Assignment 1 Solutions

Q1. $F(2,5,-10,10)$ is the number system.

(a) Smallest positive normalized number is

$$(1.0000)_2 (2^{-10}) = 2^{-10} \approx 9.77(10^{-4})$$

By stating it in this form, we understand the size

(b) Largest positive normalized number is

$$(1.1111)_2 (2^{10}) = \underbrace{\left(\sum_{i=0}^{4} 2^{-i}\right)}_{\text{sum a geometric series}} (2^{10}) = 1984. = 1.984(10^3)$$

(c) First, to get approximately the desired answers, we want to find "a" such that

$$e^a \le 1984. \quad \text{(the largest number in } F\text{)}$$

$$\Rightarrow \quad a \le \ln(1984.) \approx 7.593$$

Start with $a_1 \in F(2,5,-10,10)$ slightly too large and then decrease it.

For example, I could choose $a_1 = 7.75 = (111.11)_2$ (a number larger than 7.593 that is easy to write in base 2).

$$e^{a_1} \approx 2321.6 > 1984. \quad \text{(so } a_1 \text{ is too large).}$$

Next smaller is $a_2 = (111.10)_2 = 7.50$ and $e^{a_2} \approx 1808.0 < 1984.$

Therefore, $\boxed{a = (111.10)_2}$ is the desired answer.

Q2. The numbers in S are the following 12 numbers plus the corresponding 12 negative numbers, plus 0.

$1.00(2^{-1}) = \frac{1}{2}$; $1.01(2^{-1}) = \frac{5}{8}$; $1.10(2^{-1}) = \frac{3}{4}$; $1.11(2^{-1}) = \frac{7}{8}$

$1.00(2^{0}) = 1$; $1.01(2^{0}) = 1\frac{1}{4}$; $1.10(2^{0}) = 1\frac{1}{2}$; $1.11(2^{0}) = 1\frac{3}{4}$

$1.00(2^{1}) = 2$; $1.01(2^{1}) = 2\frac{1}{2}$; $1.10(2^{1}) = 3$; $1.11(2^{1}) = 3\frac{1}{2}$

(a) See the attached plot.

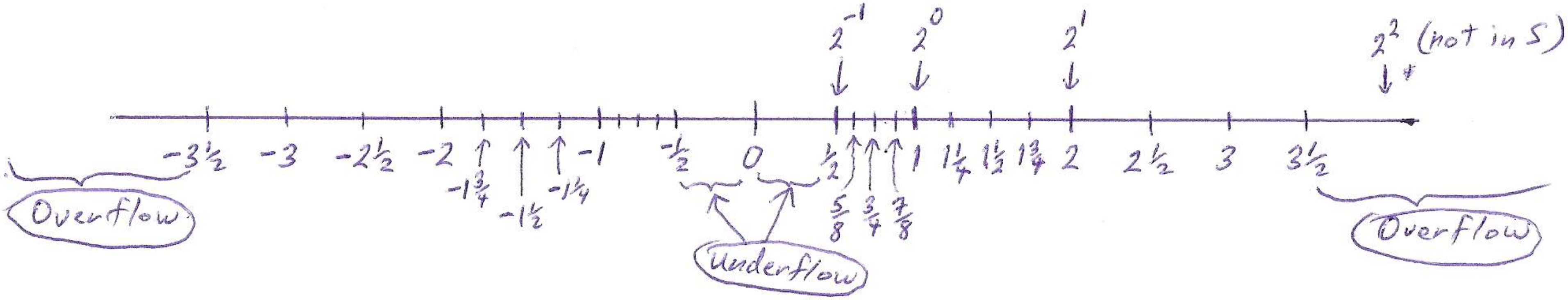
(b) See the attached plot.

(c) There are 25 numbers in S, as noted above.

(d) $\epsilon = \frac{1}{2} \beta^{1-t} = \frac{1}{2} 2^{-2} = 2^{-3}$ (or $\frac{1}{8} = 1.25 \times 10^{-1}$)

(3)

Q2(a) A plot of the numbers in S.



Note that between successive powers of the base $(2^{-1}, 2^0, 2^1, 2^2)$ there are exactly 3 numbers in S between two successive powers.

4

Question 3 In a floating point number system, we have

$$
\begin{aligned}
(a \otimes b) \oplus c &= (ab)(1+\delta_1) \oplus c \\
&= ((ab)(1+\delta_1) + c)(1+\delta_2) \\
&= ((ab+c) + ab\delta_1)(1+\delta_2) \\
&= (ab+c) + ab\delta_1 + (ab+c)\delta_2 + ab\delta_1\delta_2 \\
((a \otimes b) \oplus c) - (ab+c) &= ab\delta_1 + (ab+c)\delta_2 + ab\delta_1\delta_2 \\
&= ab\delta_1(1+\delta_2) + (ab+c)\delta_2
\end{aligned}
$$

The absolute value of the left side is less than or equal to the sum of the absolute values of the right side. So we have

$$
\begin{aligned}
|(ab+c) - ((a \otimes b) \oplus c)| = |((a \otimes b) \oplus c) - (ab+c)| &\leq |ab||\delta_1(1+\delta_2)| + |ab+c||\delta_2| \\
\frac{|(ab+c) - ((a \otimes b) \oplus c)|}{|ab+c|} &\leq \frac{|ab||\delta_1(1+\delta_2)| + |ab+c||\delta_2|}{|ab+c|} \\
&= \frac{|ab|}{|ab+c|}|\delta_1(1+\delta_2)| + |\delta_2| \\
&\leq \frac{|ab|}{|ab+c|}|\delta_1|(1+|\delta_2|) + |\delta_2|
\end{aligned}
$$

Knowing that $ab + c \neq 0$, then

$$
\frac{|(ab+c) - ((a \otimes b) \oplus c)|}{|ab+c|} \leq \frac{|ab|}{|ab+c|}\epsilon(1+\epsilon) + \epsilon
$$

$\sqcap$

(5)

Q4.

Solution ~~2~~.

(a) With $a = 1.0000$, $b = 111.11$ and $c = 1.2121$ we get the following results using the given formulas, computing in $F(10, 5, -10, 10)$.

$$\begin{aligned} r_1 &= \frac{-b + \sqrt{b^2 - 4ac}}{2a} \\ &= \frac{-111.11 + \sqrt{12345. - 4.8484}}{2.0000} \\ &= \frac{-111.11 + \sqrt{12340.}}{2.0000} \\ &= \frac{-111.11 + 111.09}{2.0000} \\ &= \frac{-.020000}{2.0000} \\ &= -.010000 \end{aligned}$$

and

$$\begin{aligned} r_2 &= \frac{-b - \sqrt{b^2 - 4ac}}{2a} \\ &= \frac{-111.11 - \sqrt{12345. - 4.8484}}{2.0000} \\ &= \frac{-111.11 - \sqrt{12340.}}{2.0000} \\ &= \frac{-111.11 - 111.09}{2.0000} \\ &= \frac{-222.20}{2.0000} \\ &= -111.10\ . \end{aligned}$$

True roots:

$r_1 = -.010910$

$r_2 = -111.10$

(to 5 significant digits)

$relerr_{r_1} = \frac{|-.010910 + .010000|}{|-.010910|} \approx 0.8\,(10^{-1})$; $relerr_{r_2} = \frac{|-111.10 + 111.10|}{|-111.10|} = 0$

(b) Rationalizing the numerator in the original formula for r_1 proceeds as follows.

$$\begin{aligned} r_1 &= \frac{(-b + \sqrt{b^2 - 4ac})}{2a} \frac{(-b - \sqrt{b^2 - 4ac})}{(-b - \sqrt{b^2 - 4ac})} \\ &= \frac{b^2 - (b^2 - 4ac)}{2a\,(-b - \sqrt{b^2 - 4ac})} \\ &= \frac{4ac}{2a\,(-b - \sqrt{b^2 - 4ac})} \\ &= \frac{2c}{-b - \sqrt{b^2 - 4ac}}\ . \end{aligned}$$

Note: r_1 has 1 sig. digit correct. r_2 has all 5 sig. digits correct.

6

(c) Using the result from part (b) and a similar result for r_2, we can derive the following formulas which avoid the cancellation problem for cases where $|b| \approx \sqrt{b^2 - 4ac}$.

Algorithm R.

if $b > 0$ then

$$r_2 = (-b - \sqrt{b^2 - 4ac}) \;/\; (2\; a)$$
$$r_1 = c \;/\; (a\; r_2)$$

else

$$r_1 = (-b + \sqrt{b^2 - 4ac}) \;/\; (2\; a)$$
$$r_2 = c \;/\; (a\; r_1)$$

(d) With $a = 1.0000$, $b = 111.11$ and $c = 1.2121$ we get the following results for the roots using Algorithm R, computing in $F(10, 5, -10, 10)$. Note that $b > 0$.

$$\begin{aligned} r_2 &= (-b - \sqrt{b^2 - 4ac}) \;/\; (2\; a) \\ &= -111.10 \\ &\ \text{(see the detailed calculation for } r_2 \text{ in part (a))} \\ r_1 &= c \;/\; (a\; r_2) \\ &= 1.2121 \;/\; (-111.10) \\ &= -.010910\ . \end{aligned}$$

By calculating with higher accuracy, one can verify that the latter results for r_1 and r_2 are fully accurate to five significant digits. In contrast, the value computed for r_1 in part (a) was accurate to only one significant digit (four digits of accuracy were lost!).

This time, $\text{relerr}_{r_1} = 0$ and $\text{relerr}_{r_2} = 0$.
I.e. all 5 sig. digits correct in r_1 and r_2.