Calcolo Numerico

1. Rappresentazione dei reali in base b ed errori di troncamento e arrotondamento

1.1. Rappresentazione dei numeri reali

Fissata una base $b > 1$ allora ogni $x \in \mathbb{R}$ si puo' scrivere come

$$x = \text{sgn}(x) \cdot c_m c_{m-1} \cdots c_1 c_0 . c_{-1} c_{-2} \cdots c_{-n} \cdots$$

$$x = \text{sgn}(x) \cdot \left\{ \underbrace{\sum_{j=0}^{m} c_j \cdot b^j}_{\text{parte intera}} + \underbrace{\sum_{j=1}^{\infty} c_{-j} \cdot b^{-j}}_{parte frazionaria} \right\}$$

dove $c_j, c_{-j} \in \{0, 1, \ldots, b-1\}$ sono le cifre della rappresentazione di x in base b e gli indici j e $-j$ corrispondono alle potenze della base (positive per la parte intera, negative per quella frazionaria).

Osserviamo che la parte intera $\in \mathbb{N}$ mentre la parte frazionaria $\in [0, 1]$ ed in generale ha ∞ cifre; tecnicamente e' una *serie* (somma infinita) di termini non negativi, *convergente*.

Esempio

$$x = +1278.3405 \cdots =$$

$$= (+1) \cdot \{1 \cdot 10^3 + 2 \cdot 10^2 + 7 \cdot 10^1 + 8 \cdot 10^0 + 3 \cdot 10^{-1} + 4 \cdot 10^{-2} \cdot 0 \cdot 10^{-3} + 5 \cdot 10^{-4} \ldots\}$$

Per far vedere che la serie della parte frazionaria e' convergente, cioe' rappresenta un numero $\in [0, 1]$ usiamo il criterio del confronto per serie a termini non negativi, infatti:

$$\forall\ c_{-j} \leq b - 1 \implies c_{-j} b^{-j} \leq (b-1) b^{-j}$$

quindi

$$\sum_{j=1}^{\infty} c_{-j} b^{-j} \leq (b-1) \sum_{j=1}^{\infty} b^{-j}.$$

Quest'ultima in particolare e' riconducibile ad una *serie geometrica* con fattore ≤ 1 e di conseguenza convergente al valore $\dfrac{1}{1 - \frac{1}{b}}$.

Dobbiamo far vedere che la parte frazionaria $\in [0, 1]$. Osserviamo subito che se tutte le cifre sono pari alla massima cifra della base, come ad esempio $0.\overline{999}$ in base 10, allora la parte frazionaria $= 1$, infatti:

$$0.\overline{999} = \sum_{j=1}^{\infty} 9 \cdot 10^{-j} = 9 \sum_{j=1}^{\infty} 10^{-j} = 9 \left(\frac{1}{1 - \frac{1}{10}} - 1 \right) = 9 \frac{1}{9} = 1.$$

Questo vale per tutte le basi.

Osservazioni

- Se il numero di cifre della parte frazionaria e' finito allora $x \in \mathbb{Q}$;
- dalla precedente osservazione otteniamo che se $x \in \mathbb{R}$ (e.g. $\sqrt{2}, \pi, \ldots$) necessariamente ha ∞ cifre della parte frazionaria, $\forall\, b > 1$;
- un numero in una base, con infinite cifre frazionarie, protrebbe avere un numero finito di cifre frazionarie in un'altra base;
- per scrivere un numero nel calcolatore ho necessariamente bisogno di una parte frazionaria finita.

Proprio questo punto e' quello di interesse, ci sono infatti due modi per ottenere un numero finito di cifre: troncando o approssimando quel numero.
Quello che ci interessa in particolare e' capire che errore commettiamo quando effettuiamo una di queste due operazioni.

1.2. Troncamento

Dato un numero x definiamone una suo *troncamento* alla n-esima cifra decimale:

$$\tilde{x}_n = \operatorname{sgn}(x) \cdot \left\{ \sum_{j=0}^{m} c_j b^j + \sum_{j=1}^{n} c_{-j} b^{-j} \right\}$$

allora $\text{errore-tr}(n) = |x - \tilde{x}| = \sum_{j=n+1}^{\infty} c_{-j} b^{-j}$ ossia il *resto*, dal momento che $c_{-j} \leq b - 1$ si ha:

$$\sum_{j=n+1}^{\infty} c_{-j} b^{-j} \leq (b-1) \sum_{j=n+1}^{\infty} b^{-j} = (b-1) \sum_{j=0}^{\infty} b^{-j} - \sum_{j=0}^{n} b^{-j} =$$

$$= (b-1) \cdot \frac{1}{1 - \frac{1}{b}} - \frac{1 - b^{-(n+1)}}{1 - \frac{1}{b}} = (b-1) \cdot \frac{b^{-(n+1)}}{\frac{b-1}{b}} = (b-1) \cdot \frac{b^{-n}}{b-1} = \frac{1}{b^n}.$$

Alla fine otteniamo che $|x - \tilde{x}| \leq \frac{1}{b^n} < \epsilon$ dove ϵ e' un fattore di *tolleranza* dell'errore, di conseguenza se vogliamo che il nostro errore sia $< \epsilon$ il nostro n dovra' essere $> \log_b \epsilon$.

1.3. Arrotondamento

Dato x un numero definiamone un suo arrotondamento a n cifre:

$$\tilde{x}_n^{\text{arr}} = (\text{parte intera}) + c_{-1} b^{-1} + c_{-2} b^{-2} + \cdots + \tilde{c}_{-n} b^{-n}$$

dove

$$\tilde{c}_{-n} = \begin{cases} c_{-n} & \text{se } c_{-(n+1)} < \frac{b}{2} \\ c_{-n} + 1 & \text{se } c_{-(n+1)} \geq \frac{b}{2} \end{cases}$$

di conseguenza abbiamo che il nostro $\text{errore-arr}(n) = |x - \tilde{x}_n^{arr}| \leq \frac{b^{-n}}{2}$.

Abbiamo quindi che l'*errore massimo* nell'arrotondamento e' *meta'* dell'errore massimo del troncamento a parita' di cifre.