Analisi Numerica I
Rappresentazione dei numeri "floating point"

Ana Alonso

ana.alonso@unitn.it

25 ottobre 2019

Esempi

L'aritmetica in MatLab non e quella dei numeri reali.

```
>> a=realmax
>> b=realmax/2
>> c=(a+b)/7
>> c=a/7+b/7
>>
>> a=1, while a>0, a=a/2, end
>>
>> a=10^20
>> a+76-a
>> a-a+76
```

Numeri "floating point"

$$\begin{aligned} 394.45 &= +0.39445\ 10^{3} \\ -0.0034 &= -0.34\ 10^{-2} \\ \pi &= +0.314159265358...\ 10^{1} \end{aligned}$$

Tutti i numeri reali, tranne lo zero, si possono rappresentare (in base β) nel seguente modo:

$$\pm 0.a_1 a_2 a_3 a_4 ...\ \beta^e \quad a_1 \neq 0 \quad 0 \leq a_r \leq \beta - 1 \quad e \in \mathbb{Z}$$

Il numero zero ha una rappresentazione a se stante.

Al calcolatore bisogna fissare

- una certa quantità finita t di cifre significative,
- un valore minimo L e un valore massimo U dell'esponente.

$$\pm 0.a_1 a_2 a_3 a_4 ... a_t\ \beta^e \qquad a_1 \neq 0 \qquad 0 \leq a_r \leq \beta - 1 \qquad L \leq e \leq U\,.$$

I numeri macchina

$$\pm 0.a_1 a_2 a_3 a_4 \ldots a_t \ \beta^e \qquad a_1 \neq 0 \qquad 0 \leq a_r \leq \beta - 1 \qquad L \leq e \leq U\,.$$

Chiamiamo $\mathbb{F}(\beta, t, L, U)$ all'insieme (finito) di questi numeri e lo zero.
Matlab usa $\mathbb{F}(2, 53, -1021, 1024)$

- Il più piccolo numero positivo: $\beta^{-1}\beta^L = \beta^{L-1}$.

  `realmin`.

- Il numero massimo:
  $\beta^U \sum_{i=1}^{t}(\beta-1)\beta^{-i} = (\beta-1)\beta^U \sum_{i=1}^{t}\beta^{-i} =$
  $(\beta-1)\beta^U \frac{\beta^{-t-1}-\beta^{-1}}{\beta^{-1}-1} = (\beta-1)\beta^U \frac{\beta^{-t}-1}{1-\beta} = (1-\beta^{-t})\beta^U$

  `realmax`.

- Distanza fra il numero 1 e il numero macchina successivo:
  $\beta^{-t}\beta = \beta^{1-t}$.

  `eps`.

Rappresentazione dei numeri reali al calcolatore

Dato il numero reale

$$x = \pm\beta^e \sum_{i=1}^{\infty} a_i \beta^{-i}$$

il suo rappresentante in $\mathbb{F}(\beta, t, L, U)$ è

$$fl(x) = \begin{cases} \pm\beta^e \sum_{i=1}^{t} a_i \beta^{-i} & \text{se } a_{t+1} < \beta/2 \\ \pm\beta^e \left(\sum_{i=1}^{t} a_i \beta^{-i} + \beta^{-t}\right) & \text{se } a_{t+1} \geq \beta/2 \end{cases}$$

L'errore relativo che si commette è limitato da

$$\frac{|x - fl(x)|}{|x|} \leq \frac{1}{2} \frac{\beta^{-t}\beta^e}{\beta^{-1}\beta^e} = \frac{1}{2}\beta^{1-t}.$$

Matrice di Hilbert

- `hilb(n)` calcola la matrice di Hilbert di dimensione n. $h_{i,j} = 1/(i+j-1)$.
- Esercizio:
  Sia `A` la matrice di Hilbert 9×9, sia `sol=ones(9,1)` e `b=A*sol` .
  - Usando i comandi di Matlab risolvere $A\mathbf{x} = \mathbf{b}$.
  - Calcolare $\frac{\|\mathbf{x}-\mathbf{sol}\|}{\|\mathbf{sol}\|}$.
  - Calcolare il numero di condizionamento di A.
  - Commentare i risultati.