MATH 501 HW1

Qilin Ye

January 30, 2021

Pseudocode Programming, p.36

Implementation:

```
s = 1.0; //starting with exponent 0 (i.e. 2^0)
  for k = 1:100
    s = s/2; //keep decreasing the exponent by 1
    t = s + 1.0;
    if t <= 1.0 //detect the first time when 1.0+eps=1.0
      s = s*2;
      k = k-1; //-1 to get exponent of eps
      break
    end
  end
```

Result: $k = 52$ and the machine epsilon $e = 2^{-52} \approx 2.2204 \cdot 10^{-16}$.

Problems from Textbook

Ex.2.1.4 Prove that 4/5 is not representable exactly on the `MARC-32`. What is the closest machine number? What is the relative round-off error involved in storing this number on the `MARC-32`?

Solution

$x := 4/5$ is not representable since $4/5 = (3/4)\cdot(1-1/16) = (.1100\,1100\ldots)_2$. The two nearby machine numbers, each with 24 bits, are $x' := (.1100\ldots\mathbf{1100})_2$ and $x'' := (.1100\ldots\mathbf{1101})_2$. Since they differ by 2^{-24} and

$$x - x' = (.1100\,1100\ldots)_2 \cdot 2^{-24} = \frac{4}{5}\cdot 2^{-24}$$

we know $x'' - x = (1-4/5)\cdot 2^{-24} = 2^{-24}/5$. Therefore $\text{fl}(x) := x'' = (.1100\ldots\mathbf{1101})_2$, and the relative round-off error is

$$\frac{|\text{fl}(x)-x|}{|x|} = \frac{2^{-24}/5}{4/5} = 2^{-26}.$$

Ex.2.1.9 Show that $\text{fl}(x^k) = x^k(1+\delta)^{k-1}$ with $|\delta| \leqslant \epsilon$, if x is a floating-point machine number in a computer with unit round-off ϵ.

Solution

Similar to the example of $\sum_{i=0}^{n} x_i$ earlier in the textbook, here we recursively define $S_n := xS_{n-1}$ with $S_1 = x$ and $S^*_{n+1} := \text{fl}(S^*_n x)$ with the exception $S^*_1 = x$ since x is itself a machine number. In addition we define

$$\begin{cases} \rho_n := \dfrac{S^*_n - S_n}{S_n} & \implies S^*_n = S_n(1+\rho_n), \text{ and} \\ \delta_n := \dfrac{S^*_{n+1} - S^*_n x}{S^*_n x} & \implies S^*_{n+1} = S^*_n x(1+\delta_n). \end{cases}$$

Then,

$$\begin{aligned} 1 + \rho_{n+1} = 1 + \frac{S^*_{n+1} - S_{n+1}}{S_{n+1}} &= \frac{S^*_{n+1}}{S_{n+1}} \\ &= \frac{S^*_n x(1+\delta_n)}{S_n x} && \text{(properties of } \delta_n \text{ and } S_{n+1}) \\ &= \frac{S_n x(1+\rho_n)(1+\delta_n)}{S_n x} && \text{(property of } \rho_n) \\ &= (1+\rho_n)(1+\delta_n). \end{aligned}$$

Therefore we have $(1+\rho_k) = (1+\rho_{k-1})(1+\delta_{k-1}) = \cdots = (1+\rho_1)\prod_{i=1}^{k-1}(1+\delta_i)$. Recall that x is a machine number so $S^*_1 = S_1 = x \implies \rho_1 = 0$. It follows that

$$\text{fl}(x^k) = x^k(1+\rho_k) = x^k(1+\delta)^{k-1} \text{ for } |\delta| \leqslant \epsilon.$$

(*Update: after finishing Ex.2.1.30, it seems like the ρ_n's and δ_n's are not necessary in this proof. One can show inductively that $fl(x) = x$ and $fl(x^n) = fl[fl(x^{n-1})x] = x^n(1+\delta)^{n-1}$. For more details about the induction, see Ex.2.1.30.*)

Ex.2.1.10 Show by examples that often $\text{fl}[\text{fl}(xy)z] \neq \text{fl}[x\,\text{fl}(yz)]$ for machine numbers x, y, and z. This phenomenon is often described informally by saying *machine multiplication is not associative.*

Solution

Consider a, b, c where b is small and c large but their exponent's product is near 0. For example, let

$$\begin{cases} x := (.10\ldots0)_2 \cdot 2^{-2} = 2^{-3}, \\ y := (.10\ldots0)_2 \cdot 2^{-127} = 2^{-128}, \text{ and} \\ z := (.11\ldots1)_2 \cdot 2^{127} = 2^{127} - 2^{103}. \end{cases}$$

It follows that $(x \cdot y) = 2^{-131}$ which causes an underflow and is therefore 0, so $\text{fl}[\text{fl}(xy)z] = 0$. On the other hand,

$$\begin{aligned}\text{fl}[x\,\text{fl}(yz)] &= \text{fl}[x\,\text{fl}(2^{-1} - 2^{-25})] \\ &= \text{fl}[2^{-3}(2^{-1} - 2^{-25})] \\ &= (.10\ldots0)_2 \cdot 2^{-4}.\end{aligned}$$

Ex.2.1.20 Let $x = 2^3 + 2^{-19} + 2^{-22}$. Find the machine numbers on `MARC-22` that are just to the right and just to the left of x. Determine $\text{fl}(x)$, the absolute error $|x - \text{fl}(x)|$, and the relative error $|x - \text{fl}(x)|/|x|$. Verify that the relative error in this case does not exceed 2^{-24}.

Solution

First we write x in normalized scientific notation:

$$2^3 + 2^{-19} + 2^{-22} = (2^{-1} + 2^{-23} + 2^{-26}) \cdot 2^4 = (.\mathbf{100}\ldots\mathbf{010}\,01)_2 \cdot 2^4 \quad \text{(bold = in first 24 terms)}$$

From this we see that the truncation would give $x' = (.100\ldots010)_2 \cdot 2^4$ whereas rounding would give $x'' = (.100\ldots011)_2 \cdot 2^4$. They differ by $2^{-24} \cdot 2^4 = 2^{-20}$. Now we determine which one is $\text{fl}(x)$:

$$x - x' = ((.01)_2 \cdot 2^{-24}) \cdot 2^4 = 2^{-22} \implies x'' - x = ((.11) \cdot 2^{-25}) \cdot 2^4 = 3 \cdot 2^{-22}.$$

Clearly in this case $\text{fl}(x) = x'$ the truncation. The absolute error is 2^{-22} as shown above, and the relative error is

$$\left|\frac{2^{-22}}{2^3 + 2^{-19} + 2^{-22}}\right| = \left|\frac{2^{-22}}{2^{-22}(2^{25} + 2^3 + 1)}\right| = \frac{1}{2^{25} + 2^3 + 1} < \frac{1}{2^{24}}.$$

Ex.2.1.24 Which of these is not necessarily true on the `MARC-32`? (Here x, y, z are machine numbers and $|\delta| \leqslant 2^{-24}$.)

(a) $\text{fl}(xy) = xy(1+\delta)$ (b) $\text{fl}(x+y) = (x+y)(1+\delta)$ (c) $\text{fl}(xy) = xy/(1+\delta)$

(d) $|\text{fl}(xy) - xy| \leqslant |xy|2^{-24}$ (e) $\text{fl}(x+y+z) = (x+y+z)(1+\delta)$.

Solution

(a) True. Since x, y are machine numbers, $\text{fl}(x) = x = x(1+\delta_x)$ and $\text{fl}(y) = y = y(1+\delta_y)$ imply $\delta_x = \delta_y = 0$. Then by definition

$$\text{fl}(xy) = \text{fl}[\text{fl}(x)\text{fl}(y)] = [x(1+\delta_x)y(1+\delta_y)](1+\delta_*) = xy(1+\delta_*).$$

(b) True. Similar to above,

$$\text{fl}(x+y) = \text{fl}[\text{fl}(x) + \text{fl}(y)] = [x(1+\delta_x) + y(1+\delta_y)](1+\delta_*) = (x+y)(1+\delta_*).$$

(c) True. By (a), $\text{fl}(xy) = xy(1+\delta_1)$ with $|\delta_1| \leqslant 2^{-24}$. Now if we simply define

$$1 + \delta := \frac{1}{1+\delta_1} \implies \delta = \frac{-\delta_1}{1+\delta_1}$$

we see $\mathrm{fl}(xy) = xy/(1+\delta)$. Indeed,

$$|\delta| = \left|\frac{\delta_1}{1+\delta_1}\right| \sim |-\delta| \leqslant 2^{-24}.$$

Remark

A second thought on this problem: Taylor expansion is not sufficient to prove the claim. It is not trivial to show that $|\delta/(1+\delta)| < \epsilon$. For example, if $\delta = -\epsilon$, we immediately see that $|-\epsilon/(1-\epsilon)| > \epsilon$. In fact, when $\delta < 0$ and $|\delta|$ is sufficiently close to ϵ, we also have "$>$" as opposed to "$<$". Since if $\delta > 0$ the inequality $|\delta/(1+\delta)| < \epsilon$ holds, we will only be focusing on cases where $\delta < 0$, specifically when $|\delta|$ is very close to ϵ.

First claim: in fact we can replace $|\delta| \leqslant \epsilon$ with the stronger statement $|\delta| < \epsilon$. Recall equation (6) on page 32, the definition of relative error:

$$\left|\frac{x - \mathrm{fl}(x)}{x}\right| \leqslant \frac{2^{m-25}}{q \cdot 2^m} = \frac{2^{-25}}{q} \leqslant \frac{2^{-25}}{1/2} = 2^{-24}.$$

Notice that the two "$\leqslant$"'s *cannot* attain "$=$"'s simultaneously. The second one requires $q = 1/2$ (so x must be a machine number), whereas the first requires x to be *precisely* between the values from chopping and from rounding up (so x cannot be a machine number). Therefore we claim that $|\delta| < \epsilon$.

The next thing to notice is that the mantissa of xy contains < 48 digits. Indeed, after normalizing both, we have mantissas (exponents simply add up so they don't matter here) $(.x_1x_2\ldots x_{24})_2$ and $(.y_1y_2\ldots y_{24})_2$. The smallest term in their product that can possibly be nonzero is $2^{-48}x_{24}y_{24}$, and the largest one is $2^{-2}x_1y_1$.

Recall we said that we will be focusing on δ's very close to $-\epsilon$. Let $r := |\delta|/\epsilon$. We want to find r such that whenever $|\delta| < r\epsilon$, the "$<$" of the original inequality holds:

$$\begin{aligned}\frac{r\epsilon}{1-r\epsilon} < \epsilon &\implies \frac{2^{-24}r}{1-2^{-24}r} < 2^{-24}\\ &\implies \frac{r}{1-2^{-24}r} < 1\\ &\implies r < 1 - 2^{-24}r\\ &\implies r < \frac{1}{1+2^{-24}}.\end{aligned}$$

Is it possible to store a 48-digit mantissa into `MARK-32` with an relative error $> \epsilon/(1+2^{-24})$? The answer is no. The largest possible relative round-off error happens when the 25^{th} to 48^{th} digits is closest to $100\ldots$ (when round-off error is maximized), i.e., when the 25^{th} and 48^{th} digits are 1 and all other digits are 0. In this case, focusing on mantissa only and ignoring the exponent, the ratio between round-off error and 2^{-25} is $1-2^{-48}/2^{-25} = 1-2^{-23}$, still less than r. Therefore no 48-digit mantissa could potentially provide a counterexample to $|\delta(1+\delta)| < \epsilon$, and thus (c) is indeed true.

(d) True. This is trivial when $xy = 0$. Otherwise, by (a), $\text{fl}(xy) - xy = xy\delta \leqslant xy2^{-24}$ and so

$$\frac{\text{fl}(xy) - xy}{xy} \leqslant 2^{-24} \implies \frac{|\text{fl}(xy) - xy|}{|xy|} \leqslant 2^{-24}$$

and the claim follows.

(e) **Not necessarily true**. Setting $y = z = x$ we see (by the theorem in the chapter) that we instead need $(1 + 3\delta)$ to bound the error.

Ex.2.1.26 Which of these is a machine number on the `MARC-32`?

(i) 10^{40} (ii) $2^{-1} + 2^{-26}$ (iii) $\frac{1}{5}$ (iv) $\frac{1}{3}$ (v) $\frac{1}{256}$

Solution

(i) No, because this number will cause an overflow ($> 10^{38}$).

(ii) No, because its mantissa in normalized scientific notation contains 26 digits.

(iii) No, because $1/5 = (3/16)/(1 - 1/16) = (.0011\ 0011\ldots)_2$, an infinite binary expansion.

(iv) No, because $1/3 = (1/4)/(1 - 1/4) = (.01\ 01\ldots)_2$, also an infinite binary expansion.

(v) Yes, obviously; $1/256 = 2^{-8} = (.100\ldots)_2 \cdot 2^{-7}$.

Ex.2.1.30 What relative round-off error is possible in computing the product of n machine numbers in `MARK-32`? How is your answer changed if n numbers are not necessarily machine numbers but are within the range of the machine?

Solution

If $x_1, \ldots, x_n$ are all machine numbers, inductively we have $\text{fl}(x_1) = x_1(1 + \delta_1)^0$ (so the relative error $\leqslant \epsilon$ and

$$\text{fl}\Big(\prod_{i=1}^{k} x_i\Big) = \text{fl}\Big[\text{fl}\Big(\prod_{i=1}^{k-1} x_i\Big) \cdot x_k\Big] = \Big(\prod_{i=1}^{k-1} x_i\Big)(1 + \delta_i)^{k-2}(x_k)(1 + \delta_{k-1}) \leqslant \Big(\prod_{i=1}^{k} x_i\Big)(1 + \tilde{\delta})^{k-1}.$$

where $\tilde{\delta} := \max\{\delta_1, \ldots, \delta_{k-1}\}$. Immediately we see $|\tilde{\delta}| \leqslant \epsilon$. The relative round-off error is therefore $|(1 + \delta)^{k-1} - 1| \sim |(n - 1)\delta| \leqslant (n - 1)\epsilon = (n - 1)2^{-24}$.

On the other hand, if $x_1, \ldots, x_n$ are not necessarily machine numbers, we consider the worst case scenario where none of them are. For the calculations below, we drop the cumbersome subscripts of

δ's — they don't matter anyway, since at the end we'll bound all of them by ϵ. Then:

$$\begin{aligned}
\mathrm{fl}(x_1) &= x_1(1+\delta) \\
\mathrm{fl}(x_1x_2) &= \mathrm{fl}[\mathrm{fl}(x_1)\mathrm{fl}(x_2)] \\
&= \mathrm{fl}[(x_1)(1+\delta)(x_2)(1+\delta)] \\
&= (x_1x_2)(1+\delta)^3 \\
\mathrm{fl}(x_1x_2x_3) &= \mathrm{fl}[\mathrm{fl}(x_1x_2)\mathrm{fl}(x_3)] \\
&= \mathrm{fl}[(x_1x_2)(1+\delta)^3(x_3)(1+\delta)] \\
&= (x_1x_2x_3)(1+\delta)^5 \\
&\dots \\
\text{Inductively, } \mathrm{fl}\Big(\prod_{k=1}^{n} x_k\Big) &= \Big(\prod_{k=1}^{n} x_k\Big)(1+\delta)^{2n-1}.
\end{aligned}$$

Therefore the relative round-off error is bounded by $|(1+\delta)^{2n-1}-1| \sim |(2n-1)\delta| \leqslant (2n-1)\epsilon = (2n-1)2^{-24}$.

Ex.2.1.31 Give examples of real numbers x and y for which $\mathrm{fl}(x \odot y) \neq \mathrm{fl}(\mathrm{fl}(x) \odot \mathrm{fl}(y))$. Illustrate all four arithmetic operations using a five-decimal machine.

Solution

WLOG assume the machine is with a decimal system.

(1) +: consider $x = y := .100004$ (both with $\cdot 10^0$ so it doesn't matter). Then,

$$\begin{aligned}
\mathrm{fl}(x+y) &= \mathrm{fl}(.200008) = .20001, \text{ but} && \text{(round up)} \\
\mathrm{fl}(\mathrm{fl}(x)+\mathrm{fl}(y)) &= \mathrm{fl}(.10000+.10000) = .20000. && \text{(chop both individually)}
\end{aligned}$$

(2) $-$: consider $x := .200006$ and $y := .100002$. Then,

$$\begin{aligned}
\mathrm{fl}(x-y) &= \mathrm{fl}(.100004) = .10000, \text{ but} && \text{(chop)} \\
\mathrm{fl}(\mathrm{fl}(x)-\mathrm{fl}(y)) &= \mathrm{fl}(.20001-.10000) = .10001. && \text{(round } x \text{ up; chop } y\text{)}
\end{aligned}$$

(3) $*$(multiplication): consider $x = y := .900005$. Then,

$$\begin{aligned}
\mathrm{fl}(xy) &= \mathrm{fl}(.81000\,9\dots) = .81001, \text{ but} && \text{(round up)} \\
\mathrm{fl}(\mathrm{fl}(x)\mathrm{fl}(y)) &= \mathrm{fl}(.90001^2) = \mathrm{fl}(.81001\,8\dots) = .81002. && \text{(round up } x, y, \,\&\, \mathrm{fl}(x)\mathrm{fl}(y)\text{)}
\end{aligned}$$

(4) $\div$: consider $x := .800004$ and $y := .899995$. Then,

$$\begin{aligned}
\mathrm{fl}(x/y) &= \mathrm{fl}(.88889\,8\dots) = .88890, \text{ but} && \text{(round up)} \\
\mathrm{fl}(\mathrm{fl}(x)/\mathrm{fl}(y)) &= \mathrm{fl}(.8/.9) = \mathrm{fl}(.88888\,8\dots) = .88889. && \text{(chop, then round *2)}
\end{aligned}$$