I. INTRODUCTORY CONCEPTS

floating-point arithmetic

Every number is represented using a (fixed, finite) number of binary digits, usually called bits. A typical implementation would represent the number in the form

$$x = \sigma \times f \times \beta^{t-p}$$

σ is the sign of the number (±1), denoted by a single bit;
f is the mantissa or fraction
β is the base of the internal number system, usually binary (β = 2) or hexadecimal (β = 16),
t is the (shifted) exponent, i.e., the value that is actually stored;
p is the shift required to recover the actual exponent.

32 bits ==> **24** bits for the fraction, **7** bits for the exponent, and a **1** bit for the sign.

Overflow vs underflow

$$-63 \leq t - p \leq 64$$

The fraction is also limited

$$0 \leq f \leq \sum_{k=1}^{24} 2^{-k} = 1 - 2^{-24}.$$

Exercise : Express x=0.1 , y=0.0039 in 32 floating-point arithmetic with binary base and calculate z=x+y

Theorem 1.1 (Taylor's Theorem with Remainder) *Let $f(x)$ have $n+1$ continuous derivatives on $[a, b]$ for some $n \geq 0$, and let $x, x_0 \in [a, b]$. Then,*

$$f(x) = p_n(x) + R_n(x)$$

for

$$p_n(x) = \sum_{k=0}^{n} \frac{(x - x_0)^k}{k!} f^{(k)}(x_0), \tag{1.1}$$

and

$$R_n(x) = \frac{1}{n!} \int_{x_0}^{x} (x - t)^n f^{(n+1)}(t)dt. \tag{1.2}$$

Moreover, there exists a point ξ_x between x and x_0 such that

$$R_n(x) = \frac{(x - x_0)^{n+1}}{(n+1)!} f^{(n+1)}(\xi_x). \tag{1.3}$$

Theorem 1.2 (Mean Value Theorem) *Let f be a given function, continuous on $[a, b]$ and differentiable on (a, b). Then there exists a point $\xi \in [a, b]$ such that*

$$f'(\xi) = \frac{f(b) - f(a)}{b - a}. \tag{1.4}$$

Theorem 1.3 (Intermediate Value Theorem) *Let $f \in C([a, b])$ be given, and assume that W is a value between $f(a)$ and $f(b)$, that is, either $f(a) \leq W \leq f(b)$, or $f(b) \leq W \leq f(a)$. Then there exists a point $c \in [a, b]$ such that $f(c) = W$.*

Theorem 1.4 (Extreme Value Theorem) *Let $f \in C([a, b])$ be given; then there exists a point $m \in [a, b]$ such that $f(m) \leq f(x)$ for all $x \in [a, b]$, and a point $M \in [a, b]$ such that $f(M) \geq f(x)$ for all $x \in [a, b]$. Moreover, f achieves its maximum and minimum values on $[a, b]$ either at the endpoints a or b, or at a critical point.*

Theorem 1.5 (Integral Mean Value Theorem) *Let f and g both be in $C([a, b])$, and assume further that g does not change sign on $[a, b]$. Then there exists a point $\xi \in [a, b]$ such that*

$$\int_a^b g(t)f(t)dt = f(\xi) \int_a^b g(t)dt. \tag{1.5}$$

Theorem 1.6 (Discrete Average Value Theorem) *Let $f \in C([a, b])$ and consider the sum*

$$S = \sum_{k=1}^{n} a_k f(x_k),$$

where each point $x_k \in [a, b]$, and the coefficients satisfy

$$a_k \geq 0, \qquad \sum_{k=1}^{n} a_k = 1.$$

Then there exists a point $\eta \in [a, b]$ such that $f(\eta) = S$, i.e.,

$$f(\eta) = \sum_{k=1}^{n} a_k f(x_k).$$

Computer language : Fortran