## 1 Machine (Floating-Point) Numbers (Section 1.3, Continued)

(Example.) Let  $x = \frac{2}{3}$ .

(1) What is the binary form of x?

The algorithm for finding the binary form of the decimal is as follows:

- Given x, multiply it by 2. If the integer part of the result is 1, set the ith bit to 1. Otherwise, set it to 0.
- If the *i*th bit is 1, subtract  $x \times 2$  by 1.
- Repeat the above until one of the following occurs:
  - You hit exactly 1, or
  - You hit 23 bits after the binary point (most likely, you'll see that the bits repeat in some way).

$$\frac{2}{3} \cdot 2 = \frac{4}{3} \ge 1 \implies 1$$

$$\frac{1}{3} \cdot 2 = \frac{2}{3} < 0 \implies 0$$

$$\frac{2}{3} \cdot 2 = \frac{4}{3} \ge 1 \implies 1$$

$$\frac{1}{3} \cdot 2 = \frac{2}{3} < 0 \implies 0$$

$$\frac{2}{3} \cdot 2 = \frac{4}{3} \ge 1 \implies 1$$

$$\frac{1}{3} \cdot 2 = \frac{2}{3} < 0 \implies 0$$

$$\frac{2}{3} \cdot 2 = \frac{4}{3} \ge 1 \implies 1$$

$$\frac{1}{3} \cdot 2 = \frac{2}{3} < 0 \implies 0$$

$$\vdots$$

This gives us the binary representation 0.1010101010.... Normalizing this gives us

(2) Find  $x_{-}$  and  $x_{+}$ .

Note that  $x_{-}$  is just what we found in the previous step, but with 23 bits to the right of the binary point,

 $(1.0101010101010101010101010)_2 \times 2^{-1}$ .

Then,  $x_+$  is

 $(1.0101010101010101010101011)_2 \times 2^{-1}$ .

(3) What is f(x)?

We now consider  $x - x_{-}$  and  $x - x_{+}$ . Here,

• For x-x,

111111111122222222333333333... 123456789012345678901234567...

This gives us

$$0.101... \times 2^{-24} = \frac{2}{3} \times 2^{-24}.$$

• For  $x - x_+$ , we know this will be negative since  $x_+ > x$  (since we're rounding up). So, generally, we'll consider  $x_+ - x$ :

$$x_{+} - x = (x_{+} - x_{-}) - (x - x_{-}) = (2^{-23} \times 2^{-1}) - (\frac{2}{3} \times 2^{-24}) = \frac{1}{3} \times 2^{-24}.$$

Note that  $(x_+ - x_-) = (2^{-23} \times 2^{-1})$  came from

so, we have  $2^{-23} \times 2^{-1}$ .

Notice that  $x_+$  has the smaller error, so  $fl(x) = x^* = x_+$ .

(4) What is the relative round-off error?

This is

$$\frac{|\mathrm{fl}(x) - x|}{|x|} = \frac{|x_+ - x|}{|x|} = \frac{\frac{1}{3} \times 2^{-24}}{\frac{2}{3}} = 2^{-25}.$$

Notice that

$$\frac{|x^* - x|}{|x|} \le 2^{-24}.$$