On Algorithms for Solving $f(x) = 0$*

MORRIS W. HIRSCH AND STEPHEN SMALE
University of California, Berkeley

Introduction

We consider the problem of finding a solution for the equation $f(x) = 0$ when $f : \mathbb{R}^n \to \mathbb{R}^n$ is given. It is rarely possible to find a formula for such a point x. Usually the best one can do is to find an algorithm which produces some x' in $\mathbb{R}^n$ within a prescribed $\varepsilon > 0$ of x or, more commonly, with $|f(x')| < \varepsilon$.

In this work we explore variants of Newton's method. Our point of view is global and probabilistic: *for most initial points x_0, the algorithm should accomplish its task.* "Most" means for an open dense subset W^* of some easily described open set W such that $W - W^*$ has measure zero.

Our perspective on Newton's method is a geometric one. Let $f(x) = y \neq 0$. Let $L = L(y) \subset \mathbb{R}^n$ be a half-line

$$L = \{ty : t > 0\}$$

and consider the set $C(x)$ which is the connected component of x in $f^{-1}(L)$. If f is sufficiently smooth, then, for most x, $C(x)$ will be a smooth 1-dimensional submanifold through x.

There are two key points about these curves $C(x)$: (i) they are closed subsets in $\mathbb{R}^n - f^{-1}(0)$, (ii) if the derivative $Df(x)$ (i.e., the matrix of partial derivatives at x) is invertible (and in other cases too), then the tangent line to $C(x)$ at x is easily computed: in fact it is parallel to the vector $Df(x)^{-1}f(x)$.

Of course there is no guarantee that a solution exists, i.e., that if $f^{-1}(0)$ is not empty. Here the first observation above is useful: *if $C(x)$ not closed in $\mathbb{R}^n$*, its *closure $\bar{C}(x)$ contains a zero of f. In fact $\bar{C}(x) - C(x) \subset f^{-1}(0)$.*

A plausible strategy is to follow one of the curves $C(x)$ hoping it will lead us to a solution.

The classical Newton–Raphson iteration in fact consists simply in numerically integrating the vectorfield $X(x) = -Df(x)^{-1}f(x)$ by the Cauchy–Euler method: $x_{i+1} = x_i + hX(x_i)$, with $h = 1$. More generally, $h \in R$ could vary, for example as a function of x_i, etc. In this lies our point of departure.

Our algorithms are based on various hypotheses that guarantee the existence of a solution, and at the same time tell us at each stage in which direction to follow the curve $C(x)$.

*This paper was presented at the Courant Institute of Mathematical Sciences Conference on Scientific Computing, April 20–22, 1977, held to inaugurate the Courant Professorship of Mathematical Sciences. Reproduction in whole or in part is permitted for any purpose of the United States Government.

Communications on Pure and Applied Mathematics, Vol. XXXII, 281–312 (1979)
 0010-3640/79/0032-0281$01.00

We give a bit of historical background. The method of Sard's theorem and the idea of following curves were used early, in Hirsch [3], to give a short proof that there is no retraction of a cell onto its boundary. The argument was not constructive and the proof went by contradiction. Scarf [8] developed later, and independently, a combinatorial algorithm for finding fixed points of a continuous map of a cell into itself. Kellog, Li, and Yorke [5] used Sard's theorem, in a way that is close to the spirit of [3], to give an algorithm for finding fixed points of maps of a cell into itself. The "Global Newton method" of Smale [9] is also based on Sard's theorem and is oriented to the problem of solving systems of equations rather than finding fixed points. This last article is closest to what we are doing here. The parallel nature of Scarf's method and the global Newton method can be seen in Eaves–Scarf [2] and [9].

Geometrical background material for this paper is presented in Section 1. Versions of the main algorithm are given in Sections 2 and 4. In Section 5 a practical algorithm is given and computational experience is reported. Section 6 may be read independently. It contains an algorithm for finding a root of a complex polynomial; the algorithm *always* works, albeit slowly. Section 7 contains other algorithms that always work; they are derived from those in Sections 2 and 4.

1. The Curves $C(x)$ and Existence of Zeroes

In this section we derive geometrical properties of the curves $C(x)$. These are used to prove an existence theorem (Theorem 1.5) which will be the basis for the algorithms of later sections.

The following notation will be used.

$\mathbb{R}^n$ denotes Euclidean n-space, the set of n-tuples $x=(x_1, \cdots, x_n)$ of real numbers. The norm of x is $|x|=(\sum_{i=1}^n x_i^2)^{1/2}$; the inner product of vectors $x, y \in \mathbb{R}^n$ is $\langle x, y\rangle = \sum x_i y_i$. The unit sphere is

$$S^{n-1}=\{X\in\mathbb{R}^n : |x|=1\}.$$

The closed unit ball is

$$D^n=\{x\in\mathbb{R}^n : |x|\leqq 1\}.$$

Throughout this section we deal with a map

$$f : \mathbb{R}^n \to \mathbb{R}^n, \qquad f(x)=(f_1(x), \cdots, f_n(x)),$$

of differentiability class C^r, where r can be a positive integer, or ∞, or ω (meaning real analytic).

The derivative of f at x is the linear map $Df(x)$ (or Df_x) from $\mathbb{R}^n$ to $\mathbb{R}^n$ defined by the matrix $[(\partial f_i/\partial x_j)(x)]$. Its determinant is the Jacobian $J(x)\in\mathbb{R}$.

The set of regular points of f is

$$\operatorname{Reg}(f)=\{x\in\mathbb{R}^n : \operatorname{rank} Df_x = n\}.$$

The set of critical points is

$$\operatorname{Crit}(f)=\mathbb{R}^n-\operatorname{Reg}(f).$$

The set we are looking for is $E=f^{-1}(0)$.

Closely related f is the C^r map

$$g : \mathbb{R}^n - E \to s^{n-1}, \qquad g(x)=|f(x)|^{-1}f(x).$$

The set of regular points of g is

$$\operatorname{Reg}(g)=\{x\in\mathbb{R}^n-E : \operatorname{rank} Dg_x = n-1\}$$

and the set of critical points of g is

$$\operatorname{Crit}(g)=(\mathbb{R}^n-E)-\operatorname{Reg}(g).$$

Let $x\in\operatorname{Reg}(g)$ and define

$$A(x)=\operatorname{Reg}(g)\cap g^{-1}(g(x)).$$

By the implicit function theorem, $A(x)$ is a 1-dimensional submanifold of class C^r, closed in $\operatorname{Reg}(g)$. Let $C(x)$ be the connected component of x in $A(x)$. Then $C(x)$, being a connected 1-manifold, is C^r diffeomorphic to either $\mathbb{R}$ or S^1. The tangent line to $C(x)$ at x is parallel to the linear subspace $\operatorname{Ker} Dg_x$.

Our first task is to orient the curves $C(x)$. This is done by assigning to each $x\in\operatorname{Reg}(g)$ a unit vector $\phi(x)$ tangent to $C(x)$. To define $\phi(x)$ first pick an ordered basis $e_1(x),\cdots,e_{n-1}(x)$ of the normal space to $\operatorname{Ker} Dg_x$, in such a way that the ordered basis

$$g(x), Dg_x e_1(x),\cdots, Dg_x e_{n-1}(x)$$

defines the *positive* orientation of $\mathbb{R}^n$ (that is, the determinant of the matrix of their coordinates is positive). Now define $\phi(x)\in\operatorname{Ker} Dg_x$ to be the unique

unit vector such that the ordered basis of $\mathbb{R}^n$,

$$\phi(x), e_1(x), \cdots, e_{n-1}(x),$$

defines the *negative* orientation of $\mathbb{R}^n$.

It is easy to verify that the resulting vector field

$$\phi : \mathrm{Reg}(g) \to \mathbb{R}^n$$

does not depend on the choice of the ordered basis $e_1(x), \cdots, e_{n-1}(x)$.

Moreover, ϕ is C^{r-1}. For from the implicit function theorem we can find, locally in $\mathrm{Reg}(g)$, a nonvanishing C^{r-1} vector field θ tangent to the curves $C(x)$. Then ϕ coincides locally with one of the unit vector fields $\theta(x)/|\theta(x)|$ or $-\theta(x)/|\theta(x)|$.

From now on the curves $C(x)$ are always oriented by ϕ.

Let $x \in \mathrm{Reg}(g)$, $f(x) = y \in \mathbb{R}^n - 0$. Denote the ray through y by

$$L(y) = \{\lambda y \in \mathbb{R}^n : \lambda > 0\}.$$

Then f maps $C(x)$ into $L(y)$. By differentiating f we obtain a C^{r-1} map $\lambda : \mathrm{Reg}(g) \to \mathbb{R}$ such that

$$Df_x \phi(x) = \lambda(x) f(x).$$

The definition of ϕ implies that *if* $J(x) \neq 0$, *then* $J(x)$ *and* $\lambda(x)$ *have opposite signs, while* $\lambda(x) = 0$ *if and only if* $J(x) = 0$.

Let $\xi(t)$ be a parameterization of $C(x)$ which moves in the ϕ direction as t increases. From the preceding remarks it follows that $|f(\xi(t)|$ decreases for t near t_0, when $J(x(t_0)) > 0$.

In particular we can choose ξ to be a solution to the initial value problem

$$\frac{d\xi}{dt} = \phi(\xi),$$

$$\xi(t_0) = x_0.$$

Then it is easy to calculate that

$$\frac{d}{dt}\bigg|_{t=t_0} |f(\xi(t))| = \lambda(x_0)|f(x0)|.$$

A most important conclusion is

LEMMA 1.1. *$|f|$ is nonincreasing along the oriented curves $C(x)$ in every open set where $J(x)\geqq 0$; and $|f|$ is strictly decreasing where $J(x)>0$.*

The following notation will be useful. If $0\leqq a<b$, define

$$E[a,b]=\{x\in\mathbf{R}^n : a\leqq|f(x)|\leqq b\}.$$

If $s>0$, define

$$E(s)=\{x\in\mathbf{R}^n : |f(x)|=s\},$$

$$E_+(s)=\{x\in\mathbf{R}^n : |f(x)|\geqq s\},$$

$$E_-(s)=\{x\in\mathbf{R}^n : |f(x)|\leqq s\}.$$

We continue to write $E=E(0)=f^{-1}(0)$.

If s is a regular value for the restriction of $|f|$ to some neighborhood of $E(s)$, then $E(s)$ is a C^r submanifold, closed in $\mathrm{Reg}(g)$, and transverse to the curves $C(x)$.

For the next three results, Lemmas 1.2–1.4, we make the following assumptions:

$$|f(x_0)|=s>0,$$

(A) $$J(x_0)>0,$$

$$J\geqq 0 \quad \text{on a neighborhood of } E(s).$$

Let $\approx$ denote diffeomorphism.

LEMMA 1.2. *$C(x_0)\cap E(s_0)=x_0$ and $C(x_0)\approx\mathbf{R}$.*

Proof: Since $J(x_0)>0$, $C(x)$ crosses $E(s)$ transversely at x_0, it cannot meet $E(s)$ again by Lemma 1.1. Therefore x_0 disconnects $C(x_0)$ and thus $C(x_0)$ cannot be diffeomorphic to a circle.

When $C(x)\approx\mathbf{R}$, then $C(x)-x$ is the disjoint union of two sets, each diffeomorphic to $\mathbf{R}$. We label their closures $C_+(x)$ and $C_-(x)$, where $\phi(x)$ points toward $C_+(x)$. Notice that $C_+(x)$ is the forward orbit of x under the flow associated to the vector field ϕ.

From Lemma 1.2 we get

LEMMA 1.3. *$|f(y)|<s$ for all $y\in C_+(x_0)-x_0$.*

Suppose $C(x)\approx\mathbf{R}$. Let $\xi:\mathbf{R}\to C(x)$ be a diffeomorphism such that $\xi(0)=x$ and $\xi([0,\infty))=C_+(x)$. By a *limit point* of $C_+(x)$ we mean a point $y\in\mathbf{R}^n$ such

that there exists a sequence $t_m \to \infty$ with

$$\lim_{m \to \infty} \xi(t_m) = y .$$

It is easy to see that *every limit point of $C_+(x)$ is either a critical point of g or a zero of f.*

LEMMA 1.4. *In addition to the Assumption* (A) *above, suppose $0 < q < s$ and*

(a) *$E[q, s]$ is compact,*

(b) *$g(x_0)$ is a regular value for $g|E[q, s]$.*

Then $C_+(x_0) \cap E(q) \neq \emptyset$.

Proof: If $C_+(x_0)$ is disjoint from $E(q)$, it must be contained in $E[q, s] - E(q)$ by connectedness. By (a), $C_+(x_0)$ has a limit point $y \in E[q, s]$. Hence either $f(y) = 0$ or else $y \in \mathrm{Crit}(g)$. But $f(y) > q > 0$; and since $g(y) = g(x_0)$ by continuity, it follows from (b) that $y \in \mathrm{Reg}(g)$. This contradiction proves the lemma.

A subset W_1 of W_2 has *full measure* if $W_2 - W_1$ has measure zero.

THEOREM 1.5. *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ and $\alpha > 0$ satisfy the following conditions:*

(a) *f is C^2 and proper,*

(b) *$J^{-1}(0)$ has measure zero.*

(c) *$J(x) \geqq 0$ if $|x| \geqq \alpha$.*

Let $0 < \varepsilon < \alpha$. Then there exists a subset

$$W(\varepsilon) \subset E_+(\alpha)$$

which is open and of full measure, such that if $x_0 \in W(\varepsilon)$, then $C(x_0) \approx \mathbb{R}$ and $C_+(x_0)$ contains a point y with $|f(y)| = \varepsilon$.

Proof: For each $\beta > \alpha$ define

$$B(\varepsilon, \beta) = g(\mathrm{Crit}(g) \cap E[\varepsilon, \beta]) .$$

This is a compact subset of S^{n-1}. By the Morse–Sard theorem (see [1], [4], [10]), $B(\varepsilon, \beta)$ has measure zero in S^{n-1} because g is C^2. Therefore, $S^{n-1} - B(\varepsilon, \beta)$ has full measure and is open in S^{n-1}. Clearly every point of $S^{n-1} - B(\varepsilon, \beta)$ is a regular value for $g \mid E[\varepsilon, \beta]$.

Since $\mathrm{Reg}(f)$ has full measure in $\mathbb{R}^n$, so has $\mathrm{Reg}(g)$ in $\mathbb{R}^n - E$. This implies that $g^{-1}(S^{n-1} - B(\varepsilon, \beta))$ has full measure in $\mathbb{R}^n - E$. It suffices to check this at points in $\mathrm{Reg}(g)$. Near such a point there are C^2 coordinates making g look like a surjective linear map $\pi : \mathbb{R}^n \to \mathbb{R}^{n-1}$. Clearly π^{-1} takes subsets of full measure in $\mathbb{R}^{n-1}$ into subsets of full measure in $\mathbb{R}^n$.

Now define $W(\varepsilon, \beta)$ to be the intersection of the three open sets

$$\{x : \alpha < |f(x)| < \beta\},$$

$$g^{-1}(S^n - B(\varepsilon, \beta)),$$

$$J^{-1}(\mathbb{R}_+).$$

It is clear that $W(\varepsilon, \beta)$ is open in $\mathbb{R}^n$ and has full measure in $E[\alpha, \beta]$.

Define

$$W(\varepsilon) = \bigcup_{\beta > \alpha} W(\varepsilon, \beta).$$

Then W is open in $\mathbb{R}^n$ and has full measure in $E_+(\alpha)$.

Suppose $x_0 \in W(\varepsilon)$. Then $x_0 \in W(\varepsilon, \beta)$ for some β with $\beta > |f(x_0)| > \alpha > \varepsilon$. Put $s = |f(x_0)|$. Then the hypotheses of Lemma 1.4 are fulfilled, so $C_+(x_0) \cap E(\varepsilon) \neq \varnothing$.

As a by-product of the proof of Theorem 1.5 it follows that $C_+(x_0)$ has no limit points in $E[\varepsilon, \alpha]$. This is used to prove the next result. Recall that a *Baire subset* of $X \subset \mathbb{R}^n$ is the intersection of countably many open dense subsets of X.

THEOREM 1.6. *Under the hypotheses of Theorem* 1.5 *there exists a Baire subset* W *of full measure in* $E_+(\alpha)$ *such that if* $x_0 \in W$, *then* $C(x_0) \approx R$ *and* $C_+(x_0)$ *leads to* $f^{-1}(0)$ *for all* $x_0 \in W$. *In particular,* $f^{-1}(0)$ *is nonempty.*

Proof: Let $W(\varepsilon)$ be as in Theorem 1.5 and put

$$W = \cap \left\{ W\left(\frac{1}{n}\right) : n = 1, 2, \cdots \right\}.$$

Then W is a Baire subset of full measure in $E_+(\alpha)$. Suppose $x_0 \in W$, $|f(x_0)| = s$. By Theorem 1.5 and the remark following its proof, $C(x_0) \approx \mathbf{R}$, and for all ε such that $0 < \varepsilon < s$ we have: $C_+(x_0) \cap E(\varepsilon) \neq \varnothing$ and $C_+(x_0)$ has no limit points in $E[\varepsilon, s]$. But $E[0, s]$ is compact so that $C_+(x_0)$ does have limit points, and they are all in $f^{-1}(0)$.

From Lemma 1.1 we easily obtain the following addition to Theorem 1.6:

ADDENDUM 1.7. *In Theorem* 1.6 *let* $y \in f^{-1}(0) \cap \operatorname{clos} C_+(x_0)$. *Then* $J(y) \geqq 0$.

The condition that $\operatorname{Reg}(f)$ has full measure is always satisfied if f is

analytic, or even if f is *almost analytic*, i.e., C^1 and analytic on the complement of a closed set of measure zero. This follows from the following lemma, taking $F = J$.

LEMMA 1.8. *Let $X \subset \mathbb{R}^n$ be a proper analytic set (i.e., $X = F^{-1}(0)$ where $F : \mathbb{R}^n \to \mathbb{R}^k$ is analytic and not identically zero). Then X has measure zero.*

Proof: We use induction on $n \geqq 1$. For $n = 1$, X is countable, hence of measure zero. Suppose inductively that $n = m + 1$ and that the lemma holds for smaller values of n. Write $\mathbb{R}^{m+1} = \mathbb{R}^m \times \mathbb{R}$. Then $X_t = X \cap (\mathbb{R}^m \times t)$ has measure $m(t)$ equal to zero for all $t \in \mathbb{R}$. By Fubini's theorem, the measure of X is $\int_{-\infty}^{\infty} m(t)\, dt = 0$.

Another proof follows from the triangulation theorem of Lojasciewicz [6] that X is the union of countably many smooth open simplexes of dimensions less than n.

In Section 4 a nondegeneracy condition will be used to make $\mathrm{Reg}(f)$ have full measure.

2. An Algorithm

In this section the existence theorem of Section 1 is converted into an algorithm for finding approximate zeroes of maps $f : \mathbb{R}^n \to \mathbb{R}^n$ satisfying certain assumptions. Examples of such maps will be given in Section 3. As an application we show how the algorithm can be adapted to fixed point problems.

Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a C^r map, $r \geqq 2$. If x is a point where the Jacobian $J(x) \neq 0$, we define the *Newton vector* at x to be

$$N(x) = -\mathrm{sgn}\, J(x)\, Df_x^{-1} f(x)\,.$$

It is clear that $N(x)$ is a positive scalar multiple of the unit vector $\phi(x)$ defined in Section 1.

Let $\rho > 0$. The *Newton transformation of length* ρ is the C^{r-1} map

$$T_\rho : \mathrm{Reg}(f) \to \mathbb{R}^n$$

defined by

$$T_\rho(x) = x + tN(x)\,, \qquad t > 0\,, \qquad |tN(x)| = \rho\,.$$

We define the *Newton step of length* ρ *from* x to be $T_\rho(x)$. By ν *successive*

Newton steps we mean the sequence $T_\rho(x), (T_\rho)^2(x), \cdots, (T_\rho)^\nu(x)$ if these are defined.

THEOREM 2.1. *Let $f: \mathbb{R}^n \to \mathbb{R}^n$ be a proper analytic map. Let $\varepsilon > 0$ be given. Suppose $J \geqq 0$ outside some compact set and $J(x)$ is not identically 0. Suppose a number $\alpha_0 \in \mathbb{R}$ is known such that $J(x) \geqq 0$ if $|f(x)| \geqq \alpha_0$ (such numbers exist). Then there is an open subset W^* of full measure of the set $E_+(\alpha_0) = \{x : |f(x)| > \alpha_0\}$ such that, for every $x_0 \in W^*$, Algorithm A, described below, is defined and stops at a point x with $|f(x)| < \varepsilon$.*

The algorithm is:

ALGORITHM A. If $|f(x_0)| < \varepsilon$, a solution has been reached; stop. Otherwise perform the following recursion on l, starting with $l = 0$: Starting at $x_0 = y_{l,0}$, take successive Newton steps $y_{l,i}$, $i = 1, 2, \cdots$, of length 2^{-l}, stopping if a solution is reached. If 4^l steps have been taken without reaching a solution, or if $J(y_{l,i}) = 0$, increase l by 1 and go back to x_0.

Here is a flow chart for the algorithm; x_0 and ε are given:

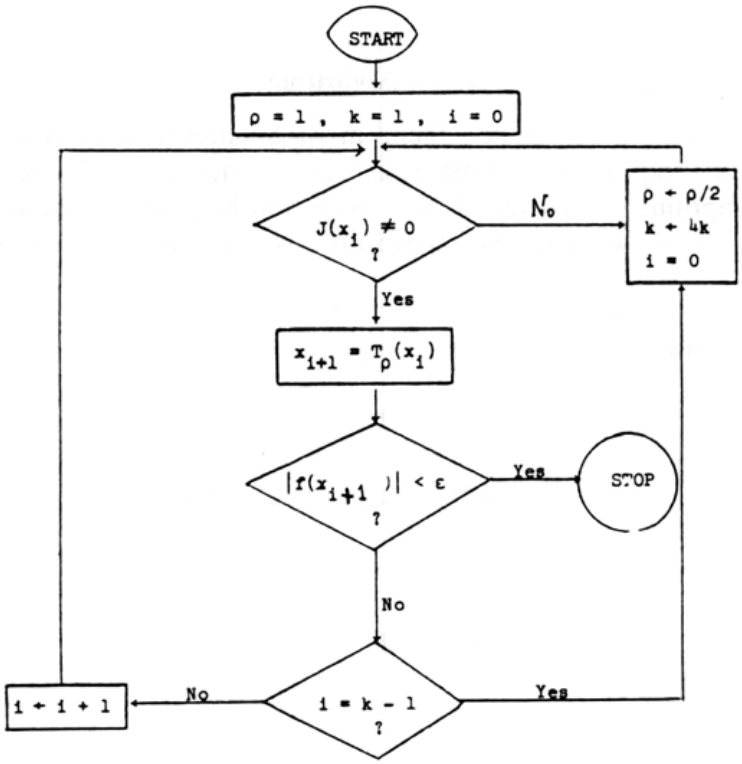

Figure 1. Flow chart for Algorithm A. x_0 and $\varepsilon > 0$ are given, $T_\rho(x) = x + tN(x)$, where $t > 0$ and $|tN(x)| = \rho$, $N(x) = -\operatorname{sgn} J(x)\, Df_x^{-1} f(x)$, $J(x) = \operatorname{Det} Df_x$.

Toward the proof, let ψ_t be the partial flow defined by the vector field ϕ. For each $x \in \mathrm{Reg}(g)$ there is a maximal half-open interval $[0, T(x))$, $0 < T(x) \leq \infty$, on which there is defined an analytic curve

$$t \mapsto \psi_t(x), \qquad 0 \leqq t < T(x),$$

which is a solution to the initial value problem

$$\frac{d}{dt}\psi_t = \phi(\psi_t),$$

$$\psi_0(x) = x.$$

The image of $[0, T(x))$ lies in $C(x)$. When $C(x) \approx \mathbf{R}$ this image is $C_+(x)$.

We extend the definition of T_ρ to points $x \in \mathrm{Reg}(g)$ by $T_\rho(x) = x + \rho\phi(x)$.

The iterated Newton transformations $(T_\rho)^k$, $k = 1, 2, \cdots$, are approximations to the maps ψ_t. They have the great advantage of being readily computable. More precisely, we have

LEMMA 2.2. *Let $p \in \mathrm{Reg}(g)$. Suppose positive numbers δ and $\tau < T(p)$ are given. Then there exist an open set $V \subset \mathrm{Reg}(g)$ with $p \in V$ and a number $\rho_* > 0$, having the following property: Suppose given an integer $\nu > 0$ and numbers $\rho_1, \cdots, \rho_\nu$ with $0 < \rho_j \leqq \rho_*$ and $\sum \rho_j = t \leqq \tau$, $j = 1, \cdots, \nu$; then both ψ_t and the composition $T_* = T_{\rho_\nu} \cdots T_{\rho_1}$ are defined on V and*

$$|\psi_t(x) - T_*(x)| < \delta$$

for all $x \in V$. Moreover, $T_* \mid V$ is a local diffeomorphism.

Proof: Everything but the last statement follows from the Cauchy–Euler approximation to solutions of the differential equation $d\psi_t/dt = \phi(\psi_t)$. The last statement follows from the fact that, as $\rho \to 0$, T_ρ converges to the identity C^1 uniformly on compact sets. For let $K \subset \mathrm{Reg}(g)$ be a neighborhood of the arc $\{\psi_s(p) : 0 \leqq s < \tau\}$ having compact closure $\bar{K}$. When V and ρ_0 are small enough, then $T_*(V) \subset K$ for all T_* defined as above; and in addition each T_ρ, $0 \leqq \rho \leqq \rho_*$, is so C^1 close to the identity in K that, by the inverse function theorem, $T_\rho \mid K$ is a local diffeomorphism. Hence, by induction on ν, T_* will be the composition of local diffeomorphisms.

We pass now to the proof of the theorem. Theorem 1.5 is applicable; therefore there exists an open set $W \subset E_+(\alpha_0)$ of full measure, such that for each $x \in W$ the oriented curve $C_+(x)$ contains a point y with $|f(y)| = \varepsilon$. To find W^* it suffices to find, for each $p \in W$, an open set $U(p) \subset W$ and an open set

$U^*(p)\subset U(p)$ of full measure, such that for all $x_0\in U^*(p)$ the algorithm is well-defined and stops. For then there is a countable set $\{p_i\}$ in W such that $W=\cup U(p_i)$; and we set $W^*=\cup U^*(P_i)$.

Fix $p\in W$ and $0<\varepsilon_1<\varepsilon$. Let $t<T(p)$ be such that $t>0$ and

$$|f(\psi_t(p))|<\varepsilon_1 .$$

Such a t exists by Theorem 1.5. By properness and continuity of f, there exists $\delta>0$ such that if $|f(z)|\leqq\varepsilon_1$ and $|z'-z|<\delta$ then $|f(z)|<\varepsilon$. Fix such a δ.

With this choice of p, t and δ let ρ_0 and V be as in the lemma. Let $U(p)\subset V$ be an open neighborhood of p such that

$$|f(\psi_t(y))|<\varepsilon_1 \quad \text{if} \quad y\in U(p) .$$

Let $m\geqq 0$ be an integer such that $2^{-m}\leqq\rho_0$ and $2^m\geqq t$. Put $\sigma=2^{-m}$ and define $U^*(p)\subset U(p)$ to be the set of x in $U(p)$ such that for all $j=0,\cdots,[\sigma/t]$ we know that

$$(T_\sigma)^j(x) \quad \text{is defined},$$

and

$$J((T_\sigma)^j(x))\neq 0 .$$

Since $J^{-1}(0)$ is an analytic variety and each $(T_\sigma)^j$ is a local diffeomorphism, it follows that $U^*(p)$ is of full measure in $U(p)$. Clearly, $U^*(p)$ is also open.

Now initialize Algorithm A at an arbitrary $x_0\in U^*(p)$. To prove that the algorithm stops it suffices to show (referring to the flow chart) that the variable k never takes value 4^{m+1}. If it does, then it must have previously taken the value 4^m. When this first happens ρ is set at $2^{-m}=\sigma$. The definition of $U^*(p)$ means that the algorithm produces the sequence

$$x_i=(T_\rho)^i(x_0), \qquad i=1,\cdots,4^m .$$

Now $[t/\rho]=[2^m t]\leqq[2^m\cdot 2^m]=4^m$ so that, at some stage of running the algorithm, $i=[t/\rho]\leqq k$. When this happens we have

$$|x_i-\psi_t(x_0)|<\delta$$

and

$$|f(\psi_t(x_0))|<\varepsilon_1 ,$$

implying that

$$|f(x_i)| < \varepsilon$$

by the choice of δ. But then the algorithm stops. Thus k cannot reach the value 4^{m+1}, so the algorithm must stop.

This argument shows that, for a given f and ε, the number of computations the algorithm makes before stopping is a locally bounded function of $x_0 \in W^*$.

The analyticity of f was not used very much—only to make sure that g is C^2 (in order to apply Theorem 1.5 of the preceding section) and to make $J^{-1}(0)$ have measure zero. Therefore, the *theorem is valid under the weaker hypothesis that f is C^2 and $J^{-1}(0)$ has measure zero.*

As an example of how it can be useful, even in analytic problems, to be able to handle nonanalytic (and even nonproper) maps, consider the problem of finding an approximate fixed point y of a map $\theta : \mathbb{R}^n \to \mathbb{R}^n$ which is analytic and satisfies

$$|\theta(x)| < |x| \quad \text{if} \quad |x| \geqq r,$$

where $r > 0$ is assumed known. Of course Brouwer's theorem, applied to the restriction of θ to the disk of radius r, ensures that a fixed point exists, but we want an algorithm for finding one.

The map

$$F : \mathbb{R}^n \to \mathbb{R}^n, \qquad f(x) = x - \theta(x),$$

is analytic but need not be proper; nor is its Jacobian necessarily greater than or equal to 0 outside a compact set. Instead of F, however, we consider the C^2 map $f : \mathbb{R} \to \mathbb{R}^n$ defined as follows. Let $\lambda : \mathbb{R}^n \to [0, 1]$ be a C^2 map such that $\lambda = 1$ on the ball of radius r, $\lambda = 0$ outside the ball of radius $2r$, and λ is analytic on $\{x : r < |x| < 2r\}$. For example,

$$\lambda(x) = \int_0^{|x|} \psi(u)\, du \Big/ \int_0^{2r} \psi(u)\, du,$$

where ψ is the C^1 bump function

$$\psi(u) = \begin{cases} (r-u)^2(2r-u)^2 & \text{if} \quad r \leqq u \leqq 2r, \\ 0 & \text{otherwise}. \end{cases}$$

Define

$$f(x)=x-\lambda(x)\theta(x).$$

If $f(x)=0$, then $|x|=|\lambda(x)|\,|\theta(x)|\leqq|\theta(x)|$ so $|x|\leqq r$ and thus $\lambda(x)=1$; hence $x=\theta(x)$ as required.

Clearly, f is C^2 and proper and its Jacobian J is at least 0 outside the ball of radius $2r$. Moreover, $J^{-1}(0)$ has measure zero, since f is analytic except on the set $\{x : |x|=r \text{ or } |x|=2r\}$ which has measure zero. Therefore the algorithm can be applied. We conclude that there is an open set W^* of full measure in $\{x : |x|>2r\}$ such that for any $\varepsilon>0$ and any $x_0\in W^*$ the algorithm will stop at some y with $|\theta(y)-y|<\varepsilon$.

The method just discussed can be formulated into a general principle:

PROPOSITION 2.3. *Let $f_*:\mathbb{R}^n\to\mathbb{R}^n$ satisfy the hypothesis of Theorem* 1.5. *Let $u:\mathbb{R}^n\to\mathbb{R}^n$ be analytic, and suppose $|u(x)|<|f_*(x)|$ for $|x|\geqq r$. Then a zero of the map $F=f_*+u$ can be found by applying Algorithm* A *to the map $f=f_*+\lambda u$, where $\lambda:\mathbb{R}^n\to[0,1]$ is as above.*

The algorithm can also be adapted to find approximate fixed points of an arbitrary analytic map $u:D^n\to D^n$. First replace u by $v:D^n\to D^n$, where $v(x)=(1-\varepsilon_1)u(x)$ with $\varepsilon_1>0$ very small. Then v maps D^n into a ball of smaller radius $1-\varepsilon_1$, and if ε_1 is small enough, then $|u(x)-x|<\varepsilon$ whenever $|v(x)-x|<\varepsilon$. (We must know enough about u to calculate such an ε_1.) Then transfer v to $\mathbb{R}^n$ by setting

$$\theta=H\circ v\circ H^{-1},$$

where

$$H:\operatorname{int}D^n\approx\mathbb{R}^n,$$

$$H(x)=(1-|x|^2)^{-1}x.$$

Notice that H maps each diameter of $\operatorname{int}D^n$ homeomorphically onto the corresponding line through the origin. It follows that there exists $r>0$ with $|\theta(x)|<|x|$ if $|x|\geqq r$. Such an r is calculable if we know ε_1. As we described above, the algorithm almost always leads to an approximate fixed point y of θ, say $|\theta(y)-y|<\delta$. If we know enough about the original map u we can calculate a δ so small that the point $x=H(y)$ satisfies $|u(x)-x|<\varepsilon$, as required.

3. Polynomial Examples

Here we give some examples of polynomial maps $f:\mathbb{R}^n\to\mathbb{R}^n$ which satisfy the hypothesis of Theorem 2.1.

EXAMPLE 1. Let $f(z)=\sum_{i=0}^{d} a_i z^i$ be a complex polynomial of degree $d>0$ with complex coefficients $a_0, \cdots, a_d \neq 0$. When we identify the complex field $\mathbf{C}$ with $\mathbf{R}^2$, f becomes a map $\mathbf{R}^2 \to \mathbf{R}^2$. It is always proper since $|f(z)|/|z|^d \to |a_d|$ as $|z| \to \infty$. Its real Jacobian at z is $|f'(z)|^2$. Thus the conditions of Theorem 2.1 are satisfied. In fact one can take $\alpha_0 = 0$ and W^* will be open and of full measure in $\mathbf{C}$ itself.

The Newton vector field is $-f(z)/f'(z)$. All but a finite set of solutions of

$$(*) \qquad \frac{dz}{dt} = \frac{-f(z)}{f'(z)}, \qquad t \in \mathbf{R},$$

lead to zeroes of f. (The others lead to zeroes of f'.) These ideas go back to Cauchy.

Notice that the right-hand side of differential equation $(*)$ can be multiplied by $|f'(z)|^2$ to obtain the desingularized equation

$$\frac{dz}{dt} = -f(z)\overline{f'(z)}.$$

We pursue this example later.

EXAMPLE 2. Let $f : \mathbf{C}^m \to \mathbf{C}^m$ be a polynomial map. Then $f=(f_1, \cdots, f_m)$, where each f_i is a polynomial in m complex variables, of degree $d_i \geqq 0$. As in Example 1, f may be considered a real polynomial map from $\mathbf{R}^{2m}$ to $\mathbf{R}^{2m}$. The Jacobian of the latter at $z \in \mathbf{C}$ is $|\det (Df(z))|^2$, where $Df(z) : \mathbf{C}^m \to \mathbf{C}^m$ denotes the complex derivative of $f : \mathbf{C}^m \to \mathbf{C}^m$ at $z \in \mathbf{C}^m$. Thus the determinant condition of Theorem 2.1 is satisfied everywhere in $\mathbf{R}^{2m}$.

For Theorem 2.1 to apply we need f to be proper. We state a fairly general condition for this.

PROPOSITION. *Let $f : \mathbf{C}^m \to \mathbf{C}^m$ (or $\mathbf{R}^m \to \mathbf{R}^m$) be a polynomial map, $f=(f_1, \cdots, f_m)$. Let $f_i = h_i + u_i$, where h_i is a homogeneous polynomial of degree $d_i > 0$ and u_i is a polynomial of degree $c_i < d_i$. If $h_1, \cdots, h_m$ have no common zero other than the origin, then f is proper.*

Proof: We consider $f : \mathbf{R}^m \to \mathbf{R}^m$, but the arguments apply equally to $\mathbf{C}^m \to \mathbf{C}^m$.

Let x_i be a sequence in $\mathbf{R}^m$ with $|x_i| \to \infty$. We must show that $|f(x_i)| \to \infty$. By taking a subsequence we may suppose $x_i/|x_i|$ converges to $x \in S^{m-1}$. By

hypothesis, there is some k with $h_k(x)\neq 0$. Then

$$f_k(x_j)=h_k(x_j)+u_k(x_j)$$

$$=|x_j|^{d_k}h_k\left(\frac{x_j}{|x_j|}\right)+u_k(x_j)$$

clearly goes to ∞ as $j\to\infty$.

When all the h_j have the same degree $d>0$, f extends to a map $F: P^m\to P^m$ of projective spaces given in homogeneous coordinates $X=[X_0,\cdots,X_m]$ by

$$F(X)=[F_0(X),\cdots,F_m(X)],$$

$$F_0(X)=X_0^d,$$

and for $j>0$,

$$F_j(X)=\begin{cases}X_0^d f_j(X_1/X_0,\cdots,X_m/X_0) & \text{if } X_0\neq 0,\\ h_j(X_1,\cdots,X) & \text{if } X_0=0.\end{cases}$$

EXAMPLE 3. We construct polynomial maps $f:\mathbb{R}^2\to\mathbb{R}^2$ as follows. Fix an integer $d\geqq 2$. The real and imaginary parts of $(x+iy)^d$ are homogeneous polynomials $u_d(x, y)$, $v_d(x, y)$ of degree d. Let $p(x, y)$, $q(x, y)$ be polynomials of lower degree and consider

$$f:\mathbb{R}^2\to\mathbb{R}^2,\qquad f=(u_d+p, v_d+q).$$

The proposition above shows that f is proper. The Jacobian J of f at $(x, y)=z$ is the sum of $d|z|^{d-1}$ and a polynomial of degree at most $d-2$. Therefore, $J(z)\to\infty$ as $|z|\to\infty$.

Is it evident that p, q may be chosen to make the Jacobian negative in bounded regions of $\mathbb{R}^2$. In fact when $d=2$, if f is not complex there will always be points where $J<0$. The set of these points is the interior of some ellipse.

A concrete example with $d=3$ is

$$f(x, y)=(x^3-3xy^2+100xy+1, 3x^2y-y^3+x+y+1).$$

In Section 5 we pursue Example 3 on the computer.

EXAMPLE 4. (Given to us by Richard Palais.) Let $d=2m+1\geqq 3$ be an

integer. Define

$$f_d : \mathbb{R}^n \to \mathbb{R}^n,$$

$$f_d(x) = |x|^{2m} x.$$

Let $p : \mathbb{R}^n \to \mathbb{R}^n$ be any polynomial map whose coordinates have degrees d. The proposition proved in Example 2 shows that the map

$$f = f_d + p : \mathbb{R}^n \to \mathbb{R}^n$$

is proper. The Jacobian of f_d at x is easily computed to be $(2m+1)|x|^{2mn}$. The Jacobian J of f is the sum of this function plus a polynomial of degrees less than $2mn$. Hence $J(x) > 0$ for large $|x|$; thus f satisfies the hypotheses of Theorem 2.1.

EXAMPLE 5. In all the preceding examples we have a polynomial map $f_* : \mathbb{R}^n \to \mathbb{R}^n$ which satisfies the conditions of Theorem 1.2, and for which there is an integer $d > 0$ with

$$\liminf_{|x| \to \infty} |f_*(x)|/|x|^d > 0.$$

Suppose $f : \mathbb{R}^n \to \mathbb{R}^n$ is a map of the form $f = f_* + u$, where $u = (u_1, \cdots, u_n)$ and each $u_i : \mathbb{R}^n \to \mathbb{R}$ is a polynomial of degree less than d. Then f is proper but its Jacobian may change sign outside every compact set: for example, $f_* = (x^3, y^3)$ and

$$f = (x^3 + x^2 + 1, y^3 + x^2 + 2).$$

Thus Theorem 2.1 does not apply to f. However, the method of Proposition 2.3 can be used provided we know $r > 0$ such that $|u(x)| < |f_*(x)|$ for $|x| \geqq r$; such an r exists. For then we can apply Algorithm A to an auxiliary map

$$F : \mathbb{R}^n \to \mathbb{R}^n, \qquad F = f_* + \lambda u,$$

when

$$\lambda : \mathbb{R}^n \to [0,1] \quad \text{is} \quad C^2,$$

$$\lambda(x)=1 \qquad \text{if} \quad |x| \leqq r,$$

$$\lambda(x)=0 \qquad \text{if} \quad |x| \geqq 2r.$$

4. Convergent Algorithms

Algorithm A of Section 2 is guaranteed (if $x_0 \in W^*$) to stop at a solution to $|f(x)|<\varepsilon$. It is desirable to have available other algorithms that are guaranteed *not* to stop, but rather to produce an infinite sequence $\{x_i\}$ which converges to a zero of f.

It is well known that the classical Newton–Raphson iteration scheme,

$$x_{i+1} = x_i - Df(x_i)^{-1} f(x_i),$$

does this *provided* x_0 is sufficiently close to a zero x_* of f such that $J(x_*) \neq 0$. This holds for C^1 maps f; Newton–Raphson iteration for C^2 maps has the stronger property of *quadratic convergence:* not only does $f(x_i) \to 0$, but also

$$|f(x_{i+1})|/|f(x_i)|^2 \to \text{constant}.$$

These standard facts are easy consequences of Taylor approximation.

In contrast to the local nature of Newton–Raphson convergence, the algorithms presented below converge to a zero of f when started at almost any x_0 sufficiently *far* from $f^{-1}(0)$. Some restrictions on f seem to be necessary for this.

We consider proper C^2 maps $f : \Omega \to \mathbb{R}^n$, where $\Omega \subset \mathbb{R}^n$ is either an open set or a compact domain with smooth boundary $\partial\Omega$.

The following nondegeneracy will be assumed:

$$\text{ND:} \begin{cases} \text{(a)} & 0 \in \mathbb{R}^n \quad \text{is a regular value of} \quad f : \Omega \to \mathbb{R}^n, \\ \text{(b)} & Df : \Omega \to L(\mathbb{R}^n, \mathbb{R}^n) \quad \text{is transverse to the singular maps.} \end{cases}$$

Condition (a) means that $J(x) \neq 0$ if $f(x)=0$. In (b) we consider the vector space $\mathcal{L} = L(\mathbb{R}^n, \mathbb{R}^n)$ of all linear maps $\mathbb{R}^n \to \mathbb{R}^n$ (or all $n \times n$ matrices). Df is a C^1 map which assigns to $x \in \Omega$ the linear map $Df(x)$ whose matrix is $[\partial f_i(x)/\partial x_j]$. For each integer r, $0 \leqq r \leqq n$, there is the subset $V(r) \subset \mathcal{L}$ of linear maps of rank r. Each $V(r)$ is an analytic submanifold of $\mathcal{L}$. Condition (b) means that Df is transverse to each submanifold $V(r)$, $r=0, \cdots, n-1$.

The following proposition is well known:

PROPOSITION 4.1. *The set of maps satisfying* ND *is open and dense in* $C^r(\Omega, \mathbb{R}^n)$, $r \geqq 2$.

Here $C^r(\Omega, \mathbb{R}^n)$ is the set of all C^r maps $\Omega \to \mathbb{R}^n$ endowed with the uniform C^r topology when Ω is compact, and the Whitney (=strong) C^r topology otherwise.

Proposition 4.1 means that ND is stable under small perturbations of f, and most maps have it. Thus in a sense ND is a mild restriction on f.

We do not use the full strength of ND but only the following consequences:

PROPOSITION 4.2. *Let f satisfy* ND. *Then,*

(a) $E = f^{-1}(0)$ *is discrete and* Df_x *is invertible for* $x \in E = f^{-1}(0)$,

(b) $f^{-1}(0)$ *is the union of a finite set of smooth submanifolds of dimensions less than n; hence $f^{-1}(0)$ is a closed set of measure zero.*

The property of f expressed in (b) of Proposition 4.2 is more easily stated than (b) of ND, and it is more easily verified in particular cases. It is not, however, a stable condition. For this reason we prefer to state the theorems of this section under the stronger hypothesis ND.

First we take $\Omega = \mathbb{R}^n$.

THEOREM 4.3. *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a proper C^2 map satisfying* ND. *Suppose also that a number $\alpha_0 \geqq 0$ is known such that $J(x) \geqq 0$ if $|f(x)| \geqq \alpha_0$. Then there is an open subset W^* of full measure in $E_+(\alpha_0) = \{x : |f(x)| > \alpha_0\}$ with the following property. When Algorithm* B *(defined below) is started at any $x_0 \in W^*$, it produces an infinite sequence $\{x_i\}$ which converges to a zero of f. Moreover, there exists i_0 such that, for $i \geqq i_0$,*

$$x_{i+1} = x_i - Df(x_i)^{-1} f(x_i) .$$

In other words, eventually and automatically the sequence proceeds by Newton–Raphson iteration.

Here is a flow chart for the algorithm:

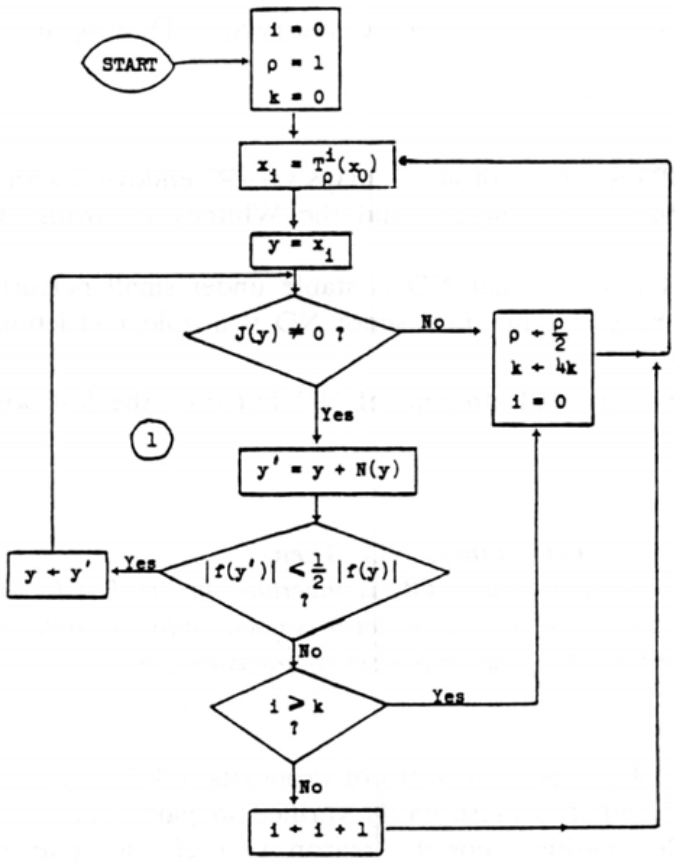

Figure 2. Flow chart for Algorithm B. x_0 is given, $N(y) = -Df_y^{-1}f(y)$.

Proof: Suppose $f(x_*)=0$, $J(x_*)>0$. It is easy to calculate that the derivative of the Newton vector field N is identity. It follows that there is a spherical neighborhood $B(x_*)$ and a number $\sigma_0>0$ such that $B(x_*)$ is invariant under all the transformations T_σ, $0<\sigma\leqq\sigma_0$, and moreover $|f(T_\sigma(x))|<|f(x)|$.

We take $B(x_*)$ so small that for all $x\in B(x_*)$ we have $x+N(x)\in B(x_*)$ and

$$|f(x+N(x))|<\tfrac{1}{2}|f(x)|\,.$$

Let $B=\cup\{B(x_*): x_*\in E, J(x_*)>0\}$. Now choose $\varepsilon>0$ so small that (i) if $|f(x)|<\varepsilon$ and $J(x)>0$, then $x\in B$ (using properness of f), and (ii) if $|f(x)|<\varepsilon$ and $|f(x+N(x))|<\frac{1}{2}(f(x))$, then $J(x)>0$. This last can be done because a

Taylor estimate yields

$$|f(x+N))| = |1 - \operatorname{sgn} J(x)| \cdot |f(x)| + o(|f(x)|^2).$$

With this choice of ε, and α_0 from Theorem 1.5, let $W^* \subset E_+(\alpha_0)$ be the full measure, open subset promised by Theorem 1.5. Fix $x_0 \in W^*$. Then there will be values (depending on x_0) $\rho = 2^{-l}$, $k = 4^l$ of the variables ρ, k in Algorithm A, where $l \geqq 0$ is an integer, such that the sequence $x_i = (T_\rho)^i(x_0)$ is defined for $i = 0, \cdots, m$ where $m \leqq k$ and $|f(x_m)| \leqq \varepsilon$.

Fix $x_0 \in W^*$. There is a smallest integer $l \geqq 0$ with the following property: for some $m \in \{0, 1, \cdots, 4^l\}$ and with $\rho = 2^{-l}$, we have $T_\rho^m(x_0) \in B$ and $|T_\rho^m(x_0)| < \varepsilon$.

Now start Algorithm B at x_0. We have to show that at some stage a value of y is reached for which the loop marked ①, which is Newton iteration, cycles forever. If this does not happen, then k must run through every integer greater than 0. If $k = 4^l$, where l is as above, then y takes the value $T_\rho^m(x_0)$, where $\rho = 2^{-l}$ and $m \leqq k$. But for such a y Newton iteration proceeds without stop, so that eventually loop ① must cycle forever. By the definition of ε and B it follows that the Newton iterates of y converge as required.

By a slight modification, Algorithm B and Theorem 4.3 can be adapted to maps defined on any nonempty open subset $\Omega \in \mathbb{R}^n$. In fact, Theorem 4.3 remains valid if $f : \Omega \to \mathbb{R}^n$ and $\alpha_0 \geqq 0$ have the properties listed in Theorem 4.3, provided Algorithm B is changed as follows: replace the query "$J(x_i) \neq 0$?" with "$x_i \in \Omega$ and $J(x_i) \neq 0$?"

An analogous change is made in Algorithm A of Section 2; one then extends Theorem 2.1 to cover maps $\Omega \to \mathbb{R}^n$. The rest of the proof follows the same pattern as that of Theorem 4.3.

In a similar way we can adapt Theorem 4.3 to maps $f : \Omega \to \mathbb{R}^n$ of a compact domain $\Omega \to \mathbb{R}^n$.

Recall the map $g : \Omega - f^{-1}(0) \to S^{n-1}$, and the unit vector field ϕ : Reg $(g) \to \mathbb{R}^n$, defined in Section 1.

THEOREM 4.4. *Let $\Omega \subset \mathbb{R}^n$ be a compact domain with smooth boundary $\partial\Omega$. (That is, $\partial\Omega$ is a compact, C^1 submanifold of $\mathbb{R}^n$, of dimension $n-1$, and Ω is the closure of a bounded component of $\mathbb{R}^n - \partial\Omega$.) Let $f : \Omega \to \mathbb{R}^n$ be a C^2 map satisfying* ND. *For all $x \in \partial\Omega$ assume that $f(x) \neq 0$, g is regular at x, and $\phi(x)$ is transverse to $\partial\Omega$ and points into Ω at x. Then there exists a neighborhood $W \subset \Omega$ of $\partial\Omega$ and an open subset of full measure $W^* \subset W - \partial\Omega$ such that if $x_0 \in W^*$, then the forward trajectory of x_0 under ϕ leads to a zero of f. Algorithm* B, *adapted as above, produces a sequence converging to a zero of f, ultimately by Newton–Raphson iteration.*

The details of the proof are left to the reader. The main observation is that if $g(x_0)$ is a regular value for g, then the forward trajectory of x_0 under the flow of ϕ cannot approach $\partial\Omega$ because of the hypothesis that $\phi \mid \partial\Omega$ points inward.

Notice that the hypothesis of Theorem 4.4 allows $J(x)$ to change sign in $\partial\Omega$.

5. A Practical Algorithm

Here we modify the previous algorithms, A and B. The new Algorithm C, similar in spirit, has proved efficient in some computer trials. Here is the flow chart:

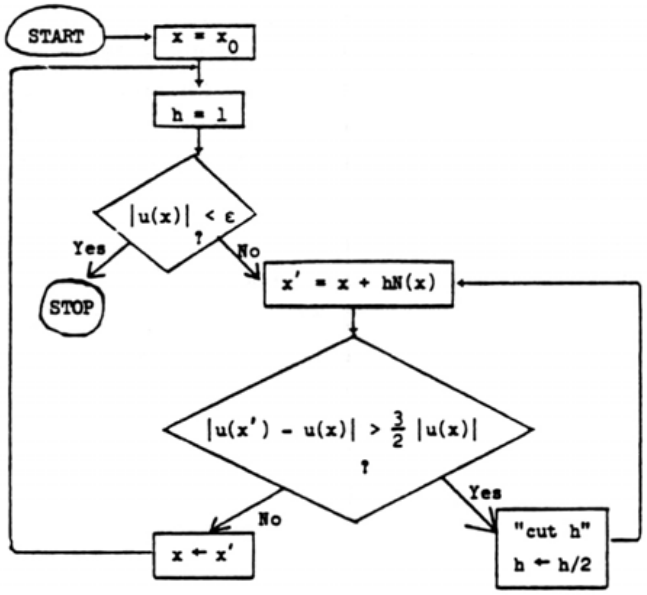

Figure 3. Flow chart for Algorithm C. To solve $|u(x)|<\varepsilon$. $u : \mathbb{R}^n \to \mathbb{R}^n$, $J(x) = \operatorname{Det} Du_x$, $N(x) = -\operatorname{sgn} J(x)\, Du_x^{-1} u(x)$. $\varepsilon > 0$ and $x_0 \in \mathbb{R}^n$ are given.

This algorithm is intended for use with maps $u : \mathbb{R}^n \to \mathbb{R}^n$ which satisfy the same conditions as before: u is proper, $J(x) \geqq 0$ outside a compact set, and J is not identically zero. There is surely no guarantee that Algorithm C will stop. It is not hard to find maps u and unbounded open sets $W \subset \mathbb{R}^n$ such that for all starting points $x_0 \in W$ the algorithm cycles forever.

On the other hand, our practical experience with Algorithm C has been excellent. Some details of this experience are now presented.

EXAMPLE 1. Here $u=(u_1, u_2): \mathbb{R}^2 \to \mathbb{R}^2$,

$$u_1(x, y) = x^3 - 3xy^2 + a_1(2x^2 + xy) + b_1 y^2 + cx + dy,$$

$$u_2(x, y) = 3x^2 y - y^3 - a_1(4xy - y^2) + a_2 x^2 + b_2.$$

The problem is to solve $|u(x, y)| < \varepsilon = 10^{-5}$, for various choices of the coefficients a_i, b_i, c, d. If a_1 is large relative to the other coefficients, then the set where $J(x) < 0$ will be big. Except for this restriction the coefficients were given arbitrary values.

On September 20, 1977, we compared Algorithm C with Newton's method, for three choices of coefficients and initial data, using a PDP 11 computer. (By "Newton's method" we mean the algorithm $x_{i+1} = x_i - Df(x_i)^{-1} f(x_i)$.) The results are as follows.

ALGORITHM C

Coefficients	Problem 1	Problem 2	Problem 3
a_1, b_1, c, d, a_2, b_2	25, 1, 2, 3, 4, 5	200, 1, 2, 3, 1, 2	25, −1, −2, −3, −4, −5
Initial (x, y)	(2, 2)	(−1, −1)	(1, 1)
Number of iterations	10	46	13
Number of cuts of h	0	2	1
Final (x, y)	(36.0454, 36.8056)	(0.5115, 197.936)	(39.0207, 38.2417)

NEWTON'S METHOD

In each problem there was no convergence after more than 100 iterations.

EXAMPLE 2. On April 2, 1975, the following map $u: \mathbb{R}^5 \to \mathbb{R}^5$ was tested on a PDP 11, to solve $|u(x)| < \varepsilon = 10^{-5}$:

$$u_1(x) = x_1^3 + a_1 x_1 x_5 + a_2,$$

$$u_2(x) = x_2^3 + b_1 x_2 x_5 + b_2,$$

$$u_3(x) = x_3^3 + c_1 x_2 x_3 + c_2 x_5,$$

$$u_4(x) = x_4^3 + d_1(x_1^2 + x_3^2) + d_2 x_5,$$

$$u_5(x) = x_5^3 + p_1 x_1 x_3 + p_2 x_2 + x_4.$$

With 14 different choices of coefficients and initial point, Algorithm C always

converged averaging about 17 iterations and 7 cuts of h. The largest number of iterations was 38 and the most times h was cut was 45.

Newton's method also solved most but not all of these problems. In one problem it was faster and in one it was equally fast; in the others it was slower, or did not converge.

Other modifications of Newton's method were tried and found to be less effective on these problems than Algorithm C.

The referee has given us the reference, Abbot and Brent, *Fast local convergence with single and multi-step methods for nonlinear equations*, J. Austr. Math. Soc., Ser. B, Vol. 19, 1975, pp. 173–199, for a discussion of and references to other algorithms on the solution of nonlinear equations.

6. A Sure-Fire Algorithm

Here we present an algorithm related to Newtonian iteration. For any nonconstant polynomial $f(z)$ it is guaranteed to produce a sequence $\{z_j\}$ converging to a zero of f, starting from any z_0.

The algorithm has several virtues. In addition to a certain robustness and naturality, the cost of computing z_{j+1} from z_j is bounded by a function of the degree n of the polynomial. Moreover, there are constants $K_n < 1$, depending only on n, with

$$\frac{|f(z_j)|}{|f(z_0)|} \leqq (K_n)^j$$

for all j.

A defect of the algorithm is that the numbers K_n, as defined below, are extremely close to 1 for large n. It is conceivable that better numbers can be found.

The Newtonian iteration scheme for solving $f(z)=0$ can be described as follows. Given $z_0, \cdots, z_j$, choose z_{j+1} to be the root of the first-order Taylor expansion of f at z_j. Thus we find the root w of

$$f(z_j)+f'(z_j)(w-z_j)=0\,,$$

so that

$$z_{j+1}=z_j-f(z_j)/f'(z_j)\,.$$

A more sophisticated strategy is to set

$$z_{j+1}=z_j-hf(z_j)/f'(z_j)\,,$$

where h is chosen so that $|f(z_{i+1})|<|f(z_i)|$. This can always be done with sufficiently small $h>0$ (if $f(z_i)\neq 0$ and $f'(z_i)\neq 0$). One can interpret this z_{i+1} as the solution w of the equation

$$f(z_i)+f'(z_i)(w-z_i)=(1-h)f(z_i).$$

A difficulty with this strategy is that one cannot make the ratio $|f(z_{i+1})|/|f(z_i)|$ *uniformly* less than 1. The root of the trouble is that $|f'(z_i)|$ may get too small.

It is natural to try to fix this by using a higher-order Taylor expansion: if $|f'(z_i)|$ is small, try the second-order Taylor expansion with the linear term deleted. In this scheme, z_{i+1} is a root w of

$$f(z_i)+\frac{f''(z_i)}{2!}(w-z_i)^2=(1-h)f(z_i).$$

Setting $\sqrt{h}=\sigma>0$ gives

$$z_{i+1}=z_i+\sigma e^{i\alpha/2}u, \qquad u^2=\frac{2!\,f(z_i)}{f''(z_i)}.$$

Pursuing this idea with higher derivatives leads to the following result.

THEOREM 6.1. *Let n be a positive integer. There are real numbers $\sigma_1,\cdots,\sigma_n$, K_n with $0<\sigma_i\leqq 1$ and $0<K_n<1$, having the following property. Let f be any complex polynomial of degree n and z any complex number. For each m such that $f^{(m)}(z)\neq 0$ let u_m be any m-th root of $m!\,f(z)/f^{(m)}(z)$ and set*

$$w_m=z+\sigma_m e^{i\alpha/m}u_m.$$

Then some m satisfies $|f(w_m)|\leqq K_n|f(z)|$.

Explicit choices of $\sigma_1,\cdots,\sigma_n, k_n$ are given below. It will be seen that the σ_i can be chosen robustly in the sense that any numbers sufficiently near them will have the same property (for perhaps a slightly larger K_n).

The algorithm for solving $f(z)=0$ begins by choosing $z_0\in\mathbf{C}$ arbitrarily. Suppose $z_0,\cdots,z_i$ have been chosen. If $f(z_i)=0$, put $z_{i+1}=z_i$. Suppose $f(z_i)\neq 0$. Put $z=z_i$ and for each m with $f^{(m)}(z)\neq 0$ define u_m and w_m as in the theorem. Let m_0 be the smallest m for which $|f(w_m)|\leqq K_n|f(z)|$ and set $z_{i+1}=w_{m_0}$.

In this way a sequence $z_0, z_1,\cdots$ is obtained; it has the property that

$$|f(z_i)|\leqq (K_n)^i|f(z_0)|.$$

Thus $|f(z_l)|\to 0$. In fact it is easy to see that the sequence $\{z_l\}$ converges to a root of f.

The σ_j come from the following result.

PROPOSITION 6.2. *Let n be a positive integer. There exist real numbers $\sigma_1, \cdots, \sigma_n, K_n$ with $0<K_n<1$ and $0<\sigma_k \leqq 1$ for $k=1, \cdots, n$, having the following property. Set*

$$h_k = \sigma_k e^{i\pi/k}, \qquad k=1, \cdots, n.$$

For any $(v_1, \cdots, v_n) \in \mathbb{C}^n - 0$ there exists $m \in \{1, \cdots, n\}$, depending on v, such that $v_m \neq 0$ and

$$|\phi_{m,v}(h_m)| < K_n,$$

where

$$\phi_{m,v}(h) = 1 + \sum_{k=1}^{n} \left(\frac{v_k}{v_m} h\right)^k.$$

Before proving the proposition we use it to prove Theorem 6.1. Let n be a positive integer. Fix the σ_k and h_k as in the proposition. Given a polynomial f of degree n let $z \in \mathbb{C}$ be arbitrary. If $f(z)=0$, take $z'=z$. Suppose $f(z)\neq 0$. Let v_k be a k-th root of

$$f^{(k)}(z)/k!\, f(z).$$

Suppose $v_m \neq 0$ and $z' = z + h_m/v_m$. Then Taylor's theorem gives

$$\begin{aligned} f(z') &= f(z) + \sum_{k=1}^{n} \frac{f^{(k)}(z)}{k!} \left(\frac{h_m}{v_m}\right)^k \\ &= f(z)\left[1 + \sum_{k=1}^{n} \frac{f^{(k)}(z)}{f(z)k!} \left(\frac{h_m}{v_m}\right)^k\right] \\ &= f(z)\left[1 + \sum_{k=1}^{n} \left(\frac{v_k}{v_m} h_m\right)^k\right]. \end{aligned}$$

By the proposition, m can be chosen so that $v_m \neq 0$ and the norm of the expression in brackets is at most K_n. This proves Theorem 6.1.

It remains to prove Proposition 6.2.

Let $i_0, \cdots, i_l$ be a finite sequence in $\{1, \cdots, n\}$, $l \geqq 1$. Call $(i_0, \cdots, i_l)$ a *cycle* if $i_0 = i_l$ and $i_k \neq i_s$ otherwise.

A real $n \times n$ matrix $\delta = [\delta_{ij}]$ is *positive on cycles* if

$$\sum_{k=1}^{l} \delta_{i_k, i_{k+1}} > 0$$

for every cycle $(i_0, \cdots, i_l)$, for all $l = 1, \cdots, n$. Note that the set of such matrices is open in the space $\mathbf{R}^{n^2}$ of all $n \times n$ matrices.

LEMMA 6.3. *Let δ as above be positive on cycles. Let $w = (w_1, \cdots, w_n) \in \mathbf{R}^n$ and define the $n \times n$ matrix*

$$W = W(w) = [w_i - w_j].$$

Then $\delta - W$ has a positive row.

Proof: If not, for each $m \in \{1, \cdots, n\}$ there is a smallest $\alpha(m) \in \{1, \cdots, n\}$ such that the $(m, \alpha(m))$ entry in $\delta - W$ is at most 0. The map α of $\{1, \cdots, n\}$ into itself has a periodic point $m \in \{1, \cdots, n\}$, of minimal period $l \in \{1, \cdots, n\}$. Then $\alpha^l(m) = m$ and $\alpha^k(m) \neq \alpha^j(m)$ if $1 \leqq k < j \leqq l$. Put $\alpha^j(m) = i_j$; then $(i_0, \cdots, i_l)$ is a cycle. We have

$$0 \geqq \sum_{k=1}^{l} (\delta - W)_{i_k, i_{k+1}}$$

$$= \sum_{k=1}^{l} \delta_{i_k, i_{k+1}} + \sum_{k=1}^{l} (w_{i_{k+1}} - w_{i_k}).$$

The last summation is 0, contradicting the assumption that δ is positive on cycles.

LEMMA 6.4. *Let n be a positive integer. Given positive numbers $\sigma_1, \cdots, \sigma_n$ define an $n \times n$ matrix $\delta = [\delta_{m,k}]$ by*

$$\delta_{m,k} = \frac{1}{k} \log \left(\frac{\sigma_m^{m-k}}{n} \right).$$

Then the σ_k can be chosen so that $\sigma_n = 1$, $0 < \sigma_k < 1$ for $k = 1, \cdots, n-1$ and δ is positive on cycles.

Proof: It suffices to consider cycles $(m_0, \cdots, m_l)$ having $m_0 < m_{k_j}$ for $j = 1, \cdots, l-1$. Positivity on cycles means

$$\sum_{j=1}^{l} \delta_{m_{j-1}, m_j} = \sum_{j=1}^{l} \frac{1}{m_j}[(m_{j-1} - m_j) \log \sigma_{m_{j-1}} - \log n] > 0$$

for every such cycle. It suffices to have

$$\frac{1}{m_1}(m_0 - m_1) \log \sigma_{m_0} > \left| \sum_{j=2}^{l} \frac{1}{m_j}(m_{j-1} - m_j) \log \sigma_{m_{j-1}} \right| + \sum_{j=1}^{l} \frac{1}{m_j} \log n,$$

or, since $1/n < |m_{j-1} - m_j|/m_j \leqq n$, to have

$$-\frac{1}{n} \log \sigma_{m_0} \geqq \sum_{j=2}^{l} n |\log \sigma_{m_{j-1}}| + \left(\sum_{j=1}^{l} \frac{1}{m_j} \right) \log n.$$

Now

$$\sum_{j=1}^{l} \frac{1}{m_j} \leqq \sum_{j=1}^{n} \frac{1}{j} < 1 + \log n,$$

because the m_j are distinct. Since $m_0 < m_j$ for $j = 1, \cdots, l-1$, it therefore suffices to choose the σ_k so that

$$(*) \qquad -\frac{1}{n} \log \sigma_k \geqq \sum_{j=k+1}^{n} n |\log \sigma_j| + (1 + \log n) \log n.$$

This can be done recursively. First put $\sigma_n = 1$. Assuming σ_j has been chosen for $n \geqq j \geqq k+1$, choose σ_k to satisfy $(*)$, or the equivalent inequality

$$\log (\sigma_k^{-1}) \geqq \sum_{j=k+1}^{n} n^2 \log (\sigma_j^{-1}) + n(1 + \log n) \log n.$$

For example, let $b = n(1 + \log n) \log n$, and put $a_{n-k} = \log (\sigma_k^{-1})/b$. It suffices to have

$$a_0 = 0,$$

$$a_k = n^2(a_0 + \cdots + a_{k-1}) + 1 \quad \text{for} \quad k = 1, \cdots, n-1.$$

This has the solution

$$a_0 = 0\,,$$

$$a_k = (n^2+1)^k \quad \text{for} \quad k = 1, \cdots, n-1\,,$$

yielding

$$\sigma_n = 1\,,$$

$$\sigma_k = n^{-[n(n^2+1)^{n-k}(1+\log n)]}$$

for $k = 1, \cdots, n-1$. If n is large, σ_1 is rather small.

We now complete the proof of Proposition 6.2. Let $\sigma_1, \cdots, \sigma_n$ be as in Lemma 6.4 and $h_1, \cdots, h_n$ as in Proposition 6.2. Define

$$K_n = \max\{1-(\sigma_m)^m/n : m = 1, \cdots, n\}\,.$$

Let $v = (v_1, \cdots, v_n) \in \mathbb{C}^n - 0$ be given and suppose at first that all $v_m \neq 0$. Then

$$|\phi_m(h_m)| \leqq |1+(h_m)^m| + \left|\sum_{\substack{k=1 \\ k\neq m}}^{n} \left(\frac{v_k}{v_m}\right)^k h_m^k\right|$$

$$\leqq 1-(\sigma_m)^m + \sum_{\substack{k=1 \\ k\neq m}}^{n} \left(\frac{|v_k|}{|v_m|}\right)^k \sigma_m^k\,.$$

We shall find an m such that

$$\left(\frac{|v_k|}{|v_m|}\right)^k (\sigma_m)^k < \frac{(\sigma_m)^m}{n}$$

for $k = 1, \cdots, n$. This will prove the proposition. Taking logarithms we see that this is equivalent to finding m for which

$$\frac{1}{k}\log\frac{(\sigma_m)^{m-k}}{n} - (\log|v_k| - \log|v_m|) > 0$$

for $k=1,\cdots,n$. Such an m exists by the lemmas: let $w_k=\log|v_k|$. Therefore, for each n-tuple v of nonzero complex numbers we have found an $m=m(v)$ such that

$$|\phi_{m,v}(h_m)|\leqq K_n\,.$$

Now let $v\in\mathbb{C}^n-0$ be arbitrary. Let $\{w_i\}$ be a sequence of vectors in $\mathbb{C}^n-0$ which converges to v, such that each component of each w_i is nonzero. Put $m(w_j)=m_j\in\{1,\cdots,n\}$. By choosing a subsequence we may assume all the m_j are equal to some $m\in\{1,\cdots,n\}$. By continuity it follows that

$$|\phi_{m,v}(h_m)|=\lim_{j\to\infty}|\phi_{m,w_j}(h_m)|\leqq K_n\,.$$

This completes the proof of Proposition 6.2.

The robustness of the σ_k and of the algorithm follows from the fact that if the $n\times n$ matrix

$$[\sigma_{m,k}]=\left[\frac{1}{k}\log\left(\frac{(\sigma_m)^{m-k}}{n}\right)\right]$$

is positive on cycles, so is any sufficiently nearby matrix.

For background, and other treatments of the problem considered here, one can see [7] as well as: Dejon and Herrin, *Constructive Aspects of the Fundamental Theorem of Algebra*, Wiley, New York, 1969.

7. Other Sure-Fire Algorithms

Algorithms A (Section 2) and B (Section 4) can be converted into algorithms which theoretically never fail. To solve $|f(x)|<\varepsilon$, for example, suppose $f:\mathbb{R}^n\to\mathbb{R}^n$ is a proper analytic map whose Jacobian J is non-negative outside a compact set and does not vanish identically. Let $y_1, y_2,\cdots$ be any sequence dense in $\mathbb{R}^n$.

ALGORITHM A$^{\#}$. For each $j=1,2,\cdots$ in succession, define a finite sequence as follows. Start Algorithm A at $x_0=y_j$; this produces a sequence $x_0,\cdots$ which we stop at x_m, $m=m(j)$, as soon as $J(x_m)=0$ or $|f(x_m)|<\varepsilon$ or $m=j$, whichever happens first.

We claim that, for some j, what happens is that $|f(x_{m(j)}|<\varepsilon$. This is because the proof of Theorem 2.1 shows that for every y in the nonempty open set $W^* \subset \mathbb{R}^n$ there is a neighborhood $V \subset W^*$ of y and an integer $\nu > 0$ with the following property: whenever Algorithm A is started at any $x_0 \in V$, it stops after at most ν steps at a solution to $|f(y)|<\varepsilon$. Since $\{y_j\}$ is dense in $\mathbb{R}^n$, there exists a j such that $y_j \in V$ and $j \geqq \nu$. Therefore when Algorithm A is started at $x_0 = y_j$, the only way the sequence $x_0, x_1, \cdots$ can stop in Algorithm $A^{\#}$ is if $|f(x_m)|<\varepsilon$ for some $m \leqq \nu$.

It is not obvious (but seems likely) that Algorithm $A^{\#}$ is usually faster than the following simple-minded algorithm: test each y_j in succession, stopping when $|f(y_j)|<\varepsilon$.

It is somewhat more practical to take $\{y_n\}$ dense in a convenient open set U contained in the set $E_+(\alpha_0)$ described in Theorem 2.1 (if such a set is known).

In a similar way one can adapt Algorithm B of Section 4 to obtain an Algorithm $B^{\#}$ with the following property: if $f : \mathbb{R}^n \to \mathbb{R}^n$ is a proper C^2 map satisfying the nondegeneracy condition ND of Section 4, and $|J| \geqq 0$ outside some compact set (which may not be explicitly known), then the successive values of x in Algorithm $B^{\#}$ form a sequence converging to a zero of f. Moreover, *after some stage the sequence proceeds* by Newton–Raphson iteration.

The algorithm depends on an arbitrary sequence $\{y_l\}$ which is dense in $\mathbb{R}^n$. Define

$$M : \mathbb{R}^n - J^{-1}(0) \to \mathbb{R}^n, \qquad M(x) = x + N(x),$$

where $N(x)$ is the Newton vector

$$N(x) = -\operatorname{sgn} J(x)\, Df_x^{-1} f(x).$$

Let M^0 be the identity map of $\mathbb{R}^n$ and by recursion let M^j be the j-th iterate of M:

$$M^j = M \circ M^{j-1} : \mathbb{R}^n - (M^{j-1})^{-1} J^{-1}(0) \to \mathbb{R}^n.$$

As usual define, for $\rho < 0$,

$$T_\rho : \mathbb{R}^n - (J^{-1}(0) \cup f^{-1}(0)) \to \mathbb{R}^n,$$

$$T_\rho(x) = x + tN(x), \qquad |tN(x)| = \rho, \qquad t > 0.$$

The algorithm has the following flow chart:

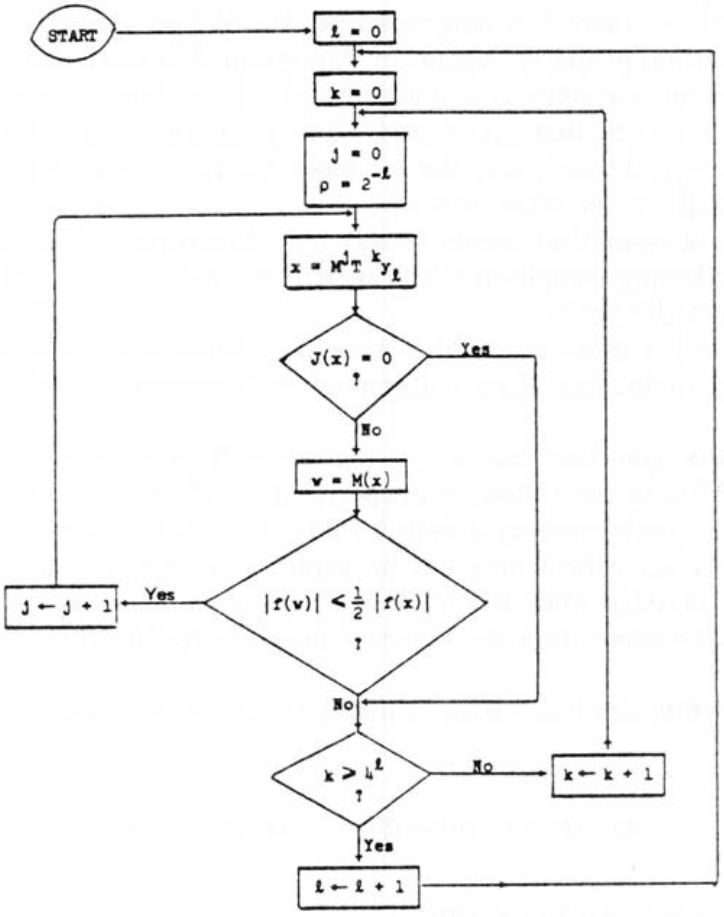

Figure 4. Flow chart for Algorithm B$^{\#}$.

The verification that Algorithm B$^{\#}$ has the property asserted above is left to the reader.

With slight modifications the algorithm can be applied to a polynomial $g(z)$ in one complex variable, of degree at least 1. To do this, first find a polynomial $f(z)$ with simple roots which divides $g(z)$. This can be done by forming the sequence $g = g_0, g_1 \cdots$, where g_k is the greatest common divisor of g_{k-1} and its derivative g'_{k-1}. Eventually degree $g_{k+1} = 0$ and we take $f = g_k$. Let S be a circle in C centered at 0 which is large enough to contain the roots of f in its interior; such a circle can be easily determined from the coefficients. Let $\{y_n\}$ be a sequence dense in S, e.g.,

$$y_n = r(\cos n\pi\alpha + i \sin n\pi\alpha),$$

where $r>0$ is the radius of S and $\alpha \in \mathbb{R}$ is irrational. Then Algorithm $B^{\#}$ will produce a sequence converging to a root of f. The proof is left to the reader.

Bibliography

[1] Abraham, R., and Robbin, J., *Transversal Mappings and Flows*, Benjamin, New York, 1967.

[2] Eaves, C., and Scarf, H., *The solution of systems of piecewise linear equations*, Math. Operations Res. 1, 1976, pp. 1–27.

[3] Hirsch, M., *A proof of the non-retractibility of a cell onto its boundary*, Proc. Amer. Math. Soc. 14, 1963, pp. 364–365.

[4] Hirsch, M., *Differential Topology*, Springer-Verlag, New York, 1976.

[5] Kellogg, R. B., Li, T. Y., and Yorke, J. A., *A constructive proof of the Brouwer fixed-point theorem and computational results*, SIAM J. Numer. Anal. 13, 1976, pp. 473–483.

[6] Lojasciewicz, S., *Triangulation of semi-analytic sets*, Ann. Scuole Norm. Sup. Pisa 3, 18, 1964, pp. 449–474.

[7] Ostrowski, A., *Solutions of Equations in Euclidean and Banach Spaces*, Academic Press, New York, 1973.

[8] Scarf, H., *The Computation of Economic Equilibria* (in collaboration with Hansen, T.), Yale U.P., New Haven, 1960.

[9] Smale, S., *A convergent process of price adjustment and global Newton methods*, J. Math. Econom. 3, 1976, pp. 107–120.

[10] Sternberg, S., *Lectures on Differential Geometry*, Prentice-Hall, Englewood Cliffs, N.J., 1964.

Received January, 1978.