

# Calcolo Numerico

Luca De Paulis

19 febbraio 2021

# INDICE

1	ARITMETICA DI MACCHINA	3
1.1	Rappresentazione dei numeri di macchina	3
1.1.1	Virgola fissa	3
1.1.2	Virgola mobile	3
1.1.3	Insieme dei numeri di macchina	6
A	APPENDICE A	7
A.1	Teoremi di contorno	7

# 1 | ARITMETICA DI MACCHINA

## 1.1 RAPPRESENTAZIONE DEI NUMERI DI MACCHINA

Il primo problema da risolvere quando si vuole fare analisi numerica è scegliere un metodo per rappresentare i numeri reali su una macchina. Infatti un generico numero reale ha potenzialmente una scrittura decimale infinita, dunque essendo le risorse disponibili in una macchina *finite* dobbiamo trovare un modo di approssimarlo.

### 1.1.1 Virgola fissa

Il primo metodo è il metodo **a virgola fissa**: in questo metodo si rappresentano tutti i numeri nella loro forma decimale normale e si considerano esattamente  $k$  cifre dopo la virgola.

Questo metodo è molto semplice e ci consente di fare operazioni elementari (come le somme o i prodotti) immediatamente ("in colonna"), tuttavia ha anche degli svantaggi evidenti, come

- il range dei numeri rappresentabili su  $n$  cifre/bit è piccolo;
- siccome il numero di bit dedicato ai numeri dopo la virgola è basso, la precisione è molto bassa e assolutamente non adeguata ad applicazioni di analisi numerica.

Per questo è stata inventata la rappresentazione in virgola mobile.

### 1.1.2 Virgola mobile

Innanzitutto dobbiamo trovare un modo standard per rappresentare un qualsiasi numero reale. Per far ciò ci viene in aiuto il seguente teorema.

#### **Teorema**      **Teorema di rappresentazione in base.**

**1.1.1**      Sia  $x \in \mathbb{R}$  e sia  $\beta \geq 2$  una base di rappresentazione.

Allora esistono e sono unici un **esponente**  $p \in \mathbb{Z}$  e una successione di **cifre**  $(d_i)_{i \in \mathbb{N}}$  (con  $d_i \in \mathbb{Z}$  per ogni  $i$ ) tali che

- (i)  $d_1 \neq 0$ ;
- (ii) per ogni  $i$  vale che  $0 \leq d_i \leq \beta - 1$ ;
- (iii) la successione  $d$  non è definitivamente uguale a  $\beta - 1$ , ovvero per ogni  $k > 0$  esiste un  $j \geq k$  tale che  $d_j \neq \beta - 1$ ;
- (iv) vale che

$$x = \text{sgn}(x) \cdot \beta^p \cdot \left( \sum_{i=1}^{\infty} d_i \beta^{-i} \right).$$

**Esempio 1.1.2.** Consideriamo ad esempio il numero reale 123. Per il [Teorema 1.1.1](#) possiamo esprimere 123 come

$$123 = +10^3 \cdot (0.123) = +10^3 \cdot (1 \cdot 10^{-1} + 2 \cdot 10^{-2} + 3 \cdot 10^{-3}).$$

Questa rappresentazione è l'unica possibile se imponiamo che  $d_1$  sia diverso da 0 e che le cifre della rappresentazione siano tutte considerate come *cifre decimali* moltiplicate per una certa potenza della base (in questo caso 3).

La rappresentazione data dal [Teorema 1.1.1](#) è essenzialmente il concetto di *notazione scientifica*: in notazione scientifica portiamo il numero reale in modo che abbia una singola cifra diversa da zero prima della virgola, mentre nella notazione data dal Teorema quella cifra diventa la prima cifra dopo la virgola. Inoltre, il Teorema ci garantisce che questa rappresentazione non è solamente valida in base 10, ma lo è in qualsiasi base (noi lo useremo in particolare in base 2).

Il fattore  $\beta^p$  viene detto **esponente**, mentre la parte decimale (corrispondente alla sommatoria) viene detta **mantissa**.

**Osservazione 1.1.1.** Le ipotesi (i) e (iii) del [Teorema](#) sono fondamentali per l'unicità della rappresentazione. In effetti

- se venisse a mancare la prima condizione avremmo che

$$123 = +10^3 \cdot (0.123) = +10^4 \cdot (0.0123) = +10^5 \cdot (0.00123) = \dots;$$

- se venisse a mancare la terza condizione (che ci dice che la successione  $d$  non può terminare con una sequenza infinita di  $(\beta - 1)$ , ovvero il numero non può finire con  $\beta - 1$  periodico) avremmo dei problemi più subdoli, che derivano dalla non esistenza del "9 periodico" (dimostrata in appendice, [Proposizione A.1.1](#)).

Infatti in qualsiasi base  $\beta$  il numero  $0.\overline{(\beta - 1)}$  è uguale a 1, quindi se ammettessimo entrambe le rappresentazioni verrebbe meno l'unicità.

**Notazione.** Useremo la notazione  $(n)_\beta$  per riferirci al numero  $n$  espresso in base  $\beta$ .

Ad esempio il numero  $(1011)_2$  si riferisce al numero

$$1 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0 = 8 + 2 + 1 = 11 \text{ (in base 10).}$$

Per rappresentare i numeri reali in macchina dobbiamo quindi risolvere ancora due problemi:

- la macchina usa (nella maggior parte dei casi) la base 2, quindi dobbiamo poter trasformare da base 10 in base 2;
- dobbiamo rappresentare i numeri infiniti in modo approssimato.

Per quanto riguarda il primo, l'algoritmo per trasformare un numero decimale (in base 10) in un numero in base 2 è il seguente:

1. si trasforma la parte intera (le cifre prima della virgola) in base 2 tramite divisioni successive per 2;
2. si trasforma la parte decimale (le cifre dopo la virgola) in base 2 tramite *moltiplicazioni* successive per 2.

**Esempio 1.1.3.** Trasformiamo 3.15 in base 2: ovviamente  $(3)_{10} = (11)_2$  e quindi rimane solo da trasformare la parte decimale.

L'algoritmo consiste nel prendere il numero decimale 0.15 e moltiplicarlo per 2 ripetutamente: se il risultato ha come cifra prima della virgola uno 0 aggiungiamo uno 0 alla rappresentazione in base 2, altrimenti un 1. Nella pratica:

$$\begin{aligned} 0.15 \cdot 2 &= \boxed{0}.30 \\ 0.30 \cdot 2 &= \boxed{0}.60 \\ 0.60 \cdot 2 &= \boxed{1}.20 \\ 0.20 \cdot 2 &= \boxed{0}.40 \\ 0.40 \cdot 2 &= \boxed{0}.80 \\ 0.80 \cdot 2 &= \boxed{1}.60 \\ 0.60 \cdot 2 &= \boxed{1}.20 \\ &\vdots \end{aligned}$$

Osserviamo che ripetendo il procedimento la sequenza (0.60, 0.20, 0.40, 0.80) si ripete all'infinito, dunque il numero in base 2 termina con le cifre 1001 ripetute periodicamente. La parte decimale in base 2 corrisponde quindi a  $(0.00\overline{1001})_2$ .

Il numero completo è quindi

$$(11.00\overline{1001})_2 = 2^2 \cdot (0.1100\overline{1001})_2.$$

Per rappresentare i numeri infiniti dobbiamo approssimare la parte decimale, considerando solo un numero finito  $t$  di cifre dopo la virgola. Per far ciò esistono due metodi (che esamineremo più nel dettaglio nel seguito):

- nel caso del **troncamento** si considerano le prime  $t$  cifre dopo la virgola e si scartano tutte le successive;
- nel caso dell'**arrotondamento** analizziamo la cifra decimale di posto  $t + 1$ :
  - se  $d_{t+1}$  appartiene all'intervallo  $\left[0, \frac{\beta}{2}\right)$  (dove  $\beta$  è la base) consideriamo semplicemente le prime  $t$  cifre;
  - altrimenti (se  $d_{t+1} \in \left[\frac{\beta}{2}, \beta\right]$ ) aggiungiamo 1 all'ultima cifra decimale (come riporto, quindi lo propaghiamo in caso sia necessario).

**Esempio 1.1.4.** Riprendiamo il numero considerato precedentemente, ovvero

$$(3.15)_{10} = 2^2 \cdot (0.1100\overline{1001})_2$$

e approssimiamolo a  $t = 8$  cifre decimali.

Nel caso del troncamento basta considerare le prime 8 cifre decimali, quindi il numero troncato è

$$2^2 \cdot (0.11001001)_2.$$

Nel caso dell'arrotondamento invece dobbiamo considerare la nona cifra decimale: espandendo la rappresentazione decimale osserviamo che la nona cifra è

$$0.11001001\underline{1}00\dots;$$

siccome  $1 \in \left[\frac{2}{2}, 2\right] = [1, 2]$  dobbiamo *approssimare per eccesso*, ottenendo

$$2^2 \cdot (0.11001010)_2.$$

### 1.1.3 Insieme dei numeri di macchina

Siccome un singolo numero può occupare una quantità fissa in memoria (tipicamente 32 o 64 bit) dobbiamo fissare dei limiti per l'esponente e per il numero di cifre decimali che possiamo usare per la rappresentazione. Osserviamo inoltre che, se le nostre risorse sono limitate, aumentando il numero di cifre decimali disponibili dobbiamo necessariamente diminuire lo spazio dedicato a memorizzare l'esponente e viceversa.

**Definizione 1.1.5** **Insieme dei numeri di macchina.** Sia  $\beta \geq 2$  una base,  $t$  il numero di cifre decimali utilizzabili e siano  $m, M \in \mathbb{Z}$  gli estremi per l'esponente (ovvero ogni esponente  $p$  rappresentabile deve appartenere a  $[-m, M]$ ). Si dice allora **insieme dei numeri di macchina** l'insieme

$$\mathcal{F}(\beta, t, m, M) := \{0\} \cup \left\{ x \in \mathbb{R} : x = \text{sgn}(x) \cdot \beta^p \sum_{i=1}^{\infty} d_i \beta^{-i}, \right. \\ \left. \begin{aligned} &d_1 \neq 0, \\ &0 \leq d_i \leq \beta - 1, \\ &-m \leq p \leq M \end{aligned} \right\}.$$

Chiamiamo  $\Omega$  il più grande numero di macchina.

# A

## APPENDICE A

### A.1 TEOREMI DI CONTORNO

**Proposizione A.1.1** **9-periodico.** *In base 10 i numeri  $0.\bar{9}$  e 1 sono uguali.*

Forniamo due diverse dimostrazioni di questa proposizione.

**Prima dimostrazione.** Dalle formule per trasformare i numeri periodici in frazioni sappiamo che  $0.\bar{9} = 9/9 = 1$ .  $\square$

**Seconda dimostrazione.** Espandendo la definizione di numero periodico otteniamo che

$$0.\bar{9} = 0.999\dots = 9 \cdot 10^{-1} + 9 \cdot 10^{-2} + 9 \cdot 10^{-3} + \dots = \sum_{i=1}^{\infty} 9 \cdot 10^{-i}.$$

Sfruttando la formula della serie geometrica si ottiene che

$$\begin{aligned} \sum_{i=1}^{\infty} 9 \cdot 10^{-i} &= 9 \cdot \sum_{i=1}^{\infty} \left(\frac{1}{10}\right)^i \\ &= 9 \cdot \left( \left( \sum_{i=0}^{\infty} \left(\frac{1}{10}\right)^i \right) - \left(\frac{1}{10}\right)^0 \right) \\ &= 9 \cdot \left( \frac{1}{1 - 1/10} - 1 \right) \\ &= 9 \cdot \left( \frac{10}{9} - 1 \right) \\ &= 9 \cdot \frac{1}{9} \\ &= 1. \end{aligned} \quad \square$$

La proposizione vale in generale in una base  $\beta$  qualsiasi ( $\beta \geq 2$ ).

**Proposizione A.1.2** *In base  $\beta$  ( $\beta \geq 2$ ) vale che  $0.\overline{(\beta-1)} = 1$ .*

**Dimostrazione.** La dimostrazione è uguale alla seconda dimostrazione della [Proposizione A.1.1](#).  $\square$