When is $x * (1/x) \neq 1$?

Alan Edelman[*]
Draft

December 7, 1994

1 The Problem

During the very first lecture, I now teach advanced students of numerical analysis the subtleties and intricacies of the IEEE standard for floating point computation [1, 2] These students are then ready to attack the first problem set at the very end of which may be found:

> PROBLEM: Find any IEEE double precision floating point number $1 < x < 2$ such that `x*(1/x)` does not yield `1` exactly. Find the smallest one either by a brute force search or by pruning the search substantially using mathematics. (The typewriter font here denotes a computation as it would be presented to a computer.)

The problem of finding the smallest one is surprisingly difficult. The students have just learned that the double precision numbers between 1 and 2 may be represented as

$$x = 1 + k\epsilon, \quad k = 1, 2, \ldots, 2^{52} - 1, \quad \epsilon = 2^{-52}.$$

Their natural inclination is to perform an incremental search on the computer counting from $k = 1$. This approach can work, but seems to take about a day on the students' workstations to reach the smallest k. As machines get faster, if precision stays the same, the problem may become too trivial, but so far only a few students found the smallest k by brute force. My class consists of graduate students of mathematics, computer science, and engineering. In the 1993 class, nobody succeeded in pruning the search space using mathematics. In the 1994 class, two students Ioanid Rosu and Dimitriy Betaneli succeeded. This note is a simplification of my original solution from 1991 and their solutions.

[*] Department of Mathematics Room 2-380, Massachusetts Institute of Technology, Cambridge, MA 02139, `edelman@math.mit.edu`

2 The IEEE Standard

We begin by pointing out that the interval $[1,2]$ fixes a convenient choice for the exponent. Floating point numbers are uniformly spaced in this interval with gap $\epsilon = 2^{-52}$. The interval $[\frac{1}{2},1]$ contains as many floating point numbers with gap $\epsilon/2$. We assume that arithmetic is in the "round to nearest mode." For our purposes this specifies that the arithmetic operations of add, subtract, multiply, and divide compute the floating point number nearest to the infinitely precise result. We also need to recall the "round to even" rule which specifies that in case of a tie, i.e., if the infinitely precise result is in the middle of two floating point numbers, the computation produces the one with least significant bit zero.

We use $\mathrm{fl}(z)$ to denote the nearest floating point number to the real number z. Therefore the computation `x*(1/x)` yields the result $\mathrm{fl}(x\mathrm{fl}(\frac{1}{x}))$.

3 The Solution

Step 1: The computed value of `x*(1/x)` $\in \{1-\frac{\epsilon}{2}, 1\}$.

Since the gap between consecutive numbers in the interval $(\frac{1}{2},1)$ is $\epsilon/2$, we have that $|\frac{1}{x} - \mathrm{fl}(\frac{1}{x})| \le \epsilon/4$ which implies that $|1 - x\mathrm{fl}(\frac{1}{x})| \le \epsilon x/4 < \epsilon/2$. Therefore

$$1 - \frac{\epsilon}{2} < x\mathrm{fl}(\tfrac{1}{x}) < 1 + \frac{\epsilon}{2}$$

from which the result follows by rounding. □

Step 2: The "round to even" rule is never invoked when computing `1/x`.
Proof: The round to even rule would be invoked if and only if $1/x$ were exactly halfway between two floating point numbers, i.e.,

$$\frac{1}{x} = \frac{1}{1+k\epsilon} = 1 - j\frac{\epsilon}{4},$$

where j is odd. This is equivalent to

$$jk = 2^{52}(4k-j).$$

However, it is impossible that jk be a multiple of 2^{52} if j has no even factors and $k < 2^{52}$. □

Step 3: k gives a solution to `x*1/x` $\neq 1$, if and only if k is in the interval $(k_-(m), k_+(m))$, where m is an integer,

$$k_-(m) = \frac{1}{4}\left(m + \sqrt{m^2 + \frac{8}{\epsilon}(m+\frac{1}{2})}\right), \text{ and } k_+(m) = \frac{1}{4}\left(m + \frac{1}{2} + \sqrt{(m+\frac{1}{2})^2 + \frac{8}{\epsilon}(m+\frac{1}{2})}\right).$$

Proof:

Define the integer m by the equation $\mathrm{fl}(\frac{1}{x}) = 1 - k\epsilon + m\frac{\epsilon}{2}$. Since the gap between numbers in $[\frac{1}{2},1]$ is $\frac{\epsilon}{2}$, we may equivalently define m as the unique integer solution to

$$\left|1 - k\epsilon + m\frac{\epsilon}{2} - \frac{1}{1+k\epsilon}\right| < \frac{\epsilon}{4}. \tag{1}$$

which implies that

$$(1+k\epsilon)(1-k\epsilon+m\frac{\epsilon}{2}) > 1-\frac{\epsilon}{4}(1+k\epsilon). \tag{2}$$

On the other hand, if `x*(1/x)`$\neq 1$ then from Step 1, we must have $x\text{fl}(\frac{1}{x}) < 1-\frac{\epsilon}{4}$ or

$$(1+k\epsilon)(1-k\epsilon+m\frac{\epsilon}{2}) < 1-\frac{\epsilon}{4}. \tag{3}$$

Therefore the quadratic inequalities (1),(2) and (3) along with the inequality $k < 2^{52}$ are necessary and sufficient conditions on the positive integers k and m for us to have a situation where `x*(1/x)`$\neq 1$. Together inequalities (2) and (3) are stronger than inequality (1) so we omit (1). The result is obtained by solving the quadratic inequalities (2) and (3) for k in terms of m. □

Conclusion: A small program reveals that the first m for which $(k_-(m), k_+(m))$ contains an integer is $m = 29$ which gives the answer

$$k = 257,736,490.$$

4 Postscript

The recently well-publicized flaw in floating point division on the Pentium chip has prompted the question whether `x/y` may simply be replaced by `x*(1/y)`. This note focuses on the example $y = x$ illustrating that this proposed fix would no longer conform to the IEEE standard.

Since writing this note, I have learned that Prof. W. Kahan at UC Berkeley has also assigned this question (along with many other IEEE "puzzles") to students in various classes.

References

[1] ANSI/IEEE Standard 754-1985. IEEE Standard for Binary Floating-Pont Arithmetic. IEEE, NY, 1985.

[2] ANSI/IEEE Standard 854-1987. A Radix-Independent Standard for Floating-Pont Arithmetic. IEEE, NY, 1987.