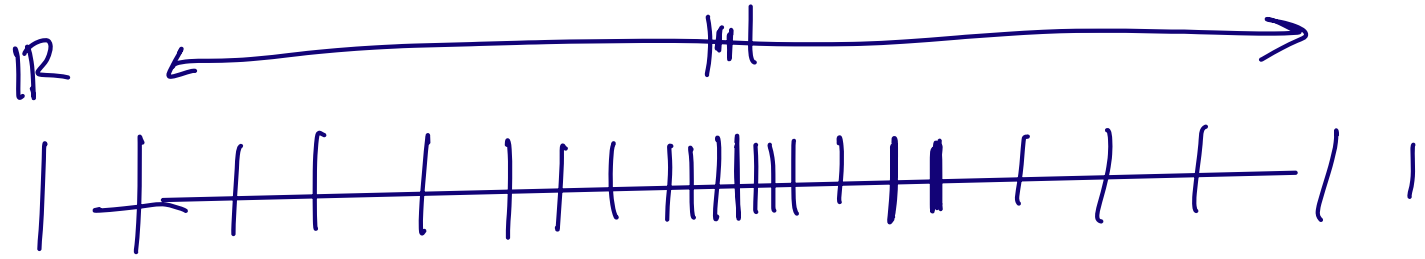
Floating Point Computation



finite precision

1.340×10^{4} 13400

floating point system

- β base
- p # of digits of precision
- $[L, U]$ exponent range $E \in [L, U]$

IEEE 754

		β	p	L	U	
SP	floats32	2	24	-126	127	(1+8+23)
DP	doubles64	2	53	-1022	1023	(1+11+52)

-126: 2, 2^{127}; $1 \ldots 254$; $0 \ldots 255$

floating point number x

$$x = \pm \left(d_0 + \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \cdots + \frac{d_{p-1}}{\beta^{p-1}} \right) \times \beta^{E}$$

$d_0 d_1 \ldots d_{p-1}$ mantissa $\quad L \le E \le U$ exponent

$0 \le d_i < \beta$

$\beta = 2 \quad p = 3 \quad L = -1 \quad U = 1$

	mantissa	E	number ?
+ −	0.00	−1	± 0
	0.00	0	± 0
	0.00	+1	± 0
	0.01	−1	0.001
	0.01	0	0.01
	0.01	+1	0.1
	0.10	−1	0.010
	0.10	0	0.10
	0.10	+1	0 1.0

1	−2
1	−1
1.	0

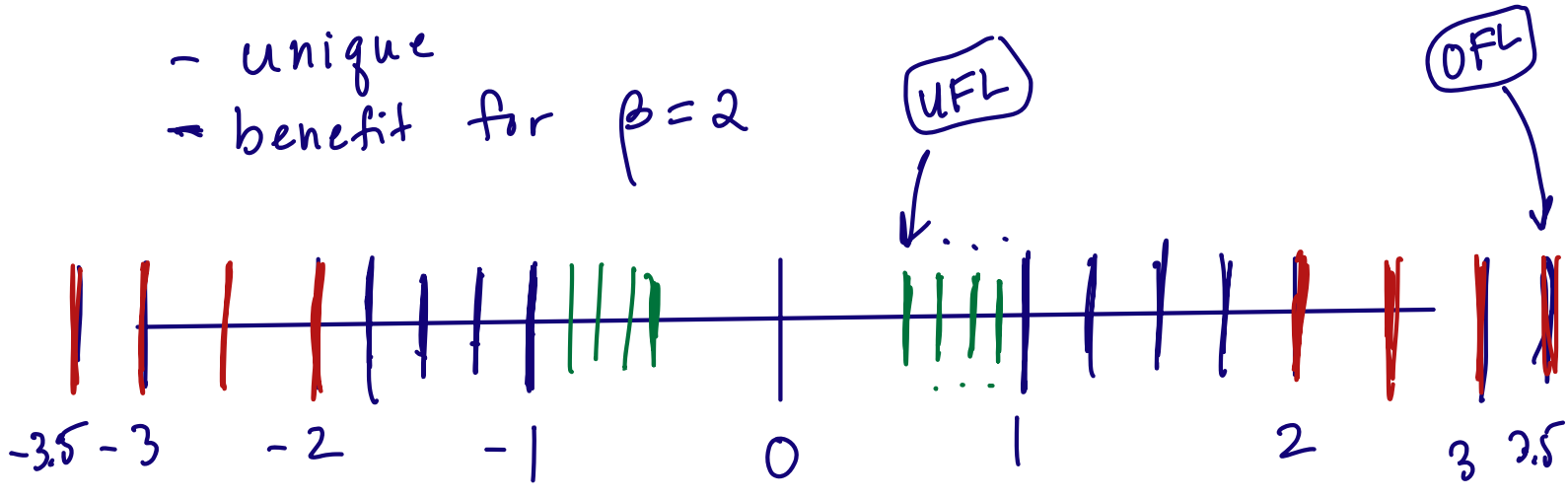
⋮

not unique — lots of duplicates!

Normalization

leading digit in mantissa not 0

- unique
- benefit for $\beta = 2$



mantissa	E	number
0.00	.	± 0
1.00	-1	± 0.1
1.00	0	± 1
1.00	+1	± 10.
1.01	-1	± 0.101
1.01	0	1.01
1.01	+1	10.1
1.10	-1	⋮
	0	
	+1	
1.11	+1	11.1

-1

0.00
0.01
0.10
0.11

$1 \times .01 \times 2^{-1}$

.001

$\frac{1}{2}$ $\frac{1}{4}$ $\frac{1}{8}$

$1.00 \times 2^{-1} = .1 = \frac{1}{2}$

$1.11 \times 2^{-1} = .111 = 7/8$

~~10.11~~

$1.00 \times 2^{0} = 1$

$.01 \times 2^{0} = \frac{1}{4}$

Overflow level $(\beta-1).(\beta-1)\ldots(\beta-1) \times \beta^U$

Underflow level $1.0\overset{p}{\ldots}0 \times \beta^L$

UFL

1.00×2^{-1}

$.01 = \frac{1}{4}$

OFL

1.11×2^1

$11.1 = 3.5$

$\left(\frac{4}{3}\right) - 1 == \left(\frac{1}{3}\right)$ ✗

$\frac{4}{3} == 1 + \frac{1}{3}$ ✓

---

4 digit decimal

$1.333(3\ldots 3\ldots) - 1$

$1.333 - 1 = 0.333 \rightarrow 3.330 \times 10^{-1}$ ✗

$\frac{1}{3} = 3.333 \times 10^{-1}$

$3.3 \times 10^{-1} \stackrel{?}{=} 3.3 \times 10^{-1}$

$1.333 == 1.333$ ✓

3.333×10^{-1}

$1. \times 10^{0}$

0.333×10^{0}

$+\ 1.0 \times 10^{0}$

1.333

$\varepsilon = 1e-10$

$x == y$

$x \neq y$
"equal"

$abs(x - y) < \varepsilon$