MATHEMATICS OF COMPUTATION, VOLUME 32, NUMBER 141
JANUARY 1978, PAGES 227–232

Stability of Rounded Off Inverses Under Iteration

By Harold G. Diamond*

Abstract. Let f be a monotone and strictly convex (or concave) function on a real interval and let g be the inverse function. Let $I(x) = x$. For ϕ a real valued function and N a positive integer let $\phi_N(x)$ denote the rounding of $\phi(x)$ to N significant figures. Let $h = g_N \circ f_N$, the composition of f_N and g_N. It is shown that

$$h \circ h \circ I_N = h \circ h \circ h \circ I_N,$$

and that equality can fail for fewer iterations.

1. Introduction. Let f denote a continuous strictly monotonic real valued function defined on some real interval $\mathcal{I}$ and let g denote the inverse function. For R a radix, N a positive integer, and ϕ any real valued function let ϕ_N denote the composition of the function $x \mapsto \phi(x)$ followed by rounding off to N significant figures in the base R.

Let $I(x) = x$, the identity function, and let $h = h^{(1)} = g_N \circ f_N$, the composition of f_N and g_N. We define $\mathcal{I}^* = \mathcal{I}^{*(1)}$, the domain of $h^{(1)}$, to be the set of $x \in \mathcal{I}$ for which $f_N(x) \in f(\mathcal{I})$. Similarly, for $k \geqslant 2$ we define $h^{(k)} = h \circ h^{(k-1)}$ on the domain $\mathcal{I}^{*(k)}$ consisting of all $x \in \mathcal{I}^{*(k-1)}$ for which $f_N \circ h^{(k-1)}(x) \in f(\mathcal{I})$.

The object of this article is to see whether h is, in some sense, an identity function on numbers having at most N significant digits. We consider the recursion

$$I_N, \quad f_N \circ I_N, \quad h \circ I_N, \quad f_N \circ h \circ I_N, \quad h^{(2)} \circ I_N, \ldots$$

and ask whether any of the equations

$$h \circ I_N = I_N, \tag{1A}$$

$$f_N \circ h \circ I_N = f_N \circ I_N, \tag{1B}$$

$$h^{(2)} \circ I_N = h \circ I_N, \tag{1C}$$

$$f_N \circ h^{(2)} \circ I_N = f_N \circ h \circ I_N, \tag{1D}$$

$$h^{(3)} \circ I_N = h^{(2)} \circ I_N, \tag{1E}$$

. . . are valid.

In concrete terms (which motivated the investigation), suppose we have a machine which very accurately performs a functional operation and then rounds off its results to N significant figures. We enter $x = I_N(x)$ and successively form

$$f_N(x), \quad g_N \circ f_N(x), \quad f_N \circ g_N \circ f_N(x), \ldots$$

Received October 27, 1975; revised September 13, 1976.
AMS (MOS) subject classifications (1970). Primary 65Q05; Secondary 65G05.
*Research supported in part by a grant from the National Science Foundation.

and ask whether, after a fixed number of steps, we get values repeated every other time.

Let us first discuss the rounding rules. To fix our ideas, from here until the statement of the theorem assume that the radix $R = 10$, the number of significant figures $N = 2$, and the interval $\mathcal{I} = \{t: .1 < t < 1\}$. Any rounding rule I_2 assigns the nearest two digit decimal in case it is unique. Specific rounding rules assign one of the two nearest values in the remaining cases. The ambiguous cases in $\mathcal{I}$ are the 90 numbers .105, .115, . . . , .985, .995.

It can happen that none of the equations (1. . .) holds. We give two examples of this phenomenon. First take the rounding rule to be

$$I_2(x) = .01[100\ x + .5] \qquad (x \in \mathcal{I}),$$

where $[t]$ = the greatest integer not exceeding t. This shifts each of the 90 ambiguous numbers upward. Let $f(t) = t + .005$. Then for $x = I_2(x) \in \mathcal{I}$ we have

$$f_2(x) = x + .01, \quad g_2(x) = x, \quad h_2(x) = x + .01.$$

This example exhibits the so-called drift phenomenon. Drift can be eliminated in the present case by changing the rounding rule to eliminate the bias toward $+\infty$. One such rounding rule is "round to even": $I_2(.105) = .10$, $I_2(.115) = .12$, $I_2(.125) = .12$, etc. Details and further references can be found in [2].

Our second example is more extreme in that drift occurs regardless of how one rounds in the ambiguous cases. Let ϕ be a function of period .01 which satisfies $\phi(0) = \phi(.01) = -.001$, $\phi(.009) = -.009$, and ϕ is linear on each of the intervals (0, .009) and (.009, .01). Let $f(t) = t + \phi(t)$ for $t \in \mathcal{I} = (.1, 1)$. For $x = I_2(x) \in \mathcal{I}$ we have

$$f_2(x) = x, \quad g_2(x) = x + .01, \quad h(x) = x + .01.$$

These examples suggest that a convexity condition is needed to obtain positive results. It is easy to see that the relation (1A) still need not be true. For example, let $f(t) = t^2$ for $0 < t < \infty$ and $x = .34$. Thus $f_2(.34) = .12$, $g_2(.12) = .35$. We shall show later that the relations (1B)–(1D) can also fail for a convex monotonic function f. However, stability is achieved for further iterates as we now show.

2. Statement of Results.

THEOREM. *Let $\mathcal{I}$ denote a real interval and f a real valued monotone function which is strictly convex or strictly concave on $\mathcal{I}$. Let R denote a radix and N a positive integer such that $R^N \geqslant 3$, and let I_N, $h^{(k)}$ and $\mathcal{I}^{*(k)}$ $(k = 1, 2, \ldots)$ be as defined above. Then* (1E) *holds for all $x \in \mathcal{I}^{*(3)}$.*

Remarks. A. The result is independent of the precise rounding rule. Indeed it is all right to round the ambiguous numbers capriciously.

B. For simplicity in stating the theorem we have used the same radix R and number of significant figures N for both the domain and range of the function f. Actually, this is unnecessary for the proof, and so our result could be restated with radices R and R' and round off to N and N' digits. We need only assume $R^N \geqslant 3$ and $R'^{N'} \geqslant 3$. See [3] for discussion of a related linear problem.

C. For some special functions a stronger theorem holds. For example, one can show that if $f(x) = 1/x$ and the radix equals 10, then (1B) holds generally and (1A) holds for each interval I of the form $[10^k, 10^{k+\frac{1}{2}}]$, where k is any integer and N is any positive integer.

3. Proof of the Theorem. We shall use the following notation throughout. Let $x = I_N(x) \in I^{*(3)}$, $y = g \circ f_N(x)$ (so that $f(y) = f_N(x)$), and $z = I_N(y) = h(x)$. We can assume that x, y and z are all distinct, for otherwise the theorem holds trivially.

We shall say that $I_N(0) = 0$. Of course we have $I_N(a) \neq 0$ if $a \neq 0$. If $f(a) = 0$ for some $a = I_N(a) \in I$, then $f_N(a) = 0$, $g \circ f_N(a) = a$, and $h(a) = a$. In this case the theorem holds. Similar remarks apply if $g(b) = 0$ for some $b = I_N(b) \in f(I)$. Thus, we can and shall exclude any occurrence of 0 except as a possible value of x in case $0 \in I^{*(3)}$.

We begin by considering the possible orderings of x, y, z. Since $z = I_N(y)$, y cannot be closer to $x = I_N(x)$ than it is to z. In other words,

$$|y - z| \leqslant |x - y|; \tag{2}$$

and the configurations $z < x < y$ and $z > x > y$ are impossible.

Now we have the simple

LEMMA. *Let f be monotone, $x = I_N(x)$, $y = g \circ f_N(x)$, and $z = I_N(y)$. Assume that z lies between x and y. Then $f_N(x) = f_N(z)$.*

Proof of the Lemma. Since f is monotone, $f(z)$ lies between $f(x)$ and $f(y)$. Since $f(x)$ rounds off to $f_N(x) = f(y)$, so does $f(z)$.

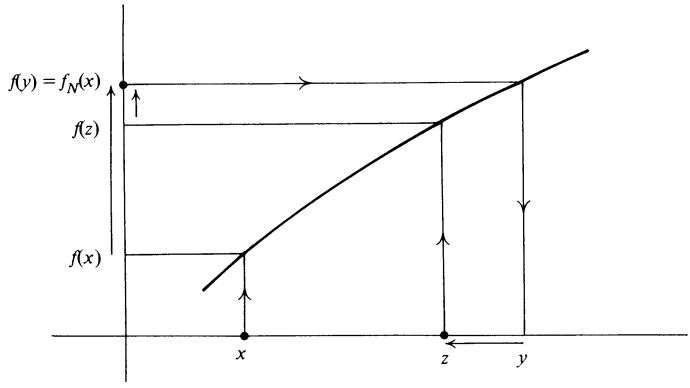

FIGURE 1

Thus, if we have $x < z < y$ or $x > z > y$, then $f_N(x) = f_N(z) = f_N \circ h(x)$. Applying $h \circ g_N$ to each side of this equation, we obtain $h^{(2)}(x) = h^{(3)}(x)$; i.e. the theorem holds in this case. (When in subsequent arguments we obtain the formula $f_N(x) = f_N(z)$, we shall simply say that the theorem is proved in that case.)

We turn our attention to the remaining possible arrangements $x < y < z$ and $x > y > z$. There are two main cases for us to consider:

$$\left|\frac{f(z) - f(y)}{z - y}\right| < \left|\frac{f(y) - f(x)}{y - x}\right| \tag{3}$$

or

$$\left|\frac{f(z)-f(y)}{z-y}\right| > \left|\frac{f(y)-f(x)}{y-x}\right| . \tag{4}$$

We remark that (3) holds if

$$x<y<z,\ f\uparrow,\ f \text{ strictly concave, or}$$
$$x<y<z,\ f\downarrow,\ f \text{ strictly convex, or}$$
$$z<y<x,\ f\uparrow,\ f \text{ strictly convex, or}$$
$$z<y<x,\ f\downarrow,\ f \text{ strictly concave,}$$

and (4) holds in the remaining four cases. It is not necessary to treat all these cases individually.

We assume first that (3) holds. Later we shall treat (4) by transforming it into (3). We have by (2) and (3) that

$$|f(z)-f(y)|<|f(y)-f(x)|, \tag{5}$$

i.e.

$$|f(z)-f_N(x)|<|f_N(x)-f(x)|.$$

This inequality implies that $f_N(z)=f_N(x)$ (and hence the theorem holds) provided that either

$$|f_N(x)|\neq R^k \quad \text{for any integer } k \tag{6}$$

or

$$|f_N(x)|=R^k \quad \text{for some } k \text{ and } |f(x)|<|f_N(x)|<|f(z)|. \tag{7}$$

Now we shall prove the theorem in case (3) holds and $|f(x)|>|f_N(x)|=R^k>|f(z)|$. We need a bit more notation. Let $\theta=f_N(z)$ and $w=g(\theta)$, i.e. $f(w)=f_N(z)$.

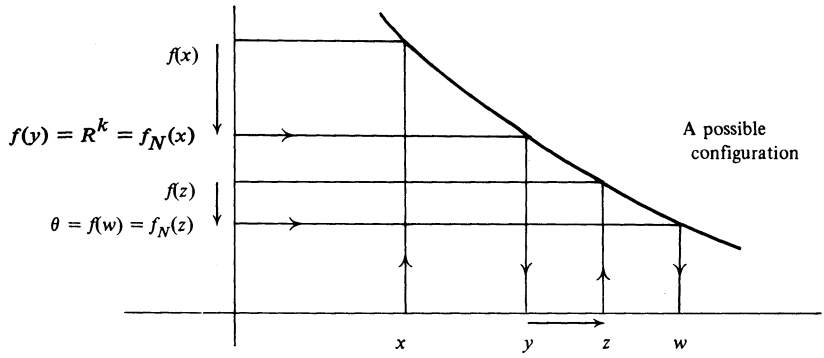

FIGURE 2

We claim that $f(x)$, $f(y)$, $f(z)$, and $f(w)$ all have the same sign. sgn $f(x)=$ sgn $f(y)$ since $f(y)=I_N(f(x))$. By (5) and the round off rules we have

$$|f(z)-f(y)|<|f(y)-f(x)|\leqslant \tfrac{1}{2}\cdot R^{k+1-N}. \tag{8}$$

Since $|f(y)|=R^k$, it follows that sgn $f(z)=$ sgn $f(y)$. For $f(w)=f_N(z)$ we use the

fact that $|f(z)| < R^k$ and the round off rules to conclude that $|f(w) - f(z)| \leqslant \frac{1}{2} \cdot R^{k-N}$. Thus

$$|f(w) - f(y)| \leqslant |f(w) - f(z)| + |f(z) - f(y)| < R^{k-N}(R + 1)/2; \tag{9}$$

and, as in the case of $f(z)$, we deduce that $f(w)$ and $f(y)$ have the same sign.

Next we give an upper bound for $|f(w)|$. We have $|f(w) - f(z)| \leqslant |f(z) - f(y)|$, since $f(z)$ rounds off to $f(w)$ rather than to $f(y)$. Also, we have $|f(y) - f(z)| = |f(y)| - |f(z)|$ since $|f(y)| > |f(z)|$ and $\operatorname{sgn} f(y) = \operatorname{sgn} f(z)$. Thus

$$|f(w)| \leqslant |f(z)| + |f(w) - f(z)| \leqslant |f(z)| + |f(y) - f(z)| = |f(y)|.$$

We conclude this part of the argument differently according to the relative size of $|f(w)|$ and $|f(z)|$. First, suppose $|f(w)| \geqslant |f(z)|$. We have $f(w)$ included between $f(z)$ and $f(y)$; and we can apply the lemma with $f(y)$, $f(z)$, $f(w)$, and g in place of x, y, z, and f, respectively. We deduce that $I_N(w) = I_N(y) = z$, i.e. $h^{(2)}(x) = h(x)$, and the theorem is proved in this case.

Now suppose that $|f(w)| < |f(z)| < |f(y)| = R^k < |f(x)|$. We set $t = I_N(w)$ and repeat the foregoing analysis with x, y, and z replaced by z, w, and t, respectively. We show that the process must now terminate by the time we reach (6).

In case t lies between z and w, then by the lemma $f_N(z) = f_N(t)$, i.e.

$$f_N \circ h(x) = f_N \circ h^{(2)}(x); \tag{10}$$

and the theorem is proved in this case.

Next suppose that t does not lie between z and w. Then as in (2), z cannot be between t and w, so w lies between z and t. Also, $f(x)$, $f(y)$, $f(z)$, and $f(w)$ lie in linear order, since their absolute values do and these numbers are all of one sign. It follows by monotonicity that x, y, z, and w and hence x, y, z, w, and t lie in linear order.

The assumed inequality (3) implies that f is steeper near x than near z. It follows that f is steeper near z than near t. Thus (3) holds with x, y and z replaced by z, w, and t, respectively.

The analogue of (5) now holds:

$$|f(t) - f_N(z)| < |f_N(z) - f(z)|.$$

It follows that $f_N(t) = f_N(z)$, i.e. (10), and hence the theorem is established, provided that

$$|f(w)| = |f_N(z)| \neq R^l \quad \text{for any integer } l. \tag{6'}$$

Now (9) and the fact that $R^N \geqslant 3$ imply that

$$|f(w)| \geqslant |f(y)| - |f(y) - f(w)| > R^k\left(1 - \frac{R+1}{2R^N}\right) \geqslant R^{k-1}.$$

Thus $R^{k-1} < |f(w)| < R^k$, and hence (6') holds. This concludes that the proof of the theorem in case f satisfies (3).

Now consider the other main case, in which (4) is assumed to hold. That is, the graph of f gets steeper as we move from x toward z.

Since y lies between x and z, $f(y)$ lies between $f(x)$ and $f(z)$ by monotonicity. By the round off of $f(z)$ we have

$$|f(w) - f(z)| = |f_N(z) - f(z)| \leqslant |f(z) - f(y)|.$$

Thus $f(y)$ lies between $f(x)$ and $f(w)$, so y lies between x and w.

It follows from (4) and the above that f has a steeper secant line over the interval between w and z than over the interval between y and z, i.e.

$$\left|\frac{f(w) - f(z)}{w - z}\right| > \left|\frac{f(z) - f(y)}{z - y}\right|.$$

It is convenient here to call $f(y) = \zeta$, $f(z) = \eta$, and $f(w) = \theta$. The preceding inequality can be rewritten as

$$\left|\frac{g(\theta) - g(\eta)}{\theta - \eta}\right| < \left|\frac{g(\eta) - g(\zeta)}{\eta - \zeta}\right|.$$

Now we have the situation of (3) with ζ, η, θ, and g in place of x, y, z, and f. It follows from the preceding argument that

$$g_N \circ f_N \circ g_N \circ f_N \circ g_N(\zeta) = g_N \circ f_N \circ g_N(\zeta)$$

or since $\zeta = f(y) = f_N(x)$, we have $h^{(3)}(x) = h^{(2)}(x)$.

The proof of the theorem is now complete in case f satisfies either of the inequalities (3) or (4).

4. An Example and a Question. We give an example of a monotone concave function for which Eqs. (1A)–(1D) fail to hold. Let $R = 10$, $N = 2$, $f(t) = 115 - 35/(t - 97)$ on $I = (98, \infty)$, and $g(t) = f^{-1}(t) = 97 + 35/(115 - t)$ for $t < 115$. We have $f_2(110) = 110$, $g_2(110) = 100$, $f_2(100) = 100$, $g_2(100) = 99$, $f_2(99) = 97$ or 98 (according to the round off rule used). Of course $g_2(97) = g_2(98) = 99$ as predicted by the theorem.

We close by posing a related problem which may have some more practical importance. Suppose one allows a "reasonable" calculational error (cf. (1)), as well as round off, for each determination of a functional value. Under what conditions does an analogue of our theorem hold?

Acknowledgement. I am indebted to the referee for correcting some errors, pressing for classification of obscurities, and for suggesting several extensions of my original version. In particular, the preceding example was found by the referee.

Department of Mathematics
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801

1. C. T. FIKE, *Computer Evaluation of Mathematical Functions*, Prentice-Hall, Englewood Cliffs, N. J., 1968.

2. JOHN F. REISER & DONALD E. KNUTH, "Evading the drift in floating point addition," *Information Processing Lett.*, v. 3, 1975, pp. 84–87.

3. DAVID W. MATULA, "The base conversion theorem," *Proc. Amer. Math. Soc.*, v. 19, 1968, pp. 716–723.