I NUMERI FINITI.
RAPPRESENTAZIONE E TRONCAMENTO

Risultato fondamentale sulla rappresentazione dei numeri reali

Ogni numero reale $\alpha \neq 0$ può esprimersi univocamente, nella base β, nella forma:

$$\alpha = \pm \left(a_1 \beta^{-1} + a_2 \beta^{-2} + a_3 \beta^{-3} + \dots \right) \beta^{\nu} =$$

$$= \pm m \beta^{\nu}$$

ove:

$\nu \in \mathbb{N}$

$\{a_i\} \in \mathbb{N}$ t.c. 1) $0 \leq a_i \leq \beta - 1$

$i = 1, 2, 3, \dots$

$a_1 \neq 0$

2) può eventualmente esistere un indice j t.c. $\forall i \geq j$ sia $a_i = 0$, MA non esiste alcun indice $k > 0$ t.c. $\forall i \geq k$ sia $a_i = \beta - 1$

- Il numero

$$m = a_1 \beta^{-1} + a_2 \beta^{-2} + a_3 \beta^{-3} + \dots$$

si definisce **mantissa** di α e

$$\beta^r$$

si definisce **parte esponente** di α, mentre l'intero

$$r$$

si definisce **esponente** di α.

- Vale:

$$\frac{1}{\beta} \leq m < 1$$

- <u>Definiamo</u>: **rappresentazione in forma scientifica** di α

$$\alpha = \pm 0,a_1 a_2 a_3 \dots \beta^r$$

cifre significative

punto radice

parte intera

I numeri finiti

I numeri reali sono, in generale, rappresentati da un numero illimitato di cifre. Nella pratica computazionale dobbiamo considerare loro approssimazioni, per le quali la mantissa m è rappresentata con un numero finito di cifre.

Fissiamo un intero $t \geq 1$; ogni numero reale α viene associato ad un numero finito $fl(\alpha)$ il quale si esprime nella forma:

$$fl(\alpha) = \begin{cases} 0 \cdot \beta^0 & , \text{ se } \alpha = 0 \\ \pm m_t \beta^r & , \text{ se } \alpha \neq 0 \end{cases}$$

dove:

$$m_t = a_1 \beta^{-1} + a_2 \beta^{-2} + \dots + a_t \beta^{-t} \quad , \quad a_1 \neq 0$$

è definito troncamento della mantissa m alla t-ma cifra.

L'arrotondamento

La quantità:

$$\left| \frac{\alpha - fl(\alpha)}{\alpha} \right|$$

è definita errore relativo di arrotondamento di α.

Teorema

Se $\alpha \neq 0$ e se $\left| \frac{\alpha - fl(\alpha)}{\alpha} \right| \leq \beta^{1-t}$, allora:

$$fl(\alpha) = \alpha (1 + \varepsilon)$$

ove: $|\varepsilon| \leq \beta^{1-t}$. Il numero $u = \beta^{1-t}$ si definisce unità di arrotondamento.

Esempio # 1

Sistema decimale : $\beta = 10$

7 cifre significative : $t = 7$

Calcolare la somma di $\alpha = 0.1234567 \cdot 10^{0}$

$\gamma = 0.6666325 \cdot 10^{4}$

$\delta = -0.6666325 \cdot 10^{4}$

I

$$fl(\gamma) + fl(\delta) = 0$$

$$fl\{fl(\gamma) + fl(\delta)\} = fl(0) = 0$$

$$fl(\alpha) + fl\{fl(\gamma) + fl(\delta)\} = \alpha + 0 = 0.1234567 \cdot 10^{0}$$

$$fl\{fl(\alpha) + fl\{fl(\gamma) + fl(\delta)\}\} = \underline{0.1234567 \cdot 10^{0}}$$

II

$$fl(\alpha) = 0.0000123 \cdot 10^{4}$$

$$fl(\gamma) = 0.6666325 \cdot 10^{4}$$

$$fl\{fl(\alpha) + fl(\gamma)\} = 0.6666448 \cdot 10^{4}$$

$$fl\{fl(\alpha) + fl(\gamma)\} + fl(\delta) = 0.6666448 \cdot 10^{4} + \\ - 0.6666325 \cdot 10^{4} = \\ = 0.0000123 \cdot 10^{4}$$

$$fl\{fl\{fl(\alpha) + fl(\gamma)\} + fl(\delta)\} = \underline{0.1230000 \cdot 10^{0}}$$

Esempio # 2

Sistema decimale: $\beta = 10$

4 cifre significative : $t = 7$

Calcolare una radice dell'equazione $x^2 - 6.433\,x + 0.009474 = 0$

$$x_2 = \frac{6.433 - \sqrt{41.383489 - 0.037896}}{2} = 0.0014731\ldots$$

$a = (6.433)^2 = 41.383489$ $\qquad fl(a) = 0.4138 \cdot 10^2$

$b = 4 \cdot 0.009474 = 0.037896$ $\qquad fl(b) = 0.3789 \cdot 10^{-1}$

$fl\{ fl(a) - fl(b) \} \equiv \gamma = 0.4134 \cdot 10^2$

$\sqrt{\gamma} = 0.64296\ldots \cdot 10^1$

$fl(\sqrt{\gamma}) = 0.6429 \cdot 10^1$

$$fl\left\{ \frac{0.6433 \cdot 10^1 - 0.6429 \cdot 10^1}{2} \right\} = 0.2000 \cdot 10^{-2}$$

In conclusione

- Per i numeri finiti non vale la proprietà associativa (cfr. Esempio #1)

- Per partendo da dati affetti da un piccolo errore relativo di arrotondamento la sottrazione tra numeri "quasi" uguali ($0.6433 \cdot 10^1$ e $0.6429 \cdot 10^1$) ha causato una grande crescita dell'errore relativo di arrotondamento sul risultato. (cfr. Esempio #2)

BEN POSIZIONE DI UN PROBLEMA

Strettamente connesso con l'arrotondamento e la rappresentazione dei numeri reali è il problema della ben posizione di un problema.

- Consideriamo il seguente sistema di due equazioni lineari nelle incognite x_1 ed x_2:

$$\begin{cases} 1.000\, x_1 + 2.000\, x_2 = 3.000 \\ 0.499\, x_1 + 1.001\, x_2 = 1.500 \end{cases}$$

L'unica soluzione è:

$$x_1 = 1.$$

$$x_2 = 1.$$

- Consideriamo ora il seguente sistema, ottenuto dal precedente applicando una piccola perturbazione sulla terza e sulla quarta cifra significativa dei due coefficienti della seconda equazione:

$$\begin{cases} 1.000\, x_1 + 2.000\, x_2 = 3.000 \\ 0.500\, x_1 + 1.002\, x_2 = 1.500 \end{cases}$$

L'unica soluzione è ora:

$$x_1 = 3.$$

$$x_2 = 0.$$

- Il problema è pertanto mal posto: perturbando di poco i dati del problema (coefficienti e termini noti, nell'esempio precedente) la soluzione è completamente diversa rispetto a quella del problema originale.

Operando con i numeri finiti i problemi mal posti sono assai frequenti.

INTEGRAZIONE NUMERICA

- Siano:

  $[a,b]$ intervallo chiuso e limitato

  $F(x)$ funzione continua in $[a,b]$

  $\{x_i\}_{i=0,\dots,n}$ punti t.c. $a = x_0 < x_1 < \dots < x_n = b$

  $\| D \| = \max \{ x_1 - x_0 , x_2 - x_1 , \dots , x_n - x_{n-1} \}$

  $\{t_i\}_{i=1,\dots,n}$ punti t.c. $t_1 \in [x_0, x_1[$, $t_2 \in [x_1, x_2[, \dots,$
  $t_n \in [x_{n-1}, x_n]$

- Definiamo: somma di Riemann relativa alla funzione $F(x)$ la quantità:

$$\sum_{i=1}^{n} (x_i - x_{i-1}) F(t_i)$$

- Definiamo: integrale di F su $[a,b]$ un numero, indicato con $\int_a^b F(x)\,dx$ t.c. $\forall \varepsilon > 0$, $\exists \delta_\varepsilon$ t.c. $\forall |x_i - x_{i-1}| < \delta_\varepsilon$ vale:

$$\left| \sum_{i=1}^{n} (x_i - x_{i-1}) F(t_i) - \int_a^b F(x)\,dx \right| < \varepsilon$$

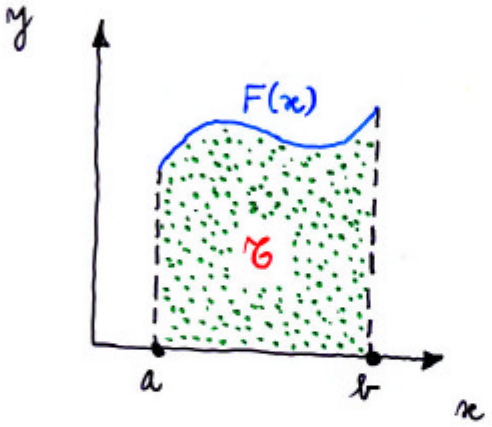
$\forall t_i \in [x_{i-1}, x_i[$.

Interpretazione geometrica dell'integrale

Supponiamo: $F(x) \geq 0 \quad , \forall\, x \in [a,b]$

Sia: $\mathcal{T} = \{ (x,y) \text{ t.c. } x \in [a,b] \,,\; y \in [0, F(x)] \}$

Allora: l'area del trapezoide $\mathcal{T}$ è espressa dal valore dell'integrale $\int_a^b F(x)\,dx$:



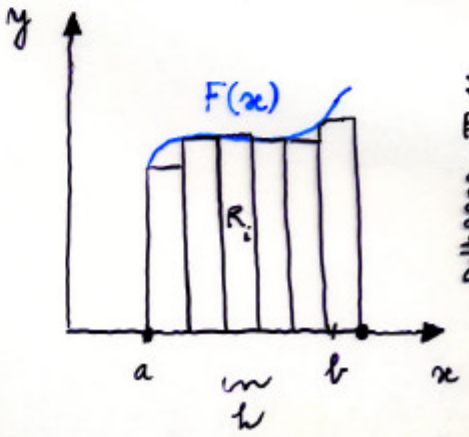
- **Formula dei rettangoli**

$$I \simeq h\,(F_0 + F_1 + \cdots F_{n-1}) \equiv R_n$$

$$F_i \equiv F(x_i)$$

$$E_n^R \equiv I - R_n = \frac{(b-a)\,h\,F'(\xi)}{2}$$

$$\xi \in [a,b]$$



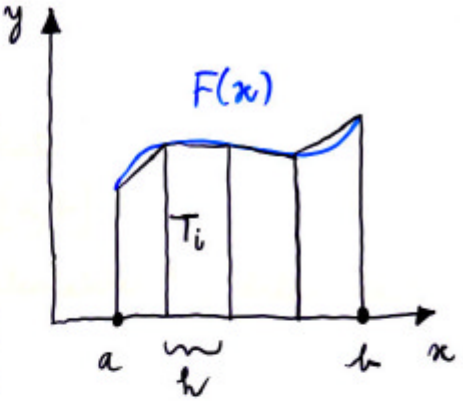
Formula dei trapezi

$$I \simeq h\left(\frac{F_0}{2} + F_1 + \cdots + \frac{F_n}{2}\right) \equiv T_n$$

$$F_i \equiv F(x_i)$$

$$E_n^T \equiv I - T_n = \frac{(b-a)\,h^2\,F''(\xi)}{12}$$

$$\xi \in [a, b]$$



APPROSSIMAZIONE DI UNA FUNZIONE

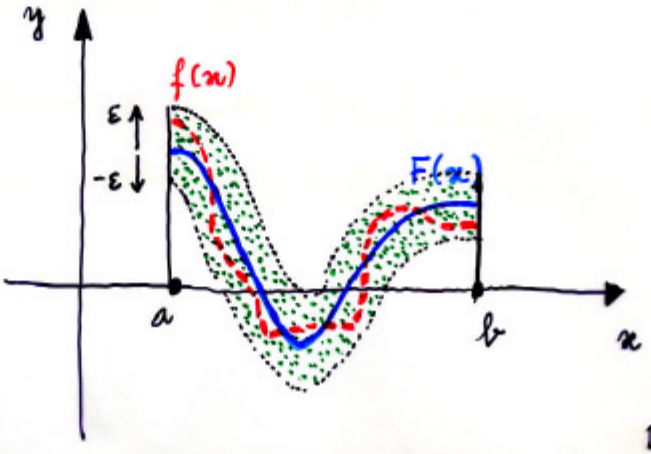
- Siano:
  - $[a, b]$ intervallo chiuso e limitato
  - $F(x)$ funzione continua in $[a, b]$
  - $f(x)$ funzione "facilmente calcolabile", data da:

$$f(x) = \lambda_0 \varphi_0(x) + \lambda_1 \varphi_1(x) + \dots + \lambda_n \varphi_n(x)$$

con $\{\varphi_i(x)\}_{i=1,\dots,n}$ sono "funzioni elementari" in $[a, b]$

- Definiamo: $f(x)$ una approssimazione di $F(x)$ in $[a, b]$ se, fissata una tolleranza $\varepsilon > 0$ risulta:

$$|F(x) - f(x)| \leq \varepsilon \quad , \forall x \in [a, b]$$



Esempio

Scegliamo come "funzioni elementari" in $[a,b]$ i monomi:

$$\varphi_0(x) = 1$$
$$\varphi_1(x) = x$$
$$\dots \quad \dots$$
$$\varphi_n(x) = x^n$$

Per il teorema di Weierstrass, assegnato $\varepsilon > 0$ è possibile determinare $n > 0$ t.c. il polinomio

$$f(x) = \lambda_0 + \lambda_1 x + \lambda_2 x^2 + \dots + \lambda_n x^n$$

sia una approssimazione di $F(x)$ in $[a,b]$.

• Formula di Lagrange

Siano assegnati nel piano cartesiano ortogonale xy $n+1$ punti distinti (x_0, y_0), (x_1, y_1), $\dots$, (x_n, y_n) appartenenti alla curva di equazione $y = F(x)$.

[Possiamo identificare i punti $\{x_i\}_{i=0,\dots,n}$ come punti di osservazione ed i valori $y_i = F(x_i)$, $i = 0, \dots, n$ come le osservazioni].

Esiste, ed è unico, un polinomio f di grado n t.c.

$$f(x_0) = F(x_0)$$
$$f(x_1) = F(x_1)$$
$$\dots \qquad \dots$$
$$f(x_n) = F(x_n)$$

espresso nella forma:

$$f(x) = F(x_0)\,l_0(x) + F(x_1)\,l_1(x) + \dots + F(x_n)\,l_n(x)$$

ove:

$$l_j(x) = \frac{(x-x_0)(x-x_1)\dots(x-x_{j-1})(x-x_{j+1})\dots(x-x_n)}{(x_j-x_0)(x_j-x_1)\dots(x_j-x_{j-1})(x_j-x_{j+1})\dots(x_j-x_n)}$$

$j = 0, 1, \dots, n$

Il polinomio $f(x)$ è definito polinomio di interpolazione di $F(x)$ secondo Lagrange; i punti $\{x_i\}_{i=0,\dots,n}$ sono definiti punti di interpolazione.

Formula di Newton

Definiamo **differenza divisa** di $F(x)$ rispetto agli argomenti x_0, x_1 la quantità

$$F[x_0, x_1] = \frac{F(x_0) - F(x_1)}{x_0 - x_1}$$

Procedendo iterativamente è possibile definire **differenza divisa** di $F(x)$ di ordine m la quantità

$$F[x_0, x_1, \dots, x_m] = \frac{F[x_0, x_1, \dots, x_{m-1}] - F[x_1, \dots, x_m]}{x_0 - x_m}$$

[**Esempio**

i	x_i	$F(x_i) = F[x_i]$		
0	0	1		
			$F[x_0, x_1] = 2$	
1	1	3		$F[x_0, x_1, x_2] = -\frac{5}{6}$
			$F[x_1, x_2] = -\frac{1}{2}$	
2	3	2		

]

Siano $F(x)$ ed $\{x_i\}_{i=0,\dots,n}$ come sopra. Il polinomio di interpolazione di $F(x)$ secondo Newton può essere espresso nella forma:

$$f(x) = F(x_0) + (x-x_0)\,F[x_0, x_1] + \dots + (x-x_0)(x-x_1)\cdots(x-x_{n-1})\,F[x_0, x_1, \dots, x_n]$$

ove $F[\cdot]$ sono le differenze divise, di vario ordine, di $F(x)$.

- **Polinomio di Taylor**

Nel caso degenere nel quale tutti gli argomenti $x_0, x_1, \dots, x_n$ coincidano con l'unico punto x_0 la formula di Newton diviene:

$$f(x) = F(x_0) + (x-x_0)\,F'(x_0) + \dots + \frac{(x-x_0)^n}{n!}\,F^{(n)}(x_0)$$

essendo infatti

$$F[\underbrace{x_0, \dots, x_0}_{n \text{ volte}}] = \frac{1}{n!}\,F^{(n)}(x_0)$$

Definiamo polinomio di Taylor il polinomio di grado n definito dalla relazione:

$$f(x) = F(x) - \frac{(x - x_0)^{n+1}}{(n+1)!} F^{(n+1)}(\xi)$$

ove: x_0 e $\xi \in]x_0, x[$ sono punti in $[a, b]$.