Note on the Round-Off Errors in Iterative Processes

By J. Descloux

Summary. This paper discusses round-off errors in iterative processes for solving equations. Let \( x_{n+1} = x_n + F(x_n) \) be a scalar iterative converging process; the different values \( x_n \) are represented in a computer with a certain precision; when \( x_n \) is close to the limit, \( F(x_n) \) is small and can perhaps be obtained easily with a higher absolute precision than \( x_n \); consequently, the addition \( x_n + F(x_n) \) will practically involve a rounding operation. Besides some general remarks, it will be shown that for a fixed-point computer an appropriate rounding method can provide a more accurate solution to the problem; analogous results are given in Appendix I for a floating-point computer; Appendix II deals with Aitken's \( \delta^2 \) process. The author is indebted to A. H. Taub for many suggestions and stimulating discussions.

1. Introduction. Let \( G^{(1)}, \ldots, G^{(m)} \) be \( m \) real functions of the real variables \( x^{(1)}, \ldots, x^{(m)} \). For any set of \( m \) numbers \( p^{(1)}, \ldots, p^{(m)} \), we shall use the vectorial notations:

\[
\mathbf{p} = (p^{(1)}, \ldots, p^{(m)});
\]

\[
|\mathbf{p}| = \sqrt{(p^{(1)})^2 + \cdots + (p^{(m)})^2}.
\]

We consider the iterative process

\[
(1) \quad x_{n+1} = G(x_n), \quad n = 0, 1, \ldots
\]

and suppose there exist a vector \( \mathbf{r} \) and a number \( b \) \((0 < b < 1)\) such that

\[
(2) \quad |G(x) - \mathbf{r}| \leq b |x - \mathbf{r}| \quad \text{for all} \quad x;
\]

the condition (2) insures the convergence of the \( x_n \)'s to \( \mathbf{r} \).

We want to realize the process (1) on a fixed-point computer under the two conditions: a) For representing each of the \( x_n^{(i)} \), we use only one “word”; we consider the content of the word as an integer; b) We may use higher precision for computing the values of the functions \( G^{(1)}, \ldots, G^{(m)} \) (or the functions \( G^{(1)}(x^{(1)}), G^{(2)}(x^{(2)}), \ldots, G^{(m)}(x^{(m)}) \)).

We distinguish two types of errors:

1) Truncation errors; even when using double precision, we cannot expect to evaluate the functions \( G^{(i)} \) exactly;
2) Round-off errors; according to condition a), the value found for \( G^{(i)} \) must be rounded to an integer.

Received June 1, 1962. This work was supported in part by the National Science Foundation.
2. Truncation Errors. Let \( H^{(1)}(x), \ldots, H^{(m)}(x) \) approximate the functions \( G^{(1)}(x), \ldots, G^{(m)}(x) \):

\[
H^{(i)}(x) = G^{(i)}(x) + \xi^{(i)}(x);
\]

\( \xi^{(i)}(x) \) is called the truncation error; it is supposed to satisfy the inequality

\[
| \xi^{(i)}(x) | \leq a^{(i)}; \quad a^{(i)} = \text{constant}.
\]

The iterative process

\[
V_{n+1} = H(V_n)
\]

is considered as an approximation of (1) and gives some information about \( r \).

**Theorem 1.** For any \( V_0 \), the sequence \( V_n \) given by (4) is bounded and all its points of accumulation \( V \) satisfy the inequality

\[
| V - r | \leq \frac{|a|}{1 - b}; \quad a = (a^{(1)}, \ldots, a^{(m)}).
\]

**Theorem 2.** The process (4) is the best possible in the following sense: for given \( a \) and \( b \), there exist \( m \) functions \( H^{(1)}(x), \ldots, H^{(m)}(x) \) for which it is impossible to find an algorithm using only \( H, a, b \), providing closer points of accumulation to \( r \) than the algorithm (4).

**Proof:** Let \( G(x) = bx + a \),

\[
H(x) = bx, \\
G'(x) = bx - a.
\]

\( H(x) \) is an approximation for both \( G(x) \) and \( G'(x) \) with limits \( r = \frac{a}{1 - b} \) and

\[
r' = \frac{-a}{1 - b}.
\]

If any sequence \( W_n \) has a point of accumulation \( W \) such that

\[
| W - r | < \frac{|a|}{1 - b},
\]

then by the triangular inequality,

\[
| W - r' | > \frac{|a|}{1 - b}
\]

and the process (4) provides in this case better information.

3. Round-off Errors. For the computer, the process (1) can be written in the form

\[
y_{n+1}^{(i)} = [G^{(i)}(y_n) + \xi^{(i)}_n]_\pi;
\]

\( y_n^{(i)} \) is an integer. \( [\ ]_\pi \) is called a rounding procedure. \( [x]_\pi \) is any integer-valued function of \( x \) satisfying the inequality:

\[
| [x]_\pi - x | < 1.
\]
We consider two particular types of rounding procedures:

1) Normal rounding: \([x]_{\text{n}} = [x + 0.5];\)

2) Anomalous rounding: \([x]_{\text{A}}: \text{ for } |x| \leq 1, \quad [x]_{\text{A}} \geq |x|;\)

\[\text{for } |x| \geq 1, \quad [x]_{\text{A}} \leq |x|.\]

**Theorem 3.** Let \(G\) and \(\xi\) satisfy equations (2) and (3). If

\[(6) \quad y_{n+1}^{(i)} = [G^{(i)}(y_n) + \xi_n]_{\text{n}}, \quad i = 1, 2 \cdots m,\]

then for any \(y_0\), there exists \(N\) such that

\[|y_n - r| \leq \frac{|a|}{1 - b} + \frac{\sqrt{m}}{2(1 - b)} \quad \text{for} \quad n > N;\]

furthermore, for given \(a\) and \(b\), there exists a function \(G\) and errors \(\xi_n\) for which the bound is attained.

Now, we restrict ourselves to the particular case \(m = 1\); i.e., the process (1) becomes scalar. Equations (1), (2), (3), and (5) can be written as:

\[(7) \quad x_{n+1} = G(x_n);\]

\[(8) \quad |G(x) - r| \leq b |x - r|;\]

\[(9) \quad y_{n+1} = [G(y_n) + \xi_n]_{\text{A}};\]

\[(10) \quad |\xi_n| \leq a;\]

**Theorem 4.** Let \(G(x)\) and \(\xi\) satisfy equations (8) and (10). If

\[(11) \quad y_{n+1} = y_n + [G(y_n) + \xi_n - y_n]_{\text{A}},\]

then for any \(y_0\), there exists \(N\) such that

\[|y_{n+1} - r| < \frac{a}{1 - b} + 1 \quad \text{for} \quad n > N.\]

Let us compare Theorem 4 with Theorem 3 for \(m = 1\). In both cases, the bounds of errors have a common part which can be recognized from Theorems 1 and 2 as provided by the truncation errors. The part due to the round-off errors is independent of \(b\) for the anomalous rounding; in particular, if \(a = 0\), the error is less than 1 and if the limit \(r\) is an integer, it is reached after a finite number of steps. When the convergence is slow, i.e., \(\beta \sim 1\), the errors can be very large for the normal rounding, even if \(a = 0\); however, if \(b < 0.5\), the normal rounding provides slightly better results than the anomalous rounding.

**Remark.** The condition (2) insures a first-order convergence for the process (1). If we assume higher convergence, i.e., if

\[|G(x) - r| \leq b |x - r|^p, \quad p > 1,\]

we get results which are quite similar, but generally not simple to formulate. Rather roughly, Theorem 4 becomes: if \(y_n\) is computed by (11), then

\[|y_{n+1} - r| < B + 1 \quad \text{for} \quad n > N,\]

where \(B\) is due to the truncation error.
4. Proofs.

**Lemma.** Let \( V_1 = G(V_0) + \xi_0 \) under assumptions (2) and (3);

a) If \( |V_0 - r| \leq \frac{|a|}{1 - b} \), then \( |V_1 - r| \leq \frac{|a|}{1 - b} \);

b) If \( |V_0 - r| > \frac{|a|}{1 - b} \), then \( |V_1 - r| < |V_0 - r| \).

**Proof.** Since \( V_1 = G(V_0) + \xi_0 \):

\[
|V_1 - r| \leq |G(V_0) - r| + |\xi_0| \leq b |V_0 - r| + |a|; 
\]

a) \( |V_0 - r| \leq \frac{|a|}{1 - b} \); we have by (12):

\[
|V_1 - r| \leq |a| \left( \frac{b}{1 - b} + 1 \right) = \frac{|a|}{1 - b}, \text{ q.e.d.}
\]

b) \( |V_0 - r| > \frac{|a|}{1 - b} \); we have by (12):

\[
|V_1 - r| \leq |V_0 - r| - (1 - b) |V_0 - r| + |a| < |V_0 - r| - |a| + |a| = |V_0 - r|, \text{ q.e.d.}
\]

**Proof of Theorem 1.** First case: There is \( N \) such that \( |V_N - r| \leq \frac{|a|}{1 - b} \); by Lemma a, the same inequality holds for all \( n > N \) and the theorem is proved.

Second case: For all \( n = 0, 1, 2, \ldots \) : \( |V_n - r| > \frac{|a|}{1 - b} \); by Lemma b, the positive sequence \( |V_n - r| \) is monotone decreasing and converges therefore to a limit of 1.

Suppose that \( l = \frac{|a|}{1 - b} + d \) where \( d > 0 \); since \( b < 1 \), there exists \( V_n \) such that \( |V_n - r| < \frac{|a|}{1 - b} + \frac{d}{b} \); by (12):

\[
|V_{n+1} - r| < \frac{b}{1 - b} |a| + d + |a| = \frac{|a|}{1 - b} + d = l,
\]

which is a contradiction.

**Proof of Theorem 3.** Since \( |[x]_n - x| \leq 0.5 \), we can write the equation (6) in the form

\[
y_{n+1}^{(i)} = G^{(i)}(y_n) + \eta_n^{(i)}, \quad i = 1, 2, \ldots m
\]

where

\[
|\eta_n^{(i)}| \leq a^{(i)} + 0.5,
\]

and therefore

\[
|n_n| \leq |a| + 0.5 \sqrt{m}.
\]
Replacing \( \xi \) by \( n \), and \( |a| \) by \( |a| + 0.5 \sqrt{m} \), we can apply Theorem 1: for any \( \epsilon \), there exists \( N \) such that

\[
|y_n - r| < \frac{|a| + 0.5 \sqrt{m}}{1 - b} + \epsilon \quad \text{for} \quad n > N;
\]

but since the \( y_n^{(i)} \)'s are integers, there exists a particular \( \epsilon \) for which the preceding inequality implies

\[
|y_n - r| \leq \frac{|a| + 0.5 \sqrt{m}}{1 - b} \quad \text{for} \quad n > N,
\]
as desired. We have still to show an example valid for every \( a \) and \( b \) where the bound of error is attained. Let

\[
G(i,(x) = bx^{(i)} - a^{(i)} - 0.5
\]

and suppose that for the particular vector \( y_0 = 0 \) we have \( \xi_0 = a \). Then

\[
y_n = 0 \quad \text{and} \quad |y_n - r| = \frac{|a| + \sqrt{m} \cdot 0.5}{1 - b} \quad \text{for} \quad n \geq 0.
\]

Proof of Theorem 4. We use the two simple properties of the anomalous rounding procedures:

1) \( x - 1 < [x]_A < x + 1; \)
2) If \( p < x < q \) and \( q - p > 1 \), then
   \[
p < q + [x - q]_A < q, \quad \text{provided that} \quad p \text{is an integer, and}
   \]
   \[
p < q + [x - q]_A < q, \quad \text{provided that} \quad q \text{is an integer.}
   \]

Since the \( y_n \)'s are integers, the theorem results from the three statements:

I If \( |y_0 - r| \leq \frac{a}{1 - b} \), then \( |y_1 - r| < \frac{a}{1 - b} + 1; \)

II If \( \frac{a}{1 - b} < |y_0 - r| < \frac{a}{1 - b} + 1 \), then \( |y_1 - r| < \frac{a}{1 - b} + 1; \)

III If \( |y_0 - r| \geq \frac{a}{1 - b} + 1 \), then \( |y_1 - r| < |y_0 - r| \).

Statement I: By Lemma a:

\[
r - \frac{a}{1 - b} \leq y_0 + G(y_0) + \xi_0 - y_0 \leq r + \frac{a}{1 - b};
\]

By property 1:

\[
r - \frac{a}{1 - b} - 1 < y_0 + [G(y_0) + \xi_0 - y_0]_A < r + \frac{a}{1 - b} + 1; \quad \text{i.e.,}
\]

\[
|y_1 - r| < r + \frac{a}{1 - b} + 1, \quad \text{q.e.d.}
\]

Statement II: We suppose \( r + \frac{a}{1 - b} < y_0 < r + \frac{a}{1 - b} + 1 \) (the proof is
analogous, when

\[ r - \frac{a}{1 - b} - 1 < y_0 < r - \frac{a}{1 - b} \]; by Lemma b:

\[ p = r - \frac{a}{1 - b} - 1 < y_0 + G(y_0) + \xi_0 - y_0 < y_0 = q; \] since

\[ y_0 > r, \quad q - p > 1 \] and we apply property 2:

\[ r - \frac{a}{1 - b} - 1 < y_0 + [G(y_0 + \xi_0 - y_0)] < y_0 < r + \frac{a}{1 - b} + 1; \] i.e.,

\[ |y_1 - r| < r + \frac{a}{1 - b} + 1, \] q.e.d.

Statement III: We suppose \( y_0 \leq r + \frac{1}{1 - b} + 1 \) (the proof is analogous when

\[ y_0 \leq r - \frac{a}{1 - b} - 1 \]; by Lemma b:

\[ p = 2r - y_0 < y_0 + G(y_0) + \xi_0 - y_0 = q; \] by property 2, since \( q - p > 1 \):

\[ 2r - y_0 < y_0 + [G(y_0) + \xi_0 - y_0] < y_0; \] i.e.,

\[ |y_1 - r| < |y_0 - r|, \] q.e.d.

APPENDIX I: Iterative Processes with a Floating-Point Computer*

Let \( r \) be a real number and \( G(x) \) be a function such that

(1) \[ |x + G(x) - r| \leq b|x - r| \] with \( 0 \leq b < 1 \) for any \( x; \)

then the sequence

(2) \[ x_{n+1} = x_n + G(x_n) \]

converges at least linearly to \( r \) for any \( x_0 \).

Suppose we want to realize (2) on a binary floating-point computer, i.e., the numbers are of the form \( \alpha \cdot 2^\beta \), where \( \alpha \) is an exact binary fraction and \( \beta \) is an integer.

A number will be called normalized if 1) \( 0.5 \leq |\alpha| < 1; \) 2) \( \alpha \) is an exact binary fraction representable by \( N \) bits and the sign; 3) \( \beta \geq -p \) (\( N \) and \( p \) are fixed numbers); furthermore there exists a real zero, representable for example by \( \alpha = 0, \beta = -p \); for greater simplicity, this zero will also be included in the class of normalized numbers.

We assume that in the realization of (2) on the computer, both \( x_n \) and \( G(x_n) \) are represented by normalized numbers; of course \( G(x) \) cannot be computed exactly in general; so we assume that value effectively computed, \( \hat{G}(x) \), satisfies the relation:

(3) \[ \hat{G}(x) = (1 + \eta)G(x) + \xi; \quad |\eta| \leq d, \quad |\xi| \leq a; \]

* A detailed discussion of the results of this appendix will be found in reference [4].
where \( \eta \) and \( \xi \) are functions of \( x \), but \( d \) and \( a \) are fixed numbers.

The effective process is given by the operation

\[
Y_{n+1} = [Y_n + \bar{G}(Y_n)]_R
\]

where \( Y_n \) and \( Y_{n+1} \) are normalized numbers; since \( Y_n + \bar{G}(Y_n) \) cannot be generally represented by a normalized number, it must be rounded as indicated by \([ \ ]_R\).

We concentrate our attention on the rounding procedure in (4) and consider two types of rounding procedures:

1) Normal rounding. \( Y_{n+1} = [Y_n + \bar{G}(Y_n)]_R \); \( Y_{n+1} \) is a normalized number such that

\[
| Y_{n+1} - (Y_n + \bar{G}(Y_n)) | = \text{minimum};
\]

when two different normalized numbers satisfy the above relation, either of them can be chosen as \( Y_{n+1} \).

2) Anomalous rounding. \( Y_{n+1} = [Y_n + \bar{G}(Y_n)]_A \); if \( \bar{G}(Y_n) \geq 0 \) let

\[
Z \text{ be the smallest normalized number such that } Z \geq Y_n + \bar{G}(Y_n),
\]

\[
W \text{ be the greatest normalized number such that } W \leq Y_n + \bar{G}(Y_n);
\]

if \( \bar{G}(Y_n) \leq 0 \) let

\[
Z \text{ be the greatest normalized number such that } Z \leq Y_n + \bar{G}(Y_n),
\]

\[
W \text{ be the smallest normalized number such that } W \geq Y_n + \bar{G}(Y_n);
\]

then

\[
[Y_n + \bar{G}(Y_n)]_A = W \quad \text{if } W \neq Y_n
\]

\[
[Y_n + \bar{G}(Y_n)]_A = Z \quad \text{if } W = Y_n.
\]

**Theorem.** a) For any \( Y_0 \), by using normal rounding in (4), there exists a finite number \( M \) such that

\[
| Y_n - r | \leq B_R = \frac{2^{-n} \mid r \mid + a(1 + 2^{-n})}{2 + 2^{-n} - (1 + d)(1 + b)(1 + 2^{-n})} \quad \text{for } n > M.
\]

b) For any \( Y_0 \), by using anomalous rounding in (4), there exists a finite number \( M \) such that

\[
| Y_n - r | < B_A = | r | 2^{-n+1} + 2^{-p-1} + \frac{a(1 + 2^{-n+1})}{2 - (1 + d)(1 + b)} \quad \text{for } n > M.
\]

If \( B_R \) or \( B_A \) is negative, it must be replaced by \(+\infty\).

In order to compare these results, first suppose \( a = 0 \). Then \( B_A \) is independent of \( b \) and \( d \) and furthermore remains very small; in case of slow convergence, i.e., when \( b \approx 1 \), \( B_R \) can become very large. The increase of magnitude of the bounds when \( a > 0 \) is almost the same for \( B_A \) and \( B_R \) for reasonable cases, so that the anomalous rounding can be considered safer than the normal rounding.

**Remarks.** 1) The relations of normal and anomalous rounding procedures are very similar in fixed-point and in floating-point arithmetics;

2) The bounds \( B_A \) and \( B_R \) are reached only in trivial cases; however, examples show that they remain realistic in every case.
APPENDIX II: Round-off Errors in Aitken's $\delta^2$ Process*

Let $G(x)$ be a real continuous function of the real variable $x$ such that the sequence $x_n$ defined by

\[ x_{n+1} = G(x_n) \]

converges to the limit $x = r$.

By Aitken's $\delta^2$ process, we define another sequence:

\[ \begin{align*}
V_{3n+1} &= G(V_{3n}) \\
V_{3n+2} &= G(V_{3n+1}) \\
V_{3n+2} &= \frac{V_{3n}V_{3n+2} - V_{3n+1}^2}{V_{3n} + V_{3n+2} - 2V_{3n+1}}
\end{align*} \tag{2} \]

Let us suppose we want to realize process (2) on a fixed-point computer with the following conditions: a) We use only one "word" for representing the $V_i$'s; we may consider the content of the word as an integer; b) We may use higher precision for computing $G(V_i)$.

We cannot expect to compute $G(V_i)$ without error; furthermore, if we are using higher precision, the result must be rounded to an integer.

Definition. A rounding procedure denoted by $[x]_R$ is any integer-valued function of the real variable $x$ satisfying the inequality:

\[ | [x]_R - x | < 1. \]

We shall use the following particular rounding procedures:

1) $[x]^\omega$: rounding away from zero; it is defined by the inequality

\[ | [x]^\omega | \geq | x | ; \]

2) $[x]^\omega$: rounding toward zero; it is defined by the inequality

\[ | [x]^\omega | \leq | x | . \]

Example. Let $G(x) = 7/8 \ x$ and $V_0 = 8$; by (2), we have

\[ \begin{align*}
V_1 &= 7 \\
V_2 &= 6,125 \\
V_3 &= 0.
\end{align*} \]

If we want to represent the $V_i$'s only by integers and if we use the normal rounding procedure, we shall find:

\[ \begin{align*}
\overline{V}_1 &= 7 \\
\overline{V}_2 &= 6 \\
\overline{V}_3 &= \infty.
\end{align*} \]

* For the proof see reference [3], part II.
This situation can be improved by using the following integer process:

\[
\begin{align*}
W_{3n+1} &= W_{3n} + [G(W_{3n}) + \xi_{3n} - W_{3n}]^x \\
W_{3n+2} &= W_{3n} + [G(W_{3n+1}) + \xi_{3n+1} - W_{3n}]^x \\
W_{3n+3} &= W_{3n} + \left( \frac{(W_{3n} - W_{3n+1})^2}{2W_{3n+1} - W_{3n} - W_{3n+2}} \right)^x.
\end{align*}
\]  

(3)

\(\xi_{3n}\) and \(\xi_{3n+1}\) are the errors of computation of \(G(W_{3n})\) and \(G(W_{3n+1})\); since the numerator and the denominator are integers, it is possible with the help of the remainder to compute \(W_{3n+3}\) without any error; if the numerator and the denominator are simultaneously equal to zero, then \(W_{3n} = W_{3n+1} = W_{3n+2}\) and we set \(W_{3n+3} = W_{3n}\).

**Theorem 1.** We suppose there exist numbers \(0 \leq b < 1\), \(0 \leq c < 1\), \(\delta \geq 0\) such that:

1) \(\lvert x_1 - r \rvert \leq b \lvert x_0 - r \rvert\)

for any \(x_0\) and \(x_1\) satisfying the relation (1);

2) \(\lvert V_2 - r \rvert \leq c \lvert V_0 - r \rvert\)

for any \(V_0\) and \(V_2\) satisfying the relations (2);

3) \(\lvert G(x) - G(y) \rvert \leq \delta \lvert x - y \rvert\)

for any \(x\) and \(y\);

4) the errors \(\xi_{3n}\) and \(\xi_{3n+1}\) in (3) satisfy the inequality

\[\lvert \xi_j \rvert \leq a \leq \frac{1}{4} \frac{(1 - b)^2(1 - c)}{(1 + c)(1 + \delta)};\]

then, for any \(W_0\) there exists a finite number \(N\) such that

\[\lvert W_{3n} - r \rvert < 1 + \frac{a}{1 - b} \text{ for } n > N.\]

**Theorem 2.** We make the assumptions:

1) The convergence of process (1) is alternating, i.e., for any \(x\)

\[0 \leq r - G(x) < x - r \text{ if } x - r > 0,\]

\[0 \leq G(x) - r < r - x \text{ if } x - r < 0.\]

\[G(x) = r \text{ if } x = r;\]

2) The errors \(\xi_{3n}\) and \(\xi_{3n+1}\) in (3) satisfy the inequality

\[\lvert \xi_j \rvert \leq a \leq \frac{1}{3},\]

where \(a\) is a fixed number; then, for any \(W_0\) there exists a finite number \(N\) such that

\[\lvert W_{3n} - r \rvert \leq 1 + a \text{ for } n > N.\]

**Remark.** Assumption (1) of Theorem 2 is sufficient for providing the conver-
gence of the $V_n$'s satisfying the equations (2) for any $V_0$. It is easy to prove the inequality:

$$| V_{2n} - r | < \frac{| V_0 - r |}{3^n}.$$ 

University of Illinois
Urbana, Illinois