Math 2301 - Summaries

Mossy

Today

Floating Point Aritmetic

Representation

Every real number x has a floating point representation.

$$x = Sb^e$$

- $S = significand$
- $b = base$
- $e = exponent$

Binary

Integer Conversion

x	50	25	12	6	3	1
$\text{mod}\ (x, 2)$	0	1	0	0	1	1

Non Integer Conversion

x	0.3125	0.625	0.25	0.5	0
$2x > 1$	0	0	1	0	1

Rounding Error Analysis

ε is the interval in the computer. A number is rounded to the nearest number able to be composed of an integer number of ε. You either round up or down unless it is exactly in between the two. In this case if d_{p-1} is odd you round up if even down.

- error in $x_* = x_* - x$
- absolute error in $x_* = |x_* - x|$
- relative error in $x_* = \frac{|x_* - x|}{|x|}$

Signifcant Digits

We define the number of significant figures to be

$$max_{d \in \mathbb{Z}} \left(\frac{|x_* - x|}{|x|} < 0.5 \times 10^{-d} \right)$$