CARPATHIAN J. MATH.
Volume **37** (2021), No. 2,
Pages 355 - 360

Online version at `https://www.carpathian.cunbm.utcluj.ro/`
Print Edition: ISSN 1584 - 2851; Online Edition: ISSN 1843 - 4401
DOI: https://doi.org/10.37193/CJM.2021.02.20

Dedicated to Prof. Ioan A. Rus on the occasion of his 85^{th} *anniversary*

On fixpoints of Higham's function

RADU T. TRÎMBIŢAŞ

ABSTRACT. We study the strange behavior in floating-point arithmetic of a function proposed by Nicholas Higham, consisting of repeated square roots extraction followed by the same number of times squaring and find its fixpoints. For IEEE standard double precision floating point numbers the fixpoints have the form

$$x \in \left\{ (1+k\mathrm{eps})^{\frac{1}{\mathrm{eps}}}, \quad k = \left[-745 : \frac{1}{2} : -\frac{1}{2}, 0 : 709 \right] \right\} \cup \{0\},$$

where eps is the machine epsilon.

1. INTRODUCTION

In [4, §1.12.2], Higham considers the behavior in floating-point arithmetic of repeated square roots extraction followed by the same number of times squaring. In [2, Problem 1.16, page 38], the authors consider a function, called `Higham`, that accepts a vector `x` as input, takes the square root 52 times, and then squares the result 52 times[1] : theoretically the result must be `x`. Here is the MATLAB code

```
function y=Higham(x)
for i=1:52
    x=sqrt(x);
end
for i=1:52
    x=x.^2;
end
y=x;
```

Then, they run the code

```
x = logspace( 0, 1, 2013 );
y = Higham( x );
plot( x, y, 'k.', x, x, '--' )
```

The result is very different to input x (see Figure 1). The reader is invited to explain the graph and as a hint, they ask the reader to find the points where $y = x$.

In the sequel, we will use the acronym FPN for **F**loating-**P**oint **N**umber (in the IEEE 754 standard, double precision). In fact, the set of FPN is $\mathbb{F} = \mathbb{F}(2, 53, -1022, 1023, true)$,

Received: 30.01.2021. In revised form: 05.06.2021. Accepted: 07.06.2021
2010 *Mathematics Subject Classification.* 65G30, 65G50, 65Y99.
Key words and phrases. *IEEE floating point numbers, floating point arithmetic, fixpoint.*
Corresponding author: Radu T. Trîmbiţaş; tradu@math.ubbcluj.ro
[1] 52 is the number of bits in the significant for a double precision floating point number (the hidden bit is not taken into account).

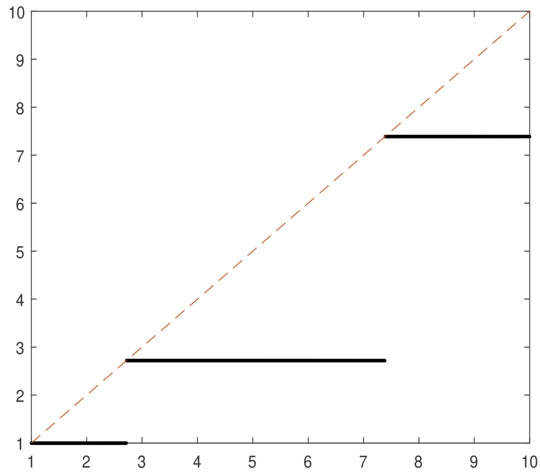

FIGURE 1. The graph of $y = x$ and $\texttt{Higham}(x)$

where 2 is the radix, 53 is the precision, -1022 is the minimum exponent, 1023 is the maximum exponent, and *true* is the value of the denormalization flag.

We wish to find the fixpoints of $\texttt{Higham} : \mathbb{F} \to \mathbb{F}$ on the set $\mathbb{F}$. These fixpoints approximate the corresponding real numbers in the usual sense of floating-point arithmetic, that is, if $\mathrm{fl}(x)$ is the floating-point representation of x, then $\mathrm{fl}(x) = x(1+\delta)$, $|\delta| < \mathrm{eps}/2$, see Section 2, and especially Fact 4.

2. SOME USEFUL FACTS ABOUT FPNS

Fact 1: eps is the distance from 1.0 to the next larger FPN, that is, $2^{-52} \approx 2.2204\mathtt{e}-16$. It is twice machine epsilon. $\mathrm{eps}(x)$ returns the positive distance from $|x|$ to the next larger floating-point number of the same precision as x.

Fact 2: $\texttt{realmax}$ is the largest FPN, $\texttt{realmax} = 2^{1023}(2-\mathrm{eps}) \approx 1.7977\mathtt{e}+308$; $\texttt{realmin}$ is the smallest normalized FPN, $\texttt{realmin} = 2^{-1022} \approx 2.2251\mathtt{e}-308$, the smallest positive FPN is $\mathrm{eps} \cdot \texttt{realmin} = 2^{-52} \cdot 2^{-1022} \approx 4.9407\mathtt{e}-324$.

Fact 3: The largest double precision FPN x for which e^x does not yield overflow is `709.7827128933839731`. In MATLAB, it is `log(realmax)`. The smallest double precision FPN for which e^x does not yield underflow is -745.1332191019411. This could be obtained with a modified variant of bisection applied to the interval $[-746, \ln(\mathrm{eps} \cdot \texttt{realmin})]$, since in MATLAB floating-point arithmetic `exp(-746)` leads to underflow. $\exp(x)$ is within the normal range (i.e. is a normalized FPN) for

$$-708.3964185322641 \le x \le 709.7827128933839731.$$

Fact 4: Axioms of Floating Point Arithmetic:

$$\mathrm{fl}(x) = x(1+\delta), \qquad |\delta| < \mathrm{eps}/2;$$
$$\mathrm{fl}(x \circledcirc y) = (x \circ y)(1+\delta), \qquad |\delta| < \mathrm{eps}/2,$$

where $\circ \in \{+, -, *, /\}$.

For details on floating-point numbers and floating-point arithmetic, see [3], [5] and [7].

3. Repeated application of square root

Case 1: $x \geq 1$. Suppose x is a FPN such that

$$e^k \leq x < e^{k+1}, \tag{3.1}$$

where $k = [\ln x]$. In exact arithmetic, taking square root 52 times is equivalent to rising to the power of eps (Fact 1). So,

$$x^{\text{eps}} \geq e^{k\text{eps}} = 1 + k\text{eps} + \frac{k^2\text{eps}^2}{2!} + \cdots \tag{3.2}$$

$$x^{\text{eps}} < e^{(k+1)} = \left(1 + (k+1)\text{eps} + \frac{(k+1)^2\text{eps}^2}{2!} + \cdots\right). \tag{3.3}$$

The error in floating point arithmetic can be evaluated as follows, using Fact 4 ($SQRT$ is the floating-point square root):

$$\begin{aligned} SQRT(x) &= x^{\frac{1}{2}}\,(1+\delta_1) \\ SQRT(SQRT(x)) &= x^{\frac{1}{4}}\,(1+\delta_1)^{\frac{1}{2}}\,(1+\delta_2) \\ &\vdots \end{aligned}$$

$$SQRT^n(x) = x^{\frac{1}{2^n}}\,(1+\delta_1)^{\frac{1}{2^{n-1}}}\,(1+\delta_2)^{\frac{1}{2^{n-2}}}\cdots(1+\delta_{n-1})^{\frac{1}{2}}\,(1+\delta_n)\,, \tag{3.4}$$

where $|\delta_k| < \text{eps}/2$. Since $n = 52$, the last error is less than

$$\left(1+\frac{\text{eps}}{2}\right)^{1+\frac{1}{2}+\cdots\frac{1}{2^{n-1}}} = \left(1+\frac{\text{eps}}{2}\right)^{2\left(1-\frac{1}{2^n}\right)} = \left(1+\frac{\text{eps}}{2}\right)^{2-2\text{eps}}. \tag{3.5}$$

(3.4) and (3.5) imply

$$SQRT^n(x) = x^{\frac{1}{2^n}}\,(1+\text{eps})^{2-2\text{eps}}. \tag{3.6}$$

From (3.2),(3.3),(3.6), we have

$$[1 + k\text{eps} + \cdots]\left(1+\frac{\text{eps}}{2}\right)^{2-2\text{eps}} \leq SQRT^{52}(x) < [1 + (k+1)\text{eps} + \cdots]\left(1+\frac{\text{eps}}{2}\right)^{2-2\text{eps}}$$

The bounds are as follows

$$\begin{aligned} [1 + k\text{eps} + \cdots]\left(1+\frac{\text{eps}}{2}\right)^{2-2\text{eps}} &= 1 + k\text{eps} + + \cdots \\ [1 + (k+1)\text{eps} + \cdots]\left(1+\frac{\text{eps}}{2}\right)^{2-2\text{eps}} &= 1 + (k+1)\text{eps} + 2\text{eps} + 2(k+1)\ln 2\text{eps}^2 + \cdots \end{aligned}$$

Neglecting the higher order terms, we eventually obtain

$$1 + k\text{eps} \leq SQRT^{52}(x) < 1 + (k+1)\,\text{eps}. \tag{3.7}$$

The differences between the bounds in (3.7) is less than eps, we have only one FPN in the corresponding interval. Using Fact 3, we conclude that $SQRT^{52}(x)$ (i.e. the value of `x` in function `Higham` after the first `for`) has the form

$$SQRT^{52}(x) = 1 + k\text{eps},\ k = 0, \ldots, 709.$$

Case 2: $0 \leq x < 1$. Suppose

$$e^{-\frac{k+1}{2}} < x \leq e^{-\frac{k}{2}}.$$

It follows

$$1-\frac{k+1}{2}\mathrm{eps}+\frac{1}{2!}\frac{(k+1)^2}{4}\mathrm{eps}^2+\cdots=e^{-\frac{k+1}{2}\mathrm{eps}}<x^{\mathrm{eps}}\leq e^{-\frac{k}{2}\mathrm{eps}}$$
$$=1-\frac{k}{2}\mathrm{eps}+\frac{1}{2!}\frac{k^2}{4}\mathrm{eps}^2+\cdots$$

The error in floating point arithmetic is the same as for Case 1 (see formula (3.5)). Using Fact 3, we obtain:

Conclusion: After the first `for` of Higham's function x has the form

$$x=1+k\mathrm{eps},\quad k=\left[-745:\frac{1}{2}:-\frac{1}{2},0:709\right] \tag{3.8}$$

4. Repeated squaring

We use the following inequalities: for $x\geq 0$ and $n\geq 0$ (see [6, pp. 266–268])

$$0\leq e^x-\left(1+\frac{x}{n}\right)^n\leq e^x\left[1-\frac{1}{\left(1+\frac{x}{n}\right)^{x/2}}\right] \tag{4.9}$$

and for $x\neq 0$ (see [1, page 84])

$$e^x>\left(1+\frac{x}{n}\right)^n>e^x\left(1+\frac{x}{n}\right)^{-x},\quad n\in\mathbb{N}, \tag{4.10}$$

From (4.9) and (4.10) it follows

$$e^x\left(1+\frac{x}{n}\right)^{-x/2}\leq\left(1+\frac{x}{n}\right)^n\leq e^x,\qquad x>0. \tag{4.11}$$

Starting from (3.8), we consider two cases for k.

Case 1. $k=0,1,\ldots,709$. Using (4.11) we obtain

$$\left(1+\frac{x+1}{n}\right)^n-\left(1+\frac{x}{n}\right)^n\geq e^{x+1}\left(1+\frac{x}{n}\right)^{-\frac{x+1}{2}}-e^x$$
$$\geq e^x\left[e\left(1+\frac{x+1}{n}\right)^{-\frac{x+1}{2}}-1\right]\geq e^x>\mathrm{eps}\left(e^x\right).$$

Now, setting $x=k$, $n=\frac{1}{\mathrm{eps}}=2^{52}$, we conclude that $(1+k\mathrm{eps})^{\frac{1}{\mathrm{eps}}}$ and $[1+(k+1)\,\mathrm{eps}]^{\frac{1}{\mathrm{eps}}}$ for $k=0,1,\ldots,709$ are distinct FPN.

Case 2. $k=\left[-745:\frac{1}{2}:-\frac{1}{2}\right]$. We apply Lagrange's theorem for function $f(x)=(1+x\mathrm{eps})^{\frac{1}{\mathrm{eps}}}$, on interval $\left[-k-\frac{1}{2},-k\right]$.

$$\left(1-\frac{k}{2}\mathrm{eps}\right)^{\frac{1}{\mathrm{eps}}}-\left(1-\frac{k+1}{2}\mathrm{eps}\right)^{\frac{1}{\mathrm{eps}}}\geq\frac{1}{2}\left[1-\left(k+\frac{1}{2}\right)\mathrm{eps}\right]^{\frac{1}{\mathrm{eps}}-1}$$
$$=\frac{1}{2}\left[1-\left(k+\frac{1}{2}\right)\mathrm{eps}\right]^{\frac{1}{\mathrm{eps}}}\left[1-\left(k+\frac{1}{2}\right)\mathrm{eps}\right]^{-1}.$$

Using (4.10), we have

$$\left(1-\frac{k}{2}\mathrm{eps}\right)^{\frac{1}{\mathrm{eps}}}-\left(1-\frac{k+1}{2}\mathrm{eps}\right)^{\frac{1}{\mathrm{eps}}}\geq\frac{1}{2}e^{-k-\frac{1}{2}}\left[1-\left(k+\frac{1}{2}\right)\mathrm{eps}\right]^{-k-\frac{1}{2}}$$
$$\geq e^{-k}\mathrm{eps},$$

since

$$\frac{1}{2}e^{-\frac{1}{2}}\left[1-\left(k+\frac{1}{2}\right)\text{eps}\right]^{-k} = \frac{1}{2\sqrt{e}\left[1-\left(k+\frac{1}{2}\right)\text{eps}\right]^{-k-\frac{1}{2}}}$$
$$= \frac{1}{2\sqrt{e}\left[1+\left(k+\frac{1}{2}\right)\text{eps}+\left(k+\frac{1}{2}\right)^2\text{eps}^2+\cdots\right]^{k+1/2}}$$
$$\geq \frac{1}{2\sqrt{e}\left[1+\left(k+\frac{1}{2}\right)\text{eps}\right]^{k+1/2}} \approx \frac{1}{2\sqrt{e}\left[1+\left(k+\frac{1}{2}\right)^2\text{eps}\right]}.$$

(We neglected higher order terms). So, $\left(1-\frac{k}{2}\text{eps}\right)^{\frac{1}{\text{eps}}}$ and $\left(1-\frac{k+1}{2}\text{eps}\right)^{\frac{1}{\text{eps}}}$ are distinct FPNs.

Conclusion: The fixpoints of Higham's function have the form

$$x = \text{fl}\left((1+k\text{eps})^{\frac{1}{\text{eps}}}\right), \quad k = \left[-745 : \frac{1}{2} : -\frac{1}{2}, 0 : 709\right], \tag{4.12}$$

and the trivial fixpoint $x = 0$.

5. Numerical checking

We performed all numerical tests in MATLAB[2].

First, we checked (3.8). Within each interval $[k, k+1)$, we generate 1000 equally spaced poins and 1000 random points x, compute $\exp(x)$ and then apply $SQRT$ 52 times. If $\exp(x)$ is within the normal range (see Fact 3), we get only one distinct value, $1 + k\text{eps}$, as (3.8) states.

As usual, the denormalized numbers are problematic (see [5, 3, 7]) . In the nonnormal range we have some exceptions, for $k \in \{-743, -742, -741, -740.5, -739, -738, -734.5, -733.5, -733, -731.5, -730\}$.

Using (4.12), the computation of fixpoints is straightforward. For fixed points within the normal range, the value $|\texttt{higham}(x) - x| = 0$. Table 1 gives some fixpoints for $k = -5 : 1/2 : -1/2$ and $k = 0 : 9$. Again, for the nonnormal range we have some exceptions, for $k \in \{-743, -742, -741, -740.5, -739, -738, -734.5, -733.5, -733, -731.5, -730\}$. Nevertheless, the absolute error is small, see Table 2.

$k = -5 : 1/2 : -1/2$	$k = 0 : 9$
0.00673794684861067	1
0.011108996496683	2.71828180818247
0.0183156383435951	7.38905598869578
0.0301973828606	20.0855367735334
0.0497870672555764	54.5981484040809
0.0820849977073196	148.41315356324
0.135335281222581	403.428787480964
0.223130157655966	1096.63311750602
0.367879438434089	2980.95780915404
0.606530657456067	8103.08314213023

TABLE 1. Fixpoints of `higham` function for $k = -5 : 1/2 : -1/2$ (left) and $k = 0 : 9$ (right)

[2] MATLAB® is a trademark of MathWorks® Inc., Natick, MA

k	fixpoints	$\lvert\mathtt{higham}(x) - x\rvert$
-743.0	$1.48219693752374e-323$	$4.94065645841247e-324$
-742.0	$3.45845952088873e-323$	$1.97626258336499e-323$
-741.0	$9.38724727098368e-323$	$5.92878775009496e-323$
-740.5	$1.53160350210786e-322$	$9.88131291682493e-323$
-739.0	$6.91691904177745e-322$	$4.44659081257122e-322$
-738.0	$1.87744945419674e-321$	$1.21540148876947e-321$
-734.5	$6.21682802162041e-320$	$4.03256380135625e-320$
-733.5	$1.68985272847082e-319$	$1.09623285499256e-319$
-733.0	$2.78608558346337e-319$	$1.80739094561645e-319$
-731.5	$1.24864222542167e-318$	$8.10015685700265e-319$
-730.0	$5.5960197156515e-318$	$3.63025108660419e-318$

TABLE 2. Fixpoints in the nonnormal range

If we try to find the fixpoints in the normal range using the fixpoint iteration with starting value $x_0 = \exp(k)$, we find the corresponding fixpoint in at most 2 iterations.

6. CONCLUSIONS

From the definition of `Higham` function, if follows $\mathtt{Higham}\,(x) = x$, for each $x \in \mathbb{R}$. Due to floating-point arithmetic on the set $\mathbb{F}$, the previous relation holds only for the points given by formula (4.12). Numerical tests confirm the formula with a few exceptions within the nonnormal range given in Table 2.

These explain completely the strange behavior of `Higham` function and the graph in Figure 1.

The results in this paper illustrates the statement that the behavior of FPNs is very different to from that of real numbers. The study of such a freak object as Higham's function should be relevant for Numerical Analysis practitioner.

REFERENCES

[1] Bullen, P., *Dictionary of Inequalities*, Second Edition, CRC, 2015
[2] Corless, R. M., Fillion, N., *A Graduate Introduction to Numerical Methods*. From the Viewpoint of Backward Error Analysis, Springer, New York Heidelberg Dordrecht London, 2013
[3] Goldberg, D., What Every Computer Scientist Should Know About Floating Point Arithmetic, *ACM Computing Surveys*, Vol 23, No 1, pp. 5–48, March 1991
[4] Higham, N. J., *Accuracy and Stability of Numerical Algorithms*, 2nd edition, SIAM, 2002
[5] Muller, J. M. et al., *Handbook of Floating-Point Arithmetic*, second edition, Birkhäuser, 2018.
[6] Mitrinović, D. S., *Analytic Inequalities*, Springer, 1970
[7] Overton, M. L., *Numerical Computing with IEEE Floating Point Arithmetic*, SIAM, 2001

DEPARTMENT OF MATHEMATICS
"BABEŞ-BOLYAI" UNIVERSITY
STR. PLOIEŞTI NO. 27, 400157, CLUJ-NAPOCA, ROMANIA
E-mail address: tradu@math.ubbcluj.ro