UNIVERSITÀ DI PISA

Dipartimento di Informatica
Corso di Laurea Triennale in Informatica

Corso 2° anno - 6 CFU

Calcolo Numerico

Professore:
Prof. Luca Germignani

Autore:
Matteo Giuntoni

Anno Accademico 2022/2023

Contents

Calcolo Numerico

Realizzato da: Giuntoni Matteo

A.A. 2022-2023

---

1 Aritmetica di Macchina

Per una macchina la scrittura $(x+y)+z \neq x+(y+z)$. Vediamo dunque che ci sono alcuni punti focali da considerare per far si che una macchina funzioni correttamente:

- Trovare uno standard per come memorizzare i numeri.
- Trovare uno standard per come manipolare i numeri.

Da questi due punti possiamo ricondurci ad un solo problema, come andare a rappresentare i numeri.

1.1 Teorema di rapresentazione

Teorema 1.1.1. Dato $x \in \mathbb{R}, x \neq 0$ esistono e sono univocamente determinati.

1. un intero $p \in \mathbb{Z}$ detto esponente della rappresentazione.
2. una successione di numeri naturiali $\{d_i\}_{i\geq 1}$ con $d_i \neq 0, 0 \leq d_i \leq B-1$ e d_i non definitvamente uguali a $B-1$, dette cire della rappresentazione; tali per cui si ha

$$x = sign(x)B^p \sum_{i=1}^{+\infty} d_i B^{-1}. \tag{1}$$

Andiamo ora ad analizzare il significato di questo teorema. Esso descrive quella che viene chiamata rappresentazione in virgola mobile, in quanto l'esponente p on è determinato in modo da avere la parte intera nulla. Le cose da considerare in questo teorema sono:

- La condizione $d_i \neq 0$ e d_i non definitivamente uguale a $B-1$ sono introdotte per garantire l'unicità delle rappresentazioni. Ad esempio:

  $$B = 10 \text{ abbiamo } 1 = +10^1(1 \cdot 10^{-1}) = +10^2(0 \cdot 10^{-1} + 1 \cdot 10^{-1})$$

  Quindi due rappresentazioni diverse per lo stesso numero, però considerando le condizioni scritte sopra la seconda non risulta accettabile perché la prima cifra è nulla.
- Il caso $x = 0$ non ammette rapresentazione normalizzata. Questa casistica viene trattata dalla macchina in un modo particolare, per questo abbiamo la condizione $x \neq$.
- Questa rapresentazione si estende anche all'insieme dei numeri complessi del tipo $z = a + ib$, utilizzando una rapresentazione come coppie di numeri reali del tipo (a, b).

Possiamo dedurre che visto che stiamo lavornano con registri di meoria di un calcolatore con memoria a numero finito, anche la quantità di cifre rapresentabili saranno a numero finito esso vinene chiamato **insieme dei numeri di macchina**.

Dal teorema di rapresentazione in base di un numero reale può avvenire assegnando delle posizioni di meoria per il segno, per l'esponente e per le cifre della rappresentazione.

Definizione 1.1.1 (Insieme die numeri di macchina)**.** *Si definisce l'insieme dei nuermi di macchina in rappresentazione floating point con t cifre, base B e range $-m, M$ l'insieme dei numeri reali.*

$$\mathbb{F}(B,t,m,M) = \{0\} \cup \{s \in \mathbb{R} \,:\, x = sign(x)B^p \sum_{i=1}^{t} d_i B^{-1}, 0 \leq d_i \leq B-1, d_1 \neq 0, -m \leq p \leq M\}$$

Si osserva in questa definizione che:

- L'insieme $\mathbb{F}$ ha cardinalità finita $N = 2B^{t-1}(B-1)(M+m+1)+1$.
- L'insieme dei numeri di macchina $\mathbb{F}(B,t,m,M)$ è simmetrico rispetto all'origine.
- Possiamo definire $\Omega = B^M(B-1)\sum_{i=1}^{t} b^{-1}$ come il più grande numero macchina e $\omega = B^{-m}B^{-1}$ come invece il più piccolo.
- Posto un $x = B^P \sum_{i=1}^{t} d_i B^{-1}$ possiamo definire il suo successiovo numero di macchina come $y = B^p(\sum_{i=1}^{t-1} d_i B^{-1} + (d_t+1)B^{-t})$. Da qui vediamo che la distanza $y - x = B^p - t$ porta i numeri ad essere non equispaziali fra di loro, quindi la distanza aumento con l'avvicinarci a Ω

Esempio 1.1.1. Facciamo ora un esempio in cui andiamo a rappresentare il numero successivo di $x = B^p \sum_{t}^{i=1} d_i B^{-1}$. Esso si può scrivere come $y = B^p\Bigg(\sum_{i=1}^{t-1} d_i B^{-1} + (d_t+1)B^{-t}\Bigg)$.

Mentre si può scrivere la distanza fra questi due valori come $y - x = B^p - t$.

E' stato fissato uno standard IEEE 754 fra gli anni 70/80, questo standard dice che, visto ci sono macchine che hanno metodi di rappresentazione diversi bisogna fissare un standard, esso appunto dice che $B = 2$ ed i registri sono a 32 o 64 bit.

Questa rapresentazione ha uno svantaggio che può sembrare minimo ma non lo è, lo 0 si rappresenta due volte con $-0, +0$. Per ovviare a questo problema si è andato ad abbandonare questa rappresentazione in esponenti ma si rapprensentato i numeri nel seguente modo: $p_1 2^0 + p_1 2^1 + \cdots + p_1 1 s^1 0$ che rappresentano numeri da 0 a $2^1 1 - 1$ quindi 2047 numeri, mentre lo 0 si può scrivere come:

- O tenendo tutti i valori a 0
- Oppure tendendo tutti i valori a 1

In entrambi i casi abbiamo un range di valori che va da $[-1022, 1024]$. A questo punto ho 2^{P-1022} numeri che la macchina rappresneta come $\pm 2^{P-1022}(0.1d_1 \ldots d_{52})$.
Impostando questo standard abbiamo $\Omega = 2^{1024}(01 \ldots 1)_2$ e $\omega = 2^{-1022}(101)_2$.

Osservazione 1.1.1. Quando $p = 0$ abbiamo i numeri che si trovano nella porzione della retta dei numeri che è compresa fra $-\omega$ ω e possiamo qui avere anche tuti 0 e quindi si introduce il caso dei numeri denormalizzati.

Se abbiamo l'esponante uguale a tutti 1, la convensione è che tutte le cifre della mantissa sono tutti uguali a 0/1 questo numero indica il $\pm\infty$ altrimenti sta a signifiare NaN (not a number). Questi valori ci permettono di gestire forme indeterminate.

1.2 Errore di rappresentazione

Quando si va a rappresentare un numero reale non nullo $x \in \mathbb{R}$ e con $x \neq 0$ si può andare a commettere degli errori di rappresentazione detto anche **errore relativo di approssimazione**, e di definisce come, prendendo un $\tilde{x} \in \mathbb{F}(B,t,m,M)$

$$\epsilon_x = \frac{\tilde{x} - x}{x} = \frac{\eta x}{x}, x \neq 0$$

Definiamo $|\epsilon_x| = \left|\frac{\tilde{x}-x}{x}\right| \leq \frac{B^{P-t}}{|x|} \leq \frac{B^{P-t}}{B^{P-1}} = B^{1-t} = u$ la u è definita come **precisione di macchina**.

Andiamo inoltre a definire le conidizioni di underflow e overflow. Dato un $x \in \mathbb{R}, x \neq 0$ abbiamo che:

1. Se $|x| < \omega$ o $|x| > \Omega$ overflow. In questo caso si va ad associare il $+\infty$.
2. Se invece $\omega \leq |x| \leq \Omega$ abbiamo underflow. In questo caso allora prendiamo una $x = B^p \sum_{i=1}^{\infty} d_i B^{-1} \longrightarrow B^p \sum_{i=1}^{t} d_i B^{-1} = \tilde{x}$ che è una approssimazione