Brief Solution to Quiz 01

Sep 29, 2020.

1. How many bits does it take to store a binary floating point number of the form $\pm 1.a_1a_2\cdots a_t \times 2^e$ with $t = 10$, $a_j \in \{0, 1\}$, $-14 \le e \le 15$? What is the distance from 1.0 to the the nearest (typo: should have been 'next larger') floating number? Explain.

   Ans:

   There are total 30 different exponents ($-14 \le e \le 15$). It takes 5 bits to give 30 or more different exponents ($2^5 = 32$). **(5 pts)** Total bits $= 1 + 10 + 5 = 16$ **(5 pts)**.

   The range of the 5-bit binary exponent $c = (b_1b_2b_3b_4b_5)_2$, $b_i = 0, 1$, is $0 \le c \le 31$. In order to cover the range $-14 \le e \le 15$, one should take $e = c - 15$, so that $e = -15$ and $e = 16$ can be reserved for underflow and overflow, respectively. With $e = c - 15$, the binary machine number of 1.0 is given by:

   $$1.0 = +(1.0)_2 \times 2^0 = 0\ \ 01111\ \ 0000000000$$

   The next larger floating point number is 0 01111 0000000001. The difference from 1.0 is 2^{-10} **(10 pts)**.

   The next smaller floating point number is 0 01110 1111111111, or

   $$(1.11\cdots 1)_2 \times 2^{-1} = 2^{-1}\left(1 + 2^{-1} + 2^{-2} + \cdots + 2^{-10}\right) = 2^{-1}\left(2 - 2^{-10}\right) = 1 - 2^{-11}$$

   The difference from 1.0 is 2^{-11}

   The distance from 1.0 to nearest floating point number is $\min(2^{-10}, 2^{-11}) = 2^{-11}$ **(Extra 10 pts)**.

2. Suppose that if $fl(y)$ is a k-digit rounding approximation to y. Show that

   $$\left|\frac{y - fl(y)}{y}\right| \le 5 \times 10^{-k}$$

   Remark: A k-digit rounding means, given $y = \pm 0.d_1d_2\cdots d_kd_{k+1}\cdots \times 10^n$, $0 \le d_i \le 9$, $d_1 > 0$, then $fl(y)$ is obtained by changing d_k to $\tilde{d}_k$ according to the value of d_{k+1}.

   Ans: If k-digit rounding arithmetic is used and

   - If $d_{k+1} \le 4$, then $fl(y) = \pm 0.d_1d_2\cdots d_k \times 10^n$. **(5pts)**

   $$\frac{|y - fl(y)|}{|y|} = \frac{|0.00\cdots 0d_{k+1}d_{k+2}\cdots|}{|0.d_1d_2\cdots d_kd_{k+1}d_{k+2}\cdots|}$$

   $$\le \frac{|0.00\cdots 04999\cdots|}{|0.10\cdots 04999\cdots|} = \frac{|4.999\cdots|}{|10\cdots 04.999\cdots|} \le \frac{|4.999\cdots|}{|10\cdots 00.000\cdots|} \le 5 \times 10^{-k}.$$

- If $d_{k+1} \geq 5$, then $fl(y) = \pm(0.d_1 d_2 \cdots (d_k + 1)) \times 10^n$.

$$\frac{|y - fl(y)|}{|y|} = \frac{|0.00\cdots 100\cdots - 0.00\cdots 0d_{k+1}d_{k+2}\cdots|}{|0.d_1d_2\cdots d_k d_{k+1}d_{k+2}\cdots|} = \frac{|1.00\cdots - 0.d_{k+1}d_{k+2}\cdots|}{|0.d_1d_2\cdots d_k d_{k+1}d_{k+2}\cdots|}\times 10^{-k}$$

$$\leq \frac{|1.00\cdots - 0.500\cdots|}{|0.10\cdots 050\cdots|} \times 10^{-k} \leq \frac{|1.00\cdots - 0.500\cdots|}{|0.10\cdots 000\cdots|} \times 10^{-k} = 5 \times 10^{-k}$$

(15 pts)

3. We showed in class the estimate of relative error resulted from $x \times y$ in terms of ε_M. Derive the corresponding result for $x \div y$, $y \neq 0$.

   Ans:

$$\frac{|x \div y - fl(fl(x) \div fl(y))|}{|x \div y|} \quad \textbf{(5 pts)} \quad = \left|\frac{x \div y - (x(1+\delta_1) \div y(1+\delta_2))(1+\delta_3)}{x \div y}\right|$$

$$= \left|\frac{x \div y - (x \div y)\frac{(1+\delta_1)}{(1+\delta_2)}(1+\delta_3)}{x \div y}\right|$$

$$= \left|1 - (1 + \delta_1 - \delta_2 + \delta_3 + O(\delta^2))\right|$$

$$\approx \left|1 - (1 + \delta_1 - \delta_2 + \delta_3)\right| \quad \textbf{(10 pts)}$$

$$\leq |\delta_1| + |\delta_2| + |\delta_3| \leq 3\epsilon_M \quad \textbf{(5 pts)}$$

4. Solve for $x^2 - 2100x + 1 = 0$ to 15 correct digits using standard double precision arithmetic. Explain how you find your answer (No explanation, no points).

   Ans: $x_1 = \frac{2100+\sqrt{2100^2-4}}{2} = 2.09999952380942e + 03$ **(3 pts)**
   $x_2 = 1/x_1$ or $\frac{2}{2100+\sqrt{2100^2-4}} = 4.76190584170225e - 04$ **(3 pts)**
   No points will be given for answer less than 15 digits.
   Avoid loss of significant digits **(4 pts)**
   Code **(10 pts)**

5. Let $A = \{(x - 100)^2 + y^2 < (40\pi)^2\}$, and $B = \{x + y > 50.1\}$. Find the number of grid points (i, j) (i, j are integers) in $A \cap B$. Write down the answer and name your code by your student ID number.

   Ans: The number of grid points in $A \cap B$ is 33475 **(10 pts)**.
   Code **(10 pts)**