BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 13, Number 2, October 1985

ON THE EFFICIENCY OF ALGORITHMS OF ANALYSIS

BY STEVE SMALE[1]

CONTENTS

CHAPTER I

1. Introduction. This is an expanded version of the Jaqueline Lewis talks I gave at Rutgers in April 1984. It is partly an exposition of recent results and new open problems. Also, some new proofs are given here. The subject is the global analysis of algorithms of linear and calculus mathematics, especially in regard to efficiency. This is part of the subject called computational complexity. However, in the past, computational complexity has usually referred to the study of algorithms for discrete problems. In what follows, the problems come from numerical analysis, operations research, and classical mathematics ("continuous" classical mathematics). It is sometimes forgotten how close numerical analysis and classical mathematics are to each other. But to confirm this relationship one can note the frequent appearance of the names of Newton, Lagrange, Gauss in numerical analysis texts, and look at *Goldstine*.

Received by the editors February 7, 1985.
1980 *Mathematics Subject Classification*. Primary 65-02.
[1]I would like to thank N. H. Kuiper and IHES for their hospitality. I would like also to acknowledge NSF support during the preparation of this paper.

0273-0979/85 $1.00 + $.25 per page

The development of computational complexity theory for continuous mathematics may well raise questions for the foundations of computer science. My point of view on this is not new. It is close to that expressed by *Von Neumann* ("The General and Logical Theory of Automata") who wrote:

> We are very far from possessing a theory of automata which deserves that name, that is, a properly mathematical-logical theory.
>
> There exists today a very elaborate system of formal logic, and, specifically, of logic as applied to mathematics. This is a discipline with many good sides, but also with certain serious weaknesses. This is not the occasion to enlarge upon the good sides, which I have certainly no intention to belittle. About the inadequacies, however, this may be said: Everybody who has worked in formal logic will confirm that it is one of the technically most refractory parts of mathematics. The reason for this is that it deals with rigid, all-or-none concepts, and has very little contact with the continuous concept of the real or of the complex number, that is, with mathematical analysis. Yet analysis is the technically most successful and best-elaborated part of mathematics. Thus formal logic is, by the nature of its approach, cut off from the best cultivated portions of mathematics, and forced onto the most difficult part of the mathematical terrain, into combinatorics.
>
> The theory of automata, of the digital, all-or-none type, as discussed up to now, is certainly a chapter in formal logic. It would, therefore, seem that it will have to share this unattractive property of formal logic. It will have to be, from the mathematical point of view, combinatorial rather than analytical.

And Von Neumann went on to say: "...a detailed, highly mathematical and more specifically analytical, theory of automata and of information is needed".

Certainly the numerical analysts have studied speed of computation. However, this has usually been in terms of a rate of convergence, or the cost in asymptotic terms. This contrasts with understanding the *total* cost, as in the subject of computational complexity.

A study of total cost for algorithms of numerical analysis yields side benefits. It forces one to consider global questions of speed of convergence, and in so doing one introduces topology and geometry in a natural way into that subject. I believe that this will have a tendency to systematize numerical analysis. This development could turn out to be comparable to the systematizing effect of dynamical systems on the subject of ordinary differential equations over the last twenty-five years.

The rest of Chapter I is organized around four theorems, named A, B, C and D. The goals and perspectives of these theorems are the same. Experience in the use of algorithms, especially with the computer in recent decades, has given rise to certain practices and beliefs. To give a deeper understanding of this culture, we try to give reasonable underlying mathematical formulations, and eventually to prove theorems, usually confirming the experience of the practitioners. Idealizations and simplifications are made, but we try to keep the essence of the observed phenomena. There is a kind of parallel in this approach to that of theoretical physics. Our primary goal is not the design of new algorithms, but we hope that this deeper understanding will eventually be constructive in that domain too.

2. On efficient zero-finding. Let me discuss an example a numerical analyst might give against the theoretization of his subject. To solve on the machine a system of nonlinear equations, the usual procedure is the following. Choose a starting point at random (but using previous experience in this choice if possible). Then apply some variation of Newton's method iteratively for a while. If that does not work, pick another starting point at random and repeat. This ad hoc procedure seems not to lend itself to the usual kind of theorizing. On the other hand a mixture of probability and global analysis might eventually yield a good understanding of this practice.

At this time Mike Shub and I (*Shub-Smale II*) have a result which shows that for a polynomial of one complex variable, this method works in fact relatively quickly; six random choices are sufficient on the average. Of course, one must spell out the appropriate modification of Newton's method, the number of iterations, etc.

In the following, I use the general idea of *Shub-Smale II*, but simplify and sharpen the result by altering the algorithm.

First the version of Newton's method, which is a little different from that of *Smale III*, *Shub-Smale I* and *II*, is specified. Suppose one is given a complex polynomial f, complex numbers, z and ω, where ω is considered as a parameter. Define $G_\omega : S \to S$, where S is the Riemann sphere, $S = \mathbf{C} \cup \infty$, by

$$G_\omega(z) = z + \frac{\omega - f(x)}{f'(z)}.$$

Thus G_0 is precisely Newton's method. If one uses Newton's method to solve $f(z) - \omega = 0$, then one obtains the iteration G_ω. Consider the problem

Prob(ε, f): Given (ε, f), $0 < \varepsilon < 1$, f a complex polynomial, find $z \in \mathbf{C}$ such that $|f(z)| < \varepsilon$.

Write $f(z) = \sum_{i=0}^{d} a_i z^i$ and suppose that $a_d = 1$, $|a_i| \leqslant 1$. In any case an easy change of variables can put f into this form.

THEOREM A. *On the average six returns of the following algorithm is sufficient to solve* Prob(ε, f) *for any* ε *and any* f (*normalized as above*).

Let $K = 98$.

Alg(ε, f):

(1) *Choose* $z_0 \in \mathbf{C}$ *satisfying* $|z_0| = 3$ *at random.*

(2) *Define* n *as the smallest integer greater than* $K(d \log 3 + |\log \varepsilon|)$ *and* $z_i = G_{\omega_i}(z_{i-1})$, $\omega_i = M^i f(z_0)$, $i = 1, \ldots, n$, *where* $M = 1 - (1/K)$.

(3) *If* $|f(z_n)| < \varepsilon$, *terminate. Otherwise return to* (1).

COROLLARY. (i) *On the average* $6K(d \log 3 + |\log \varepsilon|)$ *is a sufficient number of iterations to solve* Prob(ε, f) *by* Alg(ε, f).

(ii) *The number of arithmetic operations for the same is proportional to* $6Kd(d \log 3 + |\log \varepsilon|)$. *See* Shub-Smale II *for the counting involved.*

"On the average" in all of the above is given in terms of a probability measure on the space of all sequences of choices z_0 with $|z_0| = 3$. To obtain this measure, start with normalized Lebesgue measure on the set $\{z \in \mathbf{C} | |z| = 3\}$, and then take the infinite product measure.

The proof of Theorem A will be given in §5 of Chapter II below.

ADDED IN PROOF. I subsequently noticed that the number of iterations in Theorem A and the corollary could be reduced substantially by simply taking $|z_0| = e^3$ instead of 3 and in §5 (Chapter II) changing the condition $\textcircled{H}_{f,z} > \pi/12$ to $\textcircled{H}_{f,z} > \pi/4$. Then K comes out to about 32, the number of returns close to 2 and the average number of iterations about 2 times $32(3d + |\log \varepsilon|)$.

On the other hand the example $f(z) = z^d$ shows that one can't do better than the number of iterations being linear in d with Newton's method, even when it is globally convergent.

PROBLEM 1. Extend the result of Theorem A to two (or more) variables. *Renegar III* seems to have made a breakthrough on this problem, as this paper was being finished.

PROBLEM 2. Prove an analogous result for $|z_0| \leqslant 1$ rather than $|z_0| = 3$.

This would seem to be a more natural way to start the algorithm and one would expect a sharper estimate. However, the analysis seems difficult; see §5 of Chapter II below and *Shub-Smale II* for more on this problem.

3. On the efficiency of linear programming. We review very briefly certain recent results on the average speed of simplex type methods for the linear programming problem (LPP). One of the standard forms of this problem is

LPP: Find $x \in \mathbf{R}^n$ subject to $x \geqslant 0$, $Ax \geqslant b$ such that x minimizes $c \cdot x$.

Here (A, b, c) are the data, where A is an $m \times n$ matrix, $b \in \mathbf{R}^m$ and $c \in \mathbf{R}^n$.

The simplex method of Dantzig is a fast algorithm for exhibiting an answer to the LPP or showing that no minimum exists. In his book (*Dantzig*) on this subject he wrote (p. 160) these often-quoted lines: "Some believe that for a randomly chosen problem with fixed m (the number of constraints), the number of iterations grows in proportion to n (the number of variables)."

I proved that this was indeed the case (in *Smale IV, V*) using Dantzig's self-dual parametric variant of the simplex method (p. 245 of *Dantzig*). Here is the result in more detail.

The space of the data (A, b, c) of LPP is Cartesian space $\mathscr{D} = \mathbf{R}^{mn} \times \mathbf{R}^m \times \mathbf{R}^n$. For defining the average of functions on this space, just take the normal (or Gaussian) distribution on Cartesian space. Let $\rho_{A,b,c}$ be the number of steps of the self-dual method, defined almost everywhere on $\mathscr{D}$. Then the average of $\rho_{A,b,c}$ on $\mathscr{D}$ is defined and yields a function $\rho(m, n)$.

THEOREM B. *Fix m and $\varepsilon > 0$. Then there is a number $K > 0$ depending on m, ε and*

$$\rho(m, n) \leqslant Kn^{\varepsilon}.$$

Thus the number of steps for a fixed number of constraints grows more slowly than any prescribed root of the number of variables. Take $\varepsilon = 1$ to obtain Dantzig's conjecture, above.

PROBLEM 3. Can one estimate $\rho(m, n)$ by a polynomial function of m and n? or even a linear function?

In this direction there are the following results.

K.-H. Borgwardt since the late seventies has been working on a limited version of the LPP (one in which feasibility is built in by the special form of the constraints). He analysed a simplex-related algorithm he called "Schatteneckenalgorithmus". In *Borgwardt* he obtains an average-case polynomial bound in m and n for this problem.

More recent are results of *Adler-Megiddo*, *Adler-Karp-Shamir*, and *Todd*. They each obtained average-case bounds, quadratic in m and n, for the general LPP using another version of the simplex method. Lower bounds (quadratic) were also shown for this particular algorithm by *Adler-Megiddo*.

PROBLEM 4. Find a theory of average speed for a reasonable general class of algorithms for the LPP. This could allow one to compare different methods for efficiency. In particular, what can one say about the original simplex method as to efficiency?

I have spoken of probability measured in terms of a Gaussian distribution on the space of data. In fact, in the various works cited above, the results are proved with much gentler hypotheses made on the measure.

Some discussion of Dantzig's "self-dual parametric" method is called for. For the last couple of decades the LPP has often been considered as a special case of what is called the linear complementarity problem or LCP. The LCP may be described in this way: If M is an $N \times N$ matrix, define

$$\Phi_M: \mathbf{R}^N \to \mathbf{R}^N \quad \text{by } \Phi_M(x) = x^+ + Mx^-,$$

where x^+ is obtained from x by setting the negative coordinates zero and x^- the positive coordinates zero. Then it is easily seen that Φ_M is linear on orthants, continuous and the identity on the positive orthant.

LCP: Data (M, q), M an $N \times N$ matrix, $q \in \mathbf{R}^N$: Solve$\Phi_M(x) = q$ for $x \in \mathbf{R}^N$.

The LPP is obtained by setting

$$M = \begin{bmatrix} 0 & -A^T \\ A & 0 \end{bmatrix}, \qquad q = (c, -b),$$

and $N = n + m$.

In general the LCP unifies many of the problems of operations research. See the references in *Smale V* for a more detailed discussion of this subject.

The central algorithm for the LCP is due to Lemke and can be described by lifting back via Φ_M the segment joining q_0 to q in $\mathbf{R}^N$, where $q_0 = (1, \ldots, 1)$.

It turns out that in the context of the LCP, the self-dual algorithm coincides with Lemke's. Also, the algorithm studied by *Adler-Megiddo*, *Adler-Karp-Shamir*, and *Todd* may be interpreted as Lemke's algorithm with a different choice of q_0 (whose coordinates are powers of $\varepsilon > 0$ sufficiently small).

For me, there is a further very attractive advantage of this perspective. There is a close relationship between Newton's method and Lemke's method as can be seen in *Eaves-Scarf* and *Smale II*. There is a unity of numerical analysis and

operations research implied by this connection. In the direction of making this connection more concrete is the work of *Kuhn-Wang-Xu* and *Renegar I, II.*

There is a crucial step in the proof of Theorem B which, while easy to prove, is important conceptually. This is as follows. Suppose as above the LPP is imbedded in the LCP.

PROPOSITION. *The average, over* $(b, c) \in \mathbf{R}^m \times \mathbf{R}^n$, *number of steps to solve the* LPP *is given by*

$$\sum_H \text{measure}\{\Phi_M(H), -q_0\},$$

H a hyperorthant.

Here a hyperorthant is the intersection of a coordinate hyperplane in $\mathbf{R}^N$ with an orthant, $\{\Phi_M(H), -q_0\}$ is the convex cone in $\mathbf{R}^N$ generated by $-q_0$ and the elements of the image $\Phi_M(H)$. Recall that the Gaussian measure is used and that

$$M = \begin{bmatrix} 0 & -A^T \\ A & 0 \end{bmatrix}, \qquad q_0 = (1,\ldots,1) \in \mathbf{R}^n.$$

The proof is given in *Smale IV.*

Linear programming has part of its origins in economic theory (especially production) as in work of Leontief and Koopmans. See *Dorfman-Samuelson-Solow* and *Dantzig.* Moreover, in economic equilibrium theory, complexity of decision-making is often taken to be trivial (in contrast to practice). These factors suggest

PROBLEM 5. Relate work on the efficiency of linear programming algorithms more directly with economic theory.

There is a vast amount of literature on the simplex method and, in the last few years, on its average speed. Ron *Shamir* has written an extensive survey on the work to which I will defer, for those wishing to pursue the subject further. Also, there is the paper *Vershik-Sporyshev.*

4. On well-posed linear systems. In solving linear systems of equations, how much is the error in the input going to affect the solution typically? Von Neumann and his various co-authors, Bargmann, Goldstine, and Montgomery, dealt with this problem in three papers amounting to more than 150 pages (see *Von Neumann*). He was concerned with the question: In principle, can linear systems with a large number of variables be solved by computers?

Clearly, for nearly singular systems lack of input precision can make the output error arbitrarily large. This indicates that some kind of average result is called for. As before, for simplicity, we will use the Gaussian distribution to define a probability measure on the space of real $n \times n$ matrices.

Let us recall the notion of a condition number K_A, of a matrix A, as defined by numerical analysts. See especially *Wilkinson I* and *II*, *Forsyth-Moler*, or *Atkinson* for details. Suppose Euclidean norms are taken on Cartesian spaces and their induced operator norms on matrices.

Consider a linear system

$$Ax = b,$$

where A is an $n \times n$ matrix and $b \in \mathbf{R}^n$. One is to solve for $x \in \mathbf{R}^n$ with A and b given.

Regarding A as exact and fixed for the moment, suppose that δb is an error in input producing an error in output δx. For a linear problem it is natural to consider the relative magnitudes $\|\delta b\|/\|b\|$, $\|\delta x\|/\|x\|$. Using the equations $A(x + \delta x) = b + \delta b$, $A(\delta x) = \delta b$, this ratio is

$$\frac{\|\delta x\|/\|x\|}{\|\delta b\|/\|b\|} = \frac{\|A^{-1}(\delta b)\|}{\|\delta b\|} \frac{\|Ax\|}{\|x\|}.$$

An upper bound for this quantity over b, δb is the *condition number* K_A of A, $K_A = \|A\| \|A^{-1}\|$. Thus K_A ranges between 1 and ∞ and is large for ill-conditioned (nearly singular) matrices.

To understand something about the average error of linear systems, one is tempted to average K_A over all matrices. However, this average is infinite. On the other hand, $\log K_A$ has a finite average. But most importantly, $\log K_A$ has a direct computational interpretation, as we will see now.

To talk about precision, or the number of digits of accuracy, the use of logs is called for. This would be log base 2 for binary numbers or log base 10 for decimals. For mathematical convenience we will always use natural logs.

In speaking of the linear system $Ax = b$ with A fixed and input error δb, a reasonable series of definitions is:

DEFINITIONS:

$$\text{Input precision} = -\log \|\delta b\|,$$

$$\text{Relative input precision} = -\log(\|\delta b\|/\|b\|) = -\log\|\delta b\| + \log\|b\|,$$

$$\text{Relative output precision} = -\log(\|\delta x\|/\|x\|),$$

$$\text{Loss of precision} = \text{relative input precision} - \text{relative output precision}.$$

Thus from the definitions it follows that

$$\text{Loss of precision} = \log\left(\frac{\|A^{-1}(\delta b)\|}{\|\delta b\|} \frac{\|Ax\|}{\|x\|}\right).$$

To define a quantity L_A which depends only on the matrix A one takes the worst case of b, δb as above in the definition of condition number. Thus

$$L_A = \log K_A$$

and we have shown that L_A is the greatest loss of precision that the system $Ax = b$ can exhibit for fixed A.

The question is thus posed, what is the expected value of L_A?

I gave a lecture in my seminar in February 1984 focusing on this question and stating an explicit integral expression for this average in terms of the singular values of A. Three different estimates (upper bounds) were given to

me before the next week's meeting by L. Blum, E. Kostlan, and A. Ocneanu. Eventually this distilled into the following result:

THEOREM C. *Let $L(n)$ be the average of L_A over real $n \times n$ matrices.*
(i) KOSTLAN.

$$L(n) \leqslant 1 + \tfrac{5}{2} \log n.$$

(ii) OCNEANU. *Given any $\varepsilon > 0$, there is n_0 such that for $n > n_0$*

$$(\tfrac{2}{3} - \varepsilon)\log n \leqslant L(n).$$

The Blum estimate $L(n) \leqslant 3n \log n$ was obtained using the general theorem in *Blum-Shub* which dealt with precision for evaluating rational functions. *Ocneanu* obtained the upper bound, for any $\varepsilon > 0$,

$$L(n) \leqslant (3 + \varepsilon) \log n \quad \text{if } n > n_0(\varepsilon).$$

The estimate of $\frac{5}{2} \log n + 1$ in Theorem C is quite reasonable and even for 100 variables the average loss of precision is not badly estimated. But $L(n)$ contains in its definition a worst case; moreover, the variance is an issue. See §2 of Chapter III for these matters.

ACKNOWLEDGMENT. Conversations with L. Blum, A. Grunbaum, E. Kostlan, and A. Ocneanu were helpful to me in developing the ideas in this section.

5. On efficient approximation of integrals. In first-year calculus, three numerical methods are often given for approximating an integral of a continuous function f, say, for simplicity, on the interval $[0, 1]$. These are:

$$\text{Riemann integral,} \qquad R_h(f) = h \sum_{i=1}^{n} f(ih),$$

$$\text{Trapezoidal method,} \qquad T_h(f) = h\left[-\frac{1}{2}(f(1) + f(0)) + \sum_{i=1}^{n} f(ih)\right],$$

$$\text{Simpson's rule,} \qquad S_h(f) = \frac{h}{3}\left[f(1) + f(0) + 2\sum_{i=1}^{2n-1} f(ih) + 2\sum_{i=1}^{n} f((2i-1)h)\right].$$

Here h is the step size, so that $h = 1/n$ for the first two methods and $h = 1/2n$ for Simpson's rule, n a positive integer.

This section is concerned with the average cost of these algorithms; the practice of numerical analysts is confirmed in these cases. In fact, some simple exact formulas on the average cost (not just asymptotic cost) are produced.

This work was suggested to me as I tried to understand the theory of *Traub-Wozniakowski*. Conversations with them were especially helpful.

Besides the help of Traub and Wozniakowski, conversations with A. Calderon, P. Collet, J. Franks, M. Shub, and especially David Elworthy in Caracas, July 1984 (where I found these results) were important for me. So also were conversations with Feng Gao and Nat Smale.

An important earlier paper in this area is: F. M. Larkin, *Gaussian measure in Hilbert space and applications in numerical analysis*, Rocky Mountain J. Math. **2** (1972), 379–421..

Here is a precise statement of my conclusions. The three algorithms Riemann, Trapezoid, and Simpson define respectively maps R_h, T_h, S_h: $C^0[0,1] \to \mathbf{R}$ where $C^0[0,1]$ is the space of continuous functions on the interval $[0,1]$. Let $J: C^0[0,1] \to \mathbf{R}$ denote the integral $J(f) = \int_0^1 f(s)\,ds$.

For step size h, the error in computing $J(f)$ with the Riemann approximation is given by

$$\varepsilon_R(h,f) = |J(f) - R_h(f)|.$$

Similarly

$$\varepsilon_T(h,f) = |J(f) - T_h(f)|, \qquad \varepsilon_S(h,f) = |J(f) - S_h(f)|.$$

Of course, for $f \in C^0[0,1]$, these quantities tend to zero as $h \to \infty$. For averaging these error functions, we need a probability on function spaces. To this end, the Gaussian measure on Hilbert space is used. See §1 of Chapter III below and its references.

The Hilbert spaces of functions natural to this problem are called the Sobolev spaces $\mathscr{H}^k$ for $k = 1, 2, \ldots$. These are defined as follows for $n = 1, 2$:

Let f', f'' be the first and second derivatives of f, respectively.

$$\mathscr{H}^1 = \left\{ f \in C^0[0,1] \,\middle|\, f' \text{ is defined almost everywhere and } \int |f'|^2 < \infty \right\}.$$

The inner product on $\mathscr{H}^1$ is

$$\langle f, g \rangle_{\mathscr{H}^1} = f(0)g(0) + \int_0^1 f'(s)g'(s)\,ds.$$

Similarly

$$\mathscr{H}^2 = \left\{ f \in C^1[0,1] \,\middle|\, f'' \text{ is defined almost everywhere and } \int |f''|^2 < \infty \right\}$$

with inner product

$$\langle f, g \rangle_{\mathscr{H}^2} = f(0)g(0) + f'(0)g'(0) + \int_0^1 f''(s)g''(s)\,ds.$$

Now it is possible to average the errors displayed above and to obtain

$$\varepsilon_R^k(h) = \operatorname*{Av}_{f \in \mathscr{H}^k} \varepsilon_R(h,f),$$

$\varepsilon_T^k(h)$ and $\varepsilon_S^k(h)$ described similarly.

The cost of implementing each of these three algorithms is essentially the same for given h, and, moreover, this cost is proportional to $1/h$. To see these facts one can count the arithmetic operations in the expressions for R_h, T_h and S_h.

THEOREM D. (i) *On the average for $\mathscr{H}^1$ functions, to obtain the same accuracy, the trapezoidal method costs one half as much as the Riemann approximation. More precisely*

$$\varepsilon_R^1(h) = \sqrt{\frac{2}{\pi}}\,\frac{h}{\sqrt{3}}, \qquad \varepsilon_T^1(h) = \frac{1}{2}\varepsilon_R^1(h) \quad \text{all } h = \frac{1}{n}.$$

(ii) *On the average for $\mathscr{H}^2$ functions, Simpson's rule is much cheaper than Riemann integration. One has the precise formulae*

$$\varepsilon_R^2(h) = \sqrt{\frac{2}{\pi}}\,\frac{h}{\sqrt{6}}\left(1 + \frac{h}{2} + \frac{h^2}{10}\right)^{1/2} \quad \text{all } h = \frac{1}{n}.$$

$$\varepsilon_S^2(h) = \sqrt{\frac{2}{\pi}}\,\frac{h^2}{3\sqrt{15}} \quad \text{all } h = \frac{1}{2n}.$$

The proof can clearly be extended to give a much broader body of results. The approach here might lead to a more systematic way of analysing the cost of numerical algorithms where the mesh size h is the principal parameter, as for example, difference methods for approximating solutions of partial differential equations. The proof of Theorem D will be given in §1 of Chapter III.

CHAPTER II

1. Convergence of Newton's method. Consider a complex polynomial, $f(z) = \sum_{i=0}^{d} a_i z^i$, the a_i complex numbers. Denote by S the Riemann sphere, the complex numbers $\mathbf{C}$ with "∞" adjoined. Define a rational endomorphism (i.e. a map $z \to P(z)/Q(z)$, P, Q polynomials, of S into itself) $N_f\colon S \to S$ by

$$N_f(z) = z - \frac{f(z)}{f'(z)}.$$

Newton's method can be viewed as iterating this map starting with some $z_0 \in S$. That is, z_n is produced inductively by $z_n = N_f(z_{n-1})$, $n = 1, 2, \ldots$. We can also write $z_n = N_f^n(z_0)$, where N_f^n is the composition of N_f with itself n times. The global study (over all $z_0 \in S$) of Newton's method is thus the same as the study of N_f as a dynamical system. The early work in dynamical systems of one complex variable of Cayley, *Fatou*, and *Julia* was in large part motivated by Newton's method; see *Peitgen-Saupe-V. Haeseler*. The dynamical system point of view will become apparent as we proceed.

PROPOSITION 1. *The number $\zeta \in \mathbf{C} \subset S$ is a fixed point of N_f if and only if ζ is a zero of f. In this case, the derivative $N_f'(\zeta) = (m-1)/m$, where m is the multiplicity of that zero. Moreover, the degree of N_f is the number of distinct zeros of f.*

Here the *degree* of a rational map $T = P/Q$ of S into itself is the maximum of the degrees of polynomials P and Q, where we assume P and Q have no common factors.

The above proposition is very well known and easy to prove. The proof of the second part can be obtained by expanding f as a Taylor's series about ζ; i.e. use $f(z) = a(z-\zeta)^m + \cdots$ in the expression for N_f.

Note that always $|N_f'(\zeta)| < 1$ and it follows that there is a neighborhood U of ζ such that, for any $z \in U$, $N_f^n(z)$ is finite for all positive integers n and converges to ζ as n tends to ∞. The complex number ζ is called a sink or an attractive fixed point of N_f. The open set $B = \bigcup_{n \geq 0} N_f^{-n}(U)$ is called the *basin* of ζ.

In the general case of distinct roots of f, ζ has multiplicity one as a zero of f, $N_f'(\zeta) = 0$ and ζ is "superattractive". In the numerical analysis literature, Newton's method is said to converge quadratically. The converse of Proposition 1 was recently proved by *Saunders* (for the first time, as far as I know).

PROPOSITION 2. *Let T: $S \to S$ be a rational endomorphism such that at each fixed point ζ of T, the derivative $T'(\zeta) = (m-1)/m$ for some positive integer m. Also suppose that the degree of T equals the number of these fixed points. Then there is a polynomial f such that $T = N_f$.*

The following goes back to Cayley (see e.g. *Peitgen-Saupe-V. Haeseler*).

PROPOSITION 3. *Suppose f is a quadratic polynomial with distinct roots. Then N_f is conjugate to T: $S \to S$ with $T(\omega) = \omega^2$ by a linear fractional (or Möbius) transformation g.*

PROOF (which I learned from Gregg Saunders). Let α and β be the two roots which must go to ∞, 0 respectively under g. Let $g(z) = (z-\beta)/(z-\alpha) = \omega$. Then it can be checked that $g(N_f(z)) = \omega^2 (= T(\omega))$.

From Proposition 3, the dynamics of N_f and T are qualitatively the same. For T, there are two fixed points, ∞ and 0, both attractive. The circle $|\omega| = 1$ is invariant, and if $|\omega| < 1$, then $T^n(\omega) \to 0$; if $|\omega| > 1$, then $T^n(\omega) \to \infty$. Thus under iteration by T, every point converges to 0 or ∞ except for the unit circle. Now this circle $|\omega| = 1$ under the linear fractional transformation g corresponds to the straight line in $\mathbf{C}$ which is the perpendicular bisector of the segment between α and β. This implies that except for this line every point in $\mathbf{C}$ converges to α or β under iteration by N_f. We can say that N_f is "generally convergent" for quadratic polynomials (the case of quadratic f with coincident zeros is simpler).

For polynomials of higher degree, Newton's method is not generally convergent in any reasonable sense. To see this well-known fact I will find a polynomial f and a sink for N_f of least period two.

A *sink* α of *period* k for a rational endomorphism T is defined by the conditions $T^k(\alpha) = \alpha$ and $|(T^k)'(\alpha)| < 1$. In this case one can easily show that there is a neighborhood U of α consisting of z satisfying $(T^k)^n(z) \to \alpha$ as $n \to \infty$. So if $k > 1$ and the points $\alpha, T(\alpha), \ldots, T^{k-1}(\alpha)$ are distinct and $z \in U$, the iterates $T^i(z)$ do not converge to the fixed points of T, but asymptotically cycle about the $T^i(\alpha)$, $i = 0, \ldots, k-1$.

A similar situation will prevail for T_0 near T. Thus if N_f has a sink of least period two, then one can say fairly that N_f is *not generally convergent*. Conditions on f for 0 to be a sink for N_f of least period two will be studied to prove the well-known

PROPOSITION 4. *Newton's method is not generally convergent.*

PROOF. Let $f(z) = \sum_{i=0}^{d} a_j z^j$ and fix $a_0 = 1$, $a_1 = -1$, $a_2 = 0$ so that $N_f(0) = 1$, $N_f'(0) = 0$ and if $N_f'(1) \neq \infty$, then $(N_f^2)'(0) = 0$ by the chain rule. Thus if $N_f(1) = 0$ and $N_f'(1) \neq \infty$ then 0 is a sink of least period two (and superattractive, even). The condition $N_f(1) = 0$ is seen to be:

$$-1 + 2a_3 + 3a_4 + \cdots + (d-1)a_d = 0,$$

and the condition $N_f'(1) \neq \infty$ is satisfied if $f'(1) \neq 0$ or

$$-1 + 3a_3 + 4a_4 + \cdots + da_d \neq 0.$$

These conditions are satisfied by a dense open subset of a hyperplane of $\{(a_3,\ldots,a_d)\} = \mathbf{C}^{d-2}$ space.

As a special case let $d = 3$ and take $a_3 = 1/2$ to obtain

$$f_0(z) = \tfrac{1}{2}z^3 - z + 1.$$

Then from what we have shown, starting with points near 0, Newton's method for f_0 will approximately cycle between 0 and 1 forever. Note that this example works over the real as well as the complex numbers in the same way.

The robustness of the sink of period two will carry over to yield that N_f for f near f_0 has periodic sinks as well. Thus even on degree three polynomials Newton's method is not generally convergent. This proves Proposition 4.

In the above proof 0 and 1 were chosen for the periodic sink because of difficulty dealing with the composition $N_f \circ N_f$.

PROBLEM 6. Find more systematically the set of f whose Newton's endomorphism N_f has periodic sinks, not fixed (and thus fails to generally converge).

Proposition 4 relates to the problem discussed in §4 below: "Is there *any* purely iterative generally convergent (complex) algorithm for polynomial zero finding?" "Generally convergent" will be defined then precisely.

To make precise the idea of "close polynomials" used above, one needs to define a space of polynomials. Two ways of doing this present themselves. If f is a polynomial and λ a nonzero complex number, then f and λf have the same zeros, and the same critical points (zeros of the derivative). Moreover, $N_{\lambda_f} = N_f$, so that Newton's method is the same. Thus it is natural to identify f and λf, and consider the projective space $\mathbf{P}_d(\mathbf{C})$, complex of dimension d, of polynomials of degree less than or equal to d. To make this explicit, let $f(z) = \sum_{i=0}^{d} a_i z^i$, and identify $(n+1)$-tuples of complex numbers $(a_0,\ldots,a_d)$ and $(\lambda a_0,\ldots,\lambda a_d)$ to obtain $\mathbf{P}_d(\mathbf{C})$.

Oftentimes, it is handier to use another space, used in *Smale III* and *Shub-Smale I, II*. This is the space of all polynomials $\sum_{i=0}^{d} a_i z^i$ with $a_d = 1$ and $|a_i| \leqslant 1$, which we denote by $P_d(1)$. It is represented by a bounded polycylinder in $\mathbf{C}^d = \{(a_0,\ldots,a_{d-1})\}$.

Given any polynomial $f(x) = \sum_0^d a_i z^i$, $a_d \neq 0$, the transformation $z \to \alpha z = \omega$, $\alpha \in \mathbf{C}$, for appropriate α will transform f into $f_\alpha(\omega) = \sum_{i=0}^{d} b_i \omega^i$ with $|b_d| \geqslant |b_i|$ all i. Then further division by b_d will put the polynomial into $P_d(1)$. The roots are all changed by the factor α. This gives some justification for using $P_d(1)$.

PROBLEM 7. (Compare *Smale III*, Problem 7.) What is the probability that Newton's method will converge for a random choice of initial point? For a given polynomial f? For an average polynomial?

While there are several reasonable ways to make this into a precise mathematical problem, we will proceed using the space $P_d(1)$.

Let $f(z) = \sum_{i=0}^{d} a_i z^i$, $a_d = 1$ and $|a_i| \leqslant 1$. Denote by B_f the union of the basins of the sinks of N_f. Thus $z \in B_f$ if and only if $N_f^n(z)$ tends to a zero of f

as n goes to infinity. Recall that $D_R = \{z \in \mathbf{C} \mid |z| < R\}$, and define

$$A_f = \frac{\text{area of } B_f \cap D_2}{\text{area of } D_2}.$$

Here R is chosen to be 2 by virtue of the well-known (cf. *Henrici* or *Ostrowski*) fact that $|\zeta| < 2$ if $f(\zeta) = 0$.

This number A_f can be interpreted as the probability that Newton's method will converge for a random point in D_2. Let

$$A_d = \min_{f \in P_d(1)} A_f.$$

Of course $0 \leqslant A_d \leqslant 1$. By Proposition 3, $A_2 = 1$, and by Proposition 4,

$$A_d < 1 \text{ for } d > 2.$$

I would conjecture that $A_d > 0$, all d, but have not proved $A_3 > 0$.

PROBLEM 7.A. Estimate A_d as a function of d.

Thus A_d represents the "worst case" of the A_f. However, A_f is not continuous in f at f with multiple roots. It would seem likely that A_f becomes smaller in any neighborhood of the polynomial $f_0(z) = z^d$.

One can also average A_f over $P_d(1)$, using normalized Lebesgue measure as a probability measure to obtain $\hat{A}_d$.

PROBLEM 7.B. Estimate $\hat{A}_d$ as a function of d.

The later discussion on "approximate zeros" has some bearing on this problem.

As mentioned earlier, the global analysis of Newton's method for solving $f(\zeta) = 0$, with f a complex polynomial, is closely related to the work of Fatou and Julia on iteration of rational endomorphisms, and recent work in that area. Some of that general theory can be used to clarify the convergence question of Newton's method. The main cause of lack of convergence of N_f is due to periodic sinks other than fixed points.

The following theorem from the *Fatou, Julia* theory gives a limitation on the number of such sinks (see also *Blanchard* and *Guckenheimer*).

THEOREM 1. *The immediate basin of a periodic sink of a rational endomorphism T: $S \to S$ must contain at least one critical point of T. Moreover, if every critical point of T lies in some basin of a periodic sink, then these basins have full measure in S (and Axiom* A *is satisfied).*

Let us define these terms. If $T(\zeta) = \zeta$ and $|T'(\zeta)| < 1$, then ζ is a fixed sink. The *basin* B of ζ is the set of z such that $N_f^n(z) \to \zeta$ as $n \to \infty$. The *immediate basin* $\hat{B}_{\zeta,T}$ is the component of B containing ζ. If $T^k(\zeta) = \zeta$ and $|(T^k)'(\zeta)| < 1$ then $\{\zeta, T(\zeta), \ldots, T^{k-1}(\zeta)\}$ is a sink of period k and its immediate basin is the union

$$\bigcup_{i=0}^{k-1} \hat{B}_{T^i(\zeta),T^k}.$$

A critical point z of T is simply $z \in \mathbf{C}$ satisfying $T'(z) = 0$. By differentiating this, it is clear that there are at most $2d - 2$ critical points of T, where d is the

degree of T, and so Theorem 1 implies the existence of at most $2d - 2$ periodic sinks.

Consider now the case of Newton's method $T = N_f$. The degree of N_f is less than or equal to the degree of the polynomial f, less in case of multiple roots. The derivative of N_f is

$$N_f'(z) = \frac{f(z)f''(z)}{f'(z)^2}$$

for any z in $\mathbf{C}$. Confining our study to the case of f with distinct roots $\zeta_1,\ldots,\zeta_d$, we see that d of the $2d - 2$ critical points of N_f are the ζ_i and the remaining $d - 2$ are the *inflection points* of f, i.e. the zeros of f''. By Theorem 1, the general convergence of N_f will depend crucially on whether these inflection points lie in the basins of the zeros of f.

For example, consider the polynomial $f(z) = az^d - bz$. Here of course 0 is a zero of f, but every inflection point is 0, too. Thus by Theorem 1 there is no periodic sink except for the zeros of f, and Newton's method for f converges for a set of initial points of full measure. A more general interesting case comes from the theory of *Barna*.

THEOREM 2 (BARNA). *If f has all roots real then the inflection points of f lie in the immediate basins for N_f of the roots $\zeta_1,\ldots,\zeta_d$ of f. Moreover, except for a cantor set K of real numbers, every real number converges under N_f to a zero of f.*

The dynamics of N_f on K has been recently studied by *Saari-Urenko* and *Saunders*. Both of these references give an extended account of these questions. See also *Curry*, *Curry-Garnett-Sullivan*, *Douady*, *Martin-Hurley*, *Peitgen-Saupe-V. Haeseler*, and *Sullivan* for papers very relevant to this subject.

2. A short elementary proof of the fundamental theorem of algebra and the topology of polynomials. I would like to present a proof of the fundamental theorem of algebra which is important for complexity theory. This proof is suggested in *Smale III*. It does not use any results from topology (it is topology!). The proof is implemented as a fast algorithm in Theorem A above and the next section.

The main tool is the inverse function theorem for one variable which is used in the following form. Let f be a complex polynomial and $z \in \mathbf{C}$ with $f'(z) \neq 0$. Then there is a $\delta > 0$ and a complex analytic map

$$f_z^{-1}\colon D_\delta(f(z)) \to \mathbf{C}$$

with $f_z^{-1}(f(z)) = z$ and $f(f_z^{-1}(\omega)) = \omega$ for all $\omega \in D_\delta(f(z))$. Here $D_\delta(f(z))$ is the set of all ω such that $|\omega - f(z)| < \delta$.

Moreover, if z ranges over a closed and bounded set $K \subset \mathbf{C}$ where f' is never zero, then the corresponding $\delta > 0$ can be chosen independent of $z \in K$. The proof of this last fact can be obtained by taking a convergent subsequence as in usual compactness arguments.

Another elementary fact is assumed: If f is as above, then the derivative f' has a finite number of zeros, the *critical points* of f.

THEOREM 3 (FTA). *A complex polynomial f has a zero.*

PROOF. We use the above facts and notation. Assume $f(\theta) \neq 0$ for all critical points θ of f. Otherwise such a θ is our desired zero and we are finished.

LEMMA 1. *There is a segment l in* $\mathbf{C}$ *joining* 0 *to some number of the form* $f(z_0)$ *which meets no critical value* (*number of the form* $f(\theta)$, θ *a critical point*).

PROOF OF LEMMA 1. Let $z_1 \in \mathbf{C}$ with $f'(z_1) \neq 0$. From the inverse function theorem it follows that there are an infinite number of z_i near z_1 with $f(z_i)$ lying on distinct rays through 0. Only a finite number can meet a critical value, which proves the lemma.

LEMMA 2. *The set* $f^{-1}(l)$ *is closed and bounded.*

This follows from the fact that f is a polynomial and its behaviour near ∞ is dominated by its highest-order term.

From the lemmas and the inverse function theorem, take $\delta > 0$ so that the inverse f_z^{-1}: $D_\delta(f(z)) \to \mathbf{C}$ is defined for each $z \in f^{-1}(l)$. Let n be a positive integer with $n > |f(z_0)|/\delta$.

Define ω_i, $i = 0,\ldots,n$, by

$$\omega_i = \frac{(n-i)f(z_0)}{n}$$

and observe that $|\omega_i - \omega_{i-1}| < \delta$, for $i = 1,\ldots,n$; thus

$$\omega_i \in D_\delta(\omega_{i-1}).$$

Now one can define inductively

$$z_i = f_{z_{i-1}}^{-1}(\omega_i), \qquad i = 1,2,\ldots,n.$$

Since $f(z_i) = \omega_i, f(z_n) = \omega_n = 0$. Q.E.D.

Now a certain problem on the topology of polynomials will be considered. Let f be a complex polynomial. Newton's method is an Euler approximation to solutions of the ordinary differential equation $dz/dt = -\text{grad}|f(z)|^2$ in $\mathbf{C}$ (see *Smale III*). Let φ_t be the one-parameter group of solutions, and for a critical point θ of f define

$$W^u(\theta) = \{z \in \mathbf{C} | \varphi_t(z) \to \theta \text{ as } t \to -\infty\}.$$

Then the closure of the union of $W^u(\theta)$ over all critical points θ is an oriented planar graph which I will call Γ_f. This graph is the same as a "diagram" in *Smale I*. Its vertices are the zeros and critical points of f. This graph carries the essence of the qualitative picture of the "Newton Flow".

Figure 1 shows some examples of Γ_f, where 0 denotes a critical point and X a zero of f. Assume that the zeros of f are distinct.

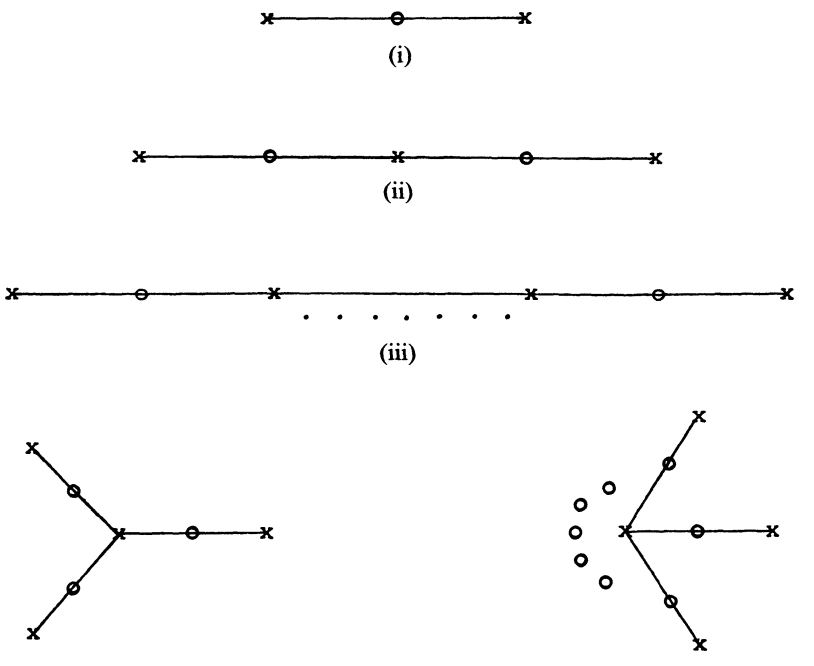

FIGURE 1. Graphs Γ_f for (i) quadratic f, real roots; (ii) cubic f, real roots; (iii) the case of *Barna*, all roots real; (iv′ and iv″) the case $f(z) = z^d - z$, for $d = 4$ and for general d.

One may prove without difficulty the following

PROPOSITION. *Suppose f is a polynomial with no two critical values lying on the same ray. Then*

(i) *Γ_f is planar (of course).*

(ii) *Γ_f is connected.*

(iii) *There are exactly two edges ending on each critical point of Γ_f. There are no edges between the critical points or between the zeros.*

(iv) *Γ_f has no cycles.*

PROBLEM 8. What (abstract) graphs Γ_f can occur for polynomials f? For polynomials with no two critical values on the same ray?

For polynomials with this last property, are the conditions of the proposition complete?

The first case of the last question is for $d = 5$. I don't know a polynomial with the graph shown in Figure 2.

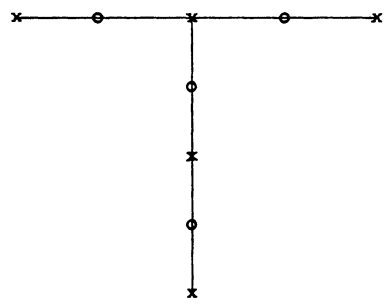

FIGURE 2

ADDED IN PROOF. See also the work of H. Th. Jongen, P. Jonker and F. Twilt, for example, *The continuous, desingularized Newton-method for meromorphic functions* Memorandum No. 501, Twente Univ. of Technology (1985).

More recently M. Shub and R. Williams have contributed to this problem.

3. Fast convergence of Newton's method. If ζ is a simple zero of a polynomial f then we have seen (Proposition 1 of §1) that $N_f(\zeta) = \zeta$ and $N_f'(\zeta) = 0$. This implies by Taylor's Theorem that $|N(\zeta) - \zeta| \leqslant c|z - \zeta|^2$ for $|z - \zeta| < \varepsilon$. Thus locally Newton's method converges fast (quadratically) to a simple zero. Yet, the constants c, ε depend on the coefficients of f in ways which often make it difficult to use this fact. The notion of "approximate zero" gives a more elegant and more useful way to deal with this fast convergence.

Motivation also comes from a different direction. Since algorithms won't in general yield exact zeros of polynomials, one seeks a replacement of the notion of a zero by a notion which will be effectively as good. "Approximate zero" is a reasonable candidate for this in view of Theorem 4, the main result of this section.

One can be concerned that multiple zeros are excluded from this framework. However, locating a multiple zero or an almost multiple zero is a badly posed problem in a certain sense. The "loss of precision" in this problem is infinite, or arbitrarily large, respectively (see §2 of Chapter 3).

Define

$$D_r(\omega_0) = \{\omega \in \mathbf{C} \mid |\omega - \omega_0| < r\} \text{ and } D_r = D_r(0).$$

Recall that if $f'(z) \neq 0$, f_z^{-1}: $D_r(f(z)) \to \mathbf{C}$ is the inverse of f, defined locally by the inverse function theorem, which sends $f(z)$ to z. Let $r = r(f_z^{-1})$ be maximal, or what is the same thing, let r be the radius of convergence of f_z^{-1} as a power series. Then

DEFINITION. The set of *approximate zeros* of a polynomial f, denoted by Ω_f, is

$$\Omega_f = \bigcup_{\substack{\zeta \\ f(\zeta)=0}} f_\zeta^{-1}(\Lambda_\zeta), r_\zeta = r\left(f_\zeta^{-1}\right), \quad \Lambda_\zeta = D_{r_\zeta/9},$$

The 9 seems ad hoc, but in fact enters naturally as we will see. The definition is justified by the next theorem.

THEOREM 4 (APPROXIMATE ZERO THEOREM). *Let z_0 be an approximate zero of f and $z_n = N_f^n(z_0)$. Then there is a $b < 1$, and*

$$|f(z_n)| < b^{2^n-1}|f(z_0)| \quad \text{all } n.$$

REMARKS. (1) Here b is defined by the condition $|f(z_0)| = b(r_\zeta/9)$, where ζ is the zero of f for which $z_0 \in f_\zeta^{-1}(\Lambda_\zeta)$. From the definition of approximate zero, the sets $f^{-1}(\Lambda_\zeta)$ must be disjoint.

(2) Versions of the notion of approximate zero and the above theorem are in *Smale III* and *Shub-Smale I*.

(3) Note the extremely rapid convergence in the estimate of the theorem.

Since there are no excess constants in the theorem, the convergence is more than asymptotically fast. If one prefers $b < \frac{1}{2}$ just replace the 9 by 18 in the definition.

(4) The result works for complex analytic functions $f: \mathcal{U} \to \mathbf{C}$ as well; here $\mathcal{U}$ is an open set of $\mathbf{C}$. In this case $r(f_z^{-1})$ is the maximal r for which f_z^{-1}: $D_r(f(z)) \to \mathcal{U}$ is defined analytically with $f_r^{-1}(f(z)) = z$ and $f(f_z^{-1}(\omega)) = \omega$ for $\omega \in D_r(f(z))$.

Now, I will give the proof of the theorem on approximate zeros. For that the following estimate is used. Let $\mathcal{U}$ be an open set of complex numbers, $z \in \mathcal{U}$, and $f: \mathcal{U} \to \mathbf{C}$ an analytic function with $f'(z) \neq 0$. Let z' denote the Newton iterate $z' = z - f(z)/f'(z)$.

THEOREM 5. *If $r > 4|f(z)|$ where $r = r(f_z^{-1})$, then*

$$\left|\frac{f(z')}{f(z)}\right| \leqslant \frac{1}{r/4|f(z)| - 1}.$$

Here the 4 is sharp.

Versions of Theorem 5 appear in *Smale III* and *Shub-Smale I*.

For the proof of Theorem 2 we need (see *Hayman*)

LOEWNER'S THEOREM. *Let g be a Schlicht function, i.e. $g(\omega) = \sum_{i=1}^{\infty} b_i\omega^i$ for $|\omega| < 1$, $b_1 = 1$ and g is one-to-one on the set of ω with $|\omega| < 1$. Let $f(z) = \sum_{i=1}^{\infty} a_i z^i$ be the inverse to g (so $fg(\omega) = \omega$ and $gf(z) = z$ for sufficiently small $|\omega|$ and $|z|$). Then*

$$|a_k| \leqslant 4^{k-1} \quad \text{all } k.$$

Actually the conclusion of Loewner's Theorem is usually stated a little differently, i.e. that

$$|a_k| \leqslant B_k, \quad B_k = 2^k \frac{1 \cdot 3 \cdots (2k-1)}{1 \cdot 2 \cdots (k+1)}, \qquad k = 1, 2, 3, \ldots.$$

In this form, Loewner shows that the B_k are the best possible, using the Koebe function for g. It follows that in the version $|a_k| \leqslant C^{k-1}$, $C = 4$ is the best possible.

An easy application of the Loewner Theorem is as follows (cf. *Smale III*).

EXTENDED LOEWNER THEOREM. *Let* $f\colon \mathbf{C} \to \mathbf{C}$ *be analytic and* $z_0 \in \mathbf{C}$. *Then*

$$\left|\frac{f^{(k)}(z_0)}{k!f'(z_0)^k}\right| \leqslant \left(\frac{4}{r}\right)^{k-1}, \qquad k = 2,3,\ldots$$

where $r = r(f_{z_0}^{-1})$ *is as above.*

For the proof we may assume $z_0 = \omega_0 = 0$, so that $f^{(k)}(z_0)/k! = a_k$ and $f'(z_0) = a_1$. Now define

$$G(u) = \frac{r}{g'(\omega_0)} g\left(\frac{u}{r}\right)$$

for $|u| < 1$ and apply the previous theorem to G. Here g is the inverse of f on $D_r(0)$.

Theorem 5 is proved using the extended Loewner Theorem. Since

$$f(z') = \sum_{i=0}^{a} \frac{f^{(k)}(z)}{k!}(z' - z)^k,$$

$$\frac{f(z')}{f(z)} = \sum_{k=2}^{\infty} (-1)^k \frac{f^{(k)}(z)}{k!} \frac{f(z)^{k-1}}{f'(z)^k}.$$

Therefore

$$\left|\frac{f(z')}{f(z)}\right| \leqslant \sum_{k=2}^{\infty} \left(\frac{4|f(z)|}{r}\right)^{k-1} \leqslant \frac{1}{r/4|f(z)| - 1},$$

which proves Theorem 5.

Still working toward the proof of Theorem 4, we need another result from Schlicht function theory. See *Hille* and *Shub-Smale I* for the very slightly generalized theorem of Koebe and Bieberbach.

KOEBE-BIEBERBACH THEOREM. *Let* $g\colon D_r \to \mathbf{C}$ *be one-to-one and analytic. Then the image* $g(D_r)$ *contains a disk of radius* $|g'(0)|r/4$ *about* $g(0)$.

Here the 4 is sharp, again by the Koebe function.

LEMMA. *Let* $f(\zeta) = 0$, $z \in f_\zeta^{-1}(\Lambda_\zeta)$ *and* $z' = z - f(z)/f'(z)$. *Then* $z' \in f_\zeta^{-1}(\Lambda_\zeta)$ *and* $|f(z')| < |f(z)|$. *Moreover,* $r \geqslant \frac{8}{9} r_\zeta$, *where* $r = r(f_z^{-1})$.

PROOF. Since $|f(z)| < r_\zeta/9$, the last sentence is immediate from the definitions. Then $r > 8|f(z)|$ as well. Thus Theorem 5 applies to yield

$$\left|\frac{f(z')}{f(z)}\right| \leqslant \frac{1}{r/4|f(z)| - 1} = \frac{4|f(z)|}{r - 4|f(z)|} < 1.$$

It remains to show that $z' \in f_\zeta^{-1}(\Lambda_\zeta)$. From what we have just proved it is sufficient to show $z' \in f_\zeta^{-1}(D_{r_\zeta})$. But by the Koebe-Bieberbach theorem, since $z' - z = -f(z)/f'(z)$, it follows that

$$|z' - z| \leqslant \frac{r_\zeta}{9|f'(z)|},$$

and the proof of the lemma is finished.

Now the proof of Theorem 4 can be finished. Let $z_0 \in f_\zeta^{-1}(\Lambda_\zeta)$ and $z_n = N_f^n(z_0)$, $n = 1, 2, \ldots$. The lemma assures us that z_n is an approximate zero, all n, and even that $z_n \in f_\zeta^{-1}(\Lambda_\zeta)$. Let $r_n = r(f_{z_n}^{-1})$. It follows from the lemma that $r_n \geqslant \frac{8}{9} r_\zeta$, all $n > 0$. Let $|f(z_0)| = b(r_\zeta/9)$, so $b < 1$.

Proceeding by induction, the case $n = 0$ of the theorem is trivial. For the general step, by Theorem 5

$$|f(z_n)| \leqslant \frac{4|f(z_{n-1})|^2}{r_{n-1} - 4|f(z_{n-1})|},$$

and by the induction hypothesis

$$\begin{aligned} |f(z_n)| &\leqslant \frac{4b^{2^{n-1}-1}|f(z_0)|b^{2^{n-1}}b(r_\zeta/9)}{r_{n-1} - 4|f(z_{n-1})|} \\ &\leqslant \frac{b^{2^n-1}4(r_\zeta/9)|f(z_0)|}{\frac{8}{9}r_\zeta - 4(r_\zeta/9)} \\ &\leqslant b^{2^n-1}|f(z_0)|. \end{aligned}$$

This proves Theorem 4.

One can imagine an algorithm for finding a zero of a typical polynomial f as follows. Choose an initial point and apply Newton's method as long as one is getting fast convergence, repeating the process if necessary. Bearing on the success of this algorithm is

PROBLEM 9. Let

$$\alpha(d) = \operatorname*{average}_{f \in P_d(1)} \frac{\text{Area}(\Omega_f \cap D_4)}{\text{Area } D_4}.$$

Estimate $\alpha(d)$ as a function of d especially from below.

Here D_4 is the disk of radius 4, but could be taken as D_1 also, for example. The average over $P_d(1)$ means the integral with respect to normalized Lebesgue measure on the space of polynomials defined in §1, Chapter 2.

I don't know if $\alpha(d) > \varepsilon > 0$, where ε is independent of d.

The following proposition, a version of one in *Shub-Smale I*, is suggestive in looking at this problem.

PROPOSITION. (i)

$$\Omega_f \supset \bigcup_{\substack{\zeta \\ f(\zeta)=0}} D_{R_\zeta}(\zeta), \quad R_\zeta = \frac{r_\zeta}{36|f'(\zeta)|},$$

and the union is disjoint.

(ii)

$$\alpha(d) \geqslant \frac{K}{d} \int_{f \in P_d(1)} \left(\frac{\rho_f}{D_f^{1/d}} \right)^2,$$

where ρ_f *is the minimum of* $|f(\theta)|$ *over the critical points* θ *of* f, D_f *is the discriminant of* f (*see* Lang) *and* $1/K = 16 \cdot 36^2$.

PROOF. Part (i) is an immediate consequence of the Koebe-Bieberbach Theorem and the definitions.

For part (ii), first note that $r_\zeta \geqslant \rho_f$ for all roots ζ of f. Next observe that $D_{R_\zeta}(\zeta) \subset D_4$ since $|\zeta| < 2$ and $R_\zeta < 2$ (the $D_{R_\zeta}(\zeta)$ are disjoint and $f \in P_d(1)$). Thus

$$\alpha(d) \geqslant K \int_{f \in P_d(1)} \rho_f^2 \sum_\zeta \frac{1}{|f'(\zeta)|^2}.$$

Now apply the inequality on arithmetic versus geometric means to obtain

$$\alpha(d) \geqslant Kd \int \frac{\rho_f^2}{\left(\prod_\zeta |f'(\zeta)|^2\right)^{1/d}}.$$

The result of the proposition follows from the identity $D_f d^d = \prod_\zeta |f'(\zeta)|$ (see *Lang*).

The function ρ_f is discussed in *Smale III*, where it is proved that

$$\mathrm{Vol}\{f \in P_d(1) | \rho_f \leqslant \alpha\} \leqslant d\alpha^2,$$

where Vol means normalized Lebesgue measure.

4. Purely iterative algorithms. Newton's method is an example of a broad class of algorithms I will call purely iterative. This concept will be formalized in the zero-finding problem for one complex polynomial.

Let $\mathscr{P}_d$ be the space of all polynomials of degree $\leqslant d$ and define

$$j\colon \mathbf{C} \times \mathscr{P}_d \to J_k$$

by

$$j(z, f) = \left(z, f(z), f'(z), \ldots, f^{(k)}(z)\right).$$

Here J_k (a "jet" space) is $\mathbf{C}^{k+2}$ representing the source and the first k derivatives. Assume for simplicity that $d \geqslant k$; then j is surjective.

The datum of a purely iterative algorithm is a rational map $F\colon J_k \to \mathbf{C}$, which will be written in the following form:

$$F(z, \xi_0, \ldots, \xi_k) = z - \frac{P(z, \xi_0, \ldots, \xi_k)}{Q(z, \xi_0, \ldots, \xi_k)},$$

where P and Q are polynomials in the $k + 2$ variables with no common factor.

A *purely iterative algorithm* is a rational endomorphism $T_f\colon \mathbf{C} \to \mathbf{C}$ depending on $f \in \mathscr{P}_d$ and having the form

$$T_f(z) = F(j(z, f))$$

for some rational map F.

Rational maps are natural in this context because they represent the primitive operations, addition, multiplication, subtraction and division, of the computer. In this case, the computer is an idealized complex computer.

In the example of Newton's method,

$$T_f = N_f, \quad k = 1, \qquad P(z, \xi_0, \xi_1) = \xi_0, \quad Q(z, \xi_0, \xi_1) = \xi_1.$$

I will say that the purely iterative algorithm T_f (defined by some F) is *generally convergent* if there is some open set $\mathscr{U}$ of full measure in $\mathbf{C} \times P_d$ such that, for $(z, f) \in \mathscr{U}$, $T_f^n(z)$ tends to a zero of f as $n \to \infty$.

The definition depends on d and k. We have seen above that for $d = 2$ ($k = 1$), Newton's method is generally convergent, and for $d > 2$, Newton's method is not generally convergent.

I conjecture a negative answer to the following:

PROBLEM 10. If $d > k + 1$, does there exist any generally convergent purely iterative algorithm?

In fact, for any $d > 2$, I know of no generally convergent purely iterative algorithm, any k, and I don't know how to prove the conjecture even for $k = 1$. The case $k = 0$ may not be so hard, perhaps using the arguments in Proposition 2 below.

It might be natural to add a homogeneity hypothesis in the problem. One could suppose that P and Q as polynomials in $(\xi_0, \ldots, \xi_k)$ are homogeneous of the same degree, so that F is defined on $\mathbf{C} \times \mathbf{P}_d(\mathbf{C})$, where $\mathbf{P}_d(\mathbf{C})$ is the projective space of polynomials. Newton's method and extensions due to Euler (see *Shub-Smale I*) are purely iterative algorithms satisfying this homogeneity condition. Even in this case it seems hard to decide the existence of a generally convergent purely iterative algorithm.

There does exist a generally convergent purely iterative algorithm in another context. That is the power method for approximating the dominant eigenvector of a matrix.

Two little propositions in the direction of the above conjecture are given.

PROPOSITION 1. *If*

$$T_f(z) = z - \frac{P(j(z, f))}{Q(j(z, f))} = F(j(z, z)), \qquad \mathscr{P}_d \xrightarrow{j} J_k \xrightarrow{F} \mathbf{C},$$

is a generally convergent purely iterative algorithm, then for each f of degree d,

$$\deg Q(j(z, f)) \geqslant \deg(P(z, f)) - 1,$$

where deg *means degree in z.*

PROOF. Otherwise $|T_f^n(z)| \to \infty$ as $n \to \infty$ (for some open set of z, f).

PROPOSITION 2. *If $d \geqslant k + 1$, $k > 0$, and*

$$T_f(z) = z - \frac{P(z, f(z), \ldots, f^k(z))}{Q(z, f(z), \ldots, f^k(z))}$$

is a generally convergent purely iterative algorithm, then $\partial P / \partial \xi_k \equiv 0$. That is to say P is independent of the last coordinate.

The proof needs some lemmas.

LEMMA 1. *If T_f is generally convergent, then $T_f(\zeta) = \zeta$ and $T_f'(\zeta) = 0$ imply $f(\zeta) = 0$.*

PROOF. If the lemma is not true there exist ζ and f such that $f(\zeta) \neq 0$ and yet ζ is a superattractive fixed point. This yields the lemma.

LEMMA 2. *Let T_f be generally convergent as above. Then there exist an integer $\rho > 0$ and polynomials α, β of variables $(z, \xi_0, \ldots, \xi_{k+1})$ such that*

$$(\xi_0 Q)^\rho \equiv \alpha P + \beta(Q - P').$$

Here $P' \equiv \sum_{i=0}^{k} \xi_{i+1} P_{\xi_i} + P_z$, using a usual partial derivative notation.

PROOF. First, by Lemma 1,

$$P/Q = 0 \text{ and } 1 - (P'Q - Q'P)/Q^2 = 0 \Rightarrow \xi_0 = 0$$

or

$$P = 0, \qquad Q - P' = 0 \Rightarrow \xi_0 Q = 0.$$

The Hilbert Nullstellensatz (see *Lang*) applies to yield a positive integer ρ and polynomials α and β as above with

$$(\xi_0 Q)^\rho \equiv \alpha P + \beta(Q - P').$$

LEMMA 3. *Let T_f be generally convergent as in Lemma 2. If $P(z, \xi_0, \ldots, \xi_k) = 0$ then*

$$[\xi_0 Q(z_0, \xi_0, \ldots, \xi_k)]^\rho P_{\xi_k}(z, \xi_0, \ldots, \xi_k) = 0.$$

PROOF. Expand α and β in powers of ξ_{k+1},

$$\alpha \equiv \sum_{i=0}^{N} \alpha_i \xi_{k+1}^i, \qquad \beta \equiv \sum_{i=0}^{M} \beta_i \xi_{k+1}^i.$$

We may assume $M = N - 1$.

In the equation of Lemma 2, the coefficient of the highest power of ξ_{k+1} is identically zero, implying

$$\alpha_N P - \beta_{N-1} P_{\xi_k} \equiv 0.$$

Suppose that $P(\bar{z}, \bar{\xi}_0, \ldots, \bar{\xi}_k) = 0$ and that $P_{\xi_k}(\bar{z}, \bar{\xi}_0, \ldots, \bar{\xi}_k) \neq 0$. Then from the previous expression it follows that $\beta_{N-1}(\bar{z}, \bar{\xi}_0, \ldots, \bar{\xi}_k) = 0$.

Similarly, from the identity

$$\alpha_{N-1} P - \beta_{N-2} P_{\xi_k} + \beta_{N-1}\left(Q - P_z - \sum_{0}^{k-1} \xi_{i+1} P_{\xi_i}\right) \equiv 0$$

it follows that $\beta_{N-2}(\bar{z}, \ldots, \bar{\xi}_k) = 0$. Continuing by induction, one obtains that $\beta(\bar{z}, \bar{\xi}_1, \ldots, \bar{\xi}_k, \xi_{k+1}) = 0$, all ξ_{k+1}. Thus, again using Lemma 2, the statement of Lemma 3 follows.

Assume one has a generally convergent purely iterative algorithm as in Proposition 2. Then by Lemma 3 there is a polynomial γ in $(z, \xi_0, \ldots, \xi_k)$ such that $(\xi_0 Q)^\rho P_{\xi_k} \equiv \gamma P$. Since Q and P have no common factors, this implies Q^ρ divides γ; thus $\xi_0^\rho P_{\xi_k} \equiv \gamma_1 P$ for some polynomial γ_1. If $k > 0$ and $\gamma_1 \not\equiv 0$ then the left side has a lower degree in ξ_k than the right side. Thus $\gamma_1 \equiv 0$ and $P_{\xi_k} \equiv 0$. This proves Proposition 2.

It is easy to find examples which satisfy the equation of Lemma 3. For instance take $k = 1$, P an arbitrary function of z and ξ_0. Let γ be an arbitrary function of (z, ξ_0, ξ_1) and $Q = \xi_1 P_{\xi_0} + P_z + \xi_0 + \gamma P$. Then this equation is satisfied with $\alpha = Q^\rho$ and $\beta = -\alpha\gamma$.

However, a generally convergent purely iterative algorithm has no periodic sink of least period two (or more). This fact has not been used, here, as it was in §1, Chapter 2, to show that Newton's method is not generally convergent.

ADDED IN PROOF. Curt McMullen in his Harvard thesis, 1985, has solved problem 10.

5. Proof of Theorem A. The goal is to prove Theorem A. First some general considerations are discussed. Recall the endomorphism G_ω: $S \to S$ defined by $G_\omega(z) = z + (\omega - f(z))/f'(z)$, $S = \mathbf{C} \cup \infty$ the Riemann sphere. One cannot expect $z_k = G_\omega(z_{k-1})$ to converge to a solution of $f(z) = \omega$. However, one can proceed as follows:

Suppose z_0 is given. For $i = 0, \ldots, n$, choose ω_i evenly spaced along the segment from $\omega_0 = f(z_0)$ to $\omega_n = \omega$. Then the algorithm defined by $z_i = G_{\omega_i}(z_{i-1})$ will converge to a solution of $f(z) = \omega$ for almost all f, provided the spacing of the ω_i is fine enough. Certain difficulties prevent us from using this more simple and natural approach. See Problem 2 and Proposition 4 at the end of this section for more on this.

Theorem A involves a small modification of the above. Now for its proof.

LEMMA 1. *Let* $z' = G_\omega(z) = z + (\omega - f(z))/f'(z)$. *Then*

$$|\omega - f(z')| \leqslant \frac{4|\omega - f(z)|^2}{r - 4|\omega - f(z)|},$$

for all z, ω *such that* $|\omega - f(z)| < r/4$. *Here* $r = r(f_z^{-1})$ *is the radius of convergence of the branch of* f^{-1} *which sends* $f(z)$ *into* z.

PROOF. Let $g(z) = f(z) - \omega$ and apply Theorem 2 of §3 (Chapter 2) to g, obtaining the lemma.

LEMMA 2. *Consider* L, M *and* $J = (1/(1 - M + L))(\sin(\pi/12) - L)$ *such that*

(1) $0 < L < \sin(\pi/12)$,
(2) $0 < M < 1$,
(3) $4 < J$,
(4) $((1 - M + L)M)(4/(J - 4)) < L$.

These inequalities are consistent. Moreover M *can be chosen less than* $1 - 1/97$ *with* $L < 1/80$.

PROOF. Note that except for (2), $M = 1$ and $L = \frac{1}{2}\sin(\pi/12)$ is a solution. Now decrease M below 1 and use continuity to prove the first statement.

For the second, let $M = 1 - \alpha\sin(\pi/12)$, $L = \beta\sin(\pi/12)$. Conditions 1–4 translate into 1′–4′:

(1′) $0 < \beta < 1$,
(2′) $0 < \alpha < (\sin(\pi/12))^{-1}$,
(3′) $0 < 1 - 4\alpha - 5\beta$,
(4′) $4(\alpha + \beta)^2 < \beta(1 - 4\alpha - 5\beta)(1 - \alpha\sin(\pi/12))$.

Consider special solutions of the form $\alpha = \beta$. Then the crucial condition (4′) specializes to

(4″) $16\alpha < (1 - 9\alpha)(1 - \alpha \sin(\pi/12))$,

and for this it is sufficient for α to satisfy

(4‴) $1/\alpha > 25 + \sin(\pi/12)$.

The rest of Lemma 2 follows by a very easy calculation. Since M is an important constant it would be good to obtain a closer approximation to the best M satisfying the conditions of Lemma 2 (not a hard problem).

Condition (H). Say that a pair (z, f), z a complex number, f a polynomial satisfies *Condition* (H) if $f(z) \neq 0$, and f_z^{-1} can be analytically extended to the sector about the ray 0 to $f(z)$ of total angle $\pi/6$ (Figure 3). This is essentially the condition $(H)_{f,z} > \pi/12$ in *Shub-Smale I, II*.

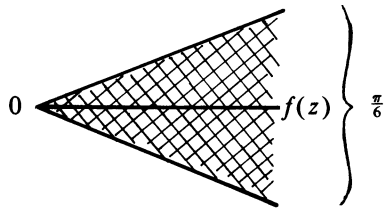

FIGURE 3

PROPOSITION 1. *Let M, L, J be as in Lemma 2, (z_0, f) satisfy* (H), *and $\omega_0 = f(z_0)$. Define $\omega_i = M^i\omega_0$. Then $z_i = f_{\omega_i}^{-1}(z_{i-1})$ is defined, finite for all $i > 0$, and*

$$
\begin{aligned}
&i \geq 0, \quad (a_i)\colon |\omega_i - f(z_i)| \leq L|\omega_i|;\\
&i > 0, \quad (b_i)\colon |\omega_i - f(z_{i-1})| \leq |\omega_{i-1}|(1 - M + L);\\
&i > 0, \quad (c_i)\colon |\omega_i - f(z_{i-1})| \leq r_{i-1}/J,
\end{aligned}
$$

where $r_{i-1} = r(f_{z_{i-1}}^{-1})$.

PROOF. (a_0) is true. Suppose that (a_{i-1}), (b_{i-1}), (c_{i-1}) are true. We will show that (b_i), (c_i) and (a_i) are true (one has to modify the argument slightly for (b_1), (c_1). We use

LEMMA 3.

$$r_{i-1} \geq (\sin(\pi/12))|\omega_{i-1}| - |f(z_{i-1}) - \omega_{i-1}|.$$

The proof of the lemma is simple trigonometry, and uses (H).

Assertion (b_i) is a consequence of the following estimate.

$$|f(z_{i-1}) - \omega_i| \leq |f(z_{i-1}) - \omega_{i-1}| + |\omega_i - \omega_{i-1}| \leq (L + 1 - M)|\omega_{i-1}|.$$

Assertion (c_i) is a consequence of Lemma 3 and the induction hypotheses as follows:

$$r_{i-1} \geqslant \left(\sin\frac{\pi}{12}\right)\left(\frac{1}{1-M+L}\right)|\omega_i - f(z_{i-1})| - \left(\frac{L}{1-M+L}\right)|\omega_i - f(z_{i-1})|$$
$$\geqslant \left(\frac{1}{1-M+L}\right)\left(\sin\frac{\pi}{12} - L\right)|\omega_i - f(z_{i-1})|.$$

Finally, (a_i) is proved as follows:

$$|f(z_i) - \omega_i| \leqslant |\omega_i - f(z_{i-1})|^2 \frac{4}{r_{i-1} - 4|\omega_i - f(z_{i-1})|}$$

by Lemma 1. By (c_i) we then obtain

$$|f(z_i) - \omega_i| \leqslant \left(\frac{4}{J-4}\right)|\omega_i - f(z_{i-1})|.$$

Then we use (b_i) to get

$$|f(z_i) - \omega_i| \leqslant \left(\frac{4}{J-4}\right)(1-M+L)\omega_{i-1} \leqslant \left(\frac{4}{J-4}\right)\left(\frac{1-M+L}{M}\right)\omega_i,$$

finishing the proof of the proposition.

From the proposition one can see how small $|f(z_n)|$ is after applying the algorithm n times. More precisely,

COROLLARY. *If* (z_0, f) *satisfy* Ⓗ *and*

$$n > \frac{\log[(1+L)|f(z_0)|/\varepsilon]}{|\log M|},$$

then $|f(z_n)| < \varepsilon$.

To see the corollary, use Proposition 1 to obtain

$$|f(z_n)| \leqslant |f(z_n) - \omega_n| + |\omega_n| \leqslant (L+1)|\omega_n| \leqslant (L+1)M^n|f(z_0)|.$$

The corollary follows.

Now the question becomes, to what extent can one expect Ⓗ to be satisfied? Mike Shub and I have dealt with this question making the (unhappy) hypothesis $|z_0| = 3$.

Let

$$V_f = \left\{ z \,|\, |z| = 3, (z, f) \text{ satisfy } Ⓗ \right\}$$

and impose the uniform (normalized Lebesgue) probability measure on $S = \{z \,|\, |z| = 3\}$.

PROPOSITION 2. *For any* f, *the measure of* V_f *is greater than* $1/6$.

This is Proposition 2 of *Shub-Smale II*.

Following this reference, let Ω be the countable product of S with itself so that if $\bar{z} \in \Omega$, $\bar{z} = (\bar{z}_1, \bar{z}_2, \bar{z}_3, \ldots)$. Impose the product measure on Ω.

Given a set $V \subset S$, let m: $\Omega \to Z^+$ be the function defined by $m(\bar{z}) =$ the first m such that $\bar{z}_i \notin V_f$ for $i < m$, but $\bar{z}_m \in V$.

PROPOSITION 3 (*See* Shub-Smale II).

$$\int_{\bar{z}\in\Omega} m(\bar{z}) = \frac{1}{\text{measure } V}.$$

Next observe that if $f(z) = \sum_0^d a_i z^i$, $a_d = 1$, $|a_i| \leqslant 1$, and $|z_0| = 3$, then

$$(*) \qquad |f(z_0)| \leqslant \frac{3^{d+1} - 1}{3 - 1} \leqslant \frac{1}{2}(3^{d+1}).$$

The proof of Theorem A is almost finished. Using the corollary of Proposition 1, Proposition 2, and Proposition 3, we need only translate the n as given in the corollary to the n of Theorem A, or

$$\frac{\log\left[(1 + L)|f(z_0)|/\varepsilon\right]}{\log M} < 98\left[|\log \varepsilon| + d \log 3\right].$$

This is easily done using (*) and the estimates on M, L in Lemma 2. This finishes the proof of Theorem A.

Note that the largest contribution to the number of steps in Theorem A is the linear factor d. This is necessary because $|z_0| = 3$, rather than, say, $|z_0| \leqslant 1$. But Proposition 2 is no longer at our disposal for small $|z_0|$. Moreover, it is possible that $|z_0| \leqslant 1$ gives a slower algorithm. Then $f(z_0)$ is more likely to be close to a ray containing a critical value, e.g. as $f(z) = z^d + 1$. This bears on Problem 2. In the same direction, the following result shows how fast this algorithm can go starting at $z_0 = 0$ in case there is a good zone of analyticity for f_0^{-1}, the branch of f^{-1} which takes $f(0)$ to 0.

PROPOSITION 4. *Suppose f is a polynomial with f_0^{-1} analytic on $N_\delta(\overline{0f(0)})$, the δ neighborhood of the segment from $f(0)$ to 0. Choose $n > 24|f(0)|/\delta$ and let $\omega_i = (n - i)f(0)/n$, $i = 0,\ldots,n$. Then $z_i = G_{\omega_i}(z_{i-1})$ is well-defined, $|f(z_{n-1})| < \delta/12$, and z_{n-1} is an approximate zero.*

PROOF. Note first that

$$|\omega_i - \omega_{i-1}| = |f(0)|/n < \delta/24.$$

By Lemma 1,

$$|f(z_i) - \omega_i| < \frac{4|f(z_{i-1}) - \omega_i|^2}{\delta - 4|f(z_{i-1}) - \omega_i|}.$$

It is sufficient to show (by induction) that

$$|f(z_i) - \omega_{i+1}| < \delta/12.$$

But using the Lemma 1 estimate,

$$\begin{aligned} |f(z_i) - \omega_{i+1}| &\leqslant |\omega_i - \omega_{i+1}| + |f(z_i) - \omega_i| \\ &< \delta/24 + \delta/24 < \delta/12. \end{aligned}$$

Q.E.D.

Proposition 4 suggests the following study. Let $\delta(f)$ be the maximum of the δ such that f_0^{-1} is analytic on $N_\delta(\overline{0f(0)})$ and let $\eta(f) = \delta(f)/|f(0)|$. What are the measure-theoretic properties of η on various spaces of polynomials? For example, estimate the volumes of $\{f|\eta(f) > \alpha\}$, as a function of α.

6. What is an algorithm?

PROBLEM 11. What is the fastest way of finding a zero of a polynomial?

This is a kind of super-problem. I would expect contributions by several mathematicians rather than a single solution. It will take a lot of thought even to find a good mathematical formulation.

In some ways, one could compare this problem with showing the existence of a zero of a polynomial. The concept of complex numbers had to be developed first. For Problem 11, one must develop the concept of algorithm to deal with the kind of mathematics involved. Consistent with the von Neumann statement quoted in the introduction, my belief is that the Turing approach to algorithms is inadequate for these purposes.

Although the definitions of such algorithms are not available at this time, my guess is that some kind of continuous or differentiable machine would be involved. In so much of the use of the digital computer, inputs are treated as real numbers and the output is a continuous function of the input. Of course a continuous machine would be an idealization of an actual machine, as is a Turing machine.

The definition of an algorithm should relate well to an actual program or flow-chart of a numerical analyst. Perhaps one could use a Random Access Machine (RAM, see *Aho-Hopcraft-Ullman*) and suppose that the registers could hold real numbers. Then one might with some care expand the list of permissible operations. There are pitfalls along the way and much thought is needed to do this right.

To be able to discuss the fastest algorithm, one has to have a definition of algorithm. I have used the word algorithm throughout this paper, yet I have not said what an algorithm is. Certainly the algorithms discussed here are not Turing machines; and to force them into the Turing machine framework would be detrimental to their analysis. It must be added that the idealizations I have suggested do not eliminate the study of round-off error. Dealing with such loss of precision is a necessary part of the program.

Problem 11 is not a clear-cut problem for various reasons. Factors which could affect the answer include dependence on the machine, whether one wants to solve one or many problems, time taken to write the program, whether polynomials have large or small degree, how the problem is presented, etc.

Dejon-Henrici is a general reference to some aspects of Problem 11.

The next problem, which was discussed in *Smale III* and *Shub-Smale II*, is still open. I restate it simply.

PROBLEM 12 (MEAN VALUE PROBLEM). Given any complex number z and polynomial f, is there a critical point θ of f such that

$$\left|\frac{f(z)-f(\theta)}{z-\theta}\right| \leqslant |f'(z)|?$$

CHAPTER III

1. Proof of Theorem D. I first give some preliminaries for the proof of Theorem D. Suppose that a Hilbert space H is given its Gaussian measure. (See *Elworthy* and *Kuo* for the meaning of this and the relationship to Wiener measure.) "Average" will refer to this measure.

PROPOSITION. *Let $L: H \to \mathbf{R}$ be a bounded linear functional. Then the average satisfies*

$$\operatorname*{Av}_{x \in H} |Lx| = \left(\frac{2}{\pi}\right)^{1/2} \|L\|.$$

PROOF. Choose $v \in H$ so that $Lx = \langle v, x \rangle$ all $x \in H$. Then $\|L\| = \|v\|$ and

$$\operatorname*{Av}_{x \in H} |Lx| = \operatorname*{Av}_{x} |\langle v, x \rangle| = \|L\| \operatorname*{Av}_{x \in H} \langle e, x \rangle,$$

where $e = v/\|v\|$. Hence (see *Elworthy*)

$$\operatorname*{Av}_{x \in H} |\langle e, x \rangle| = \left(\frac{1}{2\pi}\right)^{1/2} \int_{-\infty}^{\infty} |t| e^{-t^2/2}\, dt = \left(\frac{2}{\pi}\right)^{1/2}.$$

This yields the proposition.

We will simplify the proof of Theorem D slightly by working with the subspaces

$$\mathscr{H}_0^1 = \{ f \in \mathscr{H}^1 | f(0) = 0 \} \quad \text{and} \quad \mathscr{H}_0^2 = \{ f \in \mathscr{H}^2 | f(0) = f'(0) = 0 \}.$$

Let $J: \mathscr{H}_0^1 \to R$ be the integral, $J(f) = \int f$, and denote by $\hat{J}^{(1)} \in \mathscr{H}_0^1$ the dual. Thus

$$J(f) = \langle \hat{J}^{(1)}, f \rangle.$$

Similarly, let $\hat{J}^{(2)} \in \mathscr{H}_0^2$ denote the dual of $J: \mathscr{H}_0^2 \to R$.

Next, for $i = 1, 2$, and $0 \leqslant t \leqslant 1$, define $E_t^{(i)}: \mathscr{H}_0^i \to R$ by $E_t^{(i)}(f) = f(t)$. Let $\hat{E}_t^{(i)} \in \mathscr{H}_0^i$ be the corresponding duals of this evaluation map.

LEMMA 1.

(i) $\dfrac{d}{ds} \hat{J}^{(1)}(s) = 1 - s,$

(ii) $\dfrac{d}{ds} \hat{E}_t^{(1)}(s) = \begin{cases} 1, & s \leqslant t, \\ 0, & t \leqslant s, \end{cases}$

(iii) $\dfrac{d^2}{ds^2} \hat{J}^{(2)}(s) = \dfrac{1}{2} - s + \dfrac{1}{2} s^2,$

(iv) $\dfrac{d^2}{ds^2} \hat{E}_t^{(2)}(s) = \begin{cases} t - s, & s \leqslant t, \\ 0, & t \leqslant s. \end{cases}$

The proof is easy calculus: For (i), it amounts to checking

$$\langle \hat{J}^{(1)}, f \rangle_{\mathscr{H}^1} = \int_0^1 (1 - s) f'(s)\, ds = J(f).$$

For (ii),

$$\langle \hat{E}_t^{(1)}, f \rangle_{\mathscr{H}^1} = \int_0^t f'(s)\, ds = f(t) = E_t^{(1)}(f).$$

For (iii),

$$\langle \hat{J}^{(2)}, f \rangle_{\mathscr{H}^2} = \int_0^1 \left(\frac{1}{2} - s + \frac{1}{2} s^2\right) f''(s)\, ds = \int_0^1 f(s)\, ds = J(f).$$

For (iv),

$$\left\langle \hat{E}_t^{(2)}, f \right\rangle_{\mathscr{H}^2} = \int_0^t (t-s) f''(s)\, ds = f(t) = E_t^{(2)}(f).$$

Towards the proof of Theorem D, use the proposition to see that

$$\operatorname*{Av}_{f \in \mathscr{H}_0^1} |(J - R_h)f| = \left(\frac{2}{\pi}\right)^{1/2} \|J - R_h\|_{\mathscr{H}^1} = \left(\frac{2}{\pi}\right)^{1/2} \|\hat{J}^{(1)} - \hat{R}_h^{(1)}\|_{\mathscr{H}^1}.$$

Now

$$R_h(f) = h \sum_{i=1}^{n} f(ih) = h \sum_{i=1}^{n} E_{ih}(f),$$

and it will be shown that

$$\left\| \hat{J}^{(1)} - h \sum_{i=1}^{n} \hat{E}_{ih}^{(1)} \right\|_{\mathscr{H}^1} = \frac{h}{\sqrt{3}}.$$

By Lemma 1(ii), for $(j-1)h \leqslant s \leqslant jh$,

$$\sum_{i=1}^{n} \frac{d}{ds} \hat{E}_{ih}^{(1)}(s) = n - (j-1).$$

Thus, using Lemma 1(i),

$$\left\| \hat{J}^{(1)} - h \sum_{1}^{n} \hat{E}_{ih}^{(1)} \right\| = \left\{ \sum_{j=1}^{n} \int_{(j-1)h}^{jh} [1 - s - h(n-j+1)]^2\, ds \right\}^{1/2}.$$

But $hn = 1$, so this is

$$\left[\sum_{j=1}^{n} \frac{[s - h(j-1)]^3}{3} \Big\}_{s=(j-1)h}^{s=hj} \right]^{1/2} = \left(\frac{nh^3}{3}\right)^{1/2} = \frac{h}{\sqrt{3}},$$

which proves the first of four parts of Theorem D.

Next I will carry out the same process for the "Trapezoidal Rule". Since $T_h(f) = -\frac{1}{2}h(f(1) + f(0)) + R_h(f)$ and $f(0) = 0$,

$$\|\hat{J}^{(1)} - \hat{T}_h\|_{\mathscr{H}^1} = \left\| J^{(1)} - \hat{R}_h + \frac{h}{2} \hat{E}_1^{(1)} \right\|_{\mathscr{H}^1} = \left\{ \sum_{j=1}^{n} \int_{(j-1)h}^{jh} \left(s - hj + \frac{h}{2} \right)^2 \right\}^{1/2}$$

$$= \left\{ \frac{1}{3} \sum_{j=1}^{n} \left(\frac{h}{2}\right)^3 + \left(\frac{h}{2}\right)^3 \right\}^{1/2} = \frac{1}{2} \frac{h}{\sqrt{3}},$$

which proves the second part of Theorem D.

A couple of more lemmas help pave the way for proving the third part of Theorem D. The first is well known.

LEMMA 2.

$$\sum_{j=1}^{n} (j-1)^2 = \frac{(n-1)n(2n-1)}{6}.$$

LEMMA 3. *For* $s \in [(j-1)h, jh]$

$$\sum_{i=1}^{n} \frac{d^2}{ds^2}\hat{E}_{ih}^{(2)}(s) = \frac{h(n+1)n}{2} - \frac{hj(j-1)}{2} - s(n-j+1).$$

PROOF OF LEMMA 3. Write out the sum using Lemma 1(iv) to obtain it as

$$(jh - s) + (j+1)h - s + \cdots + nh - s.$$

The rest follows.

Let A stand for the quantity in Lemma 3. Then

$$\|\hat{J}^{(2)} - \hat{R}_h^{(2)}\|_{\mathscr{H}^2} = \left\{ \sum_{j=1}^{n} \int_{(j-1)h}^{jh} \left(\frac{1}{2} - s + \frac{1}{2}s^2 - hA \right)^2 \right\}^{1/2},$$

using Lemma 1. This quantity may be written as (using Lemma 3)

$$\frac{1}{2}\left\{ \sum_{j=1}^{n} \int_{(j-1)h}^{jh} \left[s^2 - 2s(j-1)h + (j-1)^2h^2 + \left(h^2(j-1) - h\right) \right]^2 \right\}^{1/2}.$$

One can integrate it, using $t = s - (j-1)h$; then a little calculation using Lemma 2 yields

$$\frac{h}{2\sqrt{3}}\left(1 + \frac{h}{2} + \frac{h^2}{10}\right)^{1/2},$$

which is the formula for $\varepsilon_R^2(h)$ in Theorem D.

Next,

$$S_h(f) = \frac{h}{3}\left[f(1) + 2\sum_{i=1}^{2n-1} f(ih) + 2\sum_{1}^{n} f(2(i-1)h) \right],$$

$$\frac{d^2}{ds^2}\hat{S}_h(s) = \frac{h}{3}\left[1 - s + 2\sum_{i=1}^{2n-1} \frac{d^2}{ds^2}\hat{E}_{ih}^{(2)}(s) + 2\sum_{1}^{n} \frac{d^2}{ds^2}\hat{E}_{(2i-1)h}^{(2)}(s) \right].$$

LEMMA 4.(i) *On* $(2k-2)h \leqslant s \leqslant (2k-1)h$,

$$\sum_{1}^{2n-1} \frac{d^2}{ds^2}\hat{E}_{ih}^{(2)}(s) = h(n(2n-1) - (2k-1)(k-1)) - s(2n-2k+1)$$

and

$$\sum_{i=1}^{n} \frac{d^2}{ds^2}\hat{E}_{(2i-1)h}^{(2)}(s) = h\left(n^2 - (k-1)^2\right) - s(n-k+1);$$

(ii) *on* $(2k-1)h \leqslant s \leqslant 2kh$,

$$\sum_{1}^{2n-1} \frac{d^2}{ds^2}\hat{E}_{ih}^{(2)}(s) = h(n(2n-1) - k(2k-1)) - 2s(n-k)$$

and

$$\sum_{1}^{n} \frac{d^2}{ds^2}\hat{E}_{(2i-1)h}^{(2)}(s) = h(n^2 - k^2) - s(n-k).$$

The proof is a straightforward calculation as before, keeping in mind $1 + 3 + 5 + \cdots + (2l - 1) = l^2$.

Finally, a straightforward calculation using the above expression for S_h, and Lemma 4, yields the formula

$$\varepsilon_S(h) = \left(\frac{2}{\pi}\right)^{1/2} \frac{h^2}{3\sqrt{15}},$$

finishing the proof of Theorem D.

2. Questions of precision. I will not attempt to give the proof of Theorem C, but instead refer the reader to *Kostlan* and *Ocneanu*. There are some aspects that I would like to comment on.

Since the singular values of a matrix play an important role, it is worthwhile to say what they are and how they are used. If A is any matrix (say $n \times n$ for simplicity), the *singular values* of A are the nonnegative square roots of the eigenvalues of AA^T. This makes sense because the composition of a matrix with its transpose is positive semidefinite and the eigenvalues are positive or zero. Singular values are discussed at length in *Forsyth-Moler* and *Wilkinson I*. They measure the distortion of a linear map.

Let A be an $n \times n$ real matrix and $0 \leqslant \mu_1 \leqslant \mu_2 \leqslant \cdots \leqslant \mu_n$ denote its singular values. It is easily shown that $\|A\| = \mu_n$ and $\|A^{-1}\| = 1/\mu_1$. Thus the condition number K_A of A equals μ_n/μ_1. The number L_A whose average is estimated in Theorem C therefore satisfies $L_A = \log(\mu_n/\mu_1)$. Since L_A is in fact a function of the singular values of A, the average over A is equal to an average over the space of singular values. More precisely, the following is true.

PROPOSITION.

$$L(n) = c_n \int \log\left(\frac{\mu_n}{\mu_1}\right) \prod_{i<j} (\mu_j^2 - \mu_i^2) \exp\left(-\frac{1}{2}\sum_i \mu_i^2\right) d\mu_i \cdots d\mu_n,$$

$0 \leqslant \mu_1 \leqslant \cdots \leqslant \mu_n$, *where* c_n *is the reciprocal of the same integral with the* log *factor deleted.*

The proof of the proposition uses the previous discussion together with a well-known result on the probability density for the Gaussian measure on matrices in terms of singular values. See *Kostlan* for reference and details. This proposition is the basis for the estimates given in Theorem C.

The above analysis of the system $Ax = b$ involves a worst-case hypothesis for the error in b. *Kostlan* studies the loss in precision for this system by averaging over b and the error in b. This yields zero for a given matrix A. So in this form, on the average, the amount of precision lost is the same as that gained. Thus the variance becomes crucial to estimate. *Kostlan* bounds this variance by a polynomial in n and conjectures that this quantity satisfies

$$\mathrm{Var}(A) < \pi^2/4.$$

The above comments on precision are independent of algorithms. But the limitations of precision discussed above apply to any specific algorithm. For the analysis of algorithms, precision as well as speed must be taken into account.

One can also consider the problems of precision for nonlinear problems. If f: $\mathbf{R}^n \to \mathbf{R}^n$ is a continuously differentiable map then the equation $f(x) = y$ may be looked at from two points of view: Given x find y (evaluation), or given y find x. For the first, the natural definition of condition number is $\|Df(x)\|$, where $Df(x)$ is the matrix of partial derivatives and, as usual, the norm is the operator norm. For the second problem, of solving $f(x) = y$ for x, this is replaced by $\|Df(x)^{-1}\|$. See *Wilkinson II* for discussion of these things.

Quantities reflecting the loss of precision for these two problems are thus $\log\|Df(x)\|$ and $\log\|Df(x)^{-1}\|$ respectively. For an interesting analysis of the average loss of precision in the evaluation of rational functions, see *Blum-Shub*.

If ζ is a zero of a complex polynomial f, the loss of precision for the problem of finding it is

$$L_{f,\zeta} = \log \frac{1}{|f'(\zeta)|}.$$

For the case of a double root this is infinite (for an almost double root, arbitrarily large). One has the paradoxical situation of polynomials with double roots, but no algorithm (with the best machine even to be built) can affirm it.

References

I. Adler, R. M. Karp and R. Shamir, *A simplex variant solving an $m \times d$ linear program in $O(\min(m^2, D^2))$ expected number of pivot steps*, Report UCB CSD 83/158, Computer Science Division, University of California, Berkeley (December 1983).

I. Adler and W. Megiddo, *A simplex algorithm whose average number of steps is bounded between two quadratic functions of the smaller dimension*, Report, Dec. 1983.

A. Aho, J. Hopcraft and J. Ullman, *The design and analysis of computer algorithms*, Addison-Wesley, Reading, Mass., 1979.

K. Atkinson, *An introduction to numerical analysis*, Wiley, N. Y., 1978.

B. Barna, *Über die Divergenzpunkte des Newtonschen Verfahrens zur Bestimmung von Wurzeln Algebraischer Gleichungen*. II, Publ. Math. Debrecen **4** (1956), 384–397.

P. Blanchard, *Complex analytic dynamics on the Riemann sphere*, Bull. Amer. Math. Soc. (N.S.) **11** (1984), 85–141.

L. Blum and M. Shub, *Evaluating rational functions: Infinite precision is finite cost and tractable on average*, SIAM J. Comput. (to appear).

K. H. Borgwardt, *The average number of pivot steps required by the simplex-method is polynomial*, Z. Oper. Res. **26** (1982), 157–177.

J. Curry, *On the periodic behaviour of Newton's Method for a family of real cubic polynomials*, preprint, 1983.

J. Curry, L. Garnett and D. Sullivan, *On the iteration of a rational function: Computer experiments with Newton's method*, Comm. Math. Phys. **91** (1983), 267–277.

G. Dantzig, *Linear programming and extensions*, Princeton Univ. Press, Princeton, N.J., 1963.

B. Dejon and P. Henrici, *Constructive aspects of the fundamental theorem of algebra*, Wiley, N. Y., 1969.

J. Demmel, *A numerical analyst's Jordan canonical form*, Thesis, Univ. of California, Berkeley, 1983.

R. Dorfman, P. Samuelson and R. Solow, *Linear programming and economic analysis*, McGraw-Hill, N. Y., 1958.

A. Douady, *Systèmes dynamiques holomorphes*, Séminaire Bourbaki, No. 599, 1982.

C. Eaves and H. Scarf, *The solution of systems of piecewise linear equations*, Math. Oper. Res. **1** (1976), 1–27.

D. Elworthy, *Gaussian measures on Banach spaces and manifolds*, Global Analysis and its Applications, Vol. II, International Atomic Energy Agency, Vienna, 1974, pp. 151–166.

P. Fatou, *Sur les equations fonctionnelles*, Bull. Soc. Math. France **47**, **48** (1919–1920), 161–211; 33–94; 208–314.

G. Forsyth and C. Moler, *Computer solution of linear algebraic systems*, Prentice-Hall, Englewood Cliffs, N. J., 1967.

H. Goldstine, *A history of numerical analysis, from the 16th through the 19th century*, Springer-Verlag, Berlin and New York, 1977.

J. Guckenheimer, *Endomorphisms of the Riemann sphere*, Global Analysis (Chern and Smale, eds.), Amer. Math. Soc., Providence, R. I., 1970, pp. 95–123.

W. Hayman, *Multivalent functions*, Cambridge Univ. Press, Cambridge, 1958.

P. Henrici, *Applied and computational complex analysis*, Wiley, N. Y., 1977.

E. Hille, *Analytic function theory*. II, Ginn, Boston, 1962.

G. Julia, *Mémoire sur l'itération des fonctions rationnelles*, J. Math. Pures Appl. **4** (1918), 47–245.

E. Kostlan, *Statistical complexity of numerical linear algebra*, Thesis, Univ. of Calif. Berkeley, 1985.

H. Kuhn, Z. Wang and Senlin Xu, *On the cost of computing roots of polynomials*, Math. Programming **28** (1984), 156–163.

H. H. Kuo, *Gaussian measures in Banach spaces*, Lecture Notes in Math., vol. 463, Springer-Verlag, Berlin and New York, 1975.

S. Lang, *Algebra*, Addison-Wesley, Reading, Mass., 1963.

C. Martin and R. Hurley, *Newton's algorithm and chaotic dynamical systems*, SIAM J. Math. Anal. **16** (1984), 238–252.

M. Mehta, *Random matrices and the statistical theory of energy levels*, Academic Press, N. Y., 1967.

A. Ocneanu, *On the stability of large linear systems* (to appear).

A. Ostrowski, *Solutions of equations in Euclidean and Banach spaces*, Academic Press, N. Y., 1973.

H. O. Peitgen, D. Saupe and F. v. Haeseler, *Cayley's problem and Julia sets*, Math. Intelligencer **6** (1984), 11–20.

J. Renegar **I**, *On the complexity of a piecewise linear algorithm for approximating roots of complex polynomials*, Math. Programming (to appear).

J. Renegar **II**, *On the cost of approximating all roots of a complex polynomial*, Math. Programming (to appear).

J. Renegar **III**, *On the efficiency of Newton's method in approximating all zeros of a system of complex polynomials*, preprint, Colorado State Univ., 1984.

D. Saari and J. Urenko, *Newton's method, circle maps, and chaotic motion*, Amer. Math. Monthly **91** (1984), 3–17.

G. Saunders, *Iteration of rational functions of one complex variable and basins of attractive fixed points*, Thesis, Univ. of California, Berkeley, 1984.

R. Shamir, *The efficiency of the simplex method: A survey*, preprint, Univ. of California, Berkeley, 1984.

M. Shub, *The geometry and topology of dynamical systems and algorithms for numerical problems*, notes prepared for lectures given at D.D.4 Peking University, Beijing, China, Aug.-Sept. 1983.

M. Shub and S. Smale **I**, *Computational complexity: On the geometry of polynomials and a theory of cost, Part I*, Ann. Sci. École Norm. Sup. (4) (to appear).

M. Shub and S. Smale **II**, *Computational complexity: On the geometry of polynomials and a theory of cost: Part II*, SIAM J. Computing (to appear).

S. Smale **I**, *Differentiable dynamical systems*, The Mathematics of Time, Springer-Verlag, New York and Berlin, 1980.

S. Smale **II**, *A convergent process of price adjustment and global Newton methods*, J. Math. Econom. **3** (1976), 107–120.

S. Smale **III**, *The fundamental theorem of algebra and complexity theory*, Bull. Amer. Math. Soc. (N. S.) **4** (1981), 1–36.

S. Smale **IV**, *On the average number of steps in the simplex method of linear programming*, Math. Programming **27** (1983), 241–262.

S. Smale V, *The problem of the average speed of the Simplex method*, Mathematical Programming, The State of the Art (Bonn, 1982) (Bachem et al., eds.), Springer-Verlag, Berlin and New York, 1983.

D. Sullivan, *Quasiconformal homeomorphisms and dynamics III: Topological conjugacy classes of analytic endomorphisms*, Ann. of Math. (to appear).

M. J. Todd, *Polynomial expected behavior of a pivoting algorithm for linear complementary and linear programming problems*, Technical Report No. 595, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, New York, 1983.

J. Traub and H. Wozniakowski, *Information and computation*, in Vol. 23, Advances in Computers, vol. 23 (M. C. Yovits, ed.), Academic Press, N. Y., 1984.

A. Vershik and P. Sporyshev, *An estimate of the average number of steps in the simplex method and problems in asymptotic integral geometry*, Soviet Math. Dokl. **28** (1983). (Russian)

J. Von Neumann, *Collected works*, vol. V (A. Taub, ed.), MacMillan, N. Y., 1963.

J. Wilkinson **I**, *The algebraic eigenvalue problem*, Oxford Univ. Press, Oxford, 1965.

J. Wilkinson **II**, *Rounding errors in algebraic processes*, Prentice-Hall, Englewood Cliffs, N. J., 1973.

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF CALIFORNIA, BERKELEY, CALIFORNIA 94720