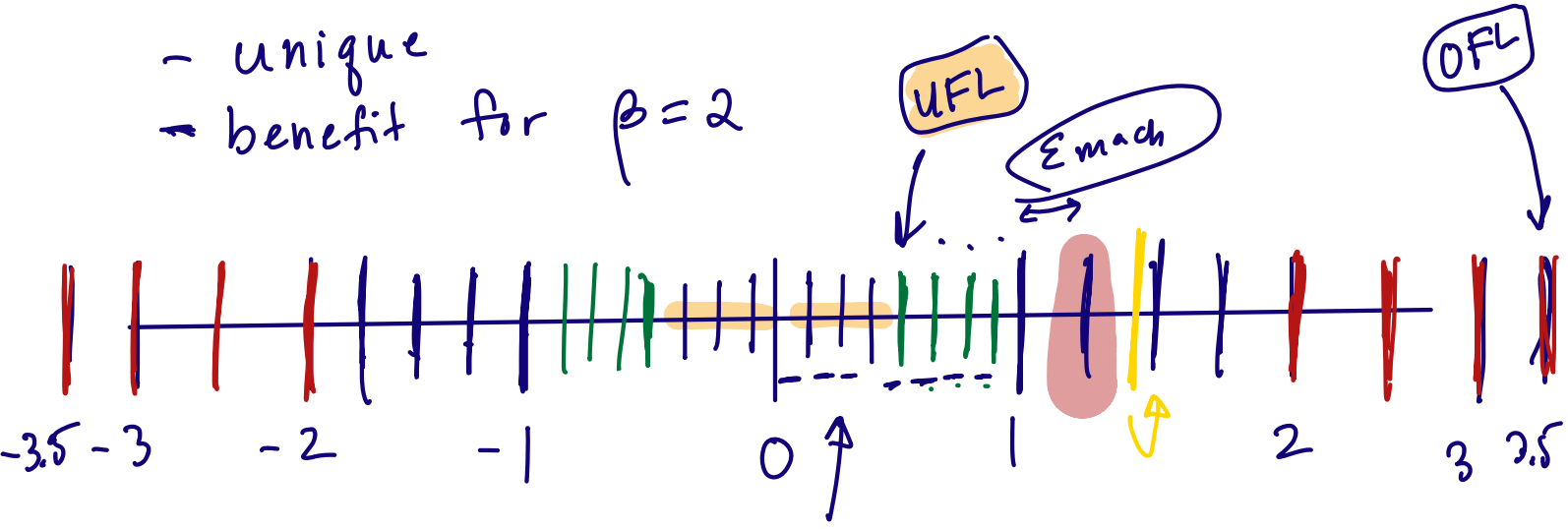
$\beta = 2 \quad p = 3 \quad L = -1 \quad U = 1$

Normalization

leading digit in mantissa not 0

- unique
- benefit for $\beta = 2$



subnormals
denormalized #'s

gradual underflow

	E
$0.0 \ldots 1$	0
$0.010 \ldots 01$	0

$E = -1, 0, 1$

$0.\underline{0}\,\underline{1} \rightarrow \frac{1}{4} \times 2^{-1} = \frac{1}{8}$

$0.10 \quad \frac{1}{2} \times 2^{-1} = \frac{1}{4}$

$0.11 = \frac{3}{8}$

1.0 ... — □

0.0 ... — 01 □

denormalized #'s precision loss

Exceptional values

Inf - #/0 — exceed OFL

NaN

$\frac{0}{0}$ $\frac{Inf}{Inf}$ $0 \cdot Inf$

$$\lim_{x \to 0} \frac{x}{x^2} = \lim_{x \to 0} \frac{1}{x} \to \infty$$

$$\lim_{x \to 0} \frac{x^2}{x} = \lim_{x \to 0} x \to 0$$

Rounding

$x \rightarrow fl(x)$

chop
round to nearest

	chop	
1.649	1.6	1.6
1.650	1.6	1.6
1.651	1.6	1.7
⋮	⋮	
1.701	1.7	1.7

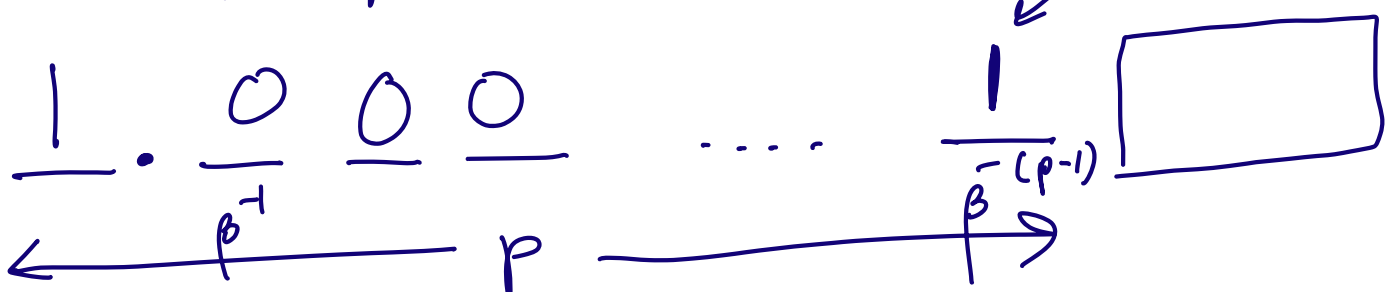
machine epsilon

$\varepsilon_{mach} = \beta^{-(p-1)}$

```
while ((1+δ)==1)
    δ *=2
```

float (32) $p=24$ $\quad 2^{-23} \approx 1.19 \times 10^{-7}$

double (64) $p=53$ $\quad 2^{-52} \approx 2.22 \times 10^{-16}$



unit roundoff

characterizes rounding error

$$\frac{\varepsilon_{mach}}{2}$$

$$fl(1.00005) = 1.001$$

$\times$

absolute error

→ relative error

$|fl(x) - x|$ absolute error

$\frac{|fl(x) - x|}{|x|}$ relative error

$fl(x) = s \cdot m \times \beta^{E}$

$$\frac{|fl(x) - x|}{|x|} \leq \frac{\frac{1}{2}\varepsilon_{mach} \times \beta^{E}}{m \times \beta^{E}}$$

if $fl(x)$ is normalized, $m \geq 1$

$$\leq \frac{1}{2}\varepsilon_{mach}$$

relative error in storing x as normalized $fl(x)$ is

$$\frac{|fl(x) - x|}{|x|} \leq \frac{\varepsilon_{mach}}{2}$$

$fl(0.0055)$ abs err 0.0005

$= 0.006$

relative err $\frac{1}{6}$ $\frac{0.0005}{0.001} = \frac{1}{6} \cdot \frac{1}{2}$

$\frac{8}{11}$ $\sim 8\%$ $\frac{1}{12}$

floats 32 $p = 24$ $\frac{\varepsilon}{2}$ 1.19×10^{-7}

double 64 $p = 53$ 2.22×10^{-16}

10^{-5} %

0.001

0.0005 ✓

rel. err ~~50%~~ 100%

Floating Point Math

+ − × ÷

x 1.235×10^5

y → 1.001×10^{-4}

12.35×10^4 | 1.235×10^5

1.001 | 0.100

1.235

0.000 ... 01001 ←

$fl(x + y) = x$

$x + y - x = y$ | $x - x + y = y$

$x - x = 0$

cancellation error

1.235217 − 1.23516 = 0.000383

1.235 − 1.236 = −0.001

---

multiplication

- overflow
- underflow

$\frac{x^2}{x} = x$

Responses to student questions after class

$1.\underline{0}\ \underline{0}\ \ldots\ \underline{0}\ \underline{1}$ $\quad p = 24$

2^{-23}

general FP system

β = base $\quad \beta = 2$

p = digits of precision

L = smallest exponent $\quad -126\ (1)$

U = largest exponent $\quad 127\ (254)$

biased exponent

0 ... 255

-127 .... 128 $\quad$ bias 127

1

$fl(x) \quad s \quad m \quad E$

$$fl(x) = S \times \underbrace{m_0 . m_1 \ldots m_{p-1}}_{m} \times \beta^{E}$$

1 $\quad$ 8 $\quad$ 23 $\quad (+1)^m \quad p = 24$

11 $\quad$ 52 $\quad +1 \quad p = 53$