CALCOLO NUMERICO E MATLAB
Aritmetica del calcolatore

Silvia Falletta

Dipartimento di Scienze Matematiche, Politecnico di Torino
silvia.falletta@polito.it

A.A. 2016/2017

Uno dei più grandi errori, da evitare quando si programma al calcolatore, è pensare che i risultati numerici ottenuti siano privi di errori o che gli errori siano trascurabili. Alcuni disastri, oramai passati alla storia dell'analisi numerica, sono dovuti proprio a un uso scorretto del calcolo numerico. Diverse pagine web riportano alcune di queste storie: si provi ad esempio

http://ta.twi.tudelft.nl/users/vuik/wi211/disasters.html

- Nel 1991, durante la prima guerra del Golfo Persico, un missile Patriot fallì l'intercettazione di un missile Scud iracheno: morirono 28 soldati e ci furono centinaia di feriti. L'errore di intercettazione fu dovuto a errori di arrotondamento nel programma che aveva lanciato il missile!
- Nel 1996, il razzo Ariane 5 esplose in aria pochi secondi dopo il lancio: l'esplosione fu causata da un errore di conversione da un sistema di rappresentazione dei numeri in virgola mobile a 64 bit a uno in virgola fissa a 16 bit!

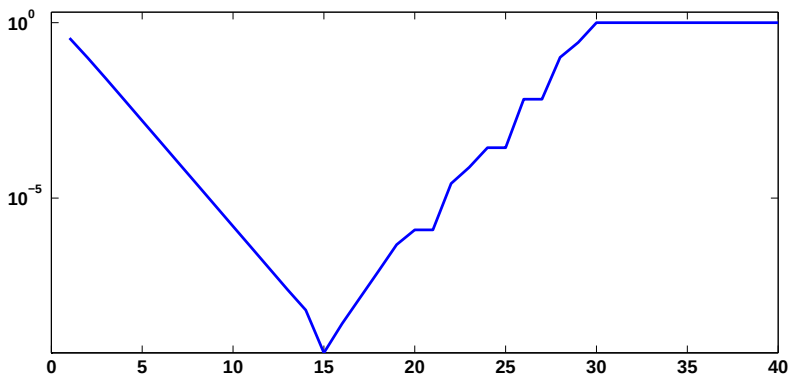
Esempio Un esempio, con conseguenze meno catastrofiche, ma comunque importanti, è costituito dal calcolo della seguente successione:

$$\begin{aligned} x_1 &= 2 \\ x_n &= 2^{n-1/2}\sqrt{1-\sqrt{1-4^{1-n}x_{n-1}^2}}, \qquad n \geq 2 \end{aligned}$$

che converge a π. Supponiamo di non essere in grado di calcolare il limite analiticamente, e vogliamo quindi implementare in Matlab un algoritmo per il calcolo del generico termine x_n. Ci aspettiamo che, per n via via sempre più grande, x_n si avvicini sempre di più al valore esatto, cioè π.

Che cosa notiamo? I valori si avvicinano sempre più a π fintanto che $n \leq 15$, poi iniziano ad aumentare.
Di seguito riportiamo il grafico dell'errore relativo. $|x_n - \pi|/\pi$

Figura : Grafico dell'errore relativo $|x_n - \pi|/\pi$ in scala logaritmica



Nel momento in cui si lavora al calcolatore, occorre tenere presente i "diversi punti di vista": noi lavoriamo con numeri in base 10 con numero finito di cifre decimali (0.1, 0.5, ...) e con numeri a infinite cifre decimali ($\sqrt{2}$, π, $\frac{1}{3}$,...). Il calcolatore invece può lavorare solo con numeri che abbiano un numero finito di cifre dopo la virgola, e utilizza la base binaria. Dobbiamo quindi tener sempre presente che:

- noi scriviamo i numeri in base 10, ma questi sono trasformati in base 2 dal calcolatore;
- il calcolatore esegue i conti in base 2 e fornisce il risultato convertendolo in base 10;
- il calcolatore lavora solo con numeri finiti.

Ciò premesso, vediamo come il calcolatore rappresenta e come lavora con i numeri.

Rappresentazione floating-point di un numero reale

Sia $N > 1$ la base del sistema di numerazione prescelto.
Si definisce **rappresentazione floating-point di un numero reale** a la seguente rappresentazione

$$a = (-1)^s p N^q, \quad s = 0, 1 \quad p \text{ reale}, q \text{ intero}$$

Esempio
Per $N = 10$, $a = 0.015 \cdot 10^{-1} = \mathbf{0.15} \cdot \mathbf{10^{-2}} = 0.0015 \cdot 10^0$.

Tale rappresentazione non è unica! Essa è univocamente determinata se si impone

$$N^{-1} \leq p < 1 \qquad (\Rightarrow \text{ per } N = 10,\ 0.1 \leq p < 1).$$

$$\Updownarrow$$

la prima cifra di p dopo il punto decimale è diversa da zero.

Se la condizione $N^{-1} \leq p < 1$ è soddisfatta allora la rappresentazione floating-point $a = (-1)^s p N^q$ si dice **normalizzata**. In tal caso, p si definisce **mantissa** di a e q **esponente** oppure **caratteristica** di a.

Esempio

Per $N = 10$, $a = 0.15 \cdot 10^{-2}$, $p = 0.15$, $q = -2$;
$a = 12.4 = 0.124 \cdot 10^2$, $p = 0.124$, $q = 2$;
$a = 0.0013 \cdot 10^4 = 0.13 \cdot 10^2$, $p = 0.13$, $q = 2$.

Fissato N, s e la coppia (p, q) della rappresentazione floating-point normalizzata $a = (-1)^s p N^q$ individuano univocamente il numero reale a. Pertanto per memorizzare a è sufficiente memorizzare il segno di a, p e q. Un calcolatore riserva per la memorizzazione di tali quantità uno spazio finito.

Ad esempio,

1) p può avere al massimo t cifre nel sistema di numerazione scelto e
2) $m \leq q \leq M$ con $m < 0$ e $M > 0$ interi.

Numeri di macchina

È evidente che non tutti i numeri reali soddisfano le condizioni dell'aritmetica fissata su di un calcolatore.
$\Rightarrow$ Non tutti i numeri reali sono esattamente rappresentabili su di un calcolatore.
Si definiscono allora **numeri di macchina** quei numeri le cui mantissa e caratteristica sono esattamente rappresentabili negli spazi a loro riservati. L'insieme dei numeri floating point è denotato con $\mathbb{F}$.

Esempio

per $N = 10$, $t = 5$, $m = -127$, $M = 128$,
$a = 1.58291 = 0.158291 \cdot 10^1$ non è un numero di macchina;
$a = 0.0038245 = 0.38245 \cdot 10^{-2}$ è un numero di macchina;
$a = 12.29 \cdot 10^{128} = 0.1229 \cdot 10^{130}$ non è un numero di macchina.

Quindi ogni numero reale, se non è un numero di macchina, deve essere approssimato con un numero di macchina a lui "vicino", se possibile. Dato un generico a, denotiamo con $\bar{a}$ la sua approssimazione di macchina, se esiste. Ci sono due tecniche per determinare $\bar{a}$ a partire da a:

i) **tecnica di troncamento**: si esclude la parte a destra della t-esima cifra;
ii) **tecnica di arrotondamento**: si aggiunge $\frac{1}{2}N^{-t}$ a p e poi si tronca il risultato alla t-esima cifra.

Esempio

For $N = 10,\ t = 5,$

$a = 0.158291 \cdot 10^1, i) \Rightarrow \bar{a} = 0.15829 \cdot 10^1;$

$a = 0.158291 \cdot 10^1, ii) \Rightarrow \bar{a} = 0.15829 \cdot 10^1;$

$a = 0.158298 \cdot 10^1, i) \Rightarrow \bar{a} = 0.15829 \cdot 10^1.$

$a = 0.158298 \cdot 10^1, ii) \Rightarrow \bar{a} = 0.15830 \cdot 10^1.$

Sia $a = (-1)^s p N^q$ con $N^{-1} \leq p < 1$.

Se $q < m$, il numero a non è rappresentabile a causa di un problema di **underflow**.

Se $q > M$, il numero a non è rappresentabile a causa di un problema di **overflow**.
In questi casi il calcolatore segnala il tipo di problema che si è verificato.

Sia $\bar{a}$ un'approssimazione del numero a, si definisce **errore assoluto** la quantità

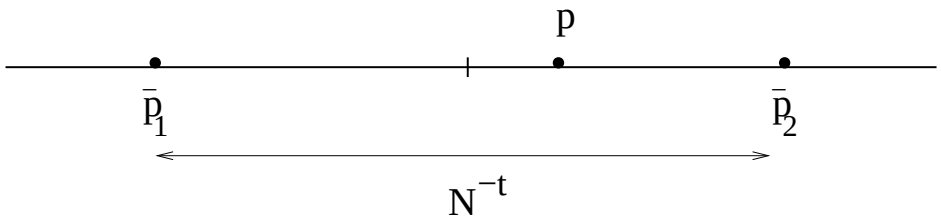
$$e_a = |a - \bar{a}|;$$

errore relativo la quantità

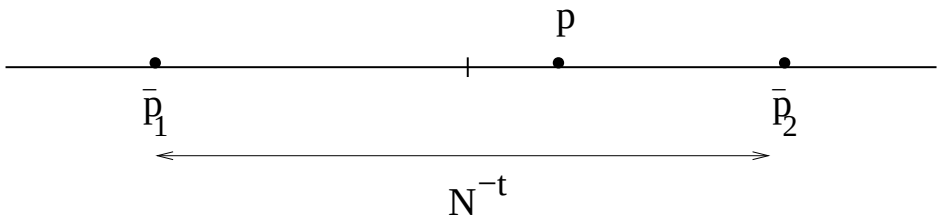
$$e_r = \frac{|a - \bar{a}|}{|a|}.$$

Vogliamo stimare gli errori assoluto e relativo che si commettono quando si approssima il numero reale a con il numero di macchina $\bar{a}$ ottenuto con le tecniche i) e ii).

Sia $a = (-1)^s p N^q$ con $N^{-1} \leq p < 1$ e $\bar{a} = (-1)^s \bar{p} N^q$, dove $\bar{p}$ è stato ottenuto applicando i) o ii) a p.
$\bar{p}$ non ha più di t cifre, e la distanza tra due mantisse di macchina consecutive è esattamente N^{-t}.



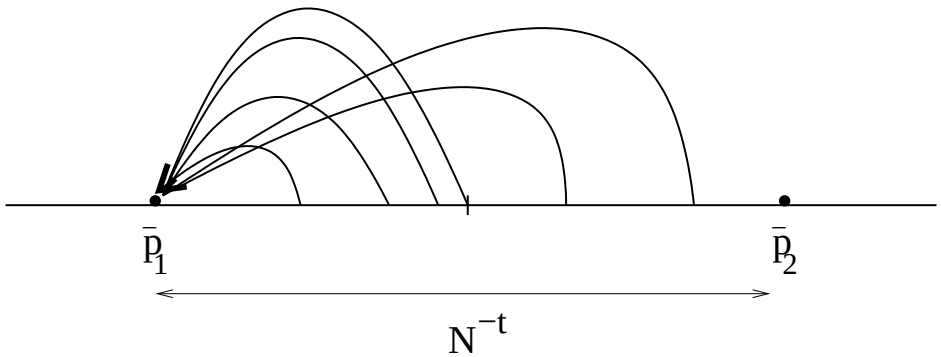
Sia $a = (-1)^s p N^q$ con $N^{-1} \leq p < 1$ e $\bar{a} = (-1)^s \bar{p} N^q$, dove $\bar{p}$ è stato ottenuto applicando i) o ii) a p.
$\bar{p}$ non ha più di t cifre, e la distanza tra due mantisse di macchina consecutive è esattamente N^{-t}.



Utilizzando la tecnica i)

- se $p \in (\bar{p}_1, \bar{p}_2)$, con $\bar{p}_2 = \bar{p}_1 + N^{-t}$, $\Rightarrow \bar{p} = \bar{p}_1$;

Sia $a = (-1)^s p N^q$ con $N^{-1} \leq p < 1$ e $\bar{a} = (-1)^s \bar{p} N^q$, dove $\bar{p}$ è stato ottenuto applicando i) o ii) a p.
$\bar{p}$ non ha più di t cifre, e la distanza tra due mantisse di macchina consecutive è esattamente N^{-t}.

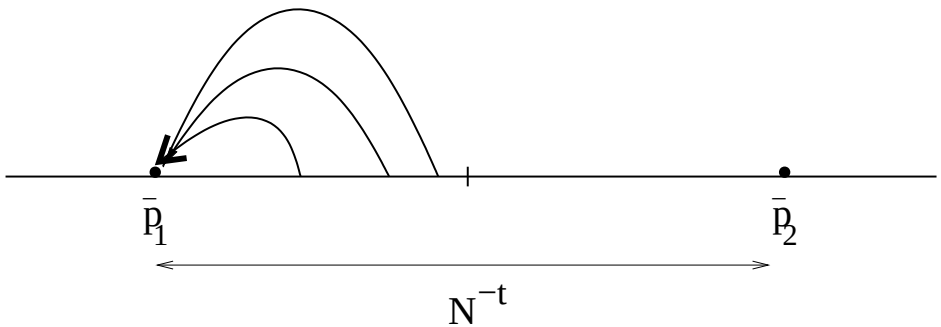


Utilizzando la tecnica i)

- se $p \in (\bar{p}_1, \bar{p}_2)$, con $\bar{p}_2 = \bar{p}_1 + N^{-t}$, $\Rightarrow \bar{p} = \bar{p}_1$;

$$\Rightarrow \qquad |p - \bar{p}| < N^{-t}.$$

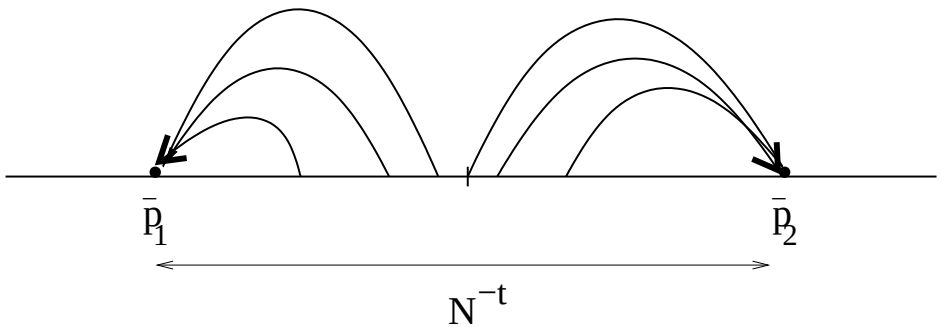
Sia $a = (-1)^s p N^q$ con $N^{-1} \leq p < 1$ e $\bar{a} = (-1)^s \bar{p} N^q$, dove $\bar{p}$ è stato ottenuto applicando i) o ii) a p.
$\bar{p}$ non ha più di t cifre, e la distanza tra due mantisse di macchina consecutive è esattamente N^{-t}.



Utilizzando la tecnica ii)

- se $p \in (\bar{p}_1, \bar{p}_1 + 1/2N^{-t})$, $\Rightarrow \bar{p} = \bar{p}_1$

Sia $a = (-1)^s p N^q$ con $N^{-1} \leq p < 1$ e $\bar{a} = (-1)^s \bar{p} N^q$, dove $\bar{p}$ è stato ottenuto applicando i) o ii) a p.
$\bar{p}$ non ha più di t cifre, e la distanza tra due mantisse di macchina consecutive è esattamente N^{-t}.



Utilizzando la tecnica ii)

- se $p \in (\bar{p}_1, \bar{p}_1 + 1/2N^{-t})$, $\Rightarrow \bar{p} = \bar{p}_1$
- se $p \in [\bar{p}_1 + 1/2N^{-t}, \bar{p}_2)$, $\Rightarrow \bar{p} = \bar{p}_2$

$$\Rightarrow \qquad |p - \bar{p}| \leq \frac{1}{2} N^{-t}.$$

Quindi, l'errore assoluto soddisfa

$$|a - \bar{a}| \begin{cases} < N^{q-t}, & \textit{per la tecnica i)} \\ \leq \frac{1}{2} N^{q-t}, & \textit{per la tecnica ii)} \end{cases},$$

l'errore relativo soddisfa (poichè $|p| \geq N^{-1}$ implica $|a| \geq N^{q-1}$),

$$\frac{|a - \bar{a}|}{|a|} \leq \frac{|a - \bar{a}|}{N^{q-1}} \begin{cases} < N^{1-t}, & \textit{per la tecnica i)} \\ \leq \frac{1}{2} N^{1-t}, & \textit{per la tecnica ii)} \end{cases}$$

La tecnica ii) è migliore della tecnica i)
i) è meno cara, ma produce un errore più grande (il doppio di quello prodotto dalla tecnica ii)).

La quantità

$$eps = \begin{cases} N^{1-t}, & \textit{per la tecnica } i) \\ \frac{1}{2}N^{1-t}, & \textit{per la tecnica } ii) \end{cases}$$

definisce la cosiddetta **precisione di macchina** oppure **errore di roundoff**. Essa è una costante caratteristica di ogni aritmetica floating-point e rappresenta una misura della massima precisione di calcolo raggiungibile.
Pertanto se $\bar{a}$ denota il numero di macchina corrispondente al numero reale a, posto $\varepsilon = (\bar{a} - a)/a$, possiamo scrivere

$$\bar{a} = a(1 + \varepsilon) \text{ con } |\varepsilon| \leq eps$$

Operazioni di macchina

Il risultato di un'operazione aritmetica tra due numeri di macchina generalmente non è un numero di macchina.

Esempio

Per $N = 10$, $t = 4$, $a_1 = 0.5823 = \bar{a}_1$, $a_2 = 0.6214 = \bar{a}_2$

$a_1 + a_2 = 1.2037 = 0.12037 \cdot 10$ non è un numero di macchina.

$\Rightarrow$ Pertanto in un calcolatore non è possibile eseguire esattamente le operazioni aritmetiche.

Abbiamo bisogno di introdurre le **operazioni di macchina**

Un' operazione di macchina associa a due numeri di macchina un terzo numero di macchina, ottenuto come segue:

1) si esegue l'operazione aritmetica esatta tra i due numeri di macchina.
2) si applica la tecnica i) (troncamento) o ii) (arrotondamento) al risultato dell'operazione esatta.

Se indichiamo con a_i, $i = 1, 2$ i valori esatti, con $\bar{a}_i$ i corrispondenti valori di macchina, con $\cdot$ l'operazione aritmetica esatta e con $\odot$ la corrispondente operazione di macchina, si ha

$$\bar{a}_1 \oplus \bar{a}_2 = \overline{\bar{a}_1 + \bar{a}_2} = (\bar{a}_1 + \bar{a}_2)(1 + \varepsilon_\oplus)$$

$$\bar{a}_1 \ominus \bar{a}_2 = \overline{\bar{a}_1 - \bar{a}_2} = (\bar{a}_1 - \bar{a}_2)(1 + \varepsilon_\ominus)$$

$$\bar{a}_1 \otimes \bar{a}_2 = \overline{\bar{a}_1 \times \bar{a}_2} = (\bar{a}_1 \times \bar{a}_2)(1 + \varepsilon_\otimes)$$

$$\bar{a}_1 \oslash \bar{a}_2 = \overline{\bar{a}_1 / \bar{a}_2} = (\bar{a}_1 / \bar{a}_2)(1 + \varepsilon_\oslash)$$

con $|\varepsilon_\odot| \leq eps$.

Esempio

Per $N = 10$, $t = 4$, $a_1 = 0.5823 = \bar{a}_1$, $a_2 = 0.6214 = \bar{a}_2$

$$\bar{a}_1 \oplus \bar{a}_2 = \begin{cases} 0.1203 \cdot 10, & \textit{per la tecnica } i) \\ 0.1204 \cdot 10, & \textit{per la tecnica } ii) \end{cases}$$

Per le operazioni di macchina rimane valida la proprietà commutativa, ma non valgono in generale le proprietà associativa e distributiva. Per esempio,

$$\bar{a}_1 \oplus \bar{a}_2 = \bar{a}_2 \oplus \bar{a}_1,$$

$$\bar{a}_1 \otimes \bar{a}_2 = \bar{a}_2 \otimes \bar{a}_1$$

$$\bar{a}_1 \oplus (\bar{a}_2 \oplus \bar{a}_3) \neq (\bar{a}_1 \oplus \bar{a}_2) \oplus \bar{a}_3$$

$$\bar{a}_1 \otimes (\bar{a}_2 \otimes \bar{a}_3) \neq (\bar{a}_1 \otimes \bar{a}_2) \otimes \bar{a}_3$$

$$\bar{a}_1 \otimes (\bar{a}_2 \oplus \bar{a}_3) \neq (\bar{a}_1 \otimes \bar{a}_2) \oplus (\bar{a}_1 \otimes \bar{a}_3)$$

$$(\bar{a}_1 \otimes \bar{a}_2) \oslash \bar{a}_3 \neq (\bar{a}_1 \oslash \bar{a}_3) \otimes \bar{a}_2$$

Pertanto, due espressioni e_1, e_2 equivalenti in aritmetica esatta, non risultano generalmente tali in aritmetica con precisione finita. Ciononostante esse saranno ugualmente considerate **equivalenti nell'aritmetica del calcolatore** quando

$$\frac{|e_1 - e_2|}{|e_1|} \approx eps \quad \left(\text{oppure} \quad \frac{|e_1 - e_2|}{|e_2|} \approx eps\right)$$

Quando si utilizza la tecnica di arrotondamento ii), la precisione di macchina può venire definita anche nel seguente modo:

$$eps = \min\{\bar{a} \in \mathbb{F},\ \bar{a} > 0 :\ 1 \oplus \bar{a} > 1\},$$

Cancellazione numerica

La **cancellazione numerica** rappresenta una delle conseguenze più gravi della rappresentazione con precisione finita dei numeri reali. Essa consiste in una perdita di cifre della mantissa che **si verifica quando si esegue un'operazione di sottrazione fra due numeri** $a_1 = p_1 N^q$ e $a_2 = p_2 N^q$ **"quasi uguali"** (per esempio $\bar{p}_1$ e $\bar{p}_2$ hanno le prime s cifre coincidenti) **ed arrotondati ai numeri di macchina** $\bar{a}_1 = \bar{p}_1 N^q$ e $\bar{a}_2 = \bar{p}_2 N^q$ (con $p_1 \neq \bar{p}_1$ e/o $p_2 \neq \bar{p}_2$).

Esempio

Per $N = 10$, $t = 5$ e tecnica ii)

in aritmetica esatta		in aritmetica finita	
a_1	$= 0.157824831$	$\bar{a}_1$	$= 0.15782$
a_2	$= 0.157348212$	$\bar{a}_2$	$= 0.15735$
$a_1 - a_2$	$= 0.000476619$	$\bar{a}_1 \ominus \bar{a}_2$	$= 0.00047$
	$= 0.476619 \cdot 10^{-3}$		$= 0.47000 \cdot 10^{-3}$

$a_1 - a_2 = 0.476619 \cdot 10^{-3}$ $\quad$ $\bar{a}_1 \ominus \bar{a}_2 = 0.47000 \cdot 10^{-3}$

Osserviamo che nella mantissa di $\bar{a}_1 \ominus \bar{a}_2$ solo le prime 2 cifre decimali sono corrette. La perdita delle restanti cifre è dovuta agli errori presenti nei due operandi, che vengono amplificati dall'operazione di sottrazione.

Ricordiamo che tutte le operazioni aritmetiche di macchina possono provocare un errore che non supera mai la precisione di macchina. Pertanto, **se i due operandi sono privi di errori, il risultato dell'operazione di sottrazione non presenta alcuna perdita di precisione**.

Talvolta, manipolando opportunamente le espressioni matematiche che definiscono un problema, è possibile evitare il fenomeno della cancellazione numerica in esso presente; quando ciò non è possibile si dice che la cancellazione è insita nel problema.

Esempio

- $y = \sqrt{x+\delta} - \sqrt{x}$, con $x > 0$. Se $|\delta| << x$ il fenomeno della cancellazione numerica si elimina nel seguente modo:
  $y = (\sqrt{x+\delta} - \sqrt{x})\ \frac{\sqrt{x+\delta}+\sqrt{x}}{\sqrt{x+\delta}+\sqrt{x}} = \frac{\delta}{\sqrt{x+\delta}+\sqrt{x}}$;
  se $\delta \approx -x$ il fenomeno non è eliminabile;
- $y = \frac{1-\cos(x)}{x^2}$, con $x \approx 0$. Il fenomeno della cancellazione numerica si elimina utilizzando $1 - \cos(x) = 2\sin^2(x/2)$:
  $y = \frac{2\sin^2(x/2)}{x^2} = \frac{1}{2}\left(\frac{\sin(x/2)}{x/2}\right)^2$;
- $y = \frac{x-\sin(x)}{\tan(x)}$, con $x \approx 0$. Il fenomeno della cancellazione numerica si elimina utilizzando il polinomio di Taylor di punto iniziale 0 e di grado opportuno:
  $y = \frac{x-(x-x^3/3!+x^5/5!-x^7/7!+\ldots)}{\tan(x)} = \frac{x^3/3!-x^5/5!+x^7/7!-\ldots}{\tan(x)}$.

Condizionamento di un problema numerico

Consideriamo il problema numerico

$$y = f(x)$$

ove x rappresenta l'input, y l'output ed f la connessione esplicita o implicita tra x ed y.

Esempio

$Ax = b$, con A, b input e x output.

$y(x) = \int_a^x g(t)dt$

Indichiamo con x i dati di input e con $\bar{x}$ una loro perturbazione, con $f(x)$ ed $f(\bar{x})$ i corrispondenti dati di output, ottenuti in precisione infinita di calcolo. Se accade che

$$\frac{||f(x) - f(\bar{x})||}{||f(x)||} \approx \frac{||x - \bar{x}||}{||x||}, \ x, f(x) \neq 0,$$

ovvero gli errori relativi sui dati di output e di input hanno lo stesso ordine di grandezza, allora il problema $y = f(x)$ si dice **ben condizionato**; altrimenti si dice **mal condizionato**.
Per studiare il condizionamento di un problema, occorre determinare relazioni del tipo

$$\frac{||f(x) - f(\bar{x})||}{||f(x)||} \approx K(f, x) \frac{||x - \bar{x}||}{||x||}$$

oppure del tipo

$$\frac{||f(x) - f(\bar{x})||}{||f(x)||} \leq K(f, x) \frac{||x - \bar{x}||}{||x||}$$

La quantità $K(f, x)$ prende il nome di **numero di condizionamento** del problema. Se $K \approx 1$ il problema è ben condizionato.

Esempio

$y = x_1 + x_2,\ x_1, x_2 \in \mathbb{R}$

$$\frac{x_1+x_2-(\bar{x}_1+\bar{x}_2)}{x_1+x_2} = \frac{x_1+x_2-x_1(1+\varepsilon_1)-x_2(1+\varepsilon_2)}{x_1+x_2} = -\frac{x_1}{x_1+x_2}\varepsilon_1 - \frac{x_2}{x_1+x_2}\varepsilon_2$$

Pertanto, le quantità $K_i = |x_i/(x_1 + x_2)|$ $i = 1, 2$ rappresentano i numeri di condizionamento del problema. Il problema è mal condizionato ($K_i \to \infty$) se $x_1 + x_2 \to 0$.

Stabilità di un algoritmo

Per **algoritmo** si intende una sequenza finita di operazioni (aritmetiche e non) che consente di ottenere l'output y^* del problema $y = f(x)$ (non necessariamente uguale ad y) a partire dall'input x.
Per giudicare la bontà di un algoritmo per la risoluzione di un problema $y = f(x)$, indichiamo con y^* la risposta fornita dall'algoritmo a partire dai dati $\bar{x}$ perturbati dall'errore di arrotondamento ed ottenuta in precisione finita di calcolo, e con $f(\bar{x})$ la soluzione del problema a partire dai dati $\bar{x}$ ed ottenuta in precisione infinita di calcolo. Se accade che

$$\frac{||f(\bar{x}) - y^*||}{||f(\bar{x})||} \approx eps, \quad f(\bar{x}) \neq 0,$$

ovvero l'errore relativo sui dati di output ha lo stesso ordine di grandezza della precisione di macchina, allora l'algoritmo si dice **numericamente stabile**; altrimenti si dice **instabile**.

Esempio L'algoritmo precedentemente descritto per il calcolo di π è instabile. L'algoritmo che si ottiene razionalizzando

$$
\begin{aligned}
x_1 &= 2 \\
x_n &= 2^{n-1/2} \sqrt{\frac{4^{1-n} x_{n-1}^2}{1 + \sqrt{1 - 4^{1-n} x_{n-1}^2}}}, \qquad n \geq 2
\end{aligned}
$$

è invece stabile.

Figura : Grafico dell'errore relativo $|x_n - \pi|/\pi$ in scala logaritmica