Noname manuscript No.
(will be inserted by the editor)

On recurrences converging to the wrong limit in finite precision and some new examples

Siegfried M. Rump

Received: date / Accepted: date

Abstract In 1989, Jean-Michel Muller gave a famous example of a recurrence where, for particular initial values, the true real iteration converges to a repellent fixed point, whereas finite precision arithmetic produces a different result, the attracting fixed point. We analyze recurrences in that spirit and remove a gap in previous arguments in the literature, that is, the recursion must be well defined. The latter is known as the Skolem problem. We identify initial values producing the limit equal to the repellent fixed point, show that in every ε-neighborhood of such initial values the recurrence is not well-defined, and characterize initial values for which the recurrence is well-defined.

We give some new examples in that spirit. For example, the correct real result, i.e., the repellent fixed point, may be correctly computed in bfloat, half, single, double, formerly extended precision (80 bit format), binary128 as well as many much higher precisions. Rounding errors may be beneficial by introducing some regularizing effect.

Keywords Recurrences · rounding errors · IEEE-754 · different precisions · bfloat · half precision (binary16) · single precision (binary32) · double precision (binary64) · extended precision (binary128) · multiple precision · Skolem problem · Pisot sequence

Mathematics Subject Classification 65G50 - 11B37

S.M. Rump
Institute for Reliable Computing,
Hamburg University of Technology,
Am Schwarzenberg-Campus 3, 21073 Hamburg, Germany,
and Visiting Professor at Waseda University,
Faculty of Science and Engineering,
3–4–1 Okubo, Shinjuku-ku, Tokyo 169–8555, Japan
E-mail: rump@tuhh.de

1 Introduction

At the ICIAM 2019 conference in Valencia the following famous recurrence was shown:

$$x_0 := 4,\ x_1 := 4.25 \quad \text{and} \quad x_{n+1} := 108 - (815 - 1500/x_{n-1})/x_n. \tag{1}$$

The true limit of this recurrence is $L = 5$, whereas in double precision (binary64) the computed limit is 100. That came as a surprise to the audience, so we decided to write this note giving the background and analysis of such recurrences.

The first example in that spirit is due to Muller [11]:

$$x_0 := 11/2,\ x_1 := 61/11 \quad \text{and} \quad x_{n+1} := 111 - (1130 - 3000/x_{n-1})/x_n. \tag{2}$$

The limit of the recurrence over the field of real numbers is $L = 6$, whereas in double precision the limit is 100. However, the initial value $x_1 := 61/11$ is not representable in binary in any precision, so that for the input data stored in the computer the limit 100 is correct.

The above example (1) was given by Kahan [9] together with an explanation on the behavior of the recurrence. In this example all input data including the initial values are representable in binary with at least 10 bits precision. In his book [12] Muller defines different initial values $x_0 := 2,\ x_1 := -4$ for his recurrence (2), also representable in binary with at least 10 bits precision, and the same behavior of the recurrence as before.

The examples have in common the attracting fixed point $L = 100$ together with a repellent fixed point $\beta = 5$ or $\beta = 6$, respectively. In Kahan's example $x_2 = \frac{76}{17}$, in Muller's first example $x_1 = \frac{61}{11}$, and in his second example $x_3 = \frac{347}{37}$. Those values have in common that they are not representable in binary, regardless of the precision. Replacing the initial value (x_{k-1}, x_k) by the computed value $(x_{k-1}, \widetilde{x}_k)$ for $k = 2, 1, 3$, respectively, it follows that the recurrence over $\mathbb{R}$, if it is well-defined, necessarily converges to the attracting fixed point $L = 100$.

An implicit assumption for that assertion is that $x_i \neq 0$ for all $i \in \mathbb{N}_0$, otherwise the recurrence is not well-defined. The problem, to identify the indices with an iterate equal to zero for a linear recurrence is known as the Skolem problem [14,6]. Instances of such problems are known to be NP-hard [2]. We will characterize the initial pairs (x_0, x_1) for which such recurrences are well-defined together with their limits.

We show that for every initial pair (x_0, x_1) with recurrence being well-defined and converging to a repellent fixed point and any ε-neighborhood of x_1 there exist x_1' in this neighborhood such that the recurrence over $\mathbb{R}$ starting with (x_0, x_1') produces $x_n = 0$ for some $n \in \mathbb{N}$.

Moreover, it is suggested in the literature that, due to the fact that some iterate is not representable in floating-point, the iteration must converge to the attracting rather than the repellent fixed point. That may not be true for the floating-point iteration due to "fortunate" rounding errors.

We give new explicit examples, starting with one where the correct value, i.e., the repellent fixed point, is produced in bfloat[1], half precision and single precision, but erroneous result in double precision (binary64) and extended precision (binary128). Other examples produce the correct limit, the repellent fixed point, in much higher precisions.

2 Analysis of recurrences

In the following we will take a closer look at recurrences of the type (1) or (2). We start the analysis in a more general setting and define

$$x_{n+1} := a + (b + c/x_{n-1})/x_n \quad \text{with} \quad a, b, c \in \mathbb{R} \tag{3}$$

for given initial values $(x_0, x_1) \in \mathbb{R}^2$. The auxiliary recurrence $y_{n+1} := x_n y_n$ for $0 \leqslant n \in \mathbb{N}$ and $y_0 := 1$ leads to

$$\frac{y_{n+2}}{y_{n+1}} = a + (b + \frac{cy_{n-1}}{y_n})\frac{y_n}{y_{n+1}}$$

provided that $x_n \neq 0$ for $0 \leqslant n \in \mathbb{N}$, so that

$$y_{n+2} = ay_{n+1} + by_n + cy_{n-1} \quad \text{for} \quad 1 \leqslant n \in \mathbb{N}. \tag{4}$$

Note that the linear recurrence (4) is always well defined. The characteristic polynomial is

$$\chi(y) = y^3 - ay^2 - by - c =: (y - \alpha)(y - \beta)(y - \gamma). \tag{5}$$

For simplicity we assume

$$|\alpha| > |\beta| > |\gamma| > 0 \quad \text{and} \quad \alpha, \beta, \gamma \in \mathbb{R}. \tag{6}$$

For all examples above and the new examples to be presented later that assumption is fulfilled. For given y_0, y_1, y_2 the recurrence (4) is characterized by a triple $(p, q, r) \in \mathbb{R}^3$ such that

$$y_n = \alpha^n p + \beta^n q + \gamma^n r \quad \text{for} \quad 0 \leqslant n \in \mathbb{N}. \tag{7}$$

The recurrence (3) is well-defined, i.e., $x_i \neq 0$ for all $i \in \mathbb{N}_0$, if and only if $y_i \neq 0$ for all $i \in \mathbb{N}$. The initial value y_0 is only a scaling leading to the same recurrence (3). Using $y_0 = p + q + r = 1$ we can rewrite (7) into

$$y_n = (\alpha^n - \gamma^n)p + (\beta^n - \gamma^n)q + \gamma^n \quad \text{for} \quad 0 \leqslant n \in \mathbb{N}. \tag{8}$$

Thus, the initial values (x_0, x_1) are coupled with (p, q) by the linear system

$$\begin{pmatrix} \alpha - \gamma & \beta - \gamma \\ \alpha^2 - \gamma^2 & \beta^2 - \gamma^2 \end{pmatrix} \begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} x_0 - \gamma \\ x_0 x_1 - \gamma^2 \end{pmatrix}. \tag{9}$$

[1] bfloat [15] uses 16 bits like half precision, but trades a larger exponent range against only 8 bit precision. It has been successfully used in deep learning and large scale networks [1,3], see also [13].

By assumption (6) the determinant $(\alpha-\beta)(\beta-\gamma)(\gamma-\alpha)$ is nonzero so that the linear system is solvable for all (x_0, x_1).

The intention of the mentioned examples is that for the specified initial values (x_0, x_1) the real recurrence (x_i) as in (3) converges to the repellent fixed point β, whereas any perturbation of (x_0, x_1) makes the recurrence (3) converge to the attracting fixed point α, the root of largest absolute value.

However, that includes the statement that the recurrence (3) is well-defined for the given initial values (x_0, x_1). Therefore we next characterize the pairs (x_0, x_1) for which that is true, i.e., the recurrence (3) is well-defined and converges to β.

Lemma 1 *Let $x_0, x_1 \in \mathbb{R}$ be given, and let the recurrence (3) with characteristic polynomial (5) satisfy (6). Then (3) is well-defined and $x_i \to \beta$ if, and only if*

$$x_0 \neq \gamma \quad \textit{and} \tag{10a}$$

$$x_1 = \beta + \gamma - \beta\gamma/x_0 \quad \textit{and} \tag{10b}$$

$$x_0 \neq \gamma - \frac{\gamma^n(\beta-\gamma)}{\beta^n - \gamma^n} \quad \textit{for all } n \geqslant 1. \tag{10c}$$

Remark 1 Note that the third condition implies $x_0 \neq 0$ for $n = 1$, and for $n = 2$ together with the second condition also $x_1 \neq 0$. Also note that for choosing $x_1 := \beta + \gamma - \beta\gamma/x_0$ the iteration, being well-defined or not, only depends on x_0.

Proof By (8),

$$\lim_{n\to\infty} x_n := \begin{cases} \alpha & \text{if, and only if, } p \neq 0, \\ \beta & \text{if, and only if, } p = 0,\ q \neq 0, \\ \gamma & \text{if, and only if, } p = q = 0, \end{cases} \tag{11}$$

where the last case is equivalent to $x_0 = x_1 = \gamma$. Since (9) determines p and q uniquely, (x_i) converges to β if, and only if, $p = 0$ and $q = \frac{x_0-\gamma}{\beta-\gamma} \neq 0$, which in turn is equivalent to

$$x_1 = \left((\beta^2-\gamma^2)q + \gamma^2\right)/x_0 = \beta + \gamma - \beta\gamma/x_0 \ \text{ and } \ x_0 \neq \gamma.$$

That means that, if (3) is well-defined, then x_i converges to β if, and only if, (10a) and (10b) are true. The recurrence (3) is well-defined if, and only if, $y_n \neq 0$ for all $n \geqslant 1$. If $p = 0$ and $q \neq 0$, that is by (8) equivalent to

$$-\gamma^n \neq (\beta^n - \gamma^n)q = \frac{(\beta^n-\gamma^n)(x_0-\gamma)}{\beta-\gamma} \quad \text{for all } n \geqslant 1.$$

Hence the recurrence (3) is well-defined if, and only if, (10c) is true. That finishes the proof. □

This shows that for (x_0, x_1) on the hyperbola H defined by $x_1 = \beta + \gamma - \beta\gamma/x_0$ the recurrence (x_i) is well-defined and converges to β except infinitely many discrete points. The accumulation point of those gaps, determined by condition (10c), is $x_0 = \gamma$, the repellent fixed point with smallest absolute value.

Next we show that the set of initial values (x_0, x_1) with $x_n = 0$ for some $n \in \mathbb{N}$, i.e., with not well-defined recurrence, form a hyperbola H_n, and the limit of those hyperbolas is H, the hyperbola of initial values for which, except for infinitely many discrete points, the recurrence converges to β.

Lemma 2 *Assume the recurrence (3) with characteristic polynomial (5) satisfies (6). For given $k \in \mathbb{N}$, denote by Z_k the set of initial values $(x_0, x_1) \in \mathbb{R}^2$ with $x_0 x_1 \neq 0, x_i \neq 0$ for $0 \leqslant i < k$, and $x_k = 0$.*

Then there exists some $k_0 \in \mathbb{N}$ such that for every $k \geqslant k_0$ the set Z_k forms a hyperbola H_k, and for $k \to \infty$ the hyperbolas H_k tend to the hyperbola H of initial pairs with limit point β of the recurrence, i.e., $x_1 = \beta + \gamma - \beta\gamma/x_0$. If (6) is sharpened into $\alpha > \beta > \gamma > 0$, then $k_0 = 2$.

Proof Set $n := k + 1$, let (x_0, x_1) be given with $x_0 x_1 \neq 0, x_i \neq 0$ for $0 \leqslant i < k$ and $x_{n-1} = x_k = 0$. Then $y_n = 0$ and $y_i \neq 0$ for $0 \leqslant i < n$, and (8) implies

$$M := \begin{pmatrix} \alpha - \gamma & \beta - \gamma & 0 \\ \alpha^n - \gamma^n & \beta^n - \gamma^n & 0 \\ \alpha^2 - \gamma^2 & \beta^2 - \gamma^2 & -x_0 \end{pmatrix} \begin{pmatrix} p \\ q \\ x_1 \end{pmatrix} = \begin{pmatrix} x_0 - \gamma \\ -\gamma^n \\ -\gamma^2 \end{pmatrix}. \tag{12}$$

The determinant of the matrix is zero if, and only if, $(\alpha - \gamma)(\beta^n - \gamma^n) = (\beta - \gamma)(\alpha^n - \gamma^n)$, i.e.,

$$f(\beta) = f(\alpha) \qquad \text{for} \quad f(x) := \frac{x^n - \gamma^n}{x - \gamma}.$$

Hence, if $\alpha > \beta > \gamma > 0$ and $x > \gamma$,

$$f(x) = x^{n-1} + x^{n-2}\gamma + \ldots + \gamma^{n-2}x + \gamma^{n-1}$$

is strictly increasing, showing that the determinant of the matrix in (12) is nonzero. Then the unique solution of the linear system is

$$x_1 = \frac{P - Q/x_0}{R} \quad \text{with} \quad \begin{cases} P = \alpha^n(\gamma^2 - \beta^2) + \beta^n(\alpha^2 - \gamma^2) + \gamma^n(\beta^2 - \alpha^2) \\ Q = \alpha^n\beta\gamma(\gamma - \beta) + \beta^n\alpha\gamma(\alpha - \gamma) + \gamma^n\alpha\beta(\beta - \alpha) \\ R = \alpha^n(\gamma - \beta) + \beta^n(\alpha - \gamma) + \gamma^n(\beta - \alpha) \end{cases}$$

forming the desired hyperbola H_k. Note that $R = -\det(M)/x_0$, so that x_1 is well-defined for $x_0 \neq 0$. With (6) but without the assumption $\alpha > \beta > \gamma > 0$, the determinant $x_0(\alpha^n(\beta - \gamma) + \beta^n(\gamma - \alpha) + \gamma^n(\alpha - \beta))$ of the matrix in (12) may vanish as for $\alpha = -7, \beta = 5, \gamma = 2$ and $n = 3$. In that case $x_3 \neq 0$ for almost all values of x_0, and the hyperbola shrinks to a point.

However, for large n the determinant tends to $x_0\alpha^n(\beta - \gamma)$ by (6), the linear system (12) is solvable for large enough n, and the assertion remains valid.

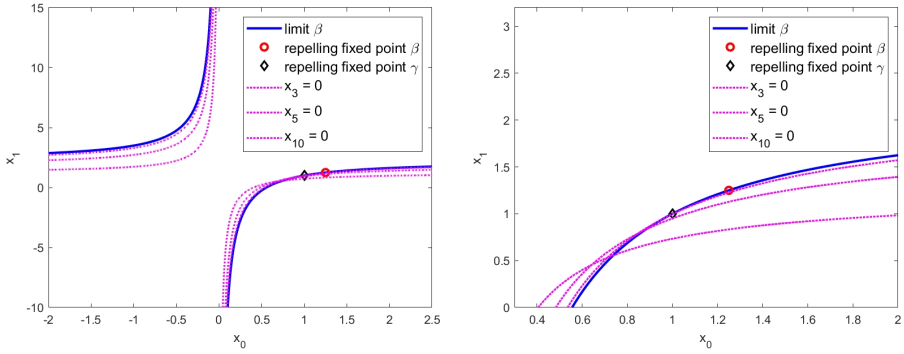
For large enough n, the recurrence is not well-defined on the hyperbola $R^{(n)}x_1 = P^{(n)} - Q^{(n)}/x_0$. For $n \to \infty$, $P^{(n)} \to \alpha^n(\gamma^2 - \beta^2)$, $Q^{(n)} \to \alpha^n\beta\gamma(\gamma - \beta)$ and $R^{(n)} \to \alpha^n(\gamma - \beta)$, which is asymptotically the hyperbola

$$x_1 = \frac{\gamma^2 - \beta^2 - \beta\gamma(\gamma - \beta)/x_0}{\gamma - \beta} = \beta + \gamma - \beta\gamma/x_0. \tag{13}$$

That is the hyperbola H of initial values (x_0, x_1) implying the limit β. □

In Figure 1 the hyperbola with initial values (x_0, x_1) converging to the limit β, and hyperbolas H_k of initial values producing a not well-defined recurrence are shown. In order to produce a better picture we use a recurrence with roots of the characteristic polynomial close together, namely $\alpha = 1.5, \beta = 1.25$ and $\gamma = 1$, but that does not change the qualitative behavior.

Fig. 1 Hyperbolas with limit β and initial values producing a not well-defined recurrence



The right picture in Figure 1 zooms near the unique initial values $(x_0, x_1) = (\gamma, \gamma)$ implying convergence to the smallest root γ. In fact, the recurrence is entirely stationary in that case.

The previous Lemma 2 and in particular (13) imply that in every ε-neighborhood of initial values (x_0, x_1) with well-defined recurrence converging to β there exists a pair of initial values with not well-defined recurrence.

Corollary 1 *Let the recurrence (3) with characteristic polynomial (5) satisfy (6). Suppose that for given initial values $(x_0, x_1) \in \mathbb{R}^2$ the recurrence (x_i) is well-defined and converges to β. Then for every $0 < \varepsilon \in \mathbb{R}$ there exists $x_1' \in \mathbb{R}$ with $|x_1' - x_1| < \varepsilon$ such that the recurrence with initial values (x_0, x_1') is not well-defined.*

Now we can verify the claims on the recurrences mentioned at the beginning. For Kahan's example (1) the roots are $(\alpha, \beta, \gamma) = (100, 5, 3)$. The recurrence with initial value x_0 is not well-defined if, and only if,

$$x_0 = 4 = \gamma - \frac{\gamma^n(\beta - \gamma)}{\beta^n - \gamma^n} = 3 - \frac{2 \cdot 3^n}{5^n - 3^n} = 3 - \frac{2}{(5/3)^n - 1} \quad \text{for some } n \geqslant 1.$$

That is obviously not possible, so taking $x_1 := \beta+\gamma-\beta\gamma/x_0 = 5+3-15/4 = 4.25$ implies $x_i \to 5 = \beta$.

For Muller's example (2) the roots are $(\alpha,\beta,\gamma) = (100,6,5)$. The recurrence with initial value x_0 is not well-defined if, and only if,

$$x_0 = \gamma - \frac{\gamma^n(\beta-\gamma)}{\beta^n-\gamma^n} = 5 - \frac{5^n}{6^n-5^n} = 5 - \frac{1}{(6/5)^n-1} \quad \text{for some } n \geqslant 1.$$

That is obviously not possible for the original value $x_0 = 11/2$ given in [11]. For $x_0 = 2$ given in [12] the recurrence is well-defined if

$$\frac{1}{(6/5)^n-1} \neq 3 \quad \text{for all } n \geqslant 1,$$

which is equivalent to $n \neq \frac{\log(4/3)}{\log(6/5)} \approx 1.58$. Hence (x_i) is well-defined and converges to $6 = \beta$.

2.1 Camouflaged convergence

It may not be visible from a floating-point iteration that an iteration is, in fact, not well-defined. Consider

$$x_0 := \frac{109225}{43691},\ x_1 := \frac{10923}{4369} \quad \text{and} \quad x_{n+1} := 56.5 + (160 - 737.5/x_{n-1})/x_n. \tag{14}$$

The roots of the characteristic polynomial are $(\alpha,\beta,\gamma) = (59,-5,2.5)$ and fulfill (6), and the initial values (x_0,x_1) satisfy $x_1 = \beta+\gamma-\beta\gamma/x_0$. According to Lemma 1 the limit is β if the recurrence is well-defined.

The result for the recurrence computed in half, single, double and infinite precision is given in Table 1, showing convergence to the attracting fixed point $\alpha = 59$.

The recurrence is constructed such that $x_0 = \gamma - \frac{\gamma^n(\beta-\gamma)}{\beta^n-\gamma^n}$ for $n = 17$, so that by Lemma 1 the real recurrence is not well-defined. That fact is camouflaged by the floating-point iteration in single and in double precision.

As for Muller's original example in [11], the initial values (x_0,x_1) in example (14) are not representable in binary floating-point, so in that respect the limit 59 computed in half, single and double precision is correct.

One may ask whether pathological examples of a recurrence exist with x_0, x_1 being exactly representable in binary in some precision and x_1 on the hyperbola $x_1 = \beta+\gamma-\beta\gamma/x_0$, but $x_0 = \gamma - \frac{\gamma^n(\beta-\gamma)}{\beta^n-\gamma^n}$ for some $n \geqslant 1$. In that case the pair (x_0,x_1) is one of the described gaps, i.e., the conditions (10a) and (10b) for convergence to β are satisfied, but by (10c) the sequence is not well-defined. We neither found such an example nor could we prove that it does not exist.

Table 1 Results for the recurrence (14).

n	half	single	double	over $\mathbb{R}$	
0	2.5019531	2.4999428	2.4999428	109225/43691	$\approx$ 2.4999428
1	2.5000000	2.5001144	2.5001144	10923/4369	$\approx$ 2.5001144
2	2.5937500	2.4997749	2.4997711	27305/10923	$\approx$ 2.4997711
3	4.4375000	2.5005341	2.5004578	13655/5461	$\approx$ 2.5004578
4	28.5000000	2.5009155	2.4990846	6825/2731	$\approx$ 2.4990846
5	56.2812500	2.5449677	2.5018315	683/273	$\approx$ 2.5018315
6	58.8750000	3.4965782	2.4963397	1705/683	$\approx$ 2.4963397
7	59	19.3815498	2.5073315	855/341	$\approx$ 2.5073314
8	59	53.8727341	2.4853823	425/171	$\approx$ 2.4853801
9	59	58.7636375	2.5294639	43/17	$\approx$ 2.5294118
10	59	58.9898109	2.4430787	105/43	$\approx$ 2.4418605
11	59	58.9995804	2.6483768	55/21	$\approx$ 2.6190476
12	59	58.9999809	2.9301292	25/11	$\approx$ 2.2727273
13	59	59	16.0674950	3	
14	59	59	50.7931126	5/3	$\approx$ 1.6666667
15	59	59	58.7463651	5	
16	59	59	58.9764139	0	
17	59	59	59.0000847		
...	...	...	...		
27	59	59	59.0000000		
28	59	59	59		
29	59	59	59		

3 Yet other pathological examples

We finally give some new examples where very small precisions yield the correct result, whereas higher precisions do not. We use the four precisions as in Table 2, three of them according to the IEEE-754 [8] floating-point standard.

Table 2 Precisions used.

name	precision in bits	exponent bits
bfloat (truncated binary16)	8	8
half precision (binary16)	11	5
single precision (binary32)	24	8
double precision (binary64)	53	11

The format "bfloat" decreases the precision of binary16 in order to increase the exponent range. It is often called truncated binary16, however, we use that format in rounding to nearest. The middle column gives the precision k in bits including the implicit 1, so that 2^{-k} is the relative rounding error unit. First, consider the recurrence

$$x_0 := -6,\ x_1 := 64 \quad \text{and} \quad x_{n+1} := 82 - (1824 - 6048/x_{n-1})/x_n. \tag{15}$$

All input data are exactly representable in 8 bits binary precision, so in bfloat and therefore in all other precisions. The roots of the characteristic equation

Table 3 Results for $x_0 := -6,\ x_1 := 64$ and $x_{n+1} := 82 - (1824 - 6048/x_{n-1})/x_n$.

n	bfloat	half	single	double	over $\mathbb{R}$
0	-6	-6	-6	-6	-6.000000
1	64	64	64	64	64.000000
2	37.750000	37.750000	37.750000	37.750000	37.750000
3	36.250000	36.187500	36.185429	36.185430	36.185430
4	36	36.031250	36.020496	36.020498	36.020498
5	36	36	36.002277	36.002276	36.002276
6	36	36	36.000256	36.000253	36.000253
7	36	36	36.000031	36.000028	36.000028
8	36	36	36.000004	36.000003	36.000003
9	36	36	36	36.000000	36.000000
10	36	36	36	36.000000	36.000000
	...	...	...	...	...
167	36	36	36	36.000456	36.000000
168	36	36	36	36.000532	36.000000
169	36	36	36	36.000620	36.000000
	...	...	...	...	...
217	36	36	36	36.867247	36.000000
218	36	36	36	36.987987	36.000000
219	36	36	36	37.121863	36.000000
	...	...	...	...	...
296	36	36	36	41.999817	36.000000
297	36	36	36	41.999843	36.000000
298	36	36	36	41.999866	36.000000
	...	...	...	...	...
442	36	36	36	42.000000	36.000000
443	36	36	36	42.000000	36.000000
444	36	36	36	42.000000	36.000000

are $(\alpha, \beta, \gamma) = (42, 36, 4)$. Thus $\beta + \gamma - \beta\gamma/x_0 = 36 + 4 + 144/6 = 64 = x_1$. By Lemma 1 the recurrence would not be well-defined if, and only if,

$$x_0 = -6 = \gamma - \frac{\gamma^n(\beta - \gamma)}{\beta^n - \gamma^n} = 4 - \frac{32}{9^n - 1} \quad \text{for some } n \geqslant 1. \tag{16}$$

That is obviously not possible.

The results for the different precisions are displayed in Table 3. Clearly the last column, iteration over $\mathbb{R}$, shows convergence to the correct limit $\beta = 36$, the repellent fixed point. The second iterate $x_2 = \frac{151}{4}$ is computed without rounding error in all mentioned precisions, but the third iterate $x_3 = \frac{5464}{151}$ is not representable in any binary precision but rounded into some $\widetilde{x}_3$.

The question whether the iteration is well-defined or not depends only on the first initial value. Hence the real iteration starting with $(x_2, \widetilde{x}_3)$ is also well-defined. Therefore, mathematically the iteration with initial values $(x_2, \widetilde{x}_3)$ converges to the attracting fixed point $\alpha = 42$. However, due to "beneficial" rounding errors the recurrence in bfloat, half and single precision converge to the correct value, the repellent fixed point $\beta = 36$.

When displayed as an integer (without trailing zeros) in Table 3, the value of the recurrence is equal to that integer. That happens in bfloat for $\widetilde{x}_4$, in half precision for $\widetilde{x}_5$, and in single precision for $\widetilde{x}_9$.

The double precision recursion becomes stationary after some 444 iterations at $\widetilde{x}_{444} \approx 42 - 2.8 \cdot 10^{-14}$, close to the attracting fixed point $\alpha = 42$.

It is sometimes suggested in the literature that, due to a rounding error in some iterate, the floating-point iteration must converge to the attracting fixed point $\alpha = 42$. In the example above that was true for double precision (binary64), but not true for smaller precisions because rounding errors may be beneficial.

One might think that increasing the precision further should yield the same erroneous result, namely suggesting convergence to the attracting fixed point $\alpha = 42$. However, that is not true. The following Table 4 shows results of

Table 4 Recurrence $x_{n+1} := 82 - (1824 - 6048/x_{n-1})/x_n$ for which $(\alpha, \beta, \gamma) = (42, 36, 4)$.

x_0	x_1	8	11	24	53	64	113
-6	64	**36**	**36**	**36**	42.0	**36**	42.0
-288	40.5	**36**	**36**	**36**	**36.0**	42	42.0
-0.5	328	**36**	**36**	**36**	**36.0**	42.0	**36.0**
-0.1875	808	**36**	**36**	**36**	**36.0**	42.0	42.0
0.5625	-216	**36**	**36**	**36**	42.0	42.0	42.0
64	37.75	**36**	**36**	**36**	42.0	**36.0**	42.0

the recurrence (15) for different initial values (x_0, x_1) and different precisions. Beyond those in Table 2 we add[2] the 80-bit format, formerly called extended precision, with 64 bits precision computed using [7], and IEEE-754 quadruple (binary128) with 113 bits precision. Using (16) one verifies that the recurrences are well-defined.

Numbers shown with a decimal point represent a floating-point number very near to α or β, close to working precision, otherwise the displayed integer is the stationary point of the floating-point iteration. Results in bold face indicate that the recurrence produces the correct limit, i.e., the repellent fixed point β.

The first line corresponds to the initial values used in Table 3. As can be seen increasing the precision to 64 bits yields the correct results, but further increasing produces the wrong but expected result, the attracting fixed point α. But that need not to be so. In the third line all but 64 bit precision produces the correct result due to beneficial rounding errors.

In all examples up to now, the repellent fixed point β is in $\mathbb{F}$. It was asked [10] by Masahide Kashiwagi from Waseda University, Tokyo, whether this is mandatory. The following final example shows that this is not the case. Consider

$$x_0 := 8,\ x_1 := -31 \quad \text{and} \quad x_{n+1} := 1.5 + (972 + 128/x_{n-1})/x_n. \tag{17}$$

All input data are representable in 8 bit binary floating-point, i.e., in bfloat and higher precisions, and the characteristic equation $y^3 - 1.5y^2 - 972y - 128$

[2] Many thanks to Kai Torben Ohlhus for performing the calculations in higher precision using MPFR [7].

has the roots

$$\alpha = 32, \quad \beta = \frac{-61 - \sqrt{3657}}{4} \approx -30.3683, \quad \gamma = \frac{-61 + \sqrt{3657}}{4} \approx -0.1317.$$

Using Lemma 1 one verifies that the recurrence is well-defined.

The repellent fixed point β is obviously not in $\mathbb{F}$, and the best we can expect is a stationary point $\tilde{\beta}$ near β. We say that the recurrence converges numerically to β in precision k bits if the relative error between $\tilde{\beta}$ and β is of the order of the relative rounding error unit 2^{-k}.

Running the recurrence (17) in different precisions of k bits,

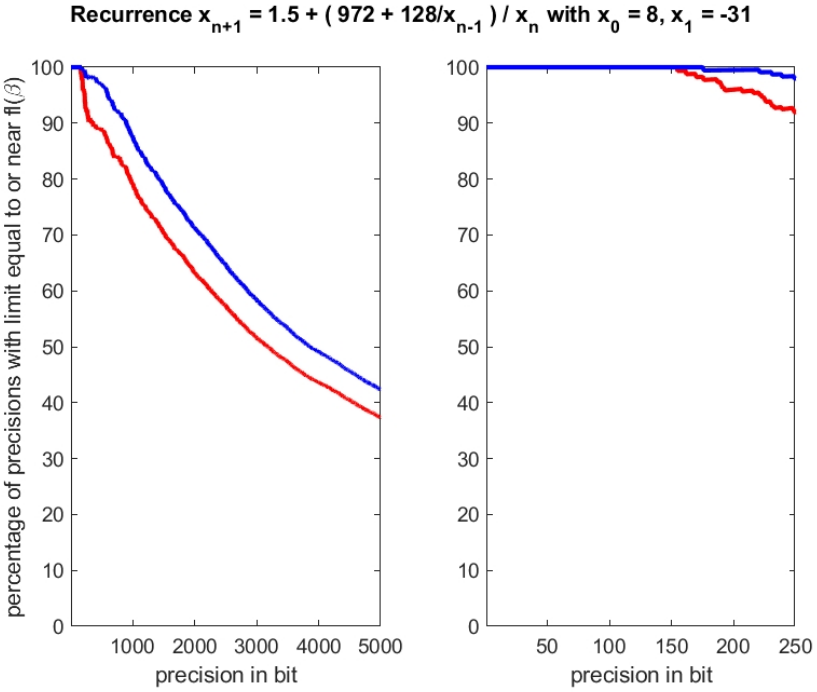
$$\text{for all} \quad k \in \{8, 9, \ldots, 227\} \quad \text{except} \quad k = 183 \tag{18}$$

it converges numerically to the repellent fixed point β. For all precisions

$$k \in \{8, 9, \ldots, 227\} \setminus \{162, 169, 177, 183, 194, 197, 198, 200, 214, 222\}$$

the stationary point was the rounded-to-nearest value of β in the given precision. Note that this includes bfloat, half, single, double, and extended precision.

Fig. 2 Percentage of precisions with numerical convergence to β.



Moreover, we ran the recurrence in all precisions from 8 up to 5000 bits. The accumulated percentage of precisions $[8, 9, \ldots, 5000]$ with numerical convergence to β is displayed in the upper curve of the left graph of Figure 2. For example, in about 90% of all precisions from 8 to 1000, or in about 60% of all precisions from 8 to 3000 numerical convergence to β was observed. The lower curve is the accumulated percentage with stationary point equal to the rounded-to-nearest value of β; this is true, for example, for about 50% of all precisions from 8 to 3000. The right graph shows the same result zoomed into precisions $[8, 9, \ldots, 250]$.

Even for very high precision recurrences of type (3) may converge numerically to β. For example, executing (17) in $17,721$ bits precision becomes stationary close to β up to working precision after some $2,262$ iterations.

As has been pointed out by Masahide Kashiwagi [10], that behavior becomes more clear when looking at the stability of the recurrence at the fixed points. Writing the recurrence (3) as

$$F(x, y) := \begin{pmatrix} y \\ a + (b + c/x)/y \end{pmatrix} \tag{19}$$

and evaluating the spectral radius of the Jacobian at the fixed points yields the results as in Table 5.

Table 5 Spectral radius of the Jacobian of (19) at the fixed points α, β and γ.

recurrence	α	β	γ
(1)	0.05	20	33.3
(2)	0.06	16.7	20
(14)	0.085	11.8	23.6
(15)	0.86	1.17	10.5
(17)	0.949	1.054	242.9

The smaller the spectral radius for the attracting fixed point α, the more we may expect a stable the recurrence. Similarly, for a spectral radius for the repellent fixed point β close to 1 instabilities are more likely. That is particularly the case for the iterations (15) and even more for (17).

We may add some interpretation of the results of the recurrence (17) in different precisions: they are correct or not depending on the point of view. Executed over $\mathbb{R}$ the recurrence converges to the repellent fixed point β, in that respect the result is correct for all precisions listed in (18) but incorrect for $k = 183$. However, $x_2 = \frac{-1883}{62}$ is not a binary floating-point number in any precision, but necessarily rounded into some $\tilde{x}_2$. The limit of the real recurrence (17) with initial values x_1 and $\tilde{x}_2$ is the attracting fixed point α, and in that respect the result for precision $k = 183$ is correct, but for all other precisions listed in (18) it is incorrect.

4 Conclusion

An analysis of recurrences based on Muller's initial example (2) is presented. Necessary and sufficient conditions are given for the recurrence being well-defined, and for convergence to a repellent fixed point. It is shown that in every ε-neighborhood of initial values x_0, x_1 with convergence to a repellent fixed point there exist initial values $x_0, \tilde{x}_1$ producing a not well-defined recurrence.

New recurrences are presented converging (correctly) to a repellent fixed point for smaller precisions such bfloat, half, single and double, but (incorrectly) not for higher precisions. Another example shows convergence to a repellent fixed point even for very high precisions like 5000 bits and more. As a result, rounding errors may be beneficial, and floating-point may have a regularizing effect [4].

5 Acknowledgment

Many thanks to to Florian Bünger, who read the manuscript very carefully and gave several valuable comments, and to Masahide Kashiwagi for fruitful discussions and suggestions. Also many thanks to Kai Torben Ohlhus for performing the calculations in higher precisions. Moreover my thanks to an anonymous referee for his or her fruitful remarks and selfless work.

References

1. M. Abadi, A.Agarwal, P. Barham, E. Brevdo E., Z. Chen Z., C. Citro, G. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Józefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, R. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *CoRR*, abs/1603.04467, 2016.
2. V.D. Blondel. The presence of a zero in an integer recurrent sequence is NP-hard to decide. *Linear Alg. Appl. (LAA)*, 351-352:91–98, 2001.
3. J. Dean, G. Corrado, R. Monga, K. Chen, D. Matthieu, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, Q. Le, and A. Ng. Large scale distributed deep networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1223–1231. Curran Associates, Inc., 2012.
4. N.J. Higham. A Multiprecision World. *SIAM News*, October, 2017.
5. P. Holoborodko. *Multiprecision Computing Toolbox for MATLAB 4.6.4.13348*. Advanpix LLC., Yokohama, Japan, 2019.
6. K. Mahler. Eine arithmetische Eigenschaft der Taylorkoeffizienten rationaler Funktionen. *Akad. Wetensch. Amsterdam*, 38:50–60, 1935.
7. L. Fousse, G. Hanrot, V. Lefèvre, P. Pélissier, and P. Zimmermann. Mpfr: A multiple-precision binary floating-point library with correct rounding. *ACM Trans. Math. Softw.*, 33(2), June 2007.
8. IEEE, New York. *ANSI/IEEE 754-2008: IEEE Standard for Floating-Point Arithmetic*, 2008.
9. W. Kahan, How futile are mindless assessments of roundoff in floating-point computations, `https://people.eecs.berkeley.edu/~wkahan/Mindless.pdf`, 2006.
10. M. Kashiwagi. private communication, 2019.

11. J.-M. Muller, Arithmétique des ordinateurs, `https://hal-ens-lyon.archives-ouvertes.fr/ensl-00086707`, 1989.
12. J.M. Muller, N. Brunie, F. de Dinechin, C.-P. Jeannerod, M. Joldes, V. Lefevre, G. Melquiond, R. Revol, and S. Torres. Handbook of Floating-Point Arithmetic. Birkhäuser, Boston, 2nd edition, 2018.
13. S. Pranesh. Low Precision Floating-Point Formats: The Wild West of Computer Arithmetic. *SIAM News*, November, 2019.
14. Th. Skolem. Einige Sätze über gewisse Reihenentwicklungen und exponentiale Beziehungen mit Anwendung auf diophantische Gleichungen. *Oslo Vid. Akad. Skrifter*, I(6):76–89, 1933.
15. G. Tagliavini, S. Mach, D. Rossi, A. Marongiu, and L. Benini. A transprecision floating-point platform for ultra-low power computing. *CoRR*, abs/1711.10374, 2017.