C. F. VAN LOAN, *Matrix Computations*, Johns Hopkins Univ. Press, Baltimore, MD, 1983.

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

9. A. GRAY, "Comparison theorems for the volumes of tubes as generalizations of the Weyl tube formula," *Topology*, v. 21, 1982, pp. 201–228.

10. P. A. GRIFFITHS, "Complex differential and integral geometry and curvature integrals associated to singularities of complex analytic varieties," *Duke Math. J.*, v. 45, 1978, pp. 427–512.

11. R. W. HAMMING, "On the distribution of numbers," *Bell System Tech. J.*, v. 49, 1970, pp. 1609–1625.

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

13. D. HOUGH, *Explaining and Ameliorating the Ill Condition of Zeros of Polynomials*, Thesis, Mathematics Department, University of California, Berkeley, CA, 1977.

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

15. W. KAHAN, *Conserving Confluence Curbs Ill-Condition*, Technical Report 6, Computer Science Dept., University of California, Berkeley, August 4, 1972.

16. T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin and New York, 1966.

17. K. KENDIG, *Elementary Algebraic Geometry*, Springer-Verlag, Berlin and New York, 1977.

18. D. E. KNUTH, *The Art of Computer Programming*, Vol. 2, Addison-Wesley, Reading, Mass., 1969.

19. E. KOSTLAN, *Statistical Complexity of Numerical Linear Algebra*, Dissertation, Math. Dept., University of California, Berkeley, 1985.

20. P. LELONG, *Fonctions plurisousharmoniques et formes différentielles positives*, Gordon and Breach, Paris, 1968.

21. A. OCNEANU, *On the Volume of Tubes About a Real Variety*, unpublished report, Mathematical Sciences Research Institute, Berkeley, 1985.

22. A. OCNEANU, *On the Loss of Precision in Solving Large Linear Systems*, Technical Report, Mathematical Sciences Research Institute, Berkeley, 1985.

23. J. RENEGAR, "On the efficiency of Newton's method in approximating all zeros of systems of complex polynomials," *Math. Oper. Res.*, v. 12, 1987, pp. 121–148.

24. J. R. RICE, "A theory of condition," *SIAM J. Numer. Anal.*, v. 3, 1966, pp. 287–310.

25. A. RUHE, "Properties of a matrix with a very ill-conditioned eigenproblem," *Numer. Math.*, v. 15, 1970, pp. 57–60.

26. L. A. SANTALÓ, *Integral Geometry and Geometric Probability*, Encyclopedia of Mathematics and Its Applications, Vol. 1, Addison-Wesley, Reading, Mass., 1976.

27. S. SMALE, "The fundamental theorem of algebra and complexity theory," *Bull. Amer. Math. Soc.* (N. S.), v. 4, 1981, pp. 1–35.

28. S. SMALE, "Algorithms for solving equations," presented at the International Congress of Mathematicians, Berkeley, 1986.

29. G. W. STEWART, "Error and perturbation bounds for subspaces associated with certain eigenvalue problems," *SIAM Rev.*, v. 15, 1973, p. 752.

30. G. STOLZENBERG, *Volumes, Limits, and Extensions of Analytic Varieties*, Lecture Notes in Math., vol. 19, Springer-Verlag, Berlin and New York, 1966.

31. P. R. THIE, "The Lelong number of a point of a complex analytic set," *Math. Ann.*, v. 172, 1967, pp. 269–312.

32. B. L. VAN DER WAERDEN, *Modern Algebra*, Vol. 1, Ungar, New York, 1953.

33. H. WEYL, "On the volume of tubes," *Amer. J. Math.*, v. 61, 1939, pp. 461–472.

34. J. H. WILKINSON, *Rounding Errors in Algebraic Processes*, Prentice-Hall, Englewood Cliffs, N.J., 1963.

35. J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, 1965.

36. J. H. WILKINSON, "Note on matrices with a very ill-conditioned eigenproblem," *Numer. Math*, v. 19, 1972, pp. 176–178.

37. J. H. WILKINSON, "On neighboring matrices with quadratic elementary divisors," *Numer. Math.*, v. 44, 1984, pp. 1–21.

38. J. H. WILKINSON, "Sensitivity of eigenvalues," *Utilitas Math.*, v. 25, 1984, pp. 5–76.