Lecture 09 SDP

$\Delta = 2^{-5} = \frac{1}{32} = 0.03125$

$\Delta = 2^{-(\text{number of frac. bits})}$

$= LSB$

quant. step. $= \Delta$

Rounding error: $e_R \in \left[-\frac{\Delta}{2}, \frac{\Delta}{2}\right]$

$8.73 \to 9$

$8.25 \to 8$ $\quad |\Delta_{FL}| \pm 0.5$

$8.5 \to 9$

$8.49 \to 8$

Truncation error: $e_T \in [-\Delta, 0]$

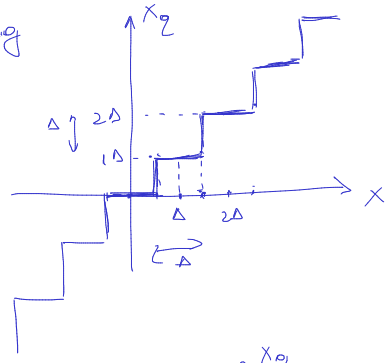
Sign-mag. truncation:

$e \in [-\Delta, 0]$, $x > 0$

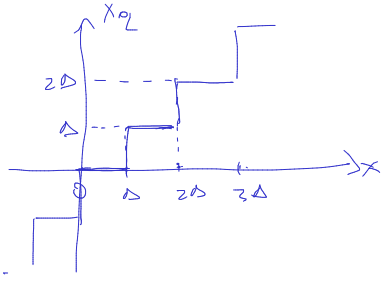
$e \in [0, \Delta]$, $x < 0$

$e_{SMT} \in [-\Delta, +\Delta]$

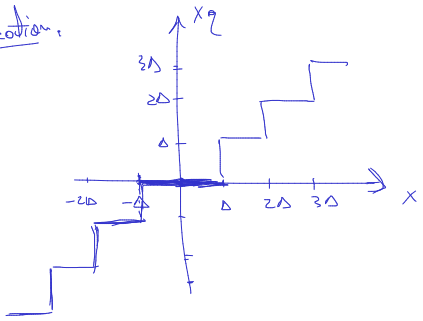
Rounding



Truncation:



Sign-mag. truncation:



Compute $\overline{x^2}$ = power of the rounding errors

$$\overline{x^2} = \int_{-\infty}^{\infty} x^2 \cdot w(x)\, dx = \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} x^2 \cdot \frac{1}{\Delta}\, dx = \frac{1}{\Delta} \cdot \frac{x^3}{3}\Big|_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} = \frac{1}{3\Delta}\left(\frac{\Delta^3}{8} + \frac{\Delta^3}{8}\right) = \frac{\Delta^2}{12}$$

$\Delta = 2^{-b}$

Truncation errors:

$$\overline{x^2} = \int_{-\Delta}^{0} x^2 \cdot \frac{1}{\Delta}\, dx = \frac{1}{\Delta} \cdot \frac{x^3}{3}\Big|_{-\Delta}^{0} = \frac{1}{\Delta}\left(0 + \frac{\Delta^3}{3}\right) = \frac{\Delta^2}{3}$$

4 times bigger than

Sign-mag. truncation

$$\overline{x^2} = \int_{-\Delta}^{\Delta} x^2 \cdot \frac{1}{2\Delta}\, dx = \frac{1}{2\Delta} \cdot \frac{x^3}{3}\Big|_{-\Delta}^{\Delta} = \frac{1}{2\Delta}\left(\frac{\Delta^3}{3} + \frac{\Delta^3}{3}\right) = \frac{\Delta^2}{3}$$

—— // ——