APPUNTI DI CALCOLO NUMERICO

Introduzione al calcolo numerico
Rappresentazione dei numeri sul calcolatore
Stabilità e condizionamento

Metodi numerici

Un fenomeno fisico può essere rappresentato attraverso un modello matematico. L'uso di un metodo numerico consente la risoluzione approssimata di tale modello matematico con l'uso di un calcolatore.
In pratica, la risoluzione di un problema con un calcolatore si riconduce all'implementazione di un algoritmo, ossia di una serie di istruzioni univocamente definite ed eseguibili in un tempo finito.

I metodi numerici servono per risolvere problemi troppo complessi, come ad esempio la risoluzione dell'integrale $\int_a^b e^{-x^2} dx$, in quanto si dimostra che non è possibile scriverne una primitiva in forma semplice, oppure problemi di cui si conosce perfettamente il metodo risolutivo, ma che risultino troppo grandi da trattare a mano, ad esempio un sistema lineare con 100 equazioni in 100 incognite.

Sistema Floating Point

Il sistema floating point (virgola mobile) è un modo di rappresentare i numeri che i calcolatori normalmente utilizzano. In sostanza, fissata una base β di un sistema di numerazione, la rappresentazione floating point di un numero reale consiste in una mantissa moltiplicata per un'opportuna potenza di β, in modo tale da non avere né parte intera, né zeri dopo la virgola.

Ad esempio con $\beta = 10$:

$$123.4567 \Rightarrow 0.1234567 * 10^3$$
$$0.00789 \Rightarrow 0.789 * 10^{-2}$$
$$0.6 \Rightarrow 0.6 * 10^0$$

L'esponente della base viene definito caratteristica.
Si può osservare che questo sistema è posizionale, infatti vale:

$$0.789 = 7 * 10^{-1} + 8 * 10^{-2} + 9 * 10^{-3}.$$

Numeri macchina

L'insieme dei numeri macchina con t cifre di mantissa, base β e range $[L,U]$ è definito come segue:

$$F(\beta,t,L,U)=\{0\}\cup\left\{x\in\Re : x=(-1)^s\,\beta^e\sum_{i=1}^{t}d_i\beta^{-i}\right\}$$

dove:

- t,β sono interi positivi, $\beta\geq 2$;
- $0\leq d_i\leq\beta-1,\ i=1,2,...,t$
- $d_1\neq 0$ rappresentazione normalizzata;
- $L\leq e\leq U$, U generalmente positivo e L negativo;
- $e=$ caratteristica
- $\sum_{i=1}^{t}d_i\beta^{-i}=$ mantissa, variabile fra β^{-1} e $1-\beta^{-t}$.

Per ogni numero reale $x\in F(\beta,t,L,U)$ si ha

$$x_{\min}=\beta^{L-1}\leq|x|\leq\beta^U\left(1-\beta^{-t}\right)=x_{\max}$$

non è cioè possibile rappresentare alcun numero minore di $x_{\min}$, a parte lo zero, senza commettere un errore di underflow. Similmente non è possibile rappresentare alcun numero maggiore di $x_{\max}$ senza commettere un errore di overflow.

Errore

Sia $x\in\Re$ e $\bar{x}$ una sua approssimazione

$$x=(-1)^S\,m\beta^e$$
$$\bar{x}=(-1)^S\,\bar{m}\beta^e.$$

Si definisce <u>errore assoluto</u> la quantità

$$E_a=|\bar{x}-x|.$$

Questo tipo di errore non porta informazioni sull'ordine di grandezza di x.

Si definisce <u>errore relativo</u> la quantità

$$E_r=\frac{|\bar{x}-x|}{x}=\frac{E_a}{x}$$

formula valida $\forall x\neq 0$.

In questo caso si hanno informazioni anche sull'ordine di grandezza di x.

Tecniche di approssimazione

Per poter rappresentare correttamente un numero reale $x = \text{sgn}(x) m \beta^e$ in numeri macchina, è necessario che esso cada nel range $[L,U]$. Se la sua mantissa ha già un numero di cifre pari a t, allora si è già in presenza di un numero macchina, altrimenti è necessario approssimare x.
I due principali metodi di approssimazione sono:

- arrotondamento: x viene approssimato con il numero macchina più vicino; in caso di equidistanza di x da entrambi i numeri macchina, maggiore e minore, si approssima il numero reale con il numero macchina che ha la cifra meno significativa della mantissa (l'ultimo a destra) pari. Questo è il metodo applicato dai calcolatori.
- troncamento: x viene approssimato al numero macchina più grande il cui valore assoluto sia minore di x.

Esempio:

Con il metodo dell'arrotondamento:

$$0.163 \Rightarrow 0.16$$
$$0.168 \Rightarrow 0.17$$
$$0.165 \Rightarrow 0.16$$
$$0.175 \Rightarrow 0.18$$

Con il metodo del troncamento:

$$0.142 \Rightarrow 0.14$$
$$0.147 \Rightarrow 0.14$$

Due mantisse macchina distano fra di loro di una quantità pari a β^{-t}. Tutti i numeri che si trovano fra due mantisse macchina devono essere approssimati, perché il calcolatore non li distingue.

È necessario ora calcolare i massimi errori che si possono commettere con i due metodi di approssimazione.

Nel caso del troncamento, dato che tutte le mantisse $m \in \left[m_1, m_1 + \beta^{-t}\right)$ sono approssimate a m_1, il massimo errore commettibile risulta essere $m - m_1 < \beta^{-t}$.

Nel caso dell'arrotondamento, invece, dato che tutte le mantisse $m \in \left(m_1 - \frac{1}{2}\beta^{-t}, m_1 + \frac{1}{2}\beta^{-t} \right]$ sono approssimate a m_1, il massimo errore commettibile risulta essere $\left|m - m_1\right| \le \frac{1}{2}\beta^{-t}$.

Stima dell'errore assoluto

Riassumendo i risultati fin qui ottenuti si ha, per quanto riguarda il troncamento,

$$\left|\bar{x} - x\right| < \beta^{e-t},$$

per quanto concerne, invece, l'arrotondamento

$$\left|\bar{x} - x\right| \leq \frac{1}{2}\beta^{e-t}.$$

Dato che $m \geq 0.10000 = \beta^{-1}$ si ha (minorazione di m)

$$\left|x\right| = m\beta^{e} \geq \beta^{-1}\beta^{e}$$

quindi

$$\text{troncamento: } \frac{\left|\bar{x} - x\right|}{\left|x\right|} \leq \frac{\left|\bar{x} - x\right|}{\beta^{e-1}} < \varepsilon_m \equiv \beta^{1-t}$$

$$\text{arrotondamento: } \frac{\left|\bar{x} - x\right|}{\left|x\right|} \leq \frac{\left|\bar{x} - x\right|}{\beta^{e-1}} \leq \varepsilon_m \equiv \frac{1}{2}\beta^{1-t}.$$

Il termine β^{1-t} viene detto epsilon di macchina e viene indicato con $eps \equiv \beta^{1-t}$.
Con il simbolo ε_m si indica la precisione di macchina, ossia la massima precisione relativa di calcolo raggiungibile sul calcolatore.
Due quantità che differiscono meno della precisione di macchina sono considerate uguali, poiché il calcolatore non le distingue.
Non ha senso cercare di determinare approssimazioni con precisione relativa minore di ε_m.

I numeri macchina possono essere rappresentati in singola precisione o in doppia precisione.

Singola precisione:

32 bit, di cui 1 per il segno, 8 per la caratteristica e 23 per la mantissa
base 2: $L = -127$, $U = 128$ e $eps = 2^{-22}$

Doppia precisione:

64 bit, di cui 1 per il segno, 11 per la caratteristica e 52 per la mantissa
base 2: $L \cong -1023$, $U = 1024$ e $eps = 2^{-52}$
base 10: $L = -308$, $U = 308$ e $eps = 10^{-16}$

Operazioni di macchina

Definiamo $fl(\cdot)$ la funzione di approssimazione. Il risultato di un'operazione di macchina è il risultato dell'operazione eseguita sui numeri macchina a cui viene applicata $fl(\cdot)$.

Esempio: somma di macchina

$$a \oplus b = fl\left(fl(a) + fl(b) \right)$$

Ogni operazione di macchina introduce un errore, prescindendo dagli errori delle approssimazioni di a e di b, definito come

$$\delta = \frac{fl(x) - x}{x}.$$

N.B.: NON vale la proprietà associativa e NON esistono relazioni di operazione inversa fra somma/sottrazione e prodotto/divisione.
Inoltre se uno dei due termini su cui effettuare l'operazione di macchina è nettamente più piccolo dell'altro, è frequente che il suo contributo venga perso.

Due espressioni si dicono equivalenti se i loro risultati distano fra loro di una quantità dello stesso ordine di grandezza di ε_m.

Cancellazione numerica

Consiste nel fenomeno della perdita di cifre significative in seguito all'operazione di sottrazione fra due numeri molto vicini fra di loro.

Esempio:

Siano

$x_1 = 0.19101972 \cdot 10^3$
$x_2 = 0.19101708 \cdot 10^3$
$\varepsilon_m = 10^{-5}$

Calcolo delle approssimazioni

$fl(x_1) = 0.191019 \cdot 10^3$
$fl(x_2) = 0.191017 \cdot 10^3$

Calcolo degli errori

$$\frac{|fl(x_1)-x_1|}{x_1} < \varepsilon_m$$
$$\frac{|fl(x_2)-x_2|}{x_2} < \varepsilon_m$$

Operazione di macchina

$$fl(x_1)\oplus[-fl(x_2)]$$
$$x_1 - x_2 = 0.264000\cdot 10^{-2}$$

Calcolo delle errore

$$\frac{|fl(x_1)\oplus[-fl(x_2)]-(x_1-x_2)|}{|(x_1-x_2)|} = 0.2424 \cong 24\% >> \varepsilon_m$$

l'errore relativo che si compie è decisamente troppo grande.

Come si spiega questo fatto?
Nel momento in cui abbiamo approssimato le mantisse di x_1 e di x_2 sono state buttate via le cifre decimali oltre la sesta (nei limiti della precisione di macchina), quindi l'incertezza si è "concentrata" sull'ultima cifra rimasta. Applicando la sottrazione la cifra significativa con incertezza risale fino, al massimo, alla prima cifra decimale. In pratica si amplifica l'errore di approssimazione.

Anche se avessimo operato applicando l'arrotondamento e non il troncamento, non avremmo ottenuto risultati migliori: si ottiene, infatti, un errore relativo pari al 13%, ossia si ottiene un errore che continua ad essere dello stesso ordine di grandezza del risultato.

Perciò l'errore non nasce dall'operazione in sé, ma dalle approssimazioni, opportunamente amplificate dalla sottrazione.

Per evitare il fenomeno di cancellazione numerica, è necessario trovare altre vie per ottenere lo stesso risultato, ossia se è possibile utilizzare forme alternative per evitare la sottrazione di macchina è bene concentrarsi su quelle.

Esempio:

Per $h \to 0$

$$f'(x_0) \cong \frac{f(x_0+h)-f(x_0)}{h} \text{ (rapporto incrementale)}$$

Quindi se h è molto piccolo si ha che $f(x_0+h)\cong f(x_0)$ e quindi si ricade nel caso precedente (cancellazione numerica). Devo cercare modi alternativi per calcolare la medesima derivata prima.

Ipotizziamo che sia $f(x)=\sin(x)$. Applicando le formule di prostaferesi è possibile trasformare la differenza di seni in un prodotto di funzioni trigonometriche:

$$\frac{\sin(x_0+h)-\sin(x_0)}{h}=\frac{2}{h}\cos\frac{2x_0+h}{2}\sin\frac{h}{2}$$

in questo modo posso ottenere lo stesso risultato evitando la cancellazione numerica, in quanto ho eliminato la sottrazione di macchina.

Stabilità di un algoritmo e condizionamento di un problema

È necessario comprendere come gli errori vengano amplificati e si propaghino dai dati al risultato di un problema, in base a come è posto il problema stesso e in base all'algoritmo utilizzato per risolverlo.

Un problema si dice ben posto se ammette una e una sola soluzione. In caso di soluzioni multiple, ad esempio con le equazioni di grado superiore al primo, è necessario specificare quale delle soluzioni si sta cercando e calcolarne una alla volta. Inoltre i risultati devono dipendere con continuità dai dati, ossia a piccole variazioni dei dati non devono corrispondere variazioni enormi dei risultati.
Se il problema non soddisfa queste richieste, allora non è ben posto.
Ci occuperemo solamente di problemi ben posti.

Consideriamo un generico problema che ha un dato d e una relazione funzionale che lega il risultato x al dato: $x=f(d)$.
Nomenclatura:

- δd è una generica perturbazione del dato in ingresso, ad esempio l'errore di approssimazione dovuto alla memorizzazione dei numeri in un calcolatore;
- $\bar{x}=f(x+\delta d)$ è la soluzione esatta del problema con ingresso $d+\delta d$
- $\tilde{x}$ è la risposta dell'algoritmo risolutivo al dato $d+\delta d$, generalmente si ha che $\tilde{x}\neq\bar{x}$.

Esempio:

Risolvere l'equazione lineare $2x=5$.

Supponiamo di introdurre una perturbazione sul termine noto, quindi il problema diventa $2\bar{x}=5.0001$; la soluzione è

$$\bar{x}=\frac{5.0001}{2}=2.500005$$

Supponiamo che, per qualche motivo, il calcolatore restituisca un numero diverso. Questo numero è, ad esempio, $\tilde{x} = 2.5001$. E si verifica che $\bar{x} \neq \tilde{x}$.

Condizionamento

È lo studio di come il problema reagisce alle perturbazioni sui dato. Cioè si vuole determinare la variazione del risultato δx avendo in ingresso una perturbazione δd del dato.

Se un δd piccolo dà δx piccoli, il problema si dice ben condizionato, altrimenti (a δd piccoli corrispondono δx grandi) si dice mal condizionato.

Numero di condizionamento

È una costante moltiplicativa K che amplifica l'errore sul risultato. Essa può essere grande o piccola. Un numero di condizionamento grande implica che una piccolissima perturbazione sui dati diviene enorme sui risultati (problema mal condizionato). Al contrario, un K piccolo significa che il problema è ben condizionato.

$$\frac{\|\delta x\|}{\|x\|} \leq K \frac{\|\delta d\|}{\|d\|}$$

$$\frac{\|\delta x\|}{\|x\|} \cong K \frac{\|\delta d\|}{\|d\|}$$

Nel primo caso la perturbazione sui risultati può arrivare fino a K volte la perturbazione sui dati; nel secondo caso, invece, la perturbazione sui risultati è circa K volte la perturbazione sui dati.

Esempio: somma di due termini

Dati:

$d_1 = a$

$d_2 = b$

Relazione funzionale:

$x = a + b$

Dati perturbati:

$\bar{a} = a + \delta a$

$\bar{b} = b + \delta b$

Risultato perturbato:

$\bar{x} = x + \delta x$

Relazione funzionale perturbata:

$\bar{x} = \bar{a} + \bar{b}$

Sostituendo:

$x + \delta x = a + \delta a + b + \delta b$

Poiché $x = a + b$ si ottiene:

$\delta x = \delta a + \delta b$

Se ci si potesse fermare qui si potrebbe dire che se δa e δb sono piccoli, anche δx lo è. Tuttavia bisogna ancora inserire il numero di condizionamento K.

$$\frac{|\bar{x} - x|}{|x|} = \frac{|\delta x|}{|x|} = \frac{|\delta a + \delta b|}{|a+b|} \leq \frac{|\delta a|}{|a+b|} + \frac{|\delta b|}{|a+b|} = \frac{|a|}{|a+b|}\frac{|\delta a|}{|a|} + \frac{|b|}{|a+b|}\frac{|\delta b|}{|b|}$$

Chiamiamo $K_a = \frac{|a|}{|a+b|}$ e $K_b = \frac{|b|}{|a+b|}$ i due coefficienti di amplificazione rispettivamente degli errori su a e su b.
Il numero di condizionamento sarà il più grande fra i coefficienti di amplificazione: $K = \max(K_a, K_b)$.

Si nota che il problema sarà sicuramente mal condizionato se le due quantità a e b sono molto simili e di segno opposto. Infatti si verifica che se $a + b \to 0$ allora $K \to \infty$.

Stabilità numerica di un algoritmo

Significa studiare come l'algoritmo amplifica gli errori.
Un algoritmo è numericamente stabile se gli errori sui risultati sono quasi dello stesso ordine di grandezza di ε_m. In formule, un algoritmo è numericamente stabile se:

$$\frac{\|\tilde{\bar{x}} - \bar{x}\|}{\|\bar{x}\|} \cong \varepsilon_m .$$

Una differenza di macchina, con cancellazione numerica, è sicuramente un algoritmo instabile.