I. Convert 477 to a normalized FPN with $\beta = 2$

We can rewrite this number into $(111011101)_2$, hence $m = 1.11011101$ and the normalized binary form is

$$477 = (1.11011101)_2 \times 2^8 \tag{1}$$

II. Convert $\frac{3}{5}$ to a normalized FPN with $\beta = 2$

$$\frac{3}{5} = (0.1001\cdots)_2 = (1.00110011\cdots) \times 2^{-1} \tag{2}$$

III. Prove $x_R - x = \beta(x - x_L)$

We can rewrite the condition as $x = 1.0 \times \beta^e = (1.0 \times \beta) \times \beta^{e-1}$. Additionally, the machine precision is $\epsilon_M = \beta^{1-p}$. Consequently,

$$x_R = (1.0 + \beta^{1-p}) \times \beta^e \tag{3}$$

$$x_L = (\beta - \beta^{1-p}) \times \beta^{e-1} \tag{4}$$

Next step is easy,

$$x_R - x = \beta^{1-p} \times \beta^e = \beta^{e-p+1} \tag{5}$$

$$x - x_L = (\beta \times 1.0 - \beta + \beta^{1-p}) \times \beta^{e-1} = \beta^{e-p} \tag{6}$$

Finally, we prove $x_R - x = \beta(x - x_L)$ successfully.

IV. Find two normalized FPNs adjcent to x and relative roundoff error

Round off $\frac{3}{5}$ into fl(x)= $x_R = (1.00110011001100110011010)_2 \times 2^{-1}$. So the two adjcent normalized FPNs are

$$x_L = (1.00110011001100110011001)_2 \times 2^{-1} \tag{7}$$

$$x_R = (1.00110011001100110011010)_2 \times 2^{-1} \tag{8}$$

As a result, the relateive error is $\epsilon = |\frac{fl(x)-x}{x}| = \frac{2^{-26}+0.6\times 2^{26}}{0.6} \approx 3.97 \times 10^{-8} = 3.97 \times 10^{-6}\%$.

V. What is the unit roundoff when drop excess bits simply

$$\epsilon_u = \epsilon_M = \beta^{1-p} = 2^{-23}$$

VI. How many bits of precision are lost in $1 - \cos\frac{1}{4}$

We can define $fl(a) = 1$ and $b =$fl($\cos\frac{1}{4}$) by

$$a = M_a \times 2^{e_a} = (1.00000000000000000000000)_2 \times 2^0 \tag{9}$$

$$b = M_b \times 2^{e_b} = (1.11111111111111101100000)_2 \times 2^{-1} \tag{10}$$

And then define $c =$fl($a - b$)= $M_c \times 2^{e_c}$, so

$$M_c = M_a - \beta^{-1} M_b \tag{11}$$

$$= (1.00000000000000000000000)_2 - (0.111111111111111101100000)_2 \tag{12}$$

$$= (0.000000000000000010100000)_2 \tag{13}$$

$$= (1.01000000000000000000000)_2 \times 2^{-17} \tag{14}$$

Namely, $c = (1.01000000000000000000000)_2 \times 2^{-17}$ and 17 bits of precision is lost.

VII. Suggest two ways to compute $1 - \cos x$

Firstly, we can use Taylor series

$$1 - \cos x = 1 - (1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots) = \sum_{i=1}^{\infty}(-1)^{i+1}\frac{x^{2i}}{(2i)!} \tag{15}$$

Secondly, we can use a trigonometric function formula $\cos 2x = 1 - 2\sin^2 x$ such that

$$1 - \cos x = 1 - (1 - 2\sin^2 \frac{x}{2}) = 2\sin^2 \frac{x}{2} \tag{16}$$

C++ programming

A. Compare three functions

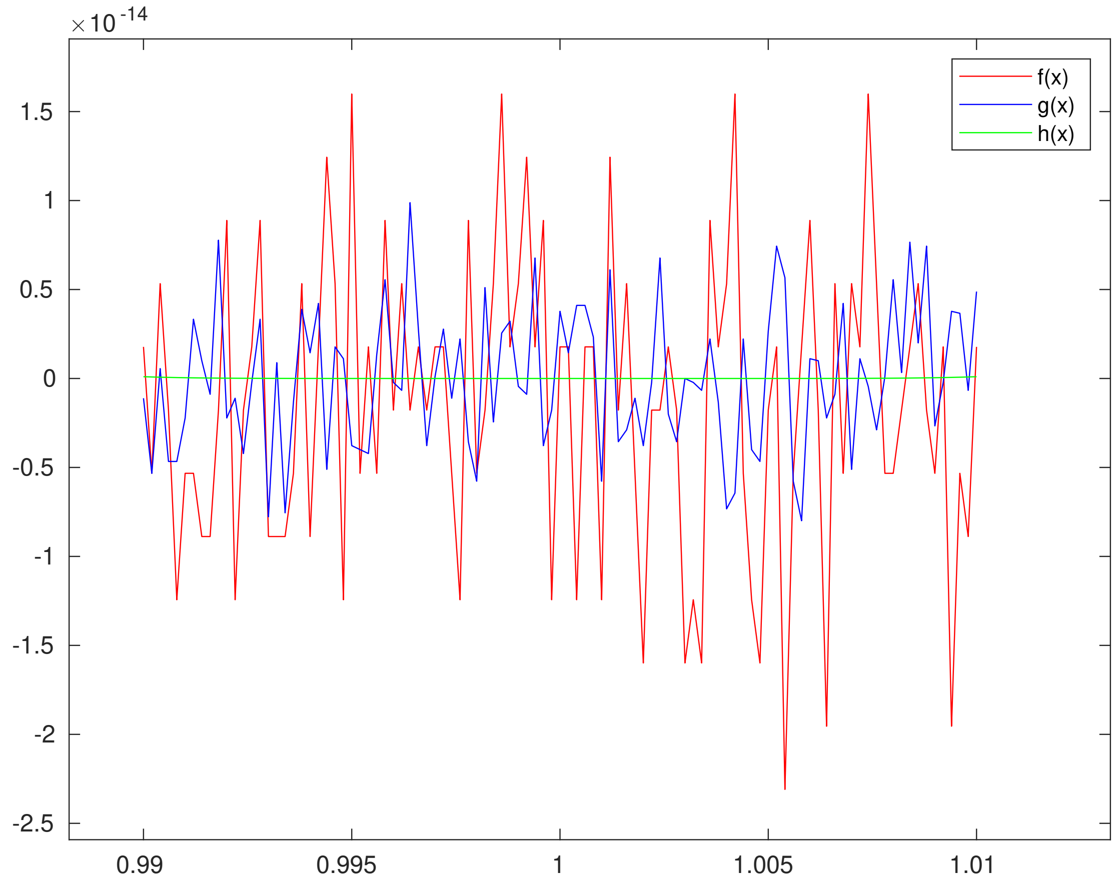

Figure 1: the difference between $f(x)$ and $g(x)$ and $h(x)$

Multiplication and division are accurate. However, addition, let say $fl(fl(x) + fl(y))$, is not accurate when $x + y \to 0$. And function $f(x)$ and $g(x)$ use addition or substraction calculation for eight times but function $h(x)$ uses substraction for only one time. As a result, function $h(x)$ is the most accurate one.

B. Consider a normalized FPN system $\mathbb{F}$

We can know the $UFL(\mathbb{F}) = 0.5$ and $OFL(\mathbb{F}) = 3.5$ easily by definition 1.10. Besides, the enumeration of elements in $\mathbb{F}$ is as following

$$1.00 \times 2^{-1}, 1.01 \times 2^{-1}, 1.10 \times 2^{-1}, 1.11 \times 2^{-1} \tag{17}$$
$$1.00 \times 2^{0}, 1.01 \times 2^{0}, 1.10 \times 2^{0}, 1.11 \times 2^{0} \tag{18}$$
$$1.00 \times 2^{1}, 1.01 \times 2^{1}, 1.10 \times 2^{1}, 1.11 \times 2^{1} \tag{19}$$
$$-1.00 \times 2^{-1}, -1.01 \times 2^{-1}, -1.10 \times 2^{-1}, -1.11 \times 2^{-1} \tag{20}$$
$$-1.00 \times 2^{0}, -1.01 \times 2^{0}, -1.1 \times 2^{0}, -1.11 \times 2^{0} \tag{21}$$
$$-1.00 \times 2^{1}, -1.01 \times 2^{1}, -1.1 \times 2^{1}, -1.11 \times 2^{1} \tag{22}$$

as well as 0. hence $\#F = 2^3 \times (1 - (-1) + 1) + 1 = 25$ consistent with corollary 1.11.

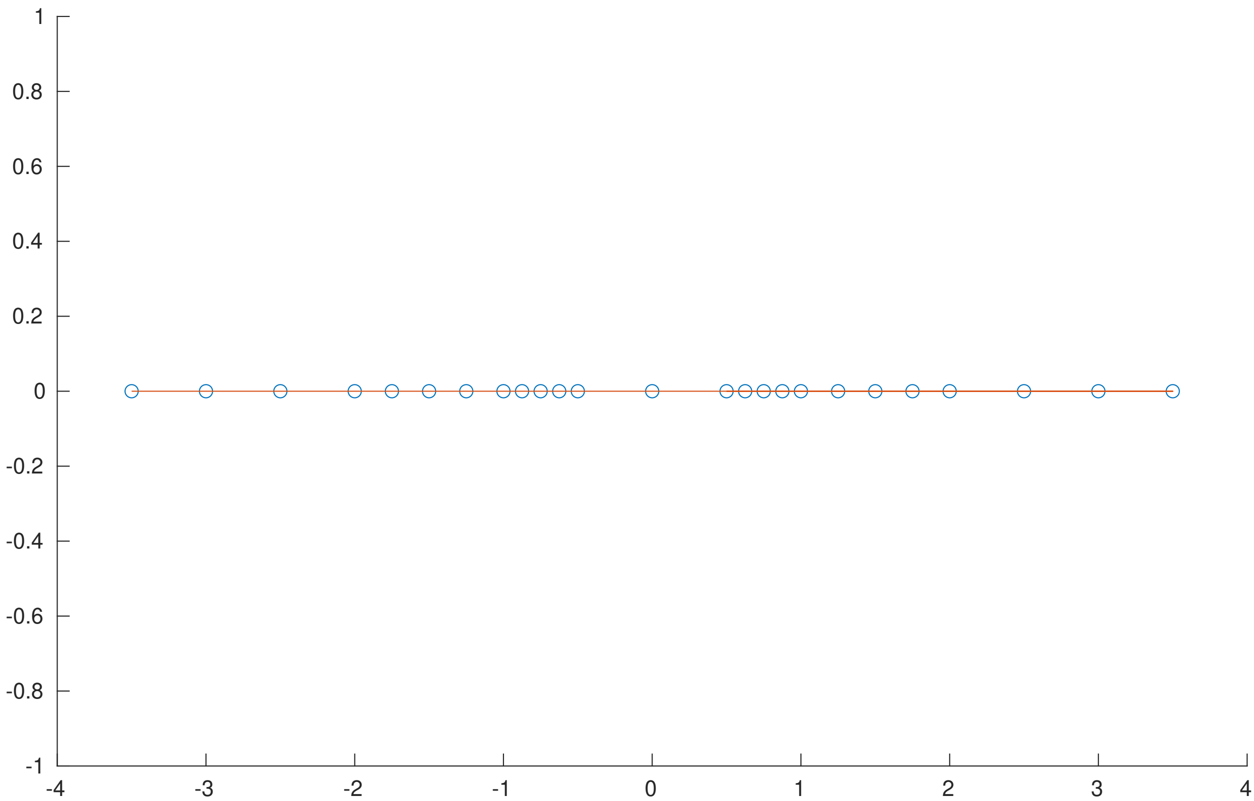

Figure 2: $\mathbb{F}$ on the real axis

Additionally, all the subnormal numbers are 0.125 0.25 0.375 -0.125 -0.25 -0.375 . Therefore, the extended $\mathbb{F}$ is as following

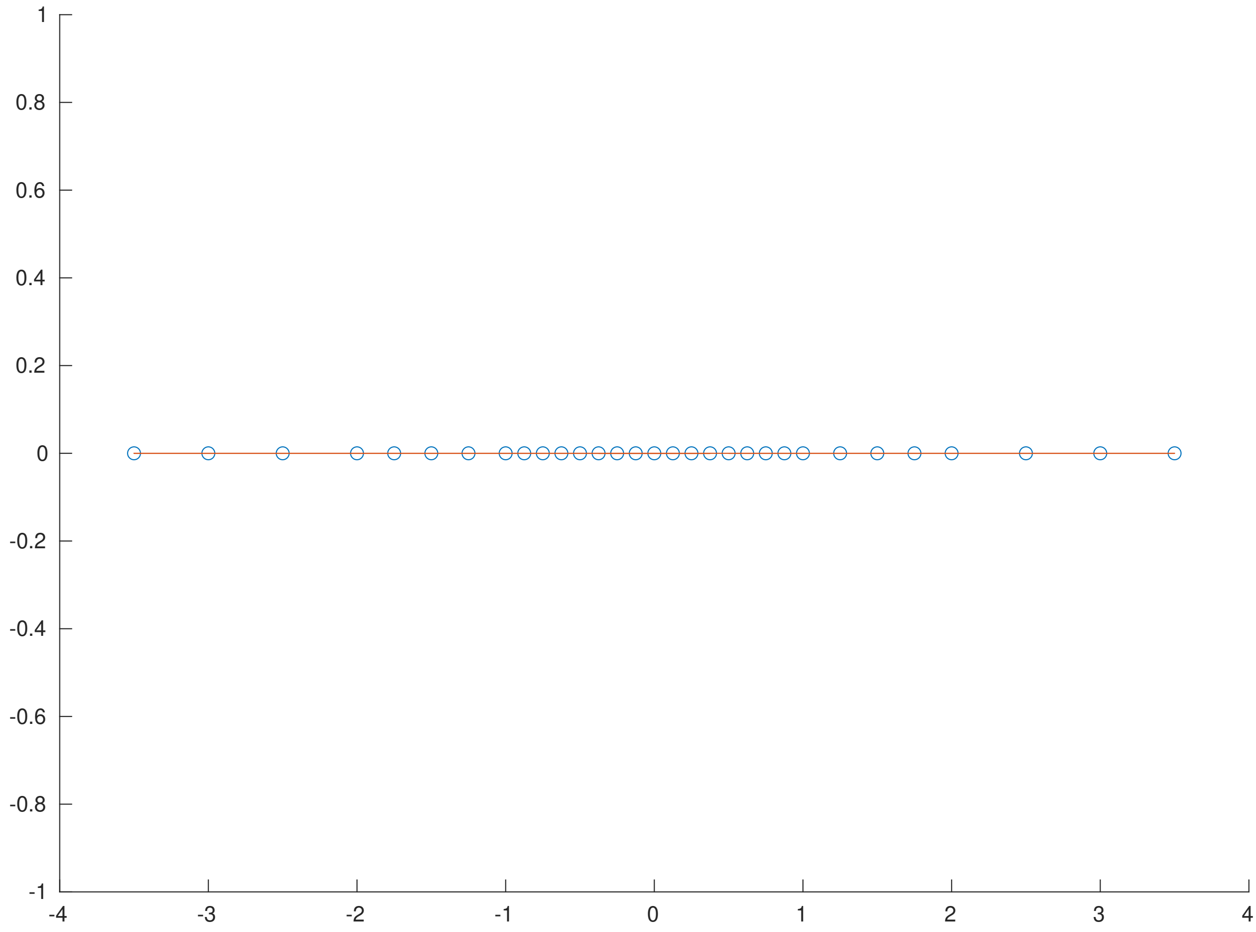

Figure 3: The *extended* $\mathbb{F}$ on the real axis