Finite Precision

Mantissa

$\pm\ .\ \boxed{1|0|0|1|1|0|1|1}$ k bits f

Exponent

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

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

f - mantissa
e - exponent
β - base (e.g., 10 or 2)

$$f = \pm \sum_{i=1}^{k} b_{-i}\, \beta^{-i}$$

Normalized
if $f \neq 0$, $b_{-1} = 1$

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

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

For infinite precision k and s would become ∞

Let $\beta = 2$

$$\pm\left(\sum_{i=1}^{k} b_{-i}\, 2^{-i}\right)\cdot 2^{e} \qquad e = \pm\sum_{i=0}^{s} s_i\, 2^{i}$$

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

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

Error between Infinite and Finite?

Absolute? Relative?

$$\pm\left(\sum_{i=1}^{\infty} b_{-i}\, 2^{-i}\right)\cdot 2^{e} \qquad e = \pm\sum_{i=0}^{\infty} s_i\, 2^{i}$$

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

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 $i \leq k$

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

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

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

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

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

$2^{-(k+1)} + 2^{-(k+2)} \ldots$

$2^{-k}\left(2^{-1} + 2^{-2} \ldots\right)$

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

$$\boxed{\begin{aligned} \text{Abs error} &\leq \left|2^{-k}\right| 2^{e} \\ &\ 2^{-k}\, 2^{e} \end{aligned}}$$

Relative Error

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

$$\frac{|x-x^*|}{x} \le \frac{|2^{-k}|\,\cancel{2^e}}{\left|\sum_{i=1}^{\infty} b_{-i}\,2^{-i}\right|\cancel{2^e}}$$

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

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

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

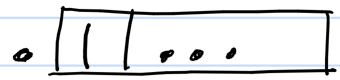
worst case error

Bound

abs error $= |x-x^*| = |x^*-x|$

$$\le |2^{-k}|\,2^e$$

Normalization constraint



smallest normalized value

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