CALCOLO DEGLI AUTOVALORI E FONDAMENTI DELLA MATEMATICA NUMERICA (La rappresentazione di macchina)

In questa lezione, ci focalizzeremo sulla rappresentazione dei numeri all'interno del calcolatore e sulle cosiddette operazioni in virgola mobile (floating point), ovvero operazioni eseguite nell'ambito dell'aritmetica discreta.
Approfondiamo quindi ora la rappresentazione dei numeri nel calcolatore. È stato più volte evidenziato come questo processo sia intrinseco all'utilizzo del calcolatore e soggetto a specifiche regole che è necessario conoscere. Analizzeremo alcune situazioni, già anticipate nelle lezioni precedenti tramite esempi, nelle quali le operazioni eseguite dal calcolatore possono, in generale, presentare potenziali instabilità. Per trattare la rappresentazione dei numeri nel calcolatore, è opportuno richiamare il processo di rappresentazione dei numeri reali, un procedimento ormai dato per scontato, ma che in realtà presuppone una rappresentazione posizionale in base decimale. Ci proponiamo di esprimere questi concetti in modo rigoroso attraverso esempi, concentrandoci in particolare sulla cosiddetta rappresentazione in base decimale.
Ricordiamo che, qualora un numero reale x venga scritto nella sua rappresentazione decimale

$$x = 2841.653,$$

dove si utilizza il punto anziché la virgola per separare la parte intera dalla parte decimale, si intende implicitamente una rappresentazione del numero x che fa riferimento alla base 10, ovvero

$$x = 2 \cdot 10^3 + 8 \cdot 10^2 + 4 \cdot 10^1 + 1 \cdot 10^0 + 6 \cdot 10^{-1} + 5 \cdot 10^{-2} + 3 \cdot 10^{-3}.$$

Nel caso specifico, per il numero 2841.653 abbiamo due migliaia, $2 \cdot 10^3$, otto centinaia, $8 \cdot 10^2$, quattro decine, $4 \cdot 10^1$, una unità, $1 \cdot 10^0$, sei decimi, $6 \cdot 10^{-1}$, cinque centesimi, $5 \cdot 10^{-2}$ e tre millesimi, $3 \cdot 10^{-3}$.
Pertanto, questa costituisce la rappresentazione autentica che dovrebbe essere utilizzata per attribuire significato alla forma convenzionale impiegata per esprimere i numeri reali. Tale rappresentazione è definita posizionale in base 10, in quanto la posizione delle singole cifre assume un significato differente a seconda del loro ordine.
È possibile normalizzare questa rappresentazione, ottenendo così una rappresentazione posizionale normalizzata, esprimendo il numero esclusivamente nella sua parte decimale

$$x = 0.2841653 \cdot 10^4,$$

Si procede quindi a rappresentare il numero come un valore reale compreso tra zero e uno, seguito da un fattore moltiplicativo di potenza di 10 che tiene conto dello spostamento delle cifre nella parte decimale. In particolare, il numero $x = 2841.653$ può essere espresso nella forma $x = 0.2841653 \cdot 10^4$. Questa è la rappresentazione posizionale normalizzata del numero di partenza x.
In generale, considerando un numero reale x della forma

$$x = \pm 0.\alpha_1\alpha_2 \dots \alpha_p\alpha_{p+1} \dots \cdot 10^q,$$

questa espressione sta a indicare che x è uguale a

$$x = \pm\left(\sum_{k=1}^{\infty} \alpha_k \cdot 10^k\right) \cdot 10^q, \qquad 0 \le \alpha_k \le 9, \alpha_1 \neq 0,$$

con l'ipotesi generale che il numero possa avere una rappresentazione decimale illimitata ($k = 1, \dots, \infty$). Questa è precisamente la natura di questa rappresentazione convenzionale in base 10.
Le cifre α_k rappresentano le singole cifre a disposizione nella rappresentazione decimale. In particolare, poiché si utilizza l'aritmetica decimale, ciascun α_k assume valori compresi tra zero e nove, ovvero $0 \le$

$\alpha_k \leq 9$. Inoltre, si assume che α_1 sia diverso da zero $\alpha_1 \neq 0$, affinché la prima cifra significativa risulti effettivamente non nulla.
Questa costituisce la rappresentazione dei numeri in base decimale. Procederemo ora all'introduzione della rappresentazione binaria e delle rappresentazioni in una base arbitraria β.
La rappresentazione binaria si ottiene sostituendo alla base 10 la base 2. Analogamente a quanto illustrato in precedenza, un numero espresso in rappresentazione binaria può essere scritto come segue

$$x = \pm\left(0.\alpha_1\alpha_2 \dots \alpha_p\alpha_{p+1} \dots\right)_\beta \cdot \beta^q, \qquad \beta \geq 2,$$

dove Il pedice β indica che si sta utilizzando una rappresentazione in una base arbitraria, non necessariamente la base 2, ma qualunque base β tale che $\beta \geq 2$. Pertanto, β può essere un qualsiasi numero intero maggiore o uguale a due. Questa rappresentazione rispetto a una base β arbitraria implica una serie della forma

$$x = \pm\left(\sum_{k=1}^{\infty} \alpha_k \cdot \beta^k\right) \cdot \beta^q, \qquad 0 \leq \alpha_k \leq \beta - 1, \alpha_1 \neq 0,$$

dove, come precedentemente, α_1 è diverso da zero, mentre i valori di α_k sono compresi tra zero e $\beta - 1$. Tale rappresentazione costituisce la forma posizionale normalizzata in una base arbitraria.
Nel caso della rappresentazione binaria, si sceglie semplicemente che la base sia pari a 2

$$\beta = 2,$$

e gli α_k saranno compresi fra zero e uno

$$\alpha_k \in \{0,1\}.$$

Abbiamo quindi in questo caso solo due cifre "decimali" a disposizione.
Nel caso in cui si utilizzi una rappresentazione diversa, come ad esempio la rappresentazione esadecimale in base 16, le cifre disponibili sono comprese tra zero e quindici. Tuttavia, poiché il valore quindici è espresso in notazione decimale, si adottano convenzioni alternative per la sua rappresentazione: si impiegano quindi le cifre decimali da zero a nove, mentre per i valori da dieci a quindici si utilizzano convenzionalmente le lettere A, B, C, D, E, F

$$\beta = 16, \qquad \alpha_k \in \{0,1,2,3,4,5,6,7,8,9, A, B, C, D, E, F\}.$$

Questi costituiscono gli elementi fondamentali per la rappresentazione esadecimale.
È ovviamente possibile convertire dalla rappresentazione binaria a quella decimale e, analogamente, dalla rappresentazione esadecimale a quella decimale. Nel caso della rappresentazione binaria, ad esempio, se si considera un numero espresso in forma binaria

$$\beta = 2, \qquad x = 10.11 = (0.1011)_2 \cdot 2^{10},$$

dove 2^{10} è espresso nella rappresentazione binaria. Si osservi che, in rappresentazione binaria, il valore 10 corrisponde al numero 2 nella rappresentazione decimale

$$(10)_2 = 1 \cdot 2^1 + 0 \cdot 2^0 = (2)_{10}.$$

Pertanto, 2^{10} in rappresentazione binaria indica che occorre spostare di due posizioni le cifre significative, ossia

$$10.11 = (0.1011)_2 \cdot 2^{10}.$$

La cosiddetta "parte decimale" (intendendo, con un abuso di notazione, la porzione che segue la virgola o, nel caso, il punto), corrisponde alla parte frazionaria del numero. In particolare,

$$(0.1011)_2 = 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3} + 1 \cdot 2^{-4} = \left(\frac{11}{16}\right)_{10}.$$

Dunque, il numero di partenza 10.11 in base 2 è il numero 2.75 in base 10

$$(0.1011)_2 \cdot 2^{10} = \left(\frac{11}{16} \cdot 4\right)_{10} = 2.75$$

In maniera simile possiamo passare dalla base esadecimale a quella decimale. Ad esempio

$$x = (0.2A0E)_{16} \cdot 16^2 = \left(\frac{2}{16} + \frac{10}{16^2} + \frac{14}{16^4}\right)_{10} \cdot 16^2 = 32 + 10 + \frac{14}{16^2} = 42.0546875.$$

Sono stati presentati esempi di rappresentazione binaria ed esadecimale poiché, molto frequentemente, i calcolatori impiegano una di queste due basi per rappresentare i propri numeri. La rappresentazione binaria e quella esadecimale risultano infatti particolarmente diffuse nell'architettura dei calcolatori, molto più rispetto alla rappresentazione decimale.

Analizziamo ora come il calcolatore gestisce i numeri reali. È stato precedentemente evidenziato che esiste la possibilità di rappresentare i numeri in una base arbitraria. Tuttavia, in generale, molti numeri presentano una rappresentazione infinita, ossia possiedono un numero illimitato di cifre dopo la virgola, costituendo una parte decimale infinita. Naturalmente, il calcolatore non è in grado di rappresentare un numero con una rappresentazione decimale infinita includendo tutte le cifre, poiché ciò implicherebbe la disponibilità di una quantità infinita di memoria per la memorizzazione. Pertanto, il calcolatore è costretto a troncare o, più precisamente, ad arrotondare i numeri, generando così la loro controparte di macchina, ovvero i cosiddetti numeri di macchina. Questa modalità di rappresentazione dei numeri da parte del calcolatore è denominata floating point (virgola mobile o punto mobile) finita, in quanto si basa sull'utilizzo di un numero finito di posizioni.

Supponiamo di considerare il numero

$$x = \pm\left(\sum_{k=1}^{\infty} \alpha_k \cdot \beta^{-k}\right) \cdot \beta^e = \pm(\alpha_1\beta^{-1} + \cdots + \alpha_{t-1}\beta^{-t+1} + \alpha_t\beta^{-t} + \cdots) \cdot \beta^e,$$

che può essere rappresentato nella rappresentazione posizionale normalizzata

$$x = \pm(0.\alpha_1\alpha_2 \ldots \alpha_{t-1}\alpha_t\alpha_{t+1} \ldots)_\beta \cdot \beta^e.$$

Definiamo $\mathrm{fl}^t(x)$ come la rappresentazione floating point con t cifre significative del numero x, nel seguente modo

$$\mathrm{fl}^t(x) = \pm(0.\alpha_1\alpha_2 \ldots \alpha_{t-1}\widetilde{\alpha_t})_\beta \cdot \beta^e,$$

dove, rispetto a x, sono stati mantenuti alcuni elementi fondamentali: β e le prime $t-1$ cifre sono rimaste invariate. Tuttavia, la t-esima cifra, indicata con $\widetilde{\alpha_t}$, può a priori differire dall'elemento originale α_t e sarà definita come segue

$$\widetilde{\alpha_t} = \begin{cases} \alpha_t & \text{se } 0 \le \alpha_{t+1} < \frac{1}{2}\beta \\ 1 + \alpha_t & \text{se } \frac{1}{2}\beta \le \alpha_{t+1} < \beta \end{cases},$$

ovvero, $\widetilde{\alpha_t}$ coincide con α_t qualora la prima cifra esclusa dalla rappresentazione sia inferiore a $\beta/2$ (nel caso decimale, ad esempio, inferiore a 5). Diversamente, sarà pari a $\alpha_t + 1$ se la prima cifra trascurata è maggiore o uguale a β/2. Pertanto, nel sistema decimale ad esempio, se la prima cifra omessa nella rappresentazione di macchina è minore di 5, si procede semplicemente al troncamento del numero; viceversa, se tale cifra è maggiore o uguale a 5, si effettua un arrotondamento incrementando la cifra nella posizione t-esima.

Nella rappresentazione floating point si impiega pertanto una rappresentazione con t cifre significative. Di conseguenza, dato un numero reale della forma seguente

$$x = \pm(0.\alpha_1\alpha_2 \ldots \alpha_{t-1}\alpha_t\alpha_{t+1} \ldots)_\beta \cdot \beta^e,$$

la sua rappresentazione floating point è

$$\mathrm{fl}^t(x) = \pm(0, \alpha_1\alpha_2 \dots \alpha_{t-1}\widetilde{\alpha_t})_\beta \cdot \beta^e,$$

e questa è la rappresentazione finita del numero reale x. La cifra $\widetilde{\alpha_t}$ è quella che si ottiene con l'operazione di arrotondamento (anche detta di round off).
Procederemo ora mostrando alcuni esempi di arrotondamento. A tal fine, utilizzeremo numeri molto semplici per comprendere come operi la logica dell'arrotondamento nel calcolatore. Consideriamo, ad esempio, il numero π

$$x = \pi = 3.14159265128 \dots .$$

Se si dovesse utilizzare una rappresentazione con sei cifre significative per il numero π, si otterrebbe la seguente approssimazione

$$\mathrm{fl}^6(\pi) = +0.314159 \cdot 10^1,$$

dove, la prima cifra da trascurare è 2, che essendo inferiore a 5 nella rappresentazione decimale viene omessa, procedendo quindi con un troncamento. Inoltre, si moltiplica per 10^1 poiché è stata effettuata una traslazione della virgola. Qualora si utilizzassero invece sette cifre significative, si otterrebbe

$$\mathrm{fl}^7(\pi) = +0.3141593 \cdot 10^1,$$

in cui l'ultima cifra significativa è 3 al posto di 2 perché la prima cifra da trascurare è 6. Questo è un esempio di arrotondamento per eccesso del numero π.
Consideriamo ora un numero periodico, espresso tramite una rappresentazione decimale illimitata.

$$x = -0.00666666666666\overline{6},$$

utilizzando sette cifre significative avremmo

$$\mathrm{fl}^7(x) = -0.6666667 \cdot 10^{-2},$$

dove, in questo caso, il termine 10^{-2} tiene conto dello spostamento della virgola, ovvero della nuova posizione assunta dalle cifre significative. La settima cifra viene arrotondata a 7 poiché la prima cifra trascurata è un 6, che è maggiore di 5. Di conseguenza, in questo caso, la rappresentazione del numero x nell'aritmetica con sette cifre significative risulta da un arrotondamento per eccesso.

Formalizziamo ora i concetti appena introdotti adottando una nomenclatura rigorosa. Definiamo numeri di macchina quei numeri che possiedono una rappresentazione finita della seguente forma

$$x = \pm(0.\alpha_1\alpha_2 \dots \alpha_{t-1}\alpha_t)_\beta \cdot \beta^e.$$

Un numero macchina x è definito da:

- un segno $\pm$,
- una mantissa, ovvero una sequenza di t cifre significative che il calcolatore è in grado di rappresentare, $\alpha_1\alpha_2 \dots \alpha_{t-1}\alpha_t$,
- una base β, intrinseca alla macchina. Inoltre, poiché ogni macchina opera con una base specifica, non è necessario specificarla ogni volta,
- un esponente e, che viene denominato caratteristica del numero.

In sintesi, un numero macchina è rappresentato come una combinazione del segno, della mantissa, della base e della caratteristica.
Possiamo concepire la rappresentazione di un numero macchina come un insieme di posizioni predisposte per la memorizzazione delle diverse componenti del numero. In particolare, vi è una cella dedicata al segno, una per ciascuna delle cifre significative della mantissa - ossia una cella per α_1, una

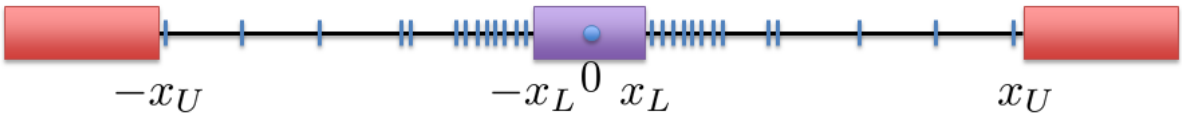

Figura 1: Rappresentazione schematica dei numeri macchina. I trattini verticali evidenziano la presenza di numeri di macchina compresi tra gli estremi indicati, x_L ed x_U.

per α_2, e così via fino a α_t - e infine una cella riservata all'esponente e. Si osservi che ciascun α_k assume valori compresi tra 0 e $\beta - 1$, in quanto la base utilizzata è β

$$0 \leq \alpha_k \leq \beta - 1.$$

L'esponente non può assumere valori compresi tra meno infinito e infinito, ma è invece limitato a un intervallo determinato dalle capacità fisiche della macchina. In altre parole, l'esponente varia entro limiti prestabiliti, dettati da vincoli di rappresentazione

$$L \leq e \leq U, \qquad L < 0, U > 0,$$

dove limite inferiore dell'esponente è indicato con $L < 0$, mentre il limite superiore è $U > 0$. La caratteristica e può quindi assumere valori interi compresi tra L e U. In sintesi, ogni numero macchina è descritto mediante un numero prefissato di cifre significative (la mantissa), un esponente compreso tra L e U (la caratteristica) e un segno.

La rappresentazione finita dei numeri all'interno del calcolatore comporta alcune problematiche. Una delle principali è la presenza di particolari intervalli, denominati zone di underflow e zone di overflow. Queste zone rappresentano limiti strutturali oltre i quali il calcolatore non è in grado di rappresentare correttamente i numeri.

Il numero più piccolo rappresentabile dal calcolatore è indicato con x_L, dove il pedice L denota il limite inferiore (lower). Tale numero è definito come

$$x_L = (0.100 \ldots 0) \cdot \beta^L = \beta^{L-1},$$

ossia si utilizza il più piccolo esponente, o più precisamente l'esponente negativo di maggiore valore assoluto, assumendo $\alpha_1 = 1$ e tutti gli altri coefficienti α_k pari a zero. Di conseguenza, x_L rappresenta il valore minimo che può essere rappresentato dal calcolatore. Quindi, per un calcolatore che impiega la base β e gli estremi inferiore e superiore L e U, il numero x_L corrisponde a β^{L-1}.

Il numero più grande rappresentabile dal calcolatore è indicato con x_U, dove il pedice U indica il limite superiore (upper)

$$x_U = (0.\gamma \ldots \gamma) \cdot \beta^U = (1 - \beta^{-t}) \cdot \beta^U.$$

In altre parole, si considera l'esponente U, il più grande possibile, mentre le cifre significative successive alla virgola assumono il valore massimo consentito, ovvero $\gamma = \beta - 1$, pertanto $0.\gamma \ldots \gamma$. Questo numero x_U corrisponde a

$$x_L = (1 - \beta^{-t}) \cdot \beta^U.$$

Il valore x_U rappresenta, dunque, il numero più grande in valore assoluto che un calcolatore possa rappresentare. Pertanto, si identificano i numeri $-x_U$, ovvero il valore negativo e positivo più elevati rappresentabili. Analogamente, il numero x_L è il valore più piccolo, in valore assoluto, rappresentabile dal calcolatore, con i corrispondenti valori $-x_L$ e x_L.

Pertanto, tutti i numeri di macchina rappresentabili si collocano negli intervalli

$$[-x_U, -x_L] \cup 0 \cup [x_L, x_U],$$

comprendendo anche il numero zero.
La Figura 1 illustra schematicamente questo tipo di rappresentazione. I trattini verticali evidenziano la presenza di numeri di macchina compresi tra gli estremi indicati. Si osserva un'accumulazione maggiore di tali numeri in prossimità dell'estremo x_L, poiché la rappresentazione risulta significativamente più densa vicino a x_L rispetto all'estremo x_U.