Lecture 13: Floating Point Arithmetic

$[1,2] = 1, 1+2^{-52}, 1+2\times2^{-52}, \ldots, 2$

largest number $= 1.79\times10^{308}$

smallest number $= 2.23\times10^{-308}$

$[2,4] = 2[1,2] = 2, 2+2^{-51}, 2+2\times2^{-51}, \ldots, 4$

$\varepsilon_{machine} = 2^{-52} \approx 2.22\times10^{-16}$

Floating point numbers F

* For all $x \in \mathbb{R}$, there exists $x' \in F$ such that
$\frac{|x-x'|}{|x|} \leq \varepsilon_{machine}$ → relative error is always $\varepsilon_{machine}$.

* Let $fl: \mathbb{R} \to F$ gives the closest floating point approximation to a real number (rounding).

* For all $x \in \mathbb{R}$, there exists ε with $|\varepsilon| \leq \varepsilon_{machine}$ such that
$$fl(x) = x(1+\varepsilon).$$

Floating point theorem

$\oplus, \ominus, \otimes, \oslash$ → floating point operation

For all $x, y \in F$, there exists ε with $|\varepsilon| < \varepsilon_{machine}$ such that
$$x \circledast y = (x * y)(1+\varepsilon)$$