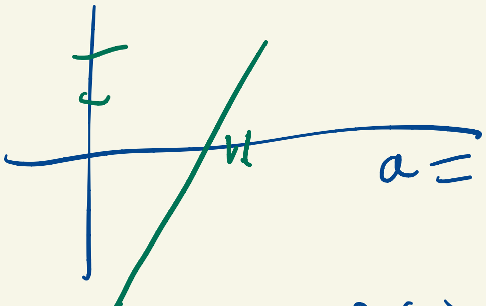
Floating Point

Math 426

University of Alaska Fairbanks

September 23, 2020

$$f(x) = f(a) + f'(a)(x-a) + \underbrace{O((x-a)^2)}_{R}$$



$a = x_*$

$$f(x) = 0 + f'(x_*)(x - x_*) + O((x-x_*)^2)$$

$$f'(x_*) = 0$$

$$\frac{f''(\xi)}{2}(x-x_*)^2$$

$$|f(x_k)| \approx |f'(x_*)| \underbrace{|x_k - x_*|}_{|e_k|} + O(e_k^2)$$

Last class

Ficitonal 8-bit floating point

1. 1 sign bit
2. 3 exponent bits (offset binary)
3. 4 mantissa bits (hidden bit representation)

$$\boxed{s}\ \underbrace{\boxed{e_1 \mid e_2 \mid e_3}}_{2^3 \text{ patterns}}\ \boxed{b_1 \mid b_2 \mid b_3 \mid b_4}$$

$s_1 e_1 e_2 e_3 b_1 b_2 b_3 b_4$ represents

$$(-1)^{s_1}(1.b_1 b_2 b_3 b_4)_2 \times 2^E$$

where $E = (e_1 e_2 e_3)_2 - \Omega$; offset $\Omega = 2^2 - 1$

$0 - (2^2 - 1) = 1 - 2^2$

2^2 negative

2^2 positive

Last class

Ficitonal 8-bit floating point

1. 1 sign bit
2. 3 exponent bits (offset binary)
3. 4 mantissa bits (hidden bit representation)

$s_1 e_1 e_2 e_3 b_1 b_2 b_3 b_4$ represents

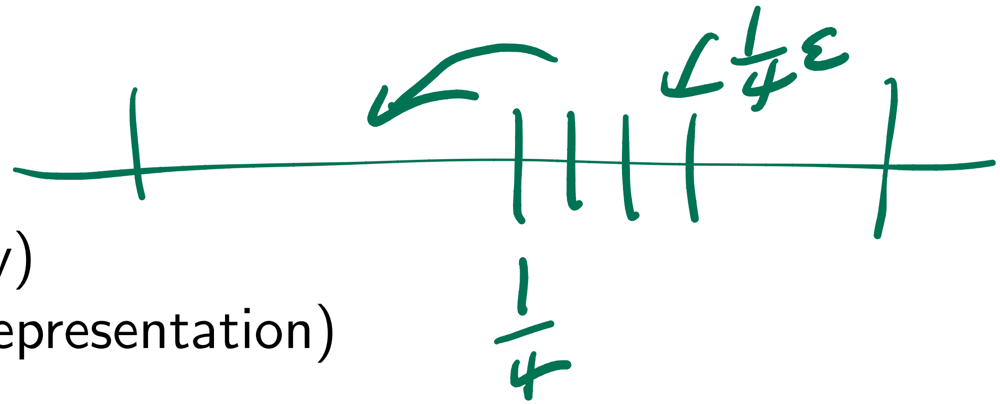
$$(-1)^{s_1}(1.b_1 b_2 b_3 b_4)_2 \times 2^E$$

where $E = (e_1 e_2 e_3)_2 - \Omega$; offset $\Omega = 2^2 - 1$

Machine ϵ: 2^{-4}

Last class

Ficitonal 8-bit floating point



1. 1 sign bit
2. 3 exponent bits (offset binary)
3. 4 mantissa bits (hidden bit representation)

$s_1 e_1 e_2 e_3 b_1 b_2 b_3 b_4$ represents

$$(-1)^{s_1}(1.b_1 b_2 b_3 b_4)_2 \times 2^E$$

where $E = (e_1 e_2 e_3)_2 - \Omega$; offset $\Omega = 2^2 - 1$

Machine ϵ: 2^{-4}

$1, \quad 1+\varepsilon$

000
001
⋮
111

except, if $e_1 e_2 e_3 = 000$:

$$(-1)^{s_1}(0.b_1 b_2 b_3 b_4)_2 \times 2^{(001)_2 - \Omega}$$

(two zeros; remainder are subnormal)

Last class

Ficitonal 8-bit floating point

1. 1 sign bit
2. 3 exponent bits (offset binary)
3. 4 mantissa bits (hidden bit representation)

$s_1 e_1 e_2 e_3 b_1 b_2 b_3 b_4$ represents

$$(-1)^{s_1}(1.b_1 b_2 b_3 b_4)_2 \times 2^E$$

where $E = (e_1 e_2 e_3)_2 - \Omega$; offset $\Omega = 2^2 - 1$

Machine ϵ: 2^{-4}

except, if $e_1 e_2 e_3 = 000$:

$$(-1)^{s_1}(0.b_1 b_2 b_3 b_4)_2 \times 2^{(001)_2 - \Omega}$$

(two zeros; remainder are subnormal) except, if $e_1 e_2 e_3 = 111$:

If $b_1 b_2 b_3 b_4 = 0000$, is $(-1)^s \infty$

Otherwise, is NaN.

IEEE 754

Single precision: 32 bits.

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

$2^8/2 = 2^7$

2^7 neg.

2^7 pos

-2^7+1

Offset $\Omega = 2^7 - 1$.

$$(-1)^s(1.b_1 \cdots b_{23})2^E$$

where $E = (e_1 \cdots e_8)_2 - \Omega$.

IEEE 754

Single precision: 32 bits.

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

Offset $\Omega = 2^7 - 1$.

$$(-1)^s(1.b_1 \cdots b_{23})2^E$$

where $E = (e_1 \cdots e_8) - \Omega$. $\boxed{\text{Machine } \epsilon?}$

$$(-1)^0(1.0\ldots0)2^0$$

$$(-1)(1.0\ldots01)2^0$$

$$\begin{array}{ccc} \uparrow & \uparrow \ \cdots & \uparrow \\ 1 & 2 & 23 \end{array}$$

$2^{-23} = \varepsilon$

IEEE 754

Single precision: 32 bits.

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

Offset $\Omega = 2^7 - 1$.

$$(-1)^s(1.b_1 \cdots b_{23})2^E$$

where $E = (e_1 \cdots e_8) - \Omega$. Machine ϵ? $2^{-23} =\approx 1.1 \times 10^{-7}$

IEEE 754

Single precision: 32 bits.

$1-(2^7-1) = 2-2^7$

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

Offset $\Omega = 2^7 - 1$.

$$(-1)^s \underbrace{(1.b_1 \cdots b_{23})}2^E$$

where $E = (e_1 \cdots e_8)_2 - \Omega$. Machine ϵ? $2^{-23} = \approx 1.1 \times 10^{-7}$

Smallest normal number?

positive

$(-1)^0 (1.0 \ldots 0) 2^E$

IEEE 754

Single precision: 32 bits.

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

Offset $\Omega = 2^7 - 1$.

$$(-1)^s(1.b_1 \cdots b_{23})2^E$$

where $E = (e_1 \cdots e_8) - \Omega$. Machine ϵ? $2^{-23} = \approx 1.1 \times 10^{-7}$

Smallest normal number?

$$2^{1-\Omega} = 2^{2-2^7} = 2^{-126} \approx 1.1 \times 10^{-38}$$

IEEE 754

Single precision: 32 bits.

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

Offset $\Omega = 2^7 - 1$.

$$(-1)^s (1.b_1 \cdots b_{23}) 2^E$$

where $E = (e_1 \cdots e_8) - \Omega$. Machine ϵ? $2^{-23} =\approx 1.1 \times 10^{-7}$

Smallest normal number?

$$2^{1-\Omega} = 2^{2-2^7} = 2^{-126} \approx 1.1 \times 10^{-38}$$

Largest number?

$$(-1)^0 \overbrace{(1.1 \cdots 1)}^{2-\varepsilon} \cdot 2^E$$

$0.0 \cdots 01$

11111110_2

100000000

$(2^8 - 2 - \Omega)$

IEEE 754

Single precision: 32 bits.

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

Offset $\Omega = 2^7 - 1$.

$$(-1)^s (1.b_1 \cdots b_{23}) 2^E$$

where $E = (e_1 \cdots e_8) - \Omega$. Machine ϵ? $2^{-23} = \approx 1.1 \times 10^{-7}$

Smallest normal number?

$$2^{1-\Omega} = 2^{2-2^7} = 2^{-126} \approx 1.1 \times 10^{-38}$$

Largest number?

$$2^{(2^8-2)-\Omega}(2-\epsilon) = 2^{2^8-2-2^7+1} = 2^{127}(2-\epsilon) \approx 2^{128} \approx 3.4 \times 10^{38}$$

IEEE 754

~~Single~~ Double precision: 64 bits.

1. 1 sign bit
2. 11 exponent bits
3. 52 mantissa bits

Offset $\Omega = 2^{10} - 1$.

$$(-1)^s (1.b_1 \cdots b_{52}) 2^E$$

where $E = (e_1 \cdots e_{11})_2 - \Omega$.

e_{11}

IEEE 754

Single precision: 64 bits.

1. 1 sign bit
2. 11 exponent bits
3. 52 mantissa bits

Offset $\Omega = 2^{10} - 1$.

$$(-1)^s(1.b_1 \cdots b_{52})2^E$$

where $E = (e_1 \cdots e_11)_2 - \Omega$. Machine ϵ?

2^{-52}

IEEE 754

Single precision: 64 bits.

1. 1 sign bit
2. 11 exponent bits
3. 52 mantissa bits

Offset $\Omega = 2^{10} - 1$.

$$(-1)^s (1.b_1 \cdots b_{52}) 2^E$$

where $E = (e_1 \cdots e_1 1)_2 - \Omega$. Machine ϵ? 2^{-52} $=\approx$ $\boxed{2.2 \times 10^{-16}}$

IEEE 754

Single precision: 64 bits.

1. 1 sign bit
2. 11 exponent bits
3. 52 mantissa bits

Offset $\Omega = 2^{10} - 1$.

$$(-1)^s(1.b_1 \cdots b_{52})2^E$$

where $E = (e_1 \cdots e_1 1)_2 - \Omega$. Machine ϵ? $2^{-52} = \approx 2.2 \times 10^{-16}$

Smallest normal number?

IEEE 754

Single precision: 64 bits.

1. 1 sign bit
2. 11 exponent bits
3. 52 mantissa bits

Offset $\Omega = 2^{10} - 1$.

$$(-1)^s(1.b_1 \cdots b_{52})2^E$$

where $E = (e_1 \cdots e_1 1)_2 - \Omega$. Machine ϵ? $2^{-52} =\approx 2.2 \times 10^{-16}$

Smallest normal number?

$$2^{1-\Omega} = 2^{2-2^{10}} = 2^{-1022} \approx 2.2 \times 10^{-308}$$

IEEE 754

Single precision: 64 bits.

1. 1 sign bit
2. 11 exponent bits
3. 52 mantissa bits

Offset $\Omega = 2^{10} - 1$.

$$(-1)^s(1.b_1 \cdots b_{52})2^E$$

where $E = (e_1 \cdots e_1 1)_2 - \Omega$. Machine ϵ? $2^{-52} = \approx 2.2 \times 10^{-16}$

Smallest normal number?

$$2^{1-\Omega} = 2^{2-2^{10}} = 2^{-1022} \approx 2.2 \times 10^{-308}$$

Largest number?

$(2-\varepsilon) \cdot 2^{(2^{11}-2)-\Omega}$

IEEE 754

Single precision: 64 bits.

1. 1 sign bit
2. 11 exponent bits
3. 52 mantissa bits

Offset $\Omega = 2^{10} - 1$.

$$(-1)^s (1.b_1 \cdots b_{52}) 2^E$$

where $E = (e_1 \cdots e_1 1)_2 - \Omega$. Machine ϵ? $2^{-52} = \approx 2.2 \times 10^{-16}$

Smallest normal number?

$$2^{1-\Omega} = 2^{2-2^{10}} = 2^{-1022} \approx 2.2 \times 10^{-308}$$

Largest number?

$$2^{(2^1 1-2)-\Omega}(2-\epsilon) = 2^{1023}(2-\epsilon) \approx 1.8 \times 10^{308}$$

Matlab demo

$$\underbrace{(1.0\cdots0}1)_2 \times 2^{-1022}/2$$

Machine epsilon: `eps`

`showfloat`: (s,e,m) bit patterns; I'll put it on the website.

$$0 \quad 0\cdots01 \quad 0 \cdots\cdots 0$$

$$1.0\cdots\underset{\substack{\uparrow \\ 52}}{0} \times 2^{-1022} \Rightarrow 2^{-1022}$$

$$2^{-1023}$$

$$0.10\cdots0 \times 2^{-1022}$$