The Geometry and Topology of Dynamical Systems

and Algorithms for Numerical Problems*

Michael Shub

In these talks I would like to broach a fairly broad range of subjects, from relationships between topology and the qualitative theory of dynamical systems to numerical analysis and computational complexity.

Recently I have become interested in applications of dynamical systems to the theory of numerical analysis and computational complexity. Normally speaking things work the other way around, we use numerical methods to approximate the solutions of differential equations with given initial conditons and for a host of other problems. Frequently the methods are iterative methods that are themselves dynamical systems. A main theme of these talks is that the study of the geometry and dynamics of these dynamical systems is useful to crucial for the understanding of the numerical methods themselves.

I'll begin by recalling some fundamental facts and examples. Let M be a compact differentiable manifold with a Riemannian metric and which perhaps has boundary. So that, for example, M can be the ball of radius r in n-dimensional Euclidian space, E^n, $B_r = \{ x \epsilon E^n \mid || x || \leq r\}$ or its boundary S_r^{n-1}, etc. Let $f: M \to R$ be a smooth function, then $V(X) = -\text{grad } f(x)$ defines a vector field on M. In the case that M has boundary we suppose that V(X) points into M along the boundary. The gradient flow of f, $\Phi_t : M \to M$ where $\frac{d}{d_t} \phi_t(x)/_{t=0} = V(X)$ is globally defined for all $t \epsilon R$ in case M has no boundary, and for all $t \geq 0$ when the boundary of M is non-empty.

Note the minus sign so that the flow flows downhill, i.e. $\frac{d}{dt}f(\phi_t(x)) = -||\text{grad } f_x||^2 \leq 0$. Morse theory proves that for an open and dense (in the C^r topology) set of functions f (called Morse functons), the Hessian

*This paper was prepared as part of lectures given at D.D.4. The work was supported by an N.S.F. grant to the author.

of f is non-singular at the critical point of f. The vector field V = - grad f then has only finitely many singularities, say $p_1, \ldots, p_m$, where $V(p_i) = 0$, and moreover, near any of the critical points p_i there is a local chart so that f has the form

$f(x) = f(p_i) - x_1^2 - x_2^2 - \ldots - x_u^2 + x_{u+1}^2 + x_{u+2}^2 + \ldots + x_n^2$, where $x = (x_1, \ldots, x_n)$.

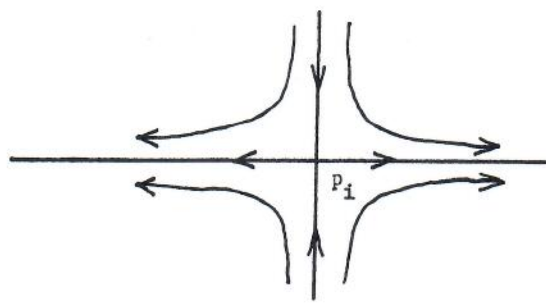
Thus for any $x \in M$, $\phi_t(x)$ converges to some p_i as $t \to +\infty$. Near p_i, - grad f takes the form

$$(+2x_1, +2x_2, \ldots, +2x_u, -2x_{u+1}, -2x_{u+2}, \ldots, -2x_n)$$

and

$$\phi_t(x_1, \ldots, x_n) = (e^{+2t}x_1, e^{+2t}x_2, \ldots, e^{+2t}x_u, e^{-2t}x_{u+1}, e^{-2t}x_{u+2}, \ldots, e^{-2t}x_n).$$

This gives the standard picture:



The points in the $(x_1, \ldots, x_u)$ space tend to p_1 as t approaches $-\infty$. Locally these are discs of dimension u called the index of the point p and $s = n - u$. These discs are denoted by $w^u_{loc}(p_i)$ and $w^s_{loc}(p_i)$ respectively, the local unstable and local stable manifolds of p_i. The set of $x \in M$ such that $\phi_t(x) \to p_i$ as $t \to \mp\infty$ is deonted by $w^{u,s}(p_i)$, the (global) unstable and stable manifolds of p_i. In the interior of $M, W^u(p_i)$ and $W^s(p_i)$ and 1-1 immersed discs of dimension u and s prespectively. The manifold M is the disjoint union of these stable manifolds, $M = \bigcup_{i=1,\ldots m} W^s(p_i)$. When there is no boundary M is also the union of the unstable manifolds $M = \bigcup_{i=1,\ldots,m} W^u(p_i)$. Now add another condition, which was introduced by Smale, that these manifolds $W^s(p_i)$, $W^u(p_i)$ are all transversal wherever they meet. The set of such f remains open and dense. The vector fields $V = -\operatorname{grad} f$ are called Morse-Smale.

Example 1

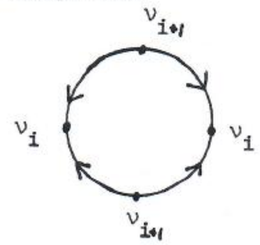
Let $f(z) = \sum_{i=0}^{d} a_i z^i$ with $a_i \in \mathbb{C}$ and $z \in \mathbb{C}$ the complex numbers be a complex polynomial of degree d such that f and f' have simple roots. We consider $\mathbb{C}$ as E^2 and let r be large enough so that B_r contains all the roots of f. Then $|f(z)|^2$ defines a Morse function on B_r and $-\operatorname{grad} |f(z)|^2$ is generally a Morse-Smale vector field.

Example 2

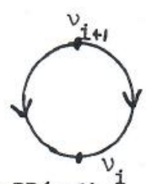
Let A be a real symmetric matrix with distinct eigenvalues $\lambda_1 < \lambda_2 < \ldots < \lambda_n$ and corresponding unit eigenvector $\nu_1 \ldots \nu_n$. Then $f(x) = \frac{1}{2} \frac{\langle x, A(x) \rangle}{\langle x, x \rangle}$ defines a Morse-function on the sphere S_1^{n-1}, here $\langle\ ,\ \rangle$ is the usual inner product in Euclidean space. The critical points of f are precisely $\pm\nu_i$ and the index of $\pm\nu_i$ is i-1. In fact on S_1^{n-1} grad $f(x) = A(x) - \langle x, A(x) \rangle X$ and one can explicitly solve $\phi_t(x) = \frac{e^{-tA}(x)}{||e^{-tA}(x)||}$ which is a Morse-Smale flow on S_1^{n-1}. The union of the unstable manifolds of $\pm\nu_1, \pm\nu_2, \ldots \pm\nu_i$ is the vector subspace spanned by $\nu_1, \ldots, \nu_i$ intersect S_1^{n-1}, while the union of the stable manifolds is the complement of the space spanned by $\nu_{i+1}, \nu_{i+2}, \ldots, \nu_3$

intersect S_1^{n-1}.

The function $f(x)$ is invariant under the indentification $x \sim -x$ on S_1^{n-1} and the flow $\phi_t(x)$ commutes with this identification. Thus f and ϕ_t induce a Morse function and a Morse-Smale flow on $S_1^{n-1}/x{\sim}{-x} = \mathbb{R}P(n-1)$ real projective (n-1) space. There is one critical point for each eigenspace corresponding to $\nu_1,\ldots,\nu_n$ of index (i-1). Thus there is one critical point for each dimension from 0 to (n-1). The intersection of the $W^u(\pm\nu_{i+1})$ and $W^s(\pm\nu_i)$ must occur in the plane of ν_i and ν_{i+1} intersect S_1^{n-1}. On this circle, the dynamics are always like



after identifying x -x on $\mathbb{R}P(n-1)$ we get



as the dynamics in the ν_i, ν_{i+1} plane in $\mathbb{R}P(n-1)$.

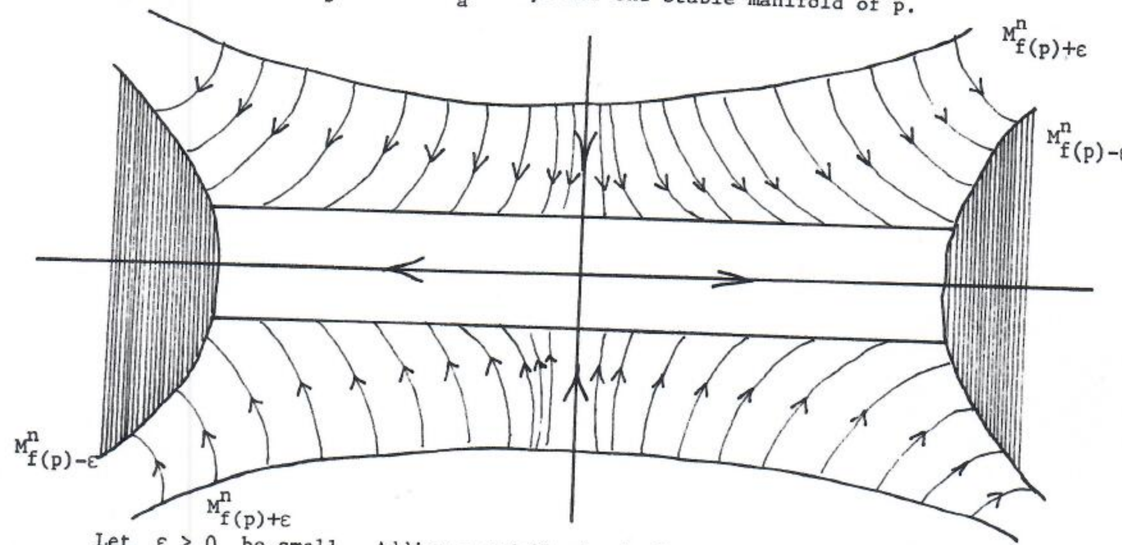
It is by now a standard result of Morse theory that passing a critical value adds a handle to the manifold. More precisely, let $f\colon M \to R$ be a Morse function. Let $M_a = f^{-1}(-\infty, a)$, so $\partial M_a = f^{-1}(a)$.

Theorem I

Suppose that $f\colon M^n \to R$ is a Morse function. If $a < b$ and $M_b^n - M_a^n$ contains exactly one critical point p of index i, then M_b^n is diffeomorphic to $M_a^n \cup_\phi S^{i-1} x D^{n-i}$ where ϕ is a diffeomorphism of $S^{i-1} x D^{n-i}$ into the boundary of M_a.

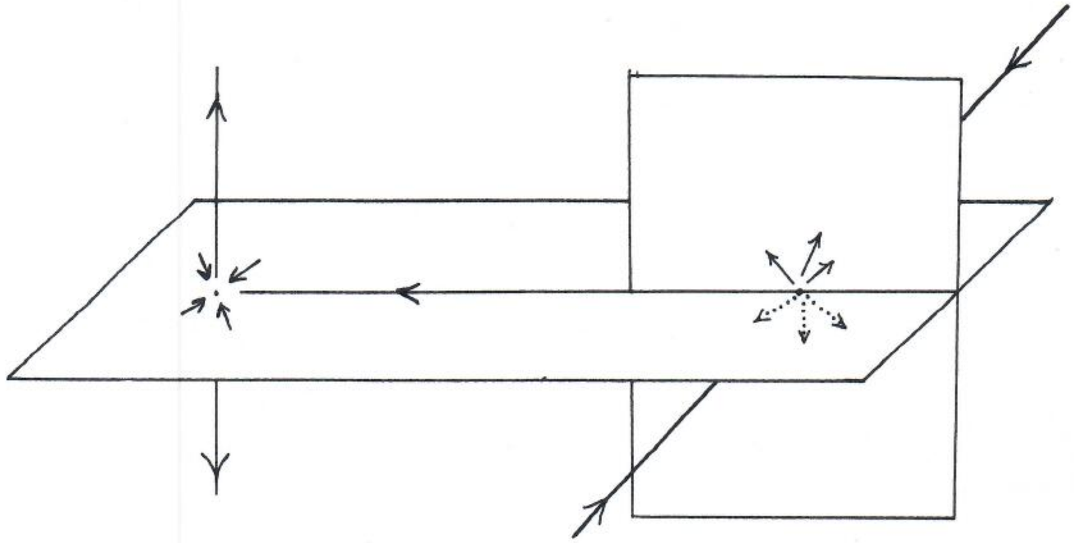
The proof of this theorem is a local argument near the critical point p.

In this form the theorem is due to Smale, see Smale 1961a. For general references on Morse theory see Bott 1982 and Milnor 1963. The gradient flow $-\text{grad } f$, pushes M_b^m down to M_a^n except for the stable manifold of p.



Let $\varepsilon > 0$ be small. Adding a neighborhood of a disc in the unstable manifold of p (which intersects $\partial M^n_{f(p)-\varepsilon}$ transversally) to $M^n_{f(p)-\varepsilon}$ produces a manifold diffeomorphic to $M^n_{f(p)+\varepsilon}$. Now since there are no singularities of f in $M_b - M_{f(p)+\varepsilon}$ or $M_{f(p)-\varepsilon} - M_a$ pushing along the solutions curves of $-$ grad f produces diffeomorphisms between M_b and $M_{f(p)+\varepsilon}$ and $M_{f(p)-\varepsilon}$ and M_a.

Smale 1961a, 1962b, 1962 exploits this structure in his work on the Poincare conjecture, h- cobordian theorem and structure of manifolds. A good exposition is given in Milnor 1965 which emphasizes the gradient approach. We turn to some of these results, which we summarize in one theorem.



Let $-$ grad f be a Morse-Smale vector field. Choose local charts for all the critical points of f so that $f(x) = f(p) - x_1^2 - \dots x_u^2 + x_{u+1}^2 + \dots + x_n^2$ for x near p. This has the effect of orienting the neighborhood of p, $W^u(p)$, $W^s(p)$ as E^n, E^u and E^s with the usual orientation. If p,q are critical points of index i+1 and i respectively then $W^u(p)$ has dimension i+1 while $W^s(q)$ has dimension n-i. The transversality hypothesis thus implies that $W^u(p) \cap W^s(q)$ consists of a finite number of orbits of the gradient flow ϕ_t, $\phi_t(m_1)$, ..., $\phi_t(m_j)$. For each m_i we may orient a basis of complementary space to $W^s(q)$ in two ways, one from the $W^u(q)$ orientation, and one that comes from adding $(-\text{ grad } f)(m_i)$ as the first element of a basis and using the $W^u(p)$ orientation. If these two orientations agree we assign +1 as the index of the intersection; if not, -1. Let $i(p,q) = \sum_{\phi_t(m_i) \subset W^u(p) \cap W^s(q)} \text{index } \phi_t(m_i)$. If $p_1,\dots,p_k$ and $q_1,\dots,q_r$ are the set of critical points of index i+1 and i respectively we let M_{i+1} be the (rxk) matrix whose (s,t) entry is $i(q_t,p_s)$.

Theorem II (Smale)

Let $f\colon M^n \to R$ be a Morse function with $-$ grad f Morse-Smale, then:

A) (Morse inequalities)

There is a finitely generated chain complex of free abelian groups $0 \to C_n \to C_{n-1} \to \cdots \to C_o \to 0$ determined by f with rank C_i equal to the number of critical points of index i and $\partial_i = M_i$ in a basis, which gives the homology of M^n.

B) (Structure of Manifolds)

Conversely, if $\Pi_1(M^n) = 0$, $n \geq 6$ and $0 \to C_n \to C_{n-1} \to \cdots \to C_0 \to 0$ is a finitely generated chain complex of free abelian groups which has as homology the homology of M, then this complex arises from a Morse function on M^n as in part A.

REMARKS: I think that this is a beautiful theorem. It serves as a prototype for theorems relating dynamics and topology. (A non-simply connected version of this theorem is proven in Maller 1980.) Part A is by far the simpler part of this theorem. Without the explicit computation of the boundary it is even more classical, and does not depend on the transversality condition. I've called Part A the Morse inequalities because they follow from the theorem with a little algebra.

Corollary 1

Let f: $M^n \to R$ be a Morse function. Let c_i be the number of critical points of f with index i and let B_i be the i^{th} Betti number with coefficients in a field F. Then one has the following inequalities:

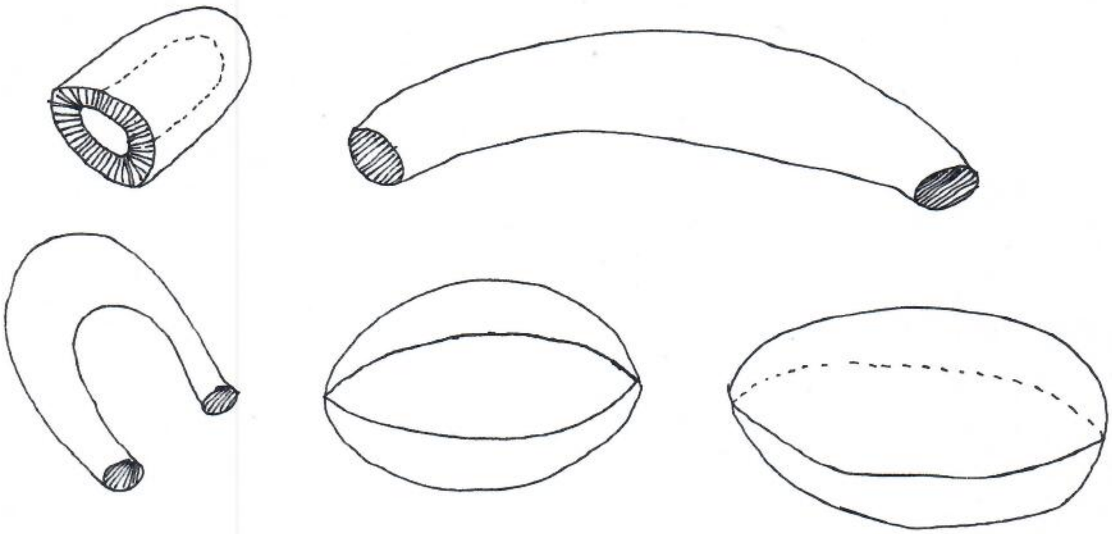
$$c_0 \geq B_0$$

$$c_1 - c_0 \geq B_1 - B_0$$

$$\sum_{k=0}^{n} (-1)^k c_k = \sum_{k=0}^{n} (-1)^k B_k$$

Proof: We can perturb f a little if necessary without changing the critical points or their indices to make the transversality hypothesis valid and thus apply Part A of the theorem. Since F is a field $C_i \otimes F$ is a vector space and we can write $C_i \otimes F \cong B_i \oplus H_i(M,F) \oplus B_{i-1}$ where $B_i \subset C_i \otimes F$ is the image $\partial_{i+1}(C_{i+1} \otimes F)$ The inequalities of the corollary are now evident.

The proof of the theorem is harder and beyond the scope of what I hope to do here, but Part A is especially instructive and I'll sketch the argument a bit. By the transversality hypothesis $W^u(p) \cap W^s(q) = \emptyset$ if index q $\geq$ index p. Thus M^n can be built first from the 0- handles followed by attachments of 1- handles followed by attachments of 2 - handles, etc.



(An i-handle is $D^i \times D^{n-i}$ which is attached by a diffeomorphism $\emptyset$ defined on $\partial D^i \times D^{n-i}$. $D^i \times 0$ is called the core disc and $0 \times D^{n-i}$ the transverse disc.) This can be seen from the proof of Theorem 1. More formally, there is a sequence of submanifolds, called a handle decomposition of M^n_1.

$\emptyset \subset M_0^n \subset M_1^n \subset \cdots \subset M_n^n = M^n$ such that $M_{i+1}^n = M_i^n \cup P_1^{i+1} \cup \cdots \cup P_s^{i+1}$ where P_k^{i+1} is a (i+1) handle. Now ..... $\to H^{i+1}(M_{i+1}^n, M_i^n) \to H^i(M_i^n, M_{i-1}^n) \to$ ..... is the complex of Theorem 2 Part A.

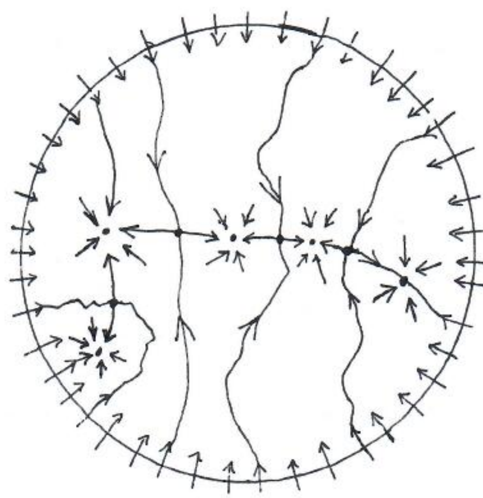
Examples

In the second example which we considered RP(n-1), C_i has rank 1 for $0 \le i \le n-1$ and $M_i = (\pm 2)$ or (0) as i is even or odd respectively, $i \neq 0$.

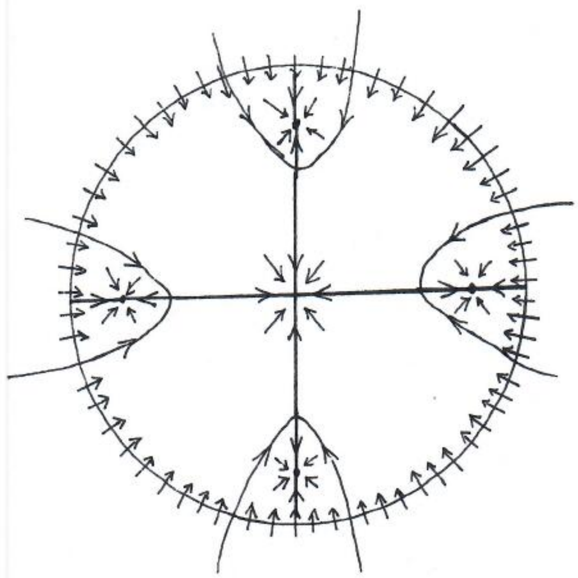
Thus

$$\begin{aligned} &H_0\ (\mathbb{R}P(n-1)) = Z \\ &H_i\ (\mathbb{R}P(n-1)) = 0 \quad \text{for } i \text{ even not } 0 \\ &H_i\ (\mathbb{R}P(n-1)) = Z_2 \quad \text{for } i \text{ odd not } (n-1) \\ &H_{n-1}(\mathbb{R}P(n-1)) = Z \quad \text{if } (n-1) \text{ is odd.} \end{aligned}$$

The first example is simpler from the Morse inequality point of view. There are d minima corresponding to the roots of f and (d-1) saddles which occur at the roots of f'. The Morse inequalities simply assert that $d - (d-1) = 1$ which is the Euler characteristic of the ball. But identifying the stable and unstable manifolds of the saddle points is a more difficult problem. Generally, that is for an open and dense set of full measure, there will be no saddle connections. That is if p and q are saddle points then $W^u(p) \cap W^s(q) = \emptyset$ and the flow will be Morse-Smale. In this case it is simple to see that the two components of the stable manifold of a saddle must both tend to infinity and the two components of the unstable manifold must tend to distinct roots. Various configurations are possible. For example:



and



are conceivable for 5th degree equations. The latter occurs for $z(z^{4-1})$.

$-\text{grad}\ |f(z)|^2 = -2f(z)\overline{f'(z)}$. If we let $\rho(z) = \frac{1}{2f'(z)\overline{f'(z)}}$ which is a positive real number we see that $-\text{grad}\ |f(z)|^2 = \rho(z)\ \frac{-f(z)}{f'(z)}$. Let $N_f(z) = N(z) = \frac{-f(z)}{f'(z)}$ be the Newton vector field and $\dot{z} = N(z)$ the Newton differential equation. Thus $-\text{grad}\ |f(z)|^2 = \rho(z)\ N(z)$, and up to reparameterization $-\text{grad}\ |f(z)|^2$ and $N(z)$ have the same orbits, that is they have the same solution curves. If we let $w = f(z)$ we see that $f'(z)\ \frac{(-f(z))}{(f'(z))} = -f(z) = -w$ and thus f maps solution curves of $\dot{z} = N(z)$ to solution curves of $\dot{w} = -w$. These latter are the half rays pointing to the origin. We state these simple geometric facts as a proposition.

<u>Proposition 1</u>

Let $f(z) = \sum_{i=0}^{d} a_i z^i$ be a complex polynomial.

a) The image by f of a solution curve of $-\text{grad}\ |f(z)|^2$ or $N_f(z)$ through the point z_0 lies on the half ray through $f(z_0)$ pointing towards the origin. If z_0 is not on the stable or unstable manifold of a critical point the image is the entire half ray. If $z_0 \in W^s(\theta)$ or $z_0 \in W^u(\theta)$ for a saddle point θ, then the image terminates at $f(\theta)$.

b) If $f(z_0) = w$ and $f'(z_0) \neq 0$ then the solution curve of $-\text{grad}\ |f(z)|^2$ or $N_f(z)$ through z_0 is the image of the half ray through w by the analytic continuation of the branch of f^{-1} taking w to z_0, $f^{-1}_{z_0}$.

c) If $f(\xi) = 0$ and $f'(\xi) \neq 0$ then the stable manifold $W^s(\xi)$ is the image by the analytic continuation f^{-1}_{ξ} which is defined on the whole complex plane minus a certain number of half lines from infinity to $f(\theta_i)$ where θ_i, $i = 1, \ldots, k$ are the critical points of f on the boundary of $W^s(\xi)$.

The proof of this elementary proposition is by the comments above and the pictures.

It is tempting to try to find a solution to the polynomial equation $f(z) = 0$ by picking a point z_0 and either:

a) Take the solution curve of $- \text{grad} \, |f(z)|^2$ through z_0. With a finite number of exceptions this curve tends to a root of $f(z)$.

b) Take the solution curve of $N(z)$ through z_0. With a finite number of exceptions this curve tends to a root of $f(z)$.

c) Take $f_{z_0}^{-1}$ of the ray $(1-h)f(z_0)$ for $0 \le h \le 1$. With the exception of a finite number of rays substituting $h = 1$ gives a zero of f.

As Smale points out in Smale 1981 one can prove the fundamental theorem of algebra by these methods. Method c) is the easiest. Moreover, many numerical methods for solving polynomial equations are intimately connected to these theoretical methods. For example, Euler's method for solving $\dot{z} = N(z)$ is $z' = z + hN(z)$, which for $h = 1$ is Newton's method $z' = z - \frac{f(z)}{f'(z)}$. Smale 1981 did an extensive study of the efficiency of the iterative methods $z' = z + hN(z)$. Then in Shub-Smale, 1982, 1983 we undertook the study of a wider class of methods.

I will digress for a moment to discuss iterative processes for the solution of a problem in general.

Let S be a topological space $U \subset S$ and $F: U \to S$ be a map. F is an iterative process. Given $X_0 \in U$ the forward orbit of X_0 is $\{X_n\}$ where $X_n = F(X_{n-1}) = F^n(x_0)$ as long as $X_{n-1} \in U$. The solutions to a problem are specified by a subset $P \subseteq S$. Given an iterative process F, the solutions of a problem P and an initial point $X_0 \in U$, then X_n converges to a solution if either $X_n \notin U$ for some n but $X_n \varepsilon P$ or if X_n is defined for all $n \varepsilon N$ and all the accumulation points of X_n are in P. An iterative process F for the solution of a problem P is called <u>locally convergent</u> if $P \subset U$ and there is a neighborhood V of P, $P \subset V \subset U$ such that for any initial point $X_0 \varepsilon V, X_n$ converges to a solution; it is globally convergent if $U = S$ and X_n converges to a solution for any initial point X_0. If S is a metric space and X_n converges to a solution then the convergence is first order or linear if there is a constant C, $0 \leq C < 1$ such that $d(X_{n+1}, P) \leq C\, d(X_n, P)$ for $n \geq 0$ and for $k>1$ that the convergence is kth order (quadratic, cubic for $k = 2,3$) if there is a constant $C \geq 0$ such that $d(X_{n+1},P) \leq Cd(X_n,P)^k$.

<u>Example</u>:

Let A be a real nxn symmetric matrix. Let $S = S_1^{n-1}$ and F be time one map of the flow $-$ grad $\left(\frac{1}{2}\frac{\langle x,A(x)\rangle}{\langle x,x\rangle}\right)$ that is $F(X) = \dfrac{e^{-A}(x)}{||e^{-A}(x)||}$

Let $P = \{x \varepsilon S_1^{n-1} | F(x) = \lambda x\}$ that is P is the eigenvectors of A. If the eigenvalues of A are distinct then P is a finite set, but if A has k equal eigenvalues then P includes the (k-1) sphere in the eigenspace of these eigenvalues. F is a globally convergent iterative process for the solution of the symmetric eigenvector problem. The convergence is linear for any initial point. All this is easy to see simply diagonalize A and compute in this system.

Of course, this method isn't practical. It involves computing e^{-A}. There is another problem as well which requires some thought. Even in the (2x2) case

with distinct eivenvalues the convergence is linear but will take longer and longer to approach P as the initial point X_0 is closer and closer but not equal to a source. I will return to this point shortly, but first examine the other example we have followed in terms of these notions.

Let $f(z)$ be a complex polynomial and ϕ_t the time t map of the solutions of the equations $\dot{z} = -\operatorname{grad} |f(z)|^2$. We consider ϕ_t defined on $\mathbb{C}$ on the Riemann sphere $\overline{\mathbb{C}} = \mathbb{C} \cup \infty = S^2$ or on B_r where r is large enough so that B_r contains all the roots of f. Let P be the set of points $\{\xi | f(\xi) = 0\}$, that is P is the roots of f. Then ϕ_t is an iterative process to find the roots of f. ϕ_t is not globally convergent but does converge for almost every initial point X_0, this follows from the discussion above. In the case that the r ots of f are all simple, this convergence is linear. Of course, this doesn't seem a practical method either since ϕ_t is not known.

Newton's method $z' = z - \frac{f(z)}{f'(z)}$ is locally convergent near the roots of f and when the roots of f are all simple Newton's method is quadratically convergent near the roots of f, in fact with a uniform constant C. This last statement follows from the fact that the roots of f are fixed points of $z' = z - \frac{f(z)}{f'(z)}$ and the derivative of $z - \frac{f(z)}{f'(z)}$ is 0 at a simple root of f.

Newton's method is a rational map of the Riemann sphere. If we ignore the trivial case d=1 and suppose that $f(z)$ is not $(z-a)^d$ for any a then $z - \frac{f(z)}{f'(z)}$ has degree ≥ 2 as a map of S^2. The dynamics of these maps have been extensively studied. Julia and Fatou in the beginning of the century and recently Sullivan, Douady, Hubbard and others have made important contributions. Clearly there can be no continuous iteration process of the sphere fixed at the roots of f (recall that there are at least two distinct ones) and globally convergent.

What should one be content with as an approximation to a solution to a problem. There are various reasonable notions. I'll list a few.

Let S be a metric space, $F \cdot U \to S$ an iterative process and $P \subset S$ the solutions to a problem. Then

1) $X_0 \in S$ in an <u>ε - solution</u> of P if $d(X_0, P) < \varepsilon$

2) $X_0 \in S$ is a first order approximate solution of P if X_n converses to a solution of P and if $d(X_0,P) < 1/2$ and $d(X_n,P) < 1/2^n$ for $n > 0$.

3) $X_0 \in S$ is a nth order approximate solution for $k > 1$ if X_n converges to a solution of P and if $d(X_0,P) < 1/2$ and $d(X_n,P) < \frac{1}{(2)^{k^n}}$ for $n > 0$.

Frequently P is defined as the zeros of a function Φ, $\Phi: S \to R$ or $\mathbb{C}$ $P = \Phi^{-1}(0)$. In this case we can speak of zeros.

1) $X_0 \in S$ is an ε - zero of Φ if $|\Phi(X_0)| < \varepsilon$.

2) $X_0 \in S$ is a first order approximate zero of Φ if X_n converges to a solution of P and if $|\Phi(X_0)| < 1/2$ and $|\Phi(X_n)| < 1/2^n$ for $n > 0$.

3) $x_0 \in S$ is a k^{th} order approximate zero of ϕ for $k > 1$ if x_n converges to a solution of P and if $|\phi(x_0)| < 1/2$ and $|\phi(x_n)| < \frac{1}{(2)^{k^n}}$ for $n > 0$.

In many ways the third alternative is the most attractive if F is not too difficult to iterate, for this with a few iterations of F the error dies rapidly and one is sure of convergence to a solution.

We may use iteration processes to design algorithms to find ε - solutions or zeros or approximate solutions or zeros of problems. Some of the issues involved are:

1) How does one find a good initial point x_0?

2) Having picked an x_0 should one stick with it stubbornly or after awhile give up and pick another?

3) Should one pick one point or several and run the iteration in parallel, stopping when one of them gives an adequate answer?

4) Fast methods are frequently not sure methods, near certain "bad"

subsets they may take very long to work.

5) What are average estimates for the work involved in solving many randomly chosen problems?

Now I return to the work of Shub-Smale 1982,1983 on algorithms for solving polynomial equations. We are to take $f^{-1}_{z_0}$ $((1-h)f(z_0))$. As long as $f'(z_0) \neq 0$, $f^{-1}_{z_0}$ is defined by a power series in h which can usually be analytically continued along the ray from $f(z_0)$ to 0. Since evaluation of an infinite power is infeasible we truncate the series at degree k in h. We write t_k to indicate the truncati

$$E_{k,h,f}\ (z_0) = t_k f^{-1}_{z_0}\ ((1-h)f(z_0))$$

$E_{k,h,f} : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ is a rational function of z. It is easy to see that $E_{1,1,f}$ is Newton's method. With h=1 and k=1,2 ... , 5 these iterations were used by Euler to solve equations.

If h=1 and z_0 is a simple root of f then $E_{k,1,f}$ (z_0) vanishes to order k in z and thus there is a neighborhood $\cup$ of z_0 such that any $x_0 \in \cup$ is a $(k+1)^{st}$ order approximate zero of f for $E_{k,1,f}$.

Much of our analysis of these iterations is based on the following theorem:

Theorem 3 (Shub-Smale, 1982)

Let $z \in \Omega \subset \mathbb{C}$ and let $g: \Omega \to \mathbb{C}$ be analytic. Suppose g_z^{-1} is defined on a disc D of radius $R(g,z)$. There are constants c_k and K_k depending on k ($c_k \approx 1$ and $K_k \simeq k$) such that if $|w-g(z)| < c_k R(g,z)$ and $gg_z{}^{-1}(w) = w$

then $$|g((t_k g_z{}^{-1})(w))-w| \leq \frac{K_k|w-g(z)|^{k+1}}{R(g,z)^k}$$

The proof of this theorem is rather technical and involves estimates on the coefficients of univalent functions and their inverses along the lines of the Bieberbach conjecture. As a rapid corollary we can give a criterion for a point to be an approximate zero.

Proposition 2

Let ξ be a simple root of the complex polynomial f and let $\rho_{f,\xi} = \min |f(\theta)|$ over critical points θ in the boundary of the stable manifold of ξ. Then f_{ξ}^{-1} is defined on the disc of radius $\rho_{f,\xi}$. Let $|w| < \min(1/2, \rho_{f,\xi}/20)$ then $f_{\xi}^{-1}(w)$ is a $(k+1)^{st}$ order approximate zero of f for $E_{k,1,f}$. Let $\rho_f = \min \rho_{f,\xi}$ over the simple roots ξ of f and 0 if f has a double root. Thus, if $|f(z_0)| < \min\left(\frac{1}{2}, \frac{\rho_f}{20}\right)$ then z_0 is a $(k+1)^{st}$ order approximate zero of f for $E_{k,1,f}$.

Proof

Let $|w| < \min\left(\frac{1}{2}, \rho_{f,\xi}/20\right)$ and $z_0 = f_{\xi}^{-1}(w)$. I claim inductively that for $n \geq 1$, $|f(z_n)|$ is monotonically decreasing and $|f(z_n)| < \min\left(\frac{1}{2k}\frac{1}{2^{(k+1)^n}}, \frac{1}{2k}\frac{1}{2^{(k+1)^n}}\rho(f,\xi)\right)$. Apply the theorem to $f(z_{n+1}) = f(t_k f_{z_n}^{-1}(0))$. Thus $|f(z_{n+1})| \leq K_k \dfrac{|f(z_n)|^{k+1}}{R(f,z_n)^k} = K_k R(f,z_n)\left(\dfrac{|f(z_n)|}{R(f,z_n)}\right)^{k+1}$.

Now consider two cases $R(f,z_n) \geq 1$ and $R(f,z_n) < 1$

$$\frac{21}{20}\rho_{f,\xi} \geq R(f,z_n) \geq \frac{19}{20}\rho_{f,\xi}$$

If $R(f,z_0) < 1$ then $|f(z_1)| < K_k \left|\dfrac{\rho_{f,\xi}/20}{\frac{19}{20}\rho_{f,\xi}}\right| = K_k\left(\frac{1}{19}\right)^{k+1} < \frac{1}{2k}\frac{1}{2^{k+1}}$

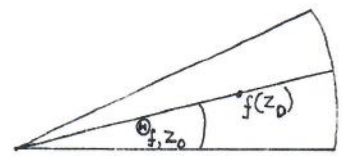
but also

$$|f(z_1)| \leq K_k |f(z_0)| \left(\frac{f(z_0)}{R(f,z_0)}\right)^k \leq \frac{K_k \rho_{f,\xi}}{20} \cdot \frac{1}{19^k} < \frac{1}{2k} \cdot \frac{1}{2^{k+1}} \rho_{f,\xi}$$

Now proceed by induction. The case $R(f,z_n) \geq 1$ is handled similarly.

Theorem 3 may also be used to estimate how well $E_{k,h,f}$ does in following the ray from $f(z_0)$ to 0 towards 0. Given f and z_0 define Θ_{f,z_0} be the largest angle less than or equal to $\pi/2$ so that $f_{z_0}^{-1}$ is defined on the open wedge of a circled centered at 0 with radius $2|f(z_0)|$ and of angle

Θ_{f,z_0} on both sides of the ray through $f(z_0)$



$f_{z_0}^{-1}$ is defined on this open wedge.

Θ_{f,z_0} is non-zero for any z_0 such that $f(z_0)$ does not lie on a ray containing a critical value.

Theorem 4 (Shub-Smale 1982)

There is a constant C_k, depending only on k such that if f(z) is a complex polynomial, if $\Theta_{f,z_0} > 0$ and $|f(z_0)| > L > 0$ then there is an h given explicitly such that

$$|f(z_n)| < L \text{ for}$$

$$n = C_k \left[\frac{\log \frac{f(z_0)}{L}}{\Theta_{f,z_0}} \right]^{k+1/k}$$

and

$$z_n = (E_{k,h,f})^n (z_0).$$

C_k decreases with increasing k to around 6. This theorem indicates that a good starting point is a point z_0 with Θ_{f,z_0} large. Let $P_d(1)$ be the set of polynomials f of degree d such that $f(z) = z^d + a_{d-1} z^{d-1} + \dots a_0$ where $|a_i| \leq 1$.

The rest of this discussion comes from Shub-Smale 1983.

Corollary 2

There exist Universal constants H,K so that for $n = K(d + |\log \varepsilon|)$, $E = E_{k,H,f}$, $f \in P_d(1)$ and $|z_0| = 3$ with $\theta_{f,z_0} \geq \pi/12$

$$|f(E^i(z_0))| < \varepsilon \text{ for some } 0 \leq i < n.$$

Now the question is, how likely is Θ_{f,z_0} to be $\geq \pi/12$. For $f \in P_d(1)$, let $V_f = \{z \mid |z| = 3 \text{ and } \theta_{f,z} > \pi/12\}$. Then using the uni orm probability measure on $S = \{z \mid |z| = 3\}$ we have:

Proposition 3 The measure of $V_f \geq 1/6$ for any f in $P_d(1)$.

The proof depends on the geometry that we have developed in Proposition 1.

Consider first the problem: Given (f,ε), $f \in P_d(1)$, $\varepsilon > 0$, produce a $z \in C$ with $|f(z)| < \varepsilon$. For this we particularize the Newton-Euler iteration scheme by choosing k and h to depend only on f and ε, in a certain way. Let

$$k = \lceil \max(\log|\log \varepsilon|, \log d) \rceil$$

where $\lceil x \rceil$ is the least integer greater than or equal to x. There are universal constants H and K, approximately 1/512 and 512 respectively. Then take

$$h = H$$

Thus with these specializations the Newton-Euler iteration scheme $E: \mathbb{C} \to \mathbb{C}$ depends only on (ε, f) and we write $E_\varepsilon = E$. With $\varepsilon > 0$ define:

Algorithm$(N\text{-}E)_\varepsilon$: Let $f \in P_d(1)$ and $n = K(d + |\log \varepsilon|)$.

(1) Choose $z_0 \in \mathbb{C}$, $|z_0| = 3$ at random and set for $i = 1,2,3,\ldots$ (an **iteration**) $z_i = E_\varepsilon(z_{i-1})$ terminating if ever $|f(z_i)| < \varepsilon$.

(2) If $i = n$, go to (1) (a cycle).

Theorem A

For each f, ε, $(N\text{-}E)_\varepsilon$ terminates with probability one and produces a z satisfying $|f(z)| < \varepsilon$. The average number of cycles is less than or equal to 6. Hence the average number of iterations is less than $6K(d + |\log \varepsilon|)$.

Here average and probability refer to the choice of the sequence of z_0 in (1) of $(N\text{-}E)_\varepsilon$.

Remark

With certainty it only takes about twice as long. In practice one can obviously do better by trying and testing $h = 1, 1/2, \ldots, H$. We haven't analyzed this. The average of the total number of arithmetic operations required is $O(d^2 + d|\log \varepsilon|)$.

In Algorithm $(N\text{-}E)_\varepsilon$ there was a random element, the choice of z_0. Now probability enters into our analysis in a second way. We average over $f \in P_d(1)$, with respect to a uniform distribution that is we normalize Lebesgue measure on $P_d(1) \subset \mathbb{C}^d = R^{2d}$. We use these probabilities since speedy algorithms are not usually infallible.

Define for each $f \in P_d(1)$

$$\varepsilon_f = \frac{1}{(2d)^{4d}} |Df|$$

where D_f is the discriminant of f (see Lang). With K as above let

$$n = K(d + |\log \varepsilon_f|).$$

Let E be the Euler-Newton iteration process with $h = H$, and $k = [\max(\log|\log \varepsilon_f|, \log d)]$, so that E depends only on f.

Algorithm N-E Let $f \in P_d(1)$, satisfy $\varepsilon_f > 0$.

(1) Set $m = 1$

(2m) Choose $z_0 \in \mathbb{C}$, $|z_0| = 3$ at random and set $z_n = E^n(z_0)$. If $|f(z_n)| < \varepsilon_f$ terminate and print: "z_n is an approximate zero."

(3) Otherwise let $m = m + 1$ and go to (2m).

<u>Theorem B</u>

Algorithm N-E terminates (and hence produces an approximate zero) with probability 1 and the average number of iterations is less than $K_1 d \log d$ where K_1 is a universal constant.

We make the probabiility considerations a bit more precise.

Let S_R^1 be the circle in C defined by $|z| = R$ and endow it with the uniform probability measure (Lebesgue measure normalized to 1). Set $R = 3$ and denote by Ω the product of S_R^1 with itself a countable number of times. Thus a point z_0 of Ω is a sequence of $\bar{z} = (\bar{z}_1, \bar{z}_2, \ldots)$ with $|z_i| = 3$. Endow Ω with the product measure as well as $P_d(1) \times \Omega$. Let T: $P_d(1) \times \Omega \to Z^+$ be defined by $T(f,z_0)$ is the first m such that $E^n(\bar{z}_m) < \varepsilon_f$.

Thus the total number of iterations of Algorithm N-E for a given f is of the form $S(f,\bar{z}) = nT(f,\bar{z})$, $n = K(d + |\log \varepsilon_f|)$. Theorem B asserts that when $\varepsilon_f > 0$, $S(f,\bar{z})$ is defined for almost all $\bar{z} \in \Omega$. Moreover $S(f) = \int_{\bar{z}\in\Omega} S(f,\bar{z})$ is defined and finite for almost all f and

$$\int_{f\in P_d(1)} S(f) \leq K_1 d \log d.$$

By Fubini's theorem, we could equally well assert that

$$\int_{(f,\bar{z})\in P_d(1) \times \Omega} S(f,\bar{z}) \leq K_1 d \log d$$

<u>Remark 1</u>

We are assuming exact arithmetic in the theory here. In general, because of the robust properties of Algorithms $(\text{N-E})_\varepsilon$ and (N-E), this is reasonable. Myong Kim is incorporating finite precision and round off errors into this theory in her thesis.

Remark 2

Our work emphasizes the theoretical side, and the understanding of classic algorithms, rather than the design of new practical algorithms. Yet the results do have some implications for the latter. For example they suggest calculating derivatives up to order $[\log d]$ and/or $[\log|\log \varepsilon|]$ could give speedier routines, especially for one complex polynomial. We haven't tested our algorithms on the machine.

Remark 3

The number of arithmetic operations in contrast to the number of iterations is $O\left(d^2(\log d)^2 \log\log d\right)$. The average result here depends on a result from Smale 1981.

Proposition 4

$$\mathrm{Vol}\{f\varepsilon P_d(1) \mid \rho_f < \alpha\} < d\alpha^2$$

Her Vol means normalized volume so that $\mathrm{Vol}(P_d(1)) = 1$ and Vol is a probability measure on $P_d(1)$.

It should be pointed out that we have taken a flexible approach. We have not insisted on sticking with our initial z_0. Average estimates in the stubborn case from Smale 1981 or Shub-Smale 1982 still yield infinite averages in the stubborn case. There are many problems remaining here, see Shub-Smale 1982, 1983.

It is perhaps instructive to work out a foolish but simple infinite average case.

For $0 \leq \varepsilon \leq 1$ let A_ε be the symmetric matrix.

$$A_\varepsilon = \begin{pmatrix} \varepsilon & 0 \\ 0 & -\varepsilon \end{pmatrix}$$

and $$F_\varepsilon(\nu) = \frac{e^{-A_\varepsilon}\nu}{\|e^{-A_\varepsilon}\nu\|}$$

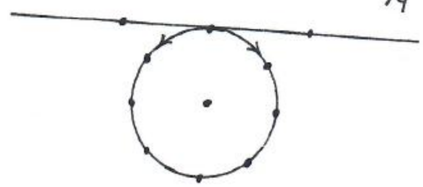
for $\nu \in S_1^1$. Then $F_\varepsilon(\nu)$ is a globally convergent iteration process to find the eigenvectors of A_ε. Let us try to use this iterative process to find approximate eigenvectors of A_ε.

Algorithm 1) Pick $v_0 \in S_1^1$ at random

2) Let $v_n = F_\varepsilon(v_{n-1})$

3) If v_n is an approximate eigenvector of A_ε stop and print "v_n is an approximate eigenvector of A_ε" ; If not go to 2.

We have $\frac{1}{2}$ a chance of picking v_0 within $\pi/4$ of $(0,\pm 1)$



Use a chart obtained from central projection onto the tangent line through $(0,1)$ and restrict attention to the positive quadrant. Thus picking a point at random between $(0,1)$ and $\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$ on S_1^1 corresponds to picking a point at random between 0 and 1 on the tangent line up to bounded distortion. Finally F_ε in this chart is simply multiplication by $e^{2\varepsilon}$. Now choose δ in $(0,1)$ at random. If ε or $\delta = 0$ then δ already represents an eigenvector, and let $N(\varepsilon,\delta) = 0$. If $\varepsilon, \delta > 0$ then let $N(\varepsilon, \delta)$ be the minimum n such that $(e^{2\varepsilon})^n \delta > 1$. To find the average number of iterations for δ to leave $(0,1)$ involves integrating $N(\varepsilon,\delta)$ over the square.

<u>Lemma</u> For $x > 1$,

$$\sum_{n=1}^{\infty} n\left(\frac{1}{x^{n-1}} - \frac{1}{x^n}\right) = \frac{x}{x-1}$$

Proof: $\displaystyle\sum_{n=1}^{\infty} n\left(\frac{1}{x^{n-1}} - \frac{1}{x^n}\right) = \sum_{n=1}^{\infty} n\left(\frac{x-1}{x^n}\right)$

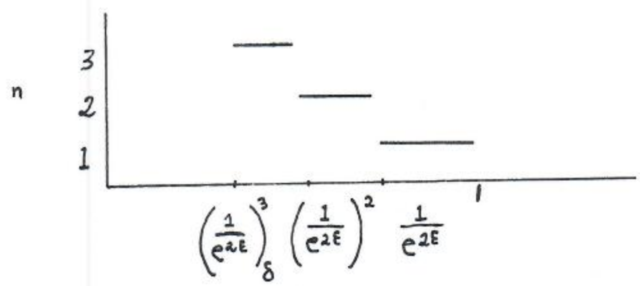
$$= \left(\frac{x-1}{x}\right)\left(\sum_{n=1}^{\infty} \frac{n}{x^{n-1}}\right)$$

$$= \left(\frac{x-1}{x}\right)\left(\frac{1}{1-U}\right)'\Bigg|_{U=\frac{1}{x}}$$

$$= \left(\frac{x-1}{x}\right)\left(\frac{1}{1-\frac{1}{x}}\right)^2$$

$$= \frac{x}{x-1}$$

Now for fixed $\varepsilon > 0$, average over δ .



$$Av_\varepsilon = \sum_{n=1}^{\infty} n\left(\frac{1}{(e^{2\varepsilon})^{n-1}} - \frac{1}{(e^{2\varepsilon})^n}\right) = \frac{e^{2\varepsilon}}{e^{2\varepsilon}-1}$$

Now integrating with respect to ε

$$\int_0^1 Av_\varepsilon \, d_\varepsilon = \int_0^1 \frac{e^{2\varepsilon}}{e^{2\varepsilon}-1} \, d\varepsilon = \frac{1}{2} \ln\left(e^{2\varepsilon}-1\right)\Big|_0^1$$

which diverges.

This example illustrates the danger of slowly repelling points. One might imagine that working still stubbornly but in parallel with two points might help matters.

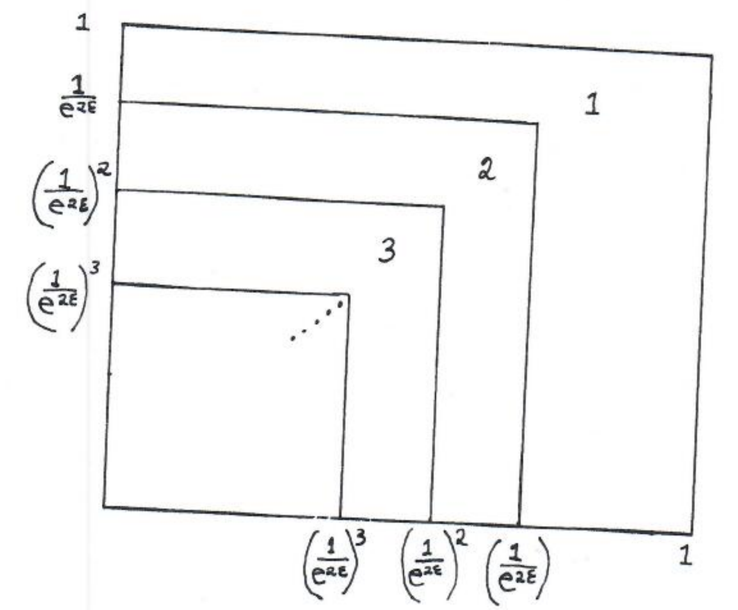
Algorithm

1) Pick $v_{0,1}$ and $v_{0,2} \in S_1^1$ at random

2) Let $v_{n,i} = F_\varepsilon(v_{n-1}, i)$ for $i = 1, 2$

3) If $v_{n,i}$ for i = 1 or 2 is an approximate eigenvector of A_ε stop and print, " $v_{n,i}$ is an approximate eigenvector of A_ε ", if not go to 2.

The comparable problem on the interval is to pick $\varepsilon, \delta_1, \delta_2 > 0$ at random define $N(\varepsilon, \delta_1, \delta_2)$ to be the minimum n such that $(e^{2\varepsilon})^n \delta_1 > 0$ or $(e^{2\varepsilon})^n \delta_2 > 0$. Now we want to integrate $N(\varepsilon, \delta_1, \delta_2)$ over the cube. For fixed $\varepsilon > 0$.



Thus $Av_\varepsilon = \sum_{n=1}^{\infty} n\left(\left(\frac{1}{e^{2\varepsilon}}\right)^{2(n-1)} - \left(\frac{1}{e^{2\varepsilon}}\right)^{2n}\right)$

$$= \frac{e^{4\varepsilon}}{e^{4\varepsilon}-1} \quad \text{and} \quad \int_0^1 \quad d\varepsilon = \int_0^1 \frac{e^{4\varepsilon}}{e^{4\varepsilon}-1}\, d\varepsilon$$

$$= \frac{1}{4} \ln\left(e^{4\varepsilon}-1\right)\Big|_0^1$$ which is still divergent.

No finite number of choices will help. It is better to stop after a certain fixed number of iterates and pick a new starting point. Blind luck is probably best.

Algorithm

1) Pick $v_0 \in S_1^1$ at random

2) If v_0 is an approximate eigenvector of A_ε stop and print, "v_0 is an approximate eigenvector of A_ε;" if not go to 2.

Since any vector within $\frac{1}{2}$ of $(1,0)$ is an approximate eigenvector for A_ε on the average π choices will produce an approximate eigenvector. If we choose randomly among a finite collection of points with at least one in each interval of length $\frac{1}{2}$ then this probabiliatie algorithm is sure to halt.

This example will illustrate many of the issues involved in the polynomial solution problem as well. Finding an ε- eigenvector is a quite different matter than an approximate eigenvector. But I'll leave this to the reader.

H. Bass, 1979, The Grothendieck Group of the Category of Abelian Group Automorphisms of Finite Order, Preprint, Columbia University.

H. Bass, 1981, Lenstra's Calculation of $G_0(R\pi)$ and Applications to Morse - Smale Diffeomorphisms in Integral Representations and Applications (Edited by K. W. Roggenkamp) Springer Lecture Notes in Mathematics Number 882, pp. 287-318.

S. Batterson, 1977, Structurally Stable Grassmann Transformations, Trans. American Mathematical Society, Vol. 231, pp. 385-404.

R. Bott, 1982, Lectures on Morse Theory Old and New, Bulletin (New Series) of the American Mathematical Society, Vol. 7, pp. 331-358.

J. Franks, 1982, Homology and Dynamical Systems, Conference Board of Mathematical Sciences Regional Conference Series in Mathematics, Number 49, American Mathematical Society, Providence, Rhode Island.

J. Franks and M. Shub, 1981, The Existence of Morse - Smale Diffeomorphisms Topology, Vol. 20, pp. 273-290.

D. Fried and M. Shub, 1979, Entropy Linearity and Chain Recurrence, Publ. Math. de e' I.H.E.S., Number 50, pp. 451-462.

B. Halpern, 1979, Morse - Smale Diffeomorphisms on Tori, Topology, Vol. 18, pp. 105-112.

M. W. Hirsch, C. C. Pugh, and M. Shub, 1977, Invariant Manifolds, Springer Lecture Notes in Mathematics Number 583, Springer Verlag, Berlin.

S. Lang, 1965, Algebra, Addison-Wesley, Reading, Mass.

H. W. Lenstra Jr., 1981, Grothendieck Groups of Abelian Group Rings, Journal of Pure and Applied Algebra, Vol. 20, pp. 173-193.

M. Maller, 1980, Fitted Diffeomorphisms of Non-Simply Connected Manifolds, Topology, Vol. 19, pp. 395-410.

M. Maller, 1981, Algebraic Problems Arising from Morse - Smale Dynamical Systems, Queens College preprint, to appear Proceedings of Rio Conference on Dynamical Systems, 1981.

M. Maller and J. Whitehead, 1981, Virtual Permutations of $\mathbb{Z}[\mathbb{Z}^n]$ Complexes, Queens College preprint, to appear Proceedings of the American Mathematical Society.

J. Milnor, 1963, Morse Theory, Annals of Mathematics Studies, Number 51, Princeton University Press, Princeton, New Jersey.

J. Milnor, 1965, Lectures on the h-Cobordism Theorem (Notes by L. Siebenmann and J. Sondow), Princeton Mathematical Notes, Princeton University Press, Princeton, New Jersey.

C. B. Moler, 1978, Three Research Problems in Numerical Linear Algebra in Proceedings of Symposia in Applied Mathematics, Vol. XXII, "Numerical Analysis" (Ed. by G. H. Golub and J. Oliger), American Mathematical Society, Providence, Rhode Island.

J. Palis and S. Smale, 1970, Structural Stability Theorems in Proceedings of Symposia in Pure Mathematics, Vol. XIV (Ed. by S. S. Chern and S. Smale), American Mathematical Society, Providence, Rhode Island.

B. N. Parlett, 1950, The Symmetric Eigenvalue Problem, Prentice Hall, Englewood Cliffs, New Jersey.

M. Shub, 1974, Dynamical Systems, Filtrations and Entropy, Bulletin of the American Mathematical Society, Vol. 80, pp. 27-41.

M. Shub and S. Smale, 1982, Computational Complexity: On the Geometry of Polynomials and a Theory of Cost, Part I, preprint.

M. Shub and S. Smale, 1983, Computational Complexity: On the Geometry of Polynomials and a Theory of Cost, Part II, preprint.

M. Shub and D. Sullivan, 1975, Homology and Dynamical Systems, Topology, Vol. 14, pp. 109-132.

S. Smale, 1961a, On Gradient Dynamical Systems, Annals of Mathematics, Vol. 74, pp. 199-206.

S. Smale, 1961b, Generalized Poincare's Conjecture in Dimension Greater than Four, Annals of Mathematics, Vol. 74, pp. 391-406.

S. Smale, 1962, On the Structure of Manifolds, American Journal of Mathematics, Vol. 84, pp. 387-399.

S. Smale, 1967, Differentiable Dynamical Systems, Bulletin of the American Mathematics Society, Vol. 73, pp. 747-817, with Notes and References in S. Smale, 1980, The Mathematics of Time, pp. 1-82, Springer, New York.

S. Smale, 1981, The Fundamental Theorem of Algebra and Complexity Theory, Bulletin (New Series) of the American Mathematical Society, Vol. 4, pp. 1-36.

B. T. Smith et al., 1976, Matrix Eigensystem Routines - Eispack Guide (2nd Edtition), Lecture Notes in Computer Science, Number 6, Springer Verlag, Berlin.

A. Vasquez, 1983, The Dynamics on Flags of a Generic Invertible Matrix, Class Notes, CUNY Graduate School.

D. S. Watkins, 1982, Understanding the QR Algorithm, SIAM Review, Vol. 24, pp. 427-439.

J. H. Wilkinson, 1966, Calculation of Eigensystems of Matrices in Numerical Analysis (Ed. by J. Walsh), Academic Press, London.

J. H. Wilkinson, 1968, Global Convergence of Tridiagonal QR Algorithm with Origin Shifts, Linear Algebra and Application vol. 1, pp. 409-420.

H. Wisniewski, Rate of Approach to Minima and Sinks - The Morse - Smale Case, Transactions of the American Mathematical Society, to appear.

Queens College and the Graduate School
of the City University of New York