Floating Point

Math 426

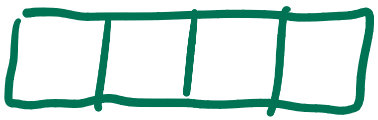
University of Alaska Fairbanks

September 21, 2020

Signed Integers (Offset Binary)

E.g., 4-bit representation of signed integers.

Unsigned:



0,1

2^4 possibilities.

$0 - 15 = 2^4 - 1$

$0000 \to 0$

$0001 \to 1$

$0010 \to 2$

$\vdots$

$1111 \to 8+4+2+1 = 15$

Signed Integers (Offset Binary)

E.g., 4-bit representation of signed integers.

Offset binary subtract off an offset (bias)

0000 → −7
0001 → −6
0010 → −5
⋮
1000 → 1
⋮
$\underbrace{1111}_{15}$ → 8

offset for 4 bits is $1-2^3$

Signed Integers (Offset Binary)

E.g., 4-bit representation of signed integers.

In most other applications, two's complement is used!

Unsigned Fixed Point Numbers

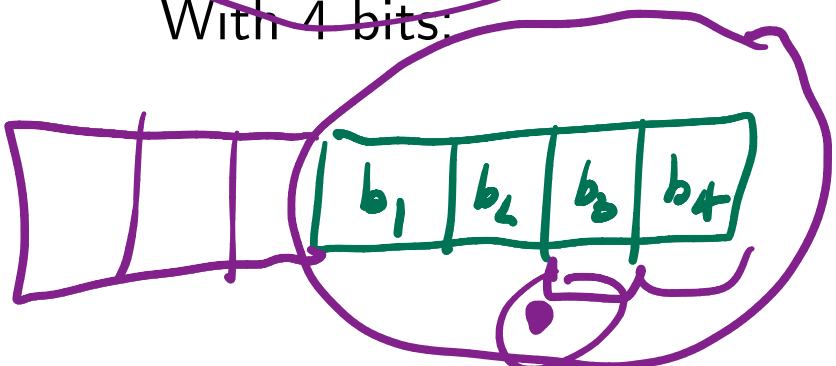
2^{-7} 1/128

$0 \le x \le 2$

$/\,2^k$

$2^2 = 4$

With 4 bits:



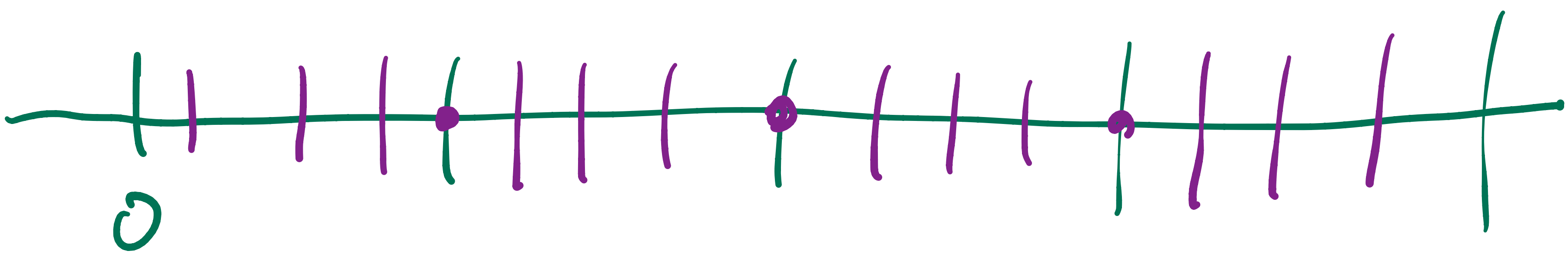
$(b_1 b_2 . b_3 b_4)_2$

smallest $(0\,0.0\,1)_2 = \frac{1}{4}$

biggest $(1\,1.1\,1)_2 = 3\,{}^3/_4$

end

4



Floating Point Numbers

in between we use

m.

$(-1)^s$

$$s \cdot m \cdot 2^E$$

- s, sign, ± 1
- m, mantissa, $1 \leq m < 2$ → fixed point number
- E, an integer, exponent

→ offset binary (signed number)

1: $s = 0$, $m = 1$, $E = 0$ $(-1)^0 \cdot 1 \cdot 2^0 = 1$

2: $s = 0$, $m = 1$, $E = 1$

Fictional 8-bit format

s	e_1	e_2	e_3	b_1	b_2	b_3	b_4

- sign: 1 bit $s=0 \to +$, $s=1 \to -$ $(-1)^s$
- exponent: 3 bits
- mantissa: 4 bits

$1 \le m < 2$

$1.000)_2$ 1

1.001_2 $1+\frac{1}{8}$

$\vdots$ $\vdots$

1.111_2 $1+7/8$

hidden bit format.

The 1 is implicit.

$(1.b_1b_2b_3b_4)_2$

$1.0000_2 = 1$

$1.0001_2 = 1+\frac{1}{16}$

$\vdots$

$1.1111_2 = 1+\frac{15}{16}$

Fictional 8-bit format

- sign: 1 bit
- exponent: 3 bits
- mantissa: 4 bits

Mantissa is in unsigned binary. Exponent is signed, offset binary.

e_1	e_2	e_3

0	0	0	-3
0	0	1	-2
0	1	0	-1
0	1	1	0
1	0	0	1
1	0	1	2
1	1	0	3
1	1	1	4

2^3 $\quad$ 2^2 pos.

2^2-1 neg.

$-(2^2-1)$ to 2^2

-3 to 4

Hidden bit representation

All numbers between 1 and 2 in base two start with a 1, so we can save a bit and gain precision by making the 1 implicit.

Hidden bit representation

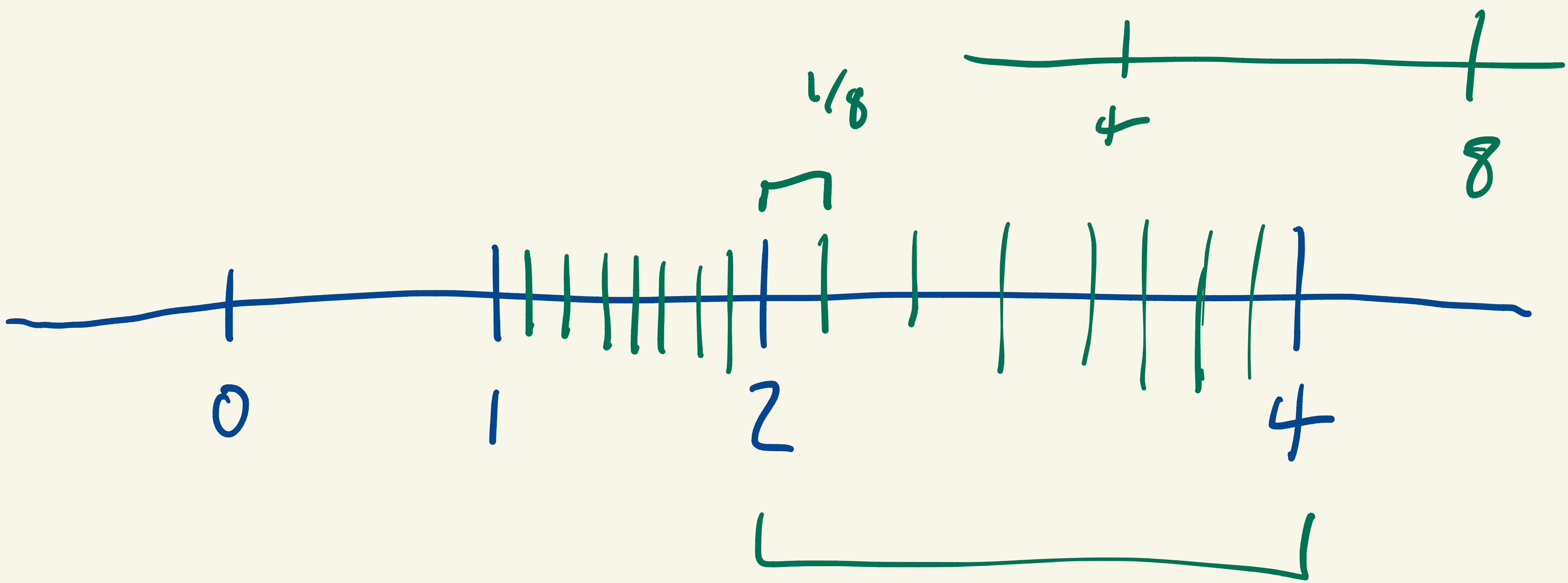
All numbers between 1 and 2 in base two start with a 1, so we can save a bit and gain precision by making the 1 implicit. All available numbers:

Available numbers:

$$\pm \left[1, 1\tfrac{1}{16}, \dots, 1\tfrac{15}{16}\right] \times \underbrace{2^{[-3,-2,\dots,4]}}_{\frac{1}{8}, \frac{1}{4}, \dots, 8, 16}$$

Available numbers:

$$\pm \underbrace{\left[1, 1\tfrac{1}{16}, \ldots, 1\tfrac{15}{16}\right]} \times \underbrace{2^{[-3,-2,\ldots,4]}}_{\frac{1}{8},\ \frac{1}{4}\ \ldots\ 8,\ 16}$$

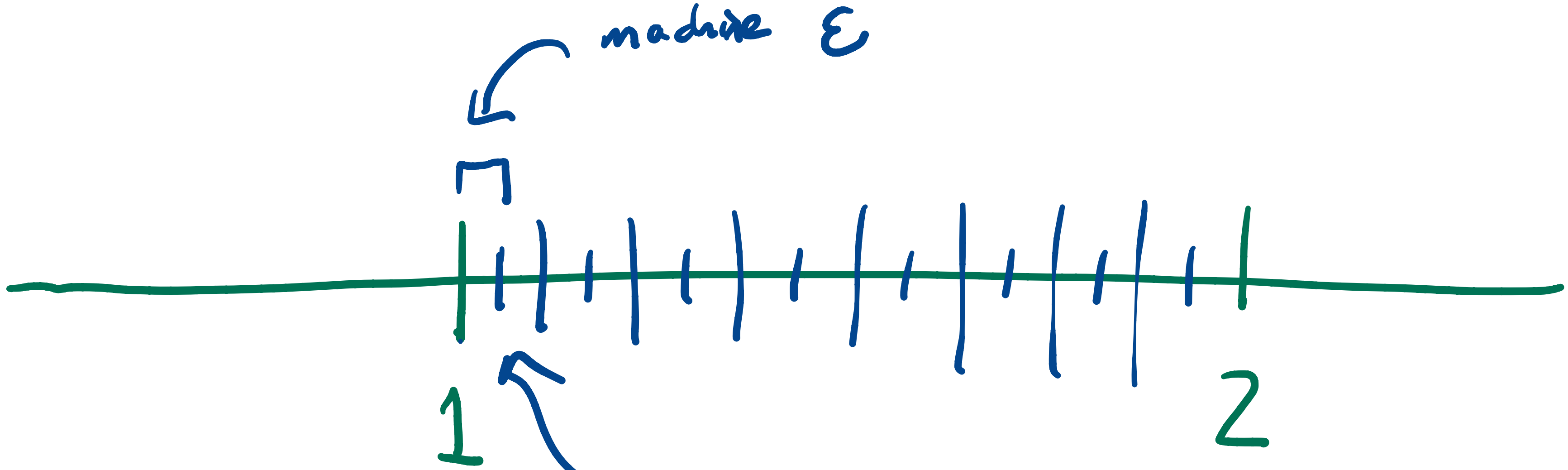


$E = 1,\ 2^1$

$s = 0$

$E = 0 \quad 2^0$

Precision and Machine Epsilon



8 bit format $\varepsilon = 1/16$

$$\left[\begin{array}{l} 1+\varepsilon \\ 1+2\varepsilon \\ \vdots \\ 2-\varepsilon \end{array}\right.$$

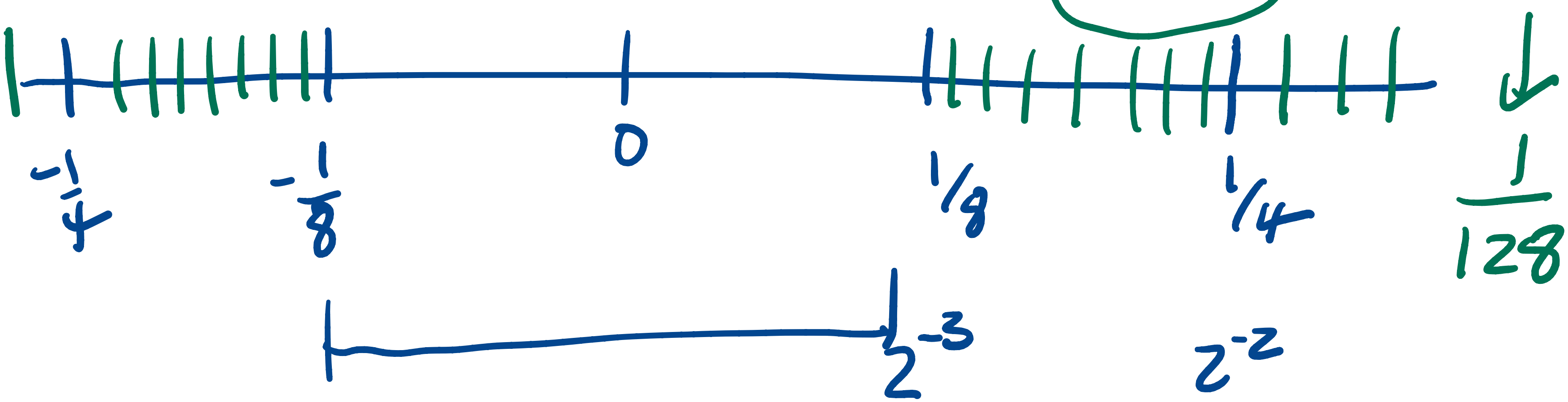
The Monster Gap Around Zero

Smallest positive number

$$\pm \underbrace{\left[1,\ 1\tfrac{1}{16},\ \ldots,\ 1\tfrac{15}{16}\right]} \times 2^{\underbrace{[-3,-2,\ \ldots,4]}_{1}}$$

$$\frac{1}{8} + \frac{1}{8}\cdot\frac{1}{16}$$

$$2^{-3}\cdot 2^{-4} \to 2^{-7}$$



Subnormal Numbers (And two zeros!)

Give up some exponents:

0 0 0 $\to$ −3	$1-2^2$	
0 0 1 $\to$ −2	⋮	$2-2^2$
⋮ ⋮	⋮	⋮
1 1 0 $\to$ 3	⋮	2^2-1
1 1 1 $\to$ 4	2^2	

Subnormal Numbers (And two zeros!)

$000 \longmapsto$ very small

$111 \longmapsto$ wierd

$$\boxed{s}\ \boxed{0\,|\,0\,|\,0}\ \boxed{0\,|\,0\,|\,0\,|\,0} \longrightarrow (-1)^s \cdot 0$$

$s = 0 \longrightarrow +0$

$s = 1 \longrightarrow -0$

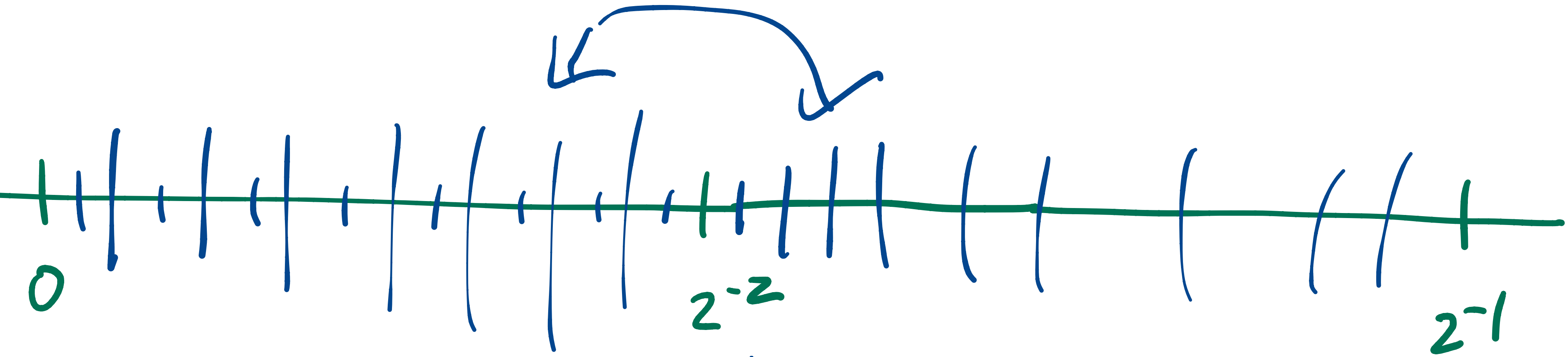
$s \quad 0\,0\,0 \quad b_1 \; b_2 \; b_3 \; b_4$

$$(-1)^s \, (0.b_1 b_2 b_3 b_4)_2 \cdot 2^{-2}$$

$1/16$

Subnormal Numbers (And two zeros!)

same step size



have less precision

subnormal numbers

steps of size $2^{-2} \cdot \varepsilon$

Infinity

$e: 111 \rightarrow \cancel{4}$

Inf

-Inf

Exponent is all 1's. Mantissa is all 0's. $\rightarrow \pm\infty$

Positive infinity: 0 111 0000

Negative infinity: 1 111 0000

$x + y$

$\infty + ?$

$\infty + 5 = \infty$

$7/\infty \rightarrow 0$

$-7/\infty \rightarrow -0$

Infinity

Exponent is all 1's. Mantissa is all 0's.

Positive infinity: 0   111   0000

Negative infinity: 1   111   0000

Infinity

Exponent is all 1's. Mantissa is all 0's.

Positive infinity: 0   111   0000

Negative infinity: 1   111   0000

$$\frac{1}{+0} \rightarrow +\mathrm{Inf}$$

$$\frac{1}{-0} \rightarrow -\mathrm{Inf}$$

Any other pattern s 111 $b_1 b_2 b_3 b_4$ is Not a Number (NaN).

$x = \mathrm{NaN}$

$$\frac{0}{0} \rightarrow \mathrm{NaN}$$

IEEE 754

Single precision: 32 bits.

1. sign: 1 bit
2. exponent: 8 bits
3. mantissa: 23 bits

Machine epsilon: $2^{-23} \approx 2.2 \times 10^{-7}$.

Smallest (normal) number: $2^{-126} \approx$