0.1 in double precision

We already know that

$$(0.1)_{10} = (0.000110011001100110011\ldots)_2$$

Let us denote the actural value stored in double precision for 0.1 by $fl(0.1)$.

In Matlab, we can easily try the following:

```
>> format hex
>> 0.1

ans =

   3fb999999999999a

>> format
```

(The last `format` will set the output format back to normal.)

In the output, `3fb999999999999a` is the Hexadecimal code of the double precision storage of 0.1. The following is a Hexadecimal(HEX) to Binary(BIN) conversion chart

HEX	0	1	2	3	4	5	6	7
BIN	0000	0001	0010	0011	0100	0101	0110	0111
HEX	8	9	a	b	c	d	e	f
BIN	1000	1001	1010	1011	1100	1101	1110	1111

Use this table to convert `3fb999999999999a` into a 64-bit binary code:

$$0011\ 1111\ 1011\ 1001\ 1001\ \cdots\ 1001\ 1001\ 1010$$

Here we use red to denote the sign bit, blue to denote the exponent, and the rest is mantissa.

1. The sign bit 0 indicates that the number is positive.
2. The exponent $c = (01111111011)_2 = (1019)_{10}$.
3. The mantissa (notice the rounding at the end)

$$\begin{aligned} f &= (0.1001100110011001100110011001100110011001100110011010)_2 \\ &= \left(\frac{2702159776422298}{4503599627370496}\right)_{10} \qquad \text{(note } 2^{52} = 4503599627370496\text{)} \\ &\approx (0.6)_{10} \end{aligned}$$

The number $\frac{2702159776422298}{4503599627370496}$ is not exactly equal to 0.6. Indeed

$$\begin{aligned} f - 0.6 &= \frac{2702159776422298}{4503599627370496} - \frac{6}{10} \\ &= \frac{27021597764222980 - 27021597764222976}{45035996273704960} \\ &= \frac{4}{45035996273704960} \\ &= \frac{4}{10} 2^{-52} \\ &= \frac{8}{10} \varepsilon \end{aligned}$$

where $\varepsilon = 2^{-53} \approx 1.11022302462516 \times 10^{-16}$ is the machine epsilon in IEEE 754 for double precision.

So finally, we end up with

$$fl(0.1) = (-1)^s 2^{c-1023}(1+f) = 2^{-4}(1.6 + 0.8\varepsilon) = \frac{1.6 + 0.8\varepsilon}{16} = 0.1 + 0.05\varepsilon$$

The absolute error is

$$|0.1 - fl(0.1)| = 0.05\varepsilon$$

and the relative error is

$$\frac{|0.1 - fl(0.1)|}{|0.1|} = 0.5\varepsilon$$

REMARK 1. Another way to compute $0.1 - fl(0.1)$ is

$$\begin{aligned} & fl(0.1) - 0.1 \\ = \quad & (0.000110011001100110011 \ldots 11010)_2 \\ & -(0.000110011001100110011 \ldots 11001100110011 \ldots)_2 \\ = \quad & (0.000000000000000000000 \ldots 00000011001100 \ldots)_2 \\ = \quad & 2^{-54} \times (0.000110011001100110011 \ldots)_2 \\ = \quad & 2^{-54} \times 0.1 = 0.05\varepsilon \end{aligned}$$