IEEE Floating Point Arithmetic

Representation of Double Precision Numbers

64 binary digits (bits) for each floating point number

$$f = \pm (1.b_1 b_2 \ldots b_{52})_2 \times 2^{(a_1 a_2 \ldots a_{11})_2}$$

Representation of Double Precision Numbers

64 binary digits (bits) for each floating point number

$$f = \pm (1.b_1 b_2 \ldots b_{52})_2 \times 2^{(a_1 a_2 \ldots a_{11})_2}$$

- 52 bits for the significand (mantissa)
- 11 bits for the exponent
- 1 bit for the sign

Representation of Double Precision Numbers

- 11 bits can be used to represent $2^{11} = 2048$ exponent values.

Representation of Double Precision Numbers

- 11 bits can be used to represent $2^{11} = 2048$ exponent values.
- $(00\ldots0)_2$ and $(11\ldots1)_2$ are reserved for special purposes.
  - $(11\ldots1)_2$ for ∞ and *NaN* (not a number *e.g.* $\infty - \infty$).

Representation of Double Precision Numbers

- 11 bits can be used to represent $2^{11} = 2048$ exponent values.
- $(00\ldots0)_2$ and $(11\ldots1)_2$ are reserved for special purposes.
  - $(11\ldots1)_2$ for ∞ and *NaN* (not a number *e.g.* $\infty - \infty$).
- The remaining 2046 exponent values represent any integer in $[-1022, 1023]$.

Representation of Double Precision Numbers

Let x be any floating point number in double precision.

$$-(1.11\ldots1)_2 \cdot 2^{1023} \leq x \leq (1.11\ldots1)_2 \cdot 2^{1023}$$

Representation of Double Precision Numbers

Let x be any floating point number in double precision.

$$-(1.11\ldots1)_2 \cdot 2^{1023} \leq x \leq (1.11\ldots1)_2 \cdot 2^{1023}$$

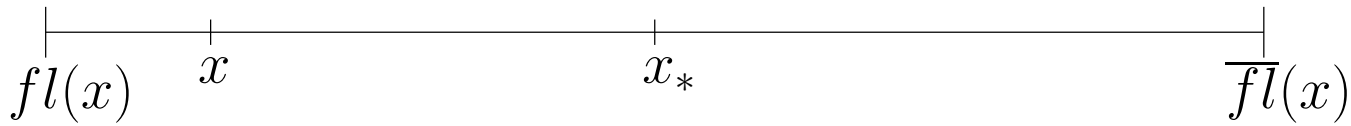
$$-\{(10.0\ldots0)_2 - (0.0\ldots1)_2\} \cdot 2^{1023} \leq x \leq \{(10.0\ldots0)_2 - (0.0\ldots1)_2\} \cdot 2^{1023}$$

$$-(2 - 2^{-52}) \cdot 2^{1023} \leq x \leq (2 - 2^{-52}) \times 2^{1023} \approx 1.8 \times 10^{308}$$

Representation of Double Precision Numbers

ϵ_{mach} (machine precision)

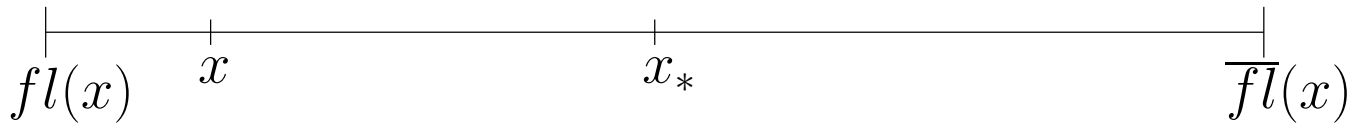
Maximal relative error due to floating point representation



Representation of Double Precision Numbers

ϵ_{mach} (machine precision)

Maximal relative error due to floating point representation

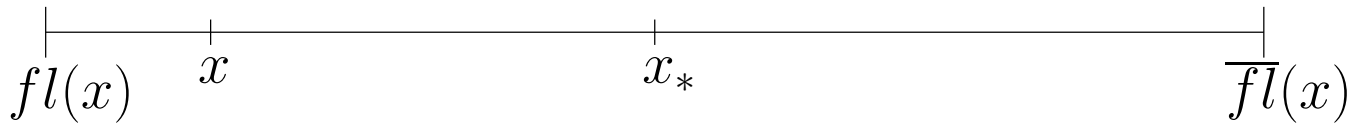


$x = s \times 2^E \in (R_{\min}, R_{\max})$

$fl(x) = \hat{s} \times 2^E$ (floating point number closest to x)

Representation of Double Precision Numbers

ϵ_{mach} (machine precision)
Maximal relative error due to floating point representation



$x = s \times 2^E \in (R_{\min}, R_{\max})$

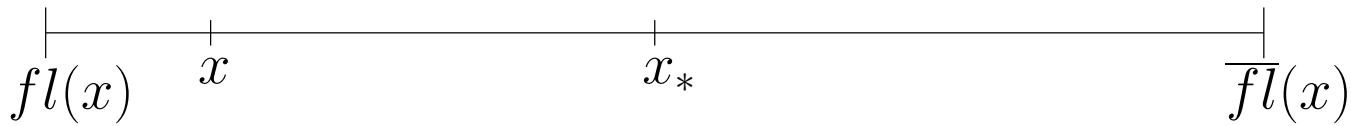
$fl(x) = \hat{s} \times 2^E$ (floating point number closest to x)

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

Representation of Double Precision Numbers

ϵ_{mach} (machine precision)

Maximal relative error due to floating point representation

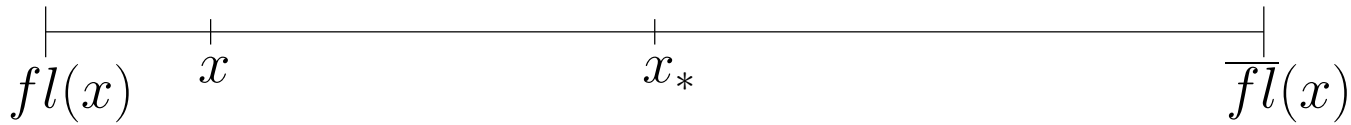


$\overline{fl}(x) = (\hat{s} + 2^{-52}) \times 2^E \quad \text{and} \quad x_* = \frac{fl(x)+\overline{fl}(x)}{2} = \left(\hat{s} + 2^{-53}\right) \times 2^E$

Representation of Double Precision Numbers

ϵ_{mach} (machine precision)

Maximal relative error due to floating point representation



$$\overline{fl}(x) = (\hat{s} + 2^{-52}) \times 2^E \quad \text{and} \quad x_* = \frac{fl(x)+\overline{fl}(x)}{2} = \left(\hat{s} + 2^{-53}\right) \times 2^E$$

Bounding relative error

$$\frac{|x - fl(x)|}{|x|} \leq \frac{|x_* - fl(x)|}{|fl(x)|} = \frac{2^{-53} \times 2^E}{\hat{s} \times 2^E} \leq \underbrace{2^{-53}}_{\epsilon_{mach}} \approx 1.11 \times 10^{-16}$$

Performing Arithmetic Operations

Arithmetic operations ($\oplus, \otimes, \ominus, \oslash$)
in double precision satisfy

$$
\begin{aligned}
x \oplus y &= fl(x + y) \\
x \ominus y &= fl(x - y) \\
x \otimes y &= fl(x \cdot y) \\
x \oslash y &= fl(x / y)
\end{aligned}
$$

where x and y are floating point numbers.

Performing Arithmetic Operations

Examples.

$1 \oplus 2^{-52} = 1 + 2^{-52}$, but $1 \oplus 2^{-54} = 1$

Performing Arithmetic Operations

Examples.

$1 \oplus 2^{-52} = 1 + 2^{-52}, \quad$ but $1 \oplus 2^{-54} = 1$

$$\begin{aligned}(1 + 2^{-52}) \otimes (2 + 2^{-51}) &= fl(2 + 2^{-51} + 2^{-51} + 2^{-103}) \\ &= fl((1 + 2^{-52} + 2^{-52} + 2^{-104}) \times 2) \\ &= 2(1 + 2^{-51})\end{aligned}$$