Math 56 Notes for lec. 3 end, 1/4/14

FLOATING-POINT REPRESENTATION

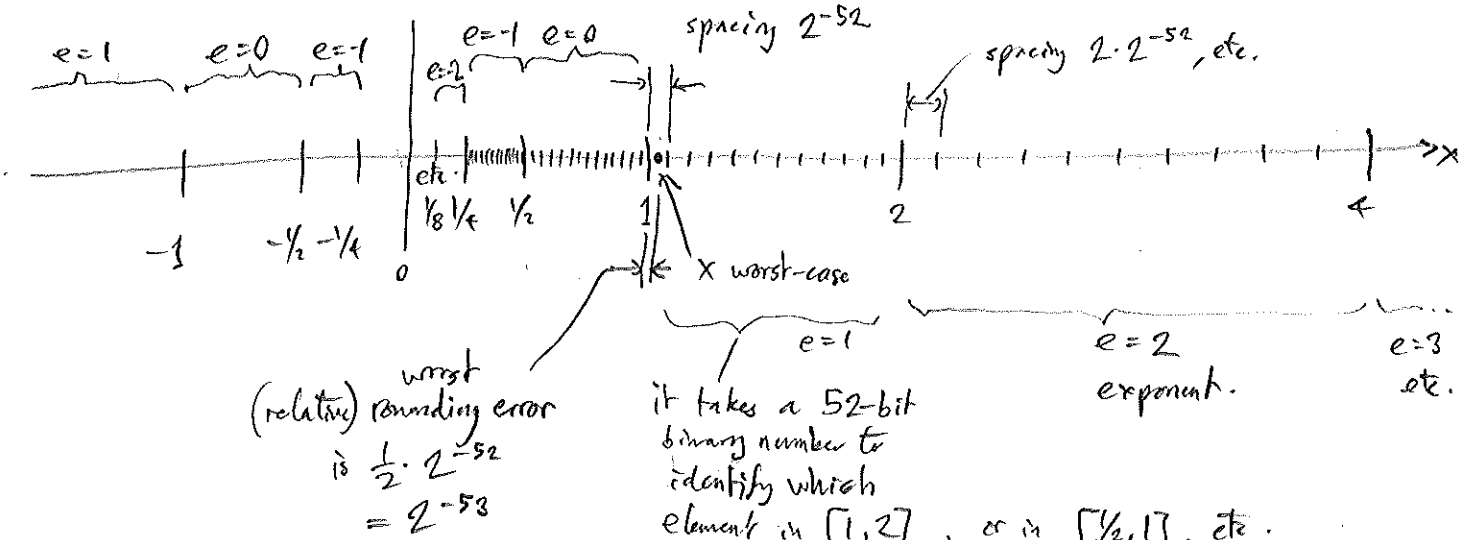
$fl(x)$ means the machine's representation of $x \in \mathbb{R}$, which is approximate.

It has small relative error in the sense that,

$$\boxed{fl(x) = (1+\varepsilon)x \quad \text{for some } |\varepsilon| \le \varepsilon_{mach}}$$

For "IEEE double precision" (the standard), $\varepsilon_{mach} = 2^{-53} \approx 1.1 \times 10^{-16}$

How is this achieved? $fl(x) \in F$, a discrete set:



worst (relative) rounding error is $\frac{1}{2} \cdot 2^{-52} = 2^{-53}$

it takes a 52-bit binary number to identify which element in $[1,2]$, or in $[\frac{1}{2},1]$, etc.

In general base β (here $\beta = 2$), and precision t (here $t = 52$)

So $x \in [\frac{1}{\beta}, 1]$ can be rep. by $fl(x) = \frac{m}{\beta^t}$ where $\beta^{t-1} \le m \le \beta^t$

(integer "mantissa")

To reach the rescaled copies of this, multiply by β^e where e = "exponent", integer.

Double-prec. devotes 11 bits to e, so $-1022 \le e \le 1023$

This means the largest & smallest sized numbers in F are $2^{1023} \approx 10^{308}$ & $2^{-1022} \approx 10^{-308}$.

The full set $F = \{0, \pm \frac{m}{\beta^t}\beta^e, \pm\infty, \text{NaN}\}$ ("not a number")

($\pm\infty$: inf.)

This is packed into 64 bits (8 bytes) as follows: