1 Notation

$\mathbb{N}$ Integers

$\mathbb{N}^0$ Non-negative integers

$\mathbb{N}^+$ Positive integers

$\lfloor a \rfloor$ Rounding towards zero, floor

$a \backslash b$ Integer division, the same as $\lfloor a/b \rfloor$

$b \mid a$ b divides a and a is not zero

$b \nmid a$ b doesn't divide a or a is zero

$[a]_N$ Rounding to nearest $10^{-N}k, k \in \mathbb{N}$, away from zero in case of uncertainty

$[\![a]\!]_N$ Rounding N significant decimal figures, away from zero in case of uncertainty

2 Binary to decimal conversion

Let we have *positive* binary floating-point number

$$x = 2^a p, \ p \in \mathbb{N}^0, a \in \mathbb{N} \tag{1}$$

We need to find decimal mantissa m and decimal exponent e such that

$$x = 10^e \times m$$
$$m \in \mathbb{N}^0$$
$$10 \nmid m$$
$$e \in \mathbb{N}$$

It is always possible to rewrite it in such a way that mantissa is not divisible by neither 2 nor 5.

$$x = 2^b 5^d q \tag{2}$$
$$q \in \mathbb{N}^0$$
$$2 \nmid q$$
$$5 \nmid q$$
$$p = 2^{b-a} 5^d q$$

Taking into account that $5^d = 10^d 2^{-d}$ we can rewrite it extracting tens exponent.

$$x = 10^d \times 2^{b-d} q \tag{3}$$
$$x = 10^b \times 5^{d-b} q \tag{4}$$

The mantissa and the exponent should be integers. Therefore final result for decimal representation of given floating-point number is

$$(m, e) = \begin{cases} (2^{b-d}q, d), & b - d \geqslant 0 \\ (5^{d-b}q, b), & b - d < 0 \end{cases}$$

$$m = 2^{\max(b-d,0)} 5^{\max(d-b,0)} q$$

$$e = \min(b, d)$$

Example 1. *For double-precision floating-point numbers following inequations are true*

$$0 \leqslant d \leqslant 22$$

$$-1\,022 \leqslant b \leqslant 1\,075$$

$$0 \leqslant q \leqslant 2^{53} - 1 = 9\,007\,199\,254\,740\,991$$

$$-1\,022 \leqslant e \leqslant 22$$

$$0 \leqslant m \leqslant (2^{53} - 1) \times 5^{1022}$$

3 Binary to decimal conversion with rounding to N decimal digits after point

Firstly let us define rounding meaning here, we denote rounded number as $[x]_N$. We define rounding on kth interval where $k \in \mathbb{N}^0$.

$$[x]_N = \frac{k}{10^N}, \ \forall x \in \begin{cases} \left[\frac{k-0.5}{10^N}, \frac{k+0.5}{10^N}\right), & k > 0 \\ \left[0, \frac{0.5}{10^N}\right), & k = 0 \end{cases}$$

We will use following known trick to compute rounded value.

$$[x]_N = \frac{\left\lfloor 10^N x + 0.5 \right\rfloor}{10^N}$$

We can consider only case $N < -e$. Otherwise rounded number will be equal to a given one. Let us denote $x' = \left\lfloor 10^N x + 0.5 \right\rfloor$ and rewrite it in terms of the powers from equation (2).

$$x' = \left\lfloor 5^{d+N} 2^{b+N} q + 0.5 \right\rfloor$$
$$d' = d + N$$
$$b' = b + N$$
$$x' = \left\lfloor 5^{d'} 2^{b'} q + 0.5 \right\rfloor \tag{5}$$

Finally this leads us to

$$x' = \begin{cases} 5^{d'}2^{b'}q, & d' \geqslant 0, b' \geqslant 0 & (6) \\[2mm] \left(5^{d'}q + 2^{-b'-1}\right) \backslash 2^{-b'}, & d' \geqslant 0, b' < 0 & (7) \\[2mm] \left(2^{b'}q + 5^{-d'}\backslash 2\right) \backslash 5^{-d'}, & d' < 0, b' \geqslant 0 & (8) \\[2mm] \left(q + 5^{-d'}2^{-b'-1}\right) \backslash \left(5^{-d'}2^{-b'}\right), & d' < 0, b' < 0 & (9) \end{cases}$$

Cases (6), (7), (9) obviously follow from (5). Let's show that (8) is correct.

Lemma 1. *The following assertion is correct for every $x \in \mathbb{N}^0$, $y \in \mathbb{N}^0$ and $q \in \mathbb{N}^0$*

$$\left\lfloor \frac{2^x q}{5^y} + \frac{1}{2} \right\rfloor = \left\lfloor \frac{2^x q - \frac{1}{2}}{5^y} + \frac{1}{2} \right\rfloor$$

Proof. It is enough to proof that $\frac{2^x q}{5^y} + \frac{1}{2} < n+1$ follows to $\frac{2^x q - \frac{1}{2}}{5^y} + \frac{1}{2} < n+1$ and $\frac{2^x q}{5^y} + \frac{1}{2} \geqslant n$ follows to $\frac{2^x q - \frac{1}{2}}{5^y} + \frac{1}{2} \geqslant n$ where $n \in \mathbb{N}^0$. Former is obvious. Let us proof latter. At first let us rewrite inequation.

$$\frac{2^x q}{5^y} + \frac{1}{2} \geqslant n$$

$$\frac{5^y}{2} \geqslant 5^y n - 2^x q$$

Note that $2 \nmid 5^y$ and right hand side is integer, this leads us to

$$\frac{5^y - 1}{2} \geqslant 5^y n - 2^x q$$

$$\frac{2^x q - \frac{1}{2}}{5^y} + \frac{1}{2} \geqslant n$$

$\blacksquare$

4 Rounding to F significant figures

Rounded mantissa m'' should satisfy following inequation.

$$0 \leqslant m'' < 10^F$$

Therefore we need to find minimal $g \in \mathbb{N}^0$ such that

$$m/10^g < 10^F - 0.5$$
$$\therefore 2m/10^g < 2 \times 10^F - 1$$
$$\therefore (2m)\backslash 10^g < 2 \times 10^F - 1$$

Now we can find g iteratively or use faster method if g is known to be big. Then we round a value to $e - g$ digits.