Feb 19, 2021 Floating point

Approx real numbers + arithmetic on computers

Error in (1) representation
(2) arithmetic

FP is like binary scientific notation

sign, fraction, exponent, base, lead digit

$-5.282 \cdot 10^{4}$

binary

$1.011 \cdot 2^{-11} = \left(1 + \frac{1}{4} + \frac{1}{8}\right) \cdot 2^{-(1+2)}$

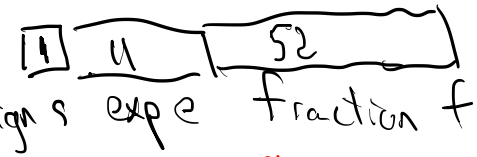
$= 0.171875$

Normalized FP assumes lead digit = 1

$\pm 1.b_1 \cdots b_5 \cdot 2^{p}$

significand

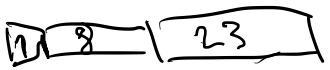
IEEE standard (64-bit)



$(-1)^{s}(1+f) \cdot 2^{e-1023}$

$2^{11} = 2048$

IEEE (32-bit)



$e - 127$

$= \text{value of } s \text{ after iterations} \quad (\hat{s}_{n-1})$

$\hat{s}_0 = 0 = s_0 \qquad \hat{s}_1 = x_1 = s_1$

$\hat{s}_k = fl\left(\left(\hat{s}_{k-1} + x_k\right)\right) = \left(\hat{s}_{k-1} + x_k\right)\cdot(1+\delta_k) \qquad |\delta_k| \le \varepsilon$

$\hat{s}_2 = (x_1 + x_2)(1+\delta_2)$

$\hat{s}_3 = (\hat{s}_2 + x_3)(1+\delta_3) = x_1(1 + \delta_1 + \delta_2 + \delta_2\delta_3) + x_2(1 + \delta_2 + \delta_2 + \delta_2\delta_3)$

$\qquad + x_3(1+\delta_3)$

$\vdots$

$\hat{s}_n = x_1\left(1 + \sum_{j=2}^{n}\delta_j\right) + x_2\left(1 + \sum_{j=2}^{n}\delta_j\right) + x_3\left(1 + \sum_{j=3}^{n}\delta_j\right) + \dots + x_n\left(1 + \sum_{j=n}^{n}\delta_j\right)$

$\tilde{x}_1 \qquad \tilde{x}_2$