Floating Point Formats

- Scientific notation:

$$\underbrace{-}_{\text{sign}}\ \underbrace{1.602}_{\text{significand}} \times \underbrace{10}_{\text{base}}{}^{\overbrace{}^{}}\underbrace{^{-19}}_{\text{exponent}}$$

- Floating point representation

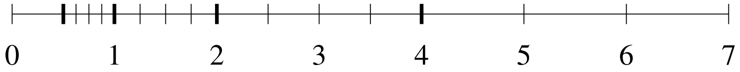
$$\pm\left(d_0 + d_1\beta^{-1} + \ldots + d_{p-1}\beta^{-(p-1)}\right)\beta^e, \quad 0 \le d_i < \beta$$

with *base* β and *precision* p

- Exponent range $[e_{\min}, e_{\max}]$
- *Normalized* if $d_0 \neq 0$ (use $e = e_{\min} - 1$ to represent 0)

Floating Point Numbers

- The gaps between adjacent numbers scale with the size of the numbers
- Relative resolution given by *machine epsilon*, $\epsilon_{\text{machine}} = .5\beta^{1-p}$
- For all x, there exists a floating point x' such that $|x - x'| \leq \epsilon_{\text{machine}}|x|$
- Example: $\beta = 2, p = 3, e_{\min} = -1, e_{\max} = 2$



Special Quantities

- $\pm\infty$ is returned when an operation overflows
- $x/\pm\infty = 0$ for any number x, $x/0 = \pm\infty$ for any nonzero number x
- Operations with infinity are defined as limits, e.g.

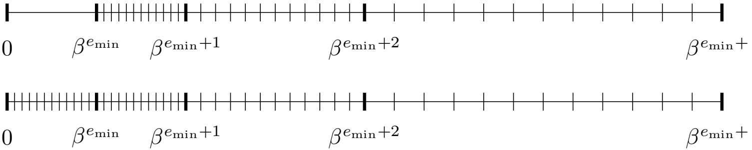
$$4 - \infty = \lim_{x \to \infty} 4 - x = -\infty$$

- NaN (Not a Number) is returned when the an operation has no well-defined finite or infinite result
- Examples: $\infty - \infty$, ∞/∞, $0/0$, $\sqrt{-1}$, $\mathrm{NaN} \odot x$

Denormalized Numbers

- With normalized significand there is a “gap” between 0 and $\beta^{e_{\min}}$
- This can result in $x - y = 0$ even though $x \neq y$, and code fragments like **if** $x \neq y$ **then** $z = 1/(x - y)$ might break
- Solution: Allow non-normalized significand when the exponent is $e_{\min}$
- This *gradual underflow* garantees that

$$x = y \Longleftrightarrow x - y = 0$$



IEEE Single Precision

- 1 sign bit, 8 exponent bits, 23 significand bits:

0	00000000	00000000000000000000000000000000
S	E	M

- Represented number:

$$(-1)^S \times 1.M \times 2^{E-127}$$

- Special cases:

	$E = 0$	$0 < E < 255$	$E = 255$
$M = 0$	± 0	Powers of 2	$\pm\infty$
$M \neq 0$	Denormalized	Ordinary numbers	NaN

IEEE Single Precision, Examples

S	E	M	Quantity
0	11111111	00000100000000000000000	NaN
1	11111111	00100010001001010101010	NaN
0	11111111	00000000000000000000000	∞
0	10000001	10100000000000000000000	$+1 \cdot 2^{129-127} \cdot 1.101 = 6.5$
0	10000000	00000000000000000000000	$+1 \cdot 2^{128-127} \cdot 1.0 = 2$
0	00000001	00000000000000000000000	$+1 \cdot 2^{1-127} \cdot 1.0 = 2^{-126}$
0	00000000	10000000000000000000000	$+1 \cdot 2^{-126} \cdot 0.1 = 2^{-127}$
0	00000000	00000000000000000000001	$+1 \cdot 2^{-126} \cdot 2^{-23} = 2^{-149}$
0	00000000	00000000000000000000000	0
1	00000000	00000000000000000000000	-0
1	10000001	10100000000000000000000	$-1 \cdot 2^{129-127} \cdot 1.101 = -6.5$
1	11111111	00000000000000000000000	$-\infty$

IEEE Floating Point Data Types

	Single precision	Double precision
Significand size (p)	24 bits	53 bits
Exponent size	8 bits	11
Total size	32 bits	64 bits
$e_{\max}$	+127	+1023
$e_{\min}$	-126	-1022
Smallest normalized	$2^{-126} \approx 10^{-38}$	$2^{-1022} \approx 10^{-308}$
Largest normalized	$2^{127} \approx 10^{38}$	$2^{1023} \approx 10^{308}$
$\epsilon_{\text{machine}}$	$2^{-24} \approx 6 \cdots 10^{-8}$	$2^{-53} \approx 10^{-16}$

Floating Point Arithmetic

- Define $\mathrm{fl}(x)$ as the closest floating point approximation to x
- By the definition of $\epsilon_{\mathrm{machine}}$, we have for the relative error:

  For all $x \in \mathbb{R}$, there exists ϵ with $|\epsilon| \leq \epsilon_{\mathrm{machine}}$
  such that $\mathrm{fl}(x) = x(1+\epsilon)$

- The result of an operation $\circledast$ using floating point numbers is $\mathrm{fl}(a \circledast b)$
- If $\mathrm{fl}(a \circledast b)$ is the nearest floating point number to $a \circledast b$, the arithmetic *rounds correctly* (IEEE does), which leads to the following property:

  For all floating point x, y, there exists ϵ with $|\epsilon| \leq \epsilon_{\mathrm{machine}}$ such that
  $x \circledast y = (x * y)(1+\epsilon)$

- Round to nearest even in the case of ties