Physica 124A (1984) 465-470
North-Holland, Amsterdam



# COMPUTER-ASSISTED PROOFS IN ANALYSIS

Oscar E. LANFORD III

IHES, 91440 Bures-sur-Yvette, France. Work supported in part by NSF Grant MCS81-07086.

Computers have a number of uses in mathematics and mathematical physics. Two which are relatively familiar are heuristic numerical exploration and determining properties of discrete objects (e.g., testing integers for primality). I will describe here an example of a less familiar use of computers: the strict verification of estimates on continuously variable quantities (real numbers) for use in the proof of qualitative statements in analysis. This report is a case study rather than a general survey; it is based on my experience working on a concrete problem, the validity of Feigenbaum's renormalization group analysis of the accumulation of period-doubling bifurcations.[2] I am confident that the strict verification of estimates by computer will have other interesting applications, but it is too early to tell how much the particular methods described here will be useful in other situations.

An argument using a computer to prove a mathematical assertion can be thought of as divided into two stages. It is first necessary to derive a sufficient condition for the validity of the assertion which can be verified by a finite computation; then to carry out the verification. The first, analytic, stage is standard mathematics, although computer exploration is likely to be used to help choose a sufficient condition which is actually true. The second, computational, stage is in principle completely mechanical, but in practice considerable thought has to go into structuring the computation so as to make it as comprehensible as possible and hence to minimize the likelihood of error.

The concrete problem to be discussed, motivated by considerations in the analysis of infinite sequences of period-doubling bifurcations,[1] is as follows: We consider the operator

$$\mathcal{T}f(x) = \frac{1}{f^2(0)} \quad f \cdot f\,(f^2(0)\,x)$$

acting on an appropriate domain in the space of mappings f of [-1,1] into itself which are even, decreasing on [0,1] (and hence have a maximum at 0), and

Presented at the VIIth INTERNATIONAL CONGRESS ON MATHEMATICAL PHYSICS, Boulder, Colorado, 1983.

which map 0 to 1. We want to show that:

1. The operator $\mathfrak{T}$ has a fixed point g, analytic on a neighborhood of [-1,1], given approximately by

$$g(x) \simeq 1 - 1.5x^2 + .1\ x^4$$

2. The spectrum of the linearization $D\mathfrak{T}(g)$ of the operator $\mathfrak{T}$ at the fixed point g is strictly inside the unit disk except for a single simple positive eigenvalue larger than one.

To prove the existence of the fixed point we use a simplified version of Newton's method. Newton's method for solving

$$\mathfrak{T}(g)-g = 0$$

gives the iteration

$$g_{n+1} = g_n - (D\mathfrak{T}(g_n) - \mathbb{1})^{-1}\ (\mathfrak{T}(g_n) - g_n)$$

We use instead an iteration of the form

$$g_{n+1} = g_n - (\Gamma - \mathbb{1})^{-1}\ (\mathfrak{T}(g_n) - g_n) \equiv \Phi(g_n) \qquad (1)$$

where $\Gamma$ is an operator (to be chosen) approximating $D\mathfrak{T}(f)$ reasonably well for f near g. It is easy to derive a sufficient condition for the convergence of the sequence $(g_n)$ obtained from (1) with a given $g_0$ as follows: In order that $\Phi$ be contractive on a ball of radius $\rho$ about $g_0$, it suffices that

$$\|D\Phi(f)\| = \|(\Gamma-\mathbb{1})^{-1}(D\mathfrak{T}(f)-\Gamma)\| \leq \kappa < 1 \quad \text{for } \|f-g_0\| \leq \rho \qquad (2)$$

and in order that $\Phi$ map this ball into itself it suffices that

$$\|(\Gamma - \mathbb{1})^{-1}\ (\mathfrak{T}(g_0) - g_0)\| \leq \rho(1 - \kappa) \qquad (3)$$

Thus, to prove the existence of a fixed point g, all we have to do is to find $g_0$, $\Gamma$, $\rho$, $\kappa$ so that (2) and (3) hold. If, furthermore, (3) is sharpened to:

$$\|(\Gamma - e^{i\theta}\,\mathbb{1})^{-1}\ (D\mathfrak{T}(f) - \Gamma)\| \leq \kappa \qquad (4)$$

for all f with $\|f - g_0\| \leq \rho$ and all real $\theta$, and if $\Gamma$ has spectrum inside the unit disk except for a single simple expanding eigenvalue, then the same will

be true for $D\mathfrak{T}(g)$. We will from now on discuss only the proof of (2) and (3); (4) is proved in much the same way.

Next we have to choose a Banach space of functions f in which the inequalities (2) and (3) are to be proved. We will work with functions f analytic in

$$\{x: |x^2 - 1| < 2.5\}$$

(Many other choices of domain of analyticity would also work.) To define the norm we will use, we first write the general mapping f as

$$f(x) = 1 + x^2 h(x^2).$$

This form builds in the assumed evenness of f and the constraint $f(0) = 1$; working with h instead of f is simply a convenient change of coordinates in the space of mappings. We next expand

$$h(z) = \sum_{n=0}^{\infty} h_n \left(\frac{z-1}{2.5}\right)^n ;$$

define a norm by

$$\|h\| = \sum_{n=0}^{\infty} |h_n| ;$$

and work in the space of h's for which this norm is finite. (Finiteness of this norm says a little more than that h is analytic on the interior of $\overline{\Omega} = \{z: |z-1| \leq 2.5\}$ and continues on the boundary and a little less than that it is analytic on a neighborhood of $\overline{\Omega}$.) The reason for choosing this norm, instead of the more obvious supremum norm, is that it is well adapted to making accurate estimates of norms of linear operators: If T is a linear operator on the space of h's equipped with the above norm, then

$$\|T\| = \sup_{n \geq 0} \|Te_n\| ,$$

where the $e_n$ are the "natural basis vectors" $\left(\frac{z-1}{2.5}\right)^n$ .

Now comes a piece of good fortune: It turns out that we can use a very simple approximate derivative $\Gamma$, namely:

$$\Gamma e_0 = 4.669\, e_0$$

$$\Gamma e_n = 0 \quad \text{for } n > 0$$

We take an explicit approximate fixed point $g_0$ obtained as the result of solving he fixed-point problem numerically with good accuracy and we choose an explicit $\rho(\simeq .01)$ and $\kappa(\simeq .9)$. Once these choices are made the inequalities (2) and (3) which will imply the existence of the fixed point are completely explicit.

Before describing how to organize the verification of these inequalities, we need to compute the operator $D\mathfrak{T}(f)$. Heuristically, this is done by replacing f by $f + \delta f$ in the formula

$$\mathfrak{T}f(x) = \frac{1}{f(1)} \; f \cdot f\,(f(1) \cdot x)$$

and extracting the terms linear in $\delta f$. This produces a sum of four terms, one for each place f appears in the expression for $\mathfrak{T}f$. These expressions must then be rewritten in terms of h and $\delta h$ (related to f, $\delta f$ by

$$f(x) = 1 + x^2 h(x^2); \quad \delta f(x) = x^2 \delta h(x^2)\ )$$

To show what the final expressions look like, we write one of the four terms:

$$(D\mathfrak{T}^{(2)}(h)e_j)(z) = \frac{\lambda}{2.5}\;(1 + \lambda^2\, zh(z))^2\,(2 + \lambda^2\, zh(\lambda^2 z))$$
$$\times\; h(\lambda^2 z)\left[\frac{\lambda^2\, z\, h(\lambda^2 z)}{2.5}\right]^{j-1},$$

$$\lambda \equiv h(1) + 1$$

(This holds for $j \geq 1$; for $j = 0$, the left-hand side vanishes.) To verify (2), we estimate the location in function space of the right-hand side of (5) in tersm of information about the location of h (and similarly for the other three terms in the expression for $D\mathfrak{T}(h)$).

These estimates are completely straightforward in principle; in practice, they quickly become unreasonably complicated if not structured carefully. To structure the computation we use the notion of rectangle in function space. By this we mean a set $\mathbf{R}$, in the Banach space of h's, of the form

$$\{h = \sum_{j=0}^{\infty} h_j \left(\frac{z-1}{2.5}\right)^j : \ell_0 \leq h_0 \leq u_0, \ldots, \quad \ell_{n-1} \leq h_{n-1} \leq u_{n-1}, \quad \sum_{j=n}^{\infty} |h_j| \leq \varepsilon\},$$

determined by 2n+1 numbers

$$\ell_0 \leq u_0, \quad \ell_1 \leq u_1, \ldots, \quad \ell_{n-1} \leq u_{n-1}, \quad \varepsilon \geq 0$$

It is straightforward and reasonably simple to work out how to do elementary operations on these rectangles. For example: Given two rectangles $\mathcal{R}_1$ and $\mathcal{R}_2$, one constructs a rectangle $\mathcal{R}_3$ (" = $\mathcal{R}_1 \times \mathcal{R}_2$") such that, whenever $h_1 \in \mathcal{R}_1$ and $h_2 \in \mathcal{R}_2$, $h_1 \cdot h_2 \in \mathcal{R}_3$. Similarly for such other operations as addition, scalar multiplication, and composition.

In terms of these elementary operations, it is also not too difficult to give a prescription for finding, given a rectangle $\mathcal{R}_0$ and an integer j, another rectangle guaranteed to contain $D\mathcal{T}(h)e_j$ for any $h \in \mathcal{R}_0$. From this rectangle, it is easy to obtain a bound on

$$\|(\Gamma - \mathbb{1})^{-1}(D\mathcal{T}(h) - \Gamma)\, e_j\|$$

which holds for all $h \in \mathcal{R}_0$. Now recall that, to prove (2), we have only to show that

$$\|(\Gamma - \mathbb{1})^{-1}(D\mathcal{T}(h)-\Gamma)\, e_j\| \leq \kappa \quad \text{for all } j.$$

The above permits us to check this inequality for any given j. There are, however, infinitely many j's to be considered, so we are not yet reduced to a finite computation. Fortunately, only finitely many j's require estimates of the above detailed kind. To see why, consider the sample term written in (5) above, and note that it has the form

$$(D\mathcal{T}^{(2)}(h)e_j)(z) = u(z) \cdot (v(z))^{j-1}$$

with u(z), v(z) independent of j. Since

$$\|u \cdot v^{j-1}\| \leq \|u\|\, \|v\|^{j-1},$$

we can deal with all large j's simply by establishing a bound of the form

$$\|v\| \leq \sigma < 1 \qquad \text{for } \|h - h_0\| \leq \rho$$

Similar analyses can be done for the other three terms in the expression for $D\mathcal{T}(h)$.

There remains one more complication. Although it is possible to program a computer to do exact arithmetic (on rational numbers), this is usually impractical and arithmetic is instead normally done to some fixed finite precision. It is therefore necessary to control the effect of round-off error

if one is to make a strict verification of (2) and (3). There is a standard technique for the automatic estimation of round-off error, known as interval arithmetic. The idea of interval arithmetic is to represent numbers "with error bars" by specifying, instead of a single finite-precision approximation for a given number, the exact end-points of an interval guaranteed to contain the number in question. One can then construct computer procedures for "doing arithmetic operations on intervals." For example: Given two intervals $[\ell_1,u_1]$ and $[\ell_2, u_2]$, with $\ell_1$, $u_1$, $\ell_2$, $u_2$ all d-digit numbers, one finds another interval $[\ell_3, u_3]$, with $\ell_3$, $u_3$ again d-digit numbers, guaranteed to contain all products $x_1 \cdot x_2$ with $x_1 \in [\ell_1, u_1]$ and $x_2 \in [\ell_2, u_2]$. (To be entirely explicit: The best possible lower bound $\ell_3$ can be obtained by forming the four exact products $\ell_1 \cdot \ell_2$, $\ell_1 \cdot u_2$, $u_1 \cdot \ell_2$, $u_1 \cdot u_2$, each of which has no more than 2d digits; picking the smallest of them; and rounding down to the next smaller n-digit number. It is not really necessary, however, to find the best possible $\ell_3$, and usually some shortcuts are taken, giving an $\ell_3$ which is a correct lower bound but not necessarily the best possible one.)

We have now described all the elements needed to construct a computer program for verifying estimates (2) and (3). As indicated, this program can be organized into a number of reasonably simple pieces. At the lowest level is a set of computer procedures (subroutines) for doing the fundamental arithmetic operations on intervals. Built on these procedures is a higher-level set of procedures for doing elementary operations on rectangles in function space. The program for verifying (2) and (3) is constructed essentially by translating formulas like (5) into a sequence of invocations of these latter procedures.

REFERENCES

1) P Collet and J P Eckmann, Iterated Maps on the Interval as Dynamical Systems (Birkhäuser, 1980).

2) O E Lanford, A computer-assisted proof of the Feigenbaum conjectures. Bull. A.M.S. (New Series) 6 (1982) 427-434.