

Finite Precision

Mantissa

$$\pm \cdot \boxed{1|0|0|1|1|0|1|1} \quad k \text{ bits} \quad f$$

Exponent

$$\pm \boxed{1|0|1|1|0|1|1} \quad s \text{ bits} \quad e$$

$$\underline{f \cdot \beta^e}$$

f - mantissa

e - exponent

β - base (e.g., 10 or 2)

$$f = \pm \sum_{i=1}^k b_{-i} \beta^{-i} \quad \text{Normalized} \\ \text{if } f \neq 0, b_{-1} = 1$$

$$e = \pm \sum_{i=0}^{s-1} s_i \beta^i$$

$$\left(\pm \sum_{i=1}^k b_{-i} \beta^{-i} \right) \cdot \beta^e \quad \text{where } e = \pm \sum_{i=0}^{s-1} s_i \beta^i$$

For infinite precision k and s would become ∞

$$\underline{\text{Let } \beta = 2}$$

$$\pm \left(\sum_{i=1}^k b_{-i} 2^{-i} \right) \cdot 2^e$$

$$e = \pm \sum_{i=0}^s s_i 2^i$$

$$s_i, b_{-i} \in \{0, 1\}$$

$$b_{-1} = 1 \text{ for } f \neq 0$$

Error between Infinite and Finite?

Absolute?

Relative?

$$\pm \left(\sum_{i=1}^{\infty} b_{-i} 2^{-i} \right) \cdot 2^e$$

$$e = \pm \sum_{i=0}^{\infty} s_i 2^i$$

The diagram shows two arrows pointing from the infinite sum in the previous block to two separate finite sums. The first arrow points to the sum from $i=1$ to k , and the second arrow points to the sum from $i=k+1$ to ∞ .

$$\left(\sum_{i=1}^k b_{-i} 2^{-i} \right) + \left(\sum_{i=k+1}^{\infty} b_{-i} 2^{-i} \right)$$

Absolute Error

$$| \text{known} - \text{result} |$$

Relative Error

$$\frac{| \text{known} - \text{result} |}{| \text{known} |}$$

Absolute Error

$$\left| \underbrace{\left(\sum_{i=1}^{\infty} b_{-i} 2^{-i} \right) \cdot 2^e}_X - \underbrace{\left(\sum_{i=1}^k b_{-i}^* 2^{-i} \right) \cdot 2^{e^*}}_{X^*} \right|$$

let $e = e^*$

$b_{-i}^* = b_{-i}$ if $i \leq k$

$$\left| \left(\sum_{i=1}^{\infty} b_{-i} 2^{-i} \right) - \left(\sum_{i=1}^k b_{-i} 2^{-i} \right) \right| \cdot 2^e$$

$$\left| \sum_{i=1}^k \cancel{(\dots)} + \sum_{i=k+1}^{\infty} (\dots) - \sum_{i=1}^k \cancel{(\dots)} \right| \cdot 2^e$$

$$\left| \sum_{i=k+1}^{\infty} b_{-i} 2^{-i} \right| 2^e$$

let $b_{-i} = 1 \quad \forall i$

$$\leq \left| \sum_{i=k+1}^{\infty} 2^{-i} \right| 2^e$$

$$\begin{cases} 2^{-(k+1)} + 2^{-(k+2)} + \dots \\ 2^{-k} (2^{-1} + 2^{-2} + \dots) \end{cases}$$

$$\left| 2^{-k} \sum_{i=1}^{\infty} 2^{-i} \right| 2^e$$

$$\text{Abs error} \leq \left| 2^{-k} \right| 2^e$$
$$2^{-k} 2^e$$

Relative Error

$$\frac{|x - x^*|}{|x|}$$

$$\frac{|x - x^*|}{x} \leq \frac{|2^{-k}| 2^e}{\left| \sum_{i=1}^{\infty} b_i 2^{-i} \right| 2^e}$$

$$\leq \frac{2^{-k}}{2^{-1}}$$

$$\leq 2^{-k} \cdot 2$$

$$\leq 2^{-k+1}$$

Bound

$$\text{abs error} = |x - x^*| = |x^* - x|$$

$$\leq |2^{-k}| 2^e$$

Normalization constraint

$$\bullet \boxed{1 \dots}$$

smallest normalized value

$$2^{-1} = \frac{1}{2}$$

worst case error