Corso di Analisi Numerica - AN410

Parte 1: introduzione e nozioni preliminari

Roberto Ferretti

ROMA TRE
UNIVERSITÀ DEGLI STUDI

DMF

- Il Calcolo Scientifico
- Questioni modellistiche ed analitiche
- Il concetto di *condizionamento* di un problema
- Questioni informatiche
- I concetti di *parametro di discretizzazione* e di *accuratezza*
- Il concetto di *stabilità* di uno schema numerico
- Operazioni in aritmetica finita

Il Calcolo Scientifico

• Nella sua accezione più generale, il *Calcolo Scientifico* è un insieme di metodologie (analitiche, fisico–matematiche, numeriche, informatiche) che portano alla descrizione quantitativa di sistemi complessi. Le motivazioni sono principalmente di due tipi:

• Simulazione: Si tratta di prevedere il comportamento di un dato processo (ad esempio, le condizioni meteorologiche)

• Progetto: Si tratta di utilizzare un modello di un dato processo per ottimizzarne le prestazioni (ad esempio, l'aerodinamica delle carrozzerie)

Modellizzazione matematica del processo

Studio della buona posizione

$\downarrow\uparrow$

Schema di approssimazione

Convergenza e stabilità dello schema

$\downarrow\uparrow$

Implementazione efficiente dello schema

• Nei *sistemi complessi* l'interazione tra aspetti modellistici, numerici ed informatici è molto stretta e le scelte fatte in uno di questi campi influenzano quelle fatte negli altri

• La *buona riuscita* della operazione di simulazione dipende in modo cruciale da una buona integrazione tra tutti questi aspetti

indice

Questioni modellistiche ed analitiche

Ogni processo può essere descritto con modelli matematici di diversa complessità, tenendo conto o meno di certi meccanismi.

Esempio: l'equazione classica (lineare) del pendolo che viene derivata sotto le ipotesi di

• Piccole oscillazioni: ciò permette di evitare una equazione nonlineare sostituendo a $\sin x$ il suo sviluppo di Taylor di primo grado

• Assenza di attriti: ciò permette di evitare il termine di primo ordine, ma se gli attriti vengono tenuti in conto in genere ciò si fa ipotizzando attriti lineari (l'attrito dell'aria è invece circa quadratico)

Lo studio della buona posizione del modello dovrebbe garantire:

• Esistenza ed unicità della soluzione: non ha senso approssimare un problema che non ha soluzione, e se ne ha più di una andrebbe caratterizzata quella di interesse

• Dipendenza continua dai dati: necessaria perché ogni schema di approssimazione eseguito da un calcolatore reale introduce perturbazioni. La soluzione deve essere stabile rispetto a queste perturbazioni

indice

Il concetto di condizionamento di un problema

Supponendo che la relazione funzionale che lega la soluzione x ai dati d del problema sia nella forma

$$x = \mathcal{F}(d) \tag{1}$$

il numero di condizionamento del problema è il rapporto tra la variazione assoluta (relativa) di x e quella di d:

$$cond = \frac{\|\mathcal{F}(d+\delta d) - \mathcal{F}(d)\|}{\|\delta d\|} \qquad \left(cond = \frac{\frac{\|\mathcal{F}(d+\delta d) - \mathcal{F}(d)\|}{\|\mathcal{F}(d)\|}}{\frac{\|\delta d\|}{\|d\|}} \right) \tag{2}$$

cioè un coefficiente di amplificazione delle perturbazioni sui dati.

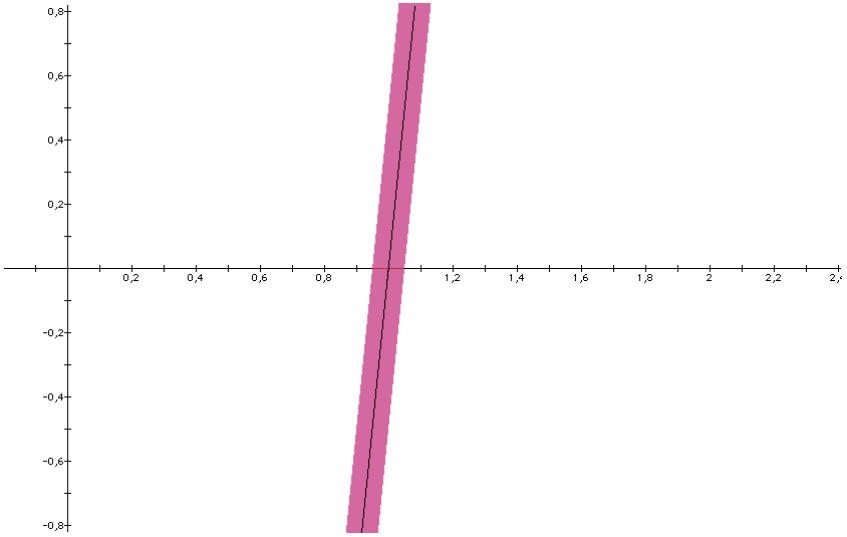
Esempio: soluzione di una equazione lineare $ax - b = 0$ al variare del termine noto b. Scrivendo la soluzione perturbata come $x + \delta x$ si ha:

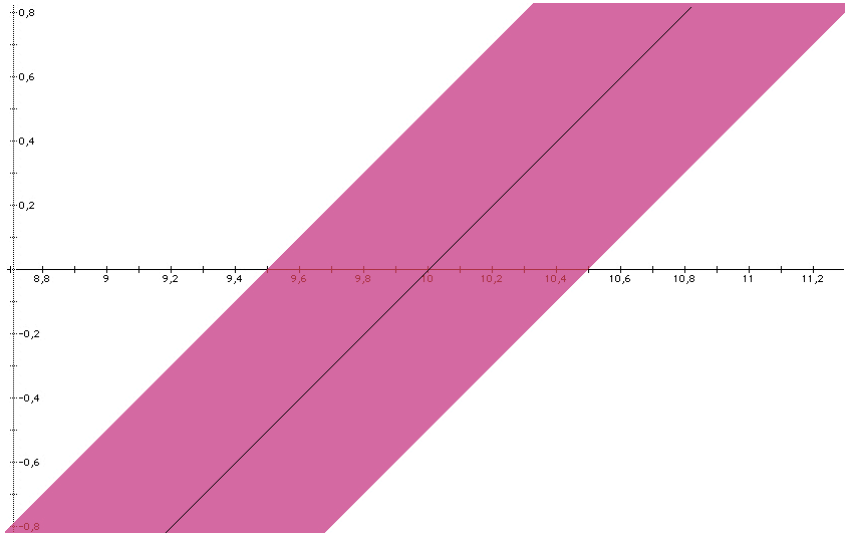
$$a(x + \delta x) = b + \delta b$$

da cui si ottiene

$$\frac{|\delta x|}{|\delta b|} = \frac{1}{|a|},$$

ovvero, come è intuitivo, che con piccoli coefficienti angolari la posizione della radice sia molto sensibile alle variazioni del termine noto.





$10x - 10 = 0$

$(\delta b \leq 0.5)$

$x - 10 = 0$

$(\delta b \leq 0.5)$

• Anche in problemi ben posti, numeri di condizionamento grandi indicano una elevata sensibilità alle perturbazioni e quindi una intrinseca difficoltà del problema ad essere approssimato accuratamente

• Anche in problemi ben posti e analiticamente ben condizionati, una ulteriore amplificazione delle perturbazioni può provenire dallo schema numerico

indice

Questioni informatiche

Nei problemi di grandi dimensioni, tipici del Calcolo Scientifico, viene data grande attenzioni alla complessità computazionale ed alla occupazione di memoria legate ad un certo algoritmo di approssimazione.

• Le scelte fatte a livello modellistico prima, numerico poi, si riflettono sulla reale calcolabilità della soluzione.

• Nel caso poi della implementazione su macchine parallele, gli algoritmi vanno di regola ripensati in questa chiave.

indice

Parametro di discretizzazione ed accuratezza

Nella grande maggioranza dei metodi numerici si possono individuare una o più grandezze che danno conto di "quanto accuratamente" si sia approssimato il problema originale, ovvero rispetto alle quali concettualmente si passa al limite. Tali grandezze vanno sotto il nome di parametri di discretizzazione.

• Di fatto non si può calcolare il limite (cioè la soluzione esatta) ma un elemento della successione di approssimazioni che sia "sufficientemente vicino" alla soluzione

• In casi semplici si può maggiorare esplicitamente la differenza tra limite e valore calcolato, in altri casi ci si accontenta di conoscere l'ordine di convergenza, che è comunque una misura di accuratezza, intesa come velocità di convergenza

• Poiché discretizzazioni più precise (metodi più accurati o calcolati più vicino al limite) presentano una maggiore complessità computazionale, l'efficienza della approssimazione è legata al compromesso tra velocità di convergenza e complessità

Esempio: approssimazione della derivata (ad esempio, di una funzione non nota in forma esplicita) mediante il rapporto incrementale. Se la funzione $f(x)$ è derivabile in x, la derivata si può ottenere come limite per $h \to 0$ del rapporto

$$\frac{f(x+h) - f(x)}{h}.$$

- h ha il ruolo di parametro di discretizzazione

- Se $f \in C^1$, non si può caratterizzare in alcun modo l'accuratezza, visto che l'unica informazione a disposizione è la convergenza di questo rapporto a $f'(x)$.

• Se però $f \in C^2$, il modulo dell'errore di discretizzazione si può calcolare (sostituendo ad $f(x+h)$ il suo sviluppo di Taylor di primo ordine) come

$$\left| f'(x) - \frac{f(x+h) - f(x)}{h} \right| = \left| f'(x) - f'(x) + \frac{f''(\xi)}{2} h \right| \leq \frac{\max_{[x,x+h]} |f''|}{2} h$$

che porta ad una maggiorazione esplicita dell'errore e ad una caratterizzazione dell'ordine di convergenza (l'errore è $O(h)$).

• Se $f \in C^3$ o ancora più regolare, non ci sono ulteriori vantaggi, e l'ordine di convergenza resta 1.

Altro modo di approssimare $f'(x)$: il rapporto incrementale centrato

$$\frac{f(x+h)-f(x-h)}{2h} = \frac{1}{2}\left(\frac{f(x+h)-f(x)}{h} + \frac{f(x)-f(x-h)}{h}\right)$$

che converge ad $f'(x)$ se $f \in C^1$.

• Esercizio: dimostrare che il rapporto incrementale centrato converge ad $f'(x)$ con ordine maggiore di 1 se $f \in C^2$, con ordine 2 se $f \in C^3$.

• Commento: questa situazione è del tutto generale e per avere una precisione elevata sono necessari sia una maggiore regolarità del problema che uno schema intrinsecamente più accurato

indice

Stabilità degli schemi numerici

Ponendo la soluzione approssimata del nostro problema nella forma:

$$\widehat{x} = \widehat{\mathcal{F}}_h(d) \tag{3}$$

chiamiamo stabilità un rapporto "basso" (uniformemente in h) tra la variazione di $\widehat{x}$ e quella di d, cioè una proprietà di buon condizionamento della funzione $\widehat{\mathcal{F}}_h$ rispetto alle perturbazioni sui dati:

- Dati provenienti da misure o da approssimazioni precedenti

- Dati affetti da errori di arrotondamento dovuti alla rappresentazione in aritmetica di macchina (perturbazione sempre presente)

Esempio: calcolo del rapporto incrementale destro di $f(x) = x^{1/3}$ in $x = 1$ con sei decimali (il valore esatto è $f'(1) = 1/3$).

h	10^{-1}	10^{-3}	10^{-5}	10^{-6}
$(1+h)^{1/3}$	1.032280	1.000333	1.000003	1.000000
$\frac{(1+h)^{1/3}-1}{h}$	0.3228	0.333	0.3	0.0

Commento: la perturbazione nel risultato non resta limitata

Un modo generale di analizzare questo fenomeno è di supporre che i due valori $f(x)$ e $f(x+h)$ siano affetti da perturbazioni δ_1, δ_2 tali che $|\delta_1|, |\delta_2| < \delta$. La differenza tra rapporto incrementale e derivata si può maggiorare come

$$\left| f'(x) - \frac{f(x+h) + \delta_1 - f(x) - \delta_2}{h} \right| \leq \cdots \leq \frac{\max_{[x,x+h]} |f''|}{2} h + \frac{2\delta}{h}$$

• Il primo di questi termini è l'errore di discretizzazione, ed è legato alla accuratezza della approssimazione.

• Il secondo termine va sotto il nome di errore di arrotondamento, ed è legato alla stabilità della approssimazione rispetto alle perturbazioni.

Esiste, a seconda dell'ordine di convergenza, una relazione ottimale tra h e la perturbazione δ che minimizza l'errore totale. Se ad esempio l'errore di discretizzazione è maggiorato da Ch^α (con $\alpha \geq 1$), l'errore totale è maggiorato da

$$\epsilon \leq Ch^\alpha + \frac{2\delta}{h}$$

ed il minimo del secondo membro si ottiene per

$$h = \left(\frac{2\delta}{C\alpha}\right)^{\frac{1}{\alpha+1}}$$

Commento: la possibilità di ottenere una approssimazione precisa dipende sia da h che da δ ed in linea di principio migliora al crescere di α

Esempio: stesso calcolo, ma con il rapporto incrementale centrato.

h	10^{-1}	10^{-3}	10^{-5}	10^{-6}
$(1+h)^{1/3}$	1.032280	1.000333	1.000003	1.000000
$(1-h)^{1/3}$	0.965489	0.999667	0.999997	1.000000
$\frac{(1+h)^{1/3}-(1-h)^{1/3}}{2h}$	0.333955	0.333	0.3	0.0

Commento: la maggiore velocità di convergenza permette di ottenere risultati migliori prima di entrare nelle condizioni di instabilità

Esercizio: effettuare gli stessi calcoli, sia con il rapporto incrementale destro (o sinistro) che con il rapporto incrementale centrato, usando per il test funzioni che siano rispettivamente C^1 ma non C^2, C^2 ma non C^3, C^3.

• Verificare che per h "molto piccolo" (a seconda della precisione con cui si lavora) si ottengono fenomeni di instabilità del risultato

• Calcolare nei vari casi l'ordine di convegenza seguendo la relazione

$$\epsilon(h) \approx Ch^\alpha \quad \Rightarrow \quad \frac{\epsilon(ah)}{\epsilon(h)} \approx a^\alpha$$

indice

Operazioni in aritmetica finita (rappresentazioni)

Gli oggetti tipici su cui si opera in Analisi Numerica sono rappresentazioni finite in virgola mobile di numeri reali, nella forma normalizzata (cioè con $d_1 \neq 0$)

$$x = \pm 0.d_1 d_2 d_3 \ldots \cdot B^p \quad \rightarrow \quad \mathrm{fl}_t(x) = \pm 0.d_1 d_2 d_3 \ldots d_{t-1} \bar{d}_t \cdot B^p$$

- Nella strategia di chopping si pone $\bar{d}_t = d_t$, mentre in quella più usuale di rounding

$$\bar{d}_t = \begin{cases} d_t & \text{se} \quad d_{t+1} < B/2 \\ d_t + 1 & \text{se} \quad d_{t+1} \geq B/2 \end{cases}$$

Standard IEEE per le rappresentazioni di macchina:

precisione	B	t	p	tipo troncamento
float	2	23	[-126,127]	rounding
double	2	52	[-1022,1023]	rounding

Esistono poi i simboli speciali `Inf` ("infinity", risultato ad esempio della divisione di un numero non nullo per zero) e `NaN` ("Not a Number", risultato di una operazione non ammissibile ma che non si può trattare come `Inf`, ad esempio il logaritmo di un numero negativo)

Oltre al fatto che i numeri rappresentabili in questo modo sono un insieme discreto, esistono anche due situazioni in cui il numero reale x non si può rappresentare neanche in modo approssimato:

- Overflow: l'esponente p è maggiore del massimo esponente rappresentabile ed x viene rappresentato come `Inf` (in precisione singola $|x| > 2^{127} \sim 10^{38}$, in precisione doppia $|x| > 2^{1023} \sim 10^{308}$)

- Underflow: l'esponente p è minore del minimo esponente rappresentabile ed x viene rappresentato come ± 0 (in precisione singola $|x| < 2^{-126} \sim 10^{-38}$, in precisione doppia $|x| < 2^{-1022} \sim 10^{-308}$)

L'entità degli errori di arrotondamento associati ad una rappresentazione in virgola mobile si caratterizza in modo naturale tramite l'errore relativo.

Si indica come errore di macchina il massimo errore relativo associato alla rappresentazione $\mathrm{fl}_t(x)$:

$$\epsilon_m = \max \left| \frac{x - \mathrm{fl}_t(x)}{x} \right|$$

o, in altro modo, il più piccolo numero per cui si abbia

$$\mathrm{fl}_t(1 + \epsilon_m) > 1$$

Poiché si ha

$$|x - \mathrm{fl}_t(x)| \leq \begin{cases} B^{p-t} & \text{(chopping)} \\ \frac{1}{2}B^{p-t} & \text{(rounding)} \end{cases}$$

ed inoltre (poiché in una rappresentazione normalizzata $d_1 \neq 0$), $|x| \geq B^{p-1}$, si ottiene per l'errore di macchina

$$\epsilon_m = \max \left| \frac{x - \mathrm{fl}_t(x)}{x} \right| = \begin{cases} B^{1-t} & \text{(chopping)} \\ \frac{1}{2}B^{1-t} & \text{(rounding)} \end{cases}$$

Nello standard IEEE l'errore macchina è quindi $\epsilon_m = 2^{-23} \sim 10^{-7}$ in precisione singola, $\epsilon_m = 2^{-52} \sim 10^{-16}$ in precisione doppia (ciò corrisponde a circa sette cifre decimali esatte nel primo caso, sedici nel secondo)

Operazioni in aritmetica finita (errori)

Data l'operazione $x \cdot y$ in aritmetica esatta, il risultato della operazione corrispondente $x \odot y$ in aritmetica finita a virgola mobile è

$$x \odot y = \mathrm{fl}_t\Big(\mathrm{fl}_t(x) \cdot \mathrm{fl}_t(y)\Big)$$

• Si può analizzare più semplicemente l'effetto degli arrotondamenti sul risultato supponendo che sia una operazione in aritmetica esatta, effettuata tra due numeri affetti da perturbazioni relative ϵ_x, ϵ_y tali che $|\epsilon_x|, |\epsilon_y| \leq \epsilon_m$. L'errore relativo sul risultato viene stimato come

$$\left|\frac{x \odot y - x \cdot y}{x \cdot y}\right| \approx \left|\frac{x(1+\epsilon_x) \cdot y(1+\epsilon_y) - x \cdot y}{x \cdot y}\right|$$

- Somma: l'errore relativo è dato da:

$$\left|\frac{x(1+\epsilon_x)+y(1+\epsilon_y)-(x+y)}{x+y}\right| = \left|\frac{x\epsilon_x+y\epsilon_y}{x+y}\right| \leq \frac{\epsilon_m(|x|+|y|)}{|x+y|}$$

- Prodotto: trascurando il termine ϵ_m^2 rispetto ad ϵ_m si ha:

$$\left|\frac{x(1+\epsilon_x)y(1+\epsilon_y)-xy}{xy}\right| = \left|\frac{xy(\epsilon_x+\epsilon_y+\epsilon_x\epsilon_y)}{xy}\right| \lesssim 2\epsilon_m$$

- Quoziente: trascurando ϵ_m rispetto ad 1 si ha:

$$\left|\frac{\frac{x(1+\epsilon_x)}{y(1+\epsilon_y)}-\frac{x}{y}}{\frac{x}{y}}\right| = \left|\frac{(1+\epsilon_x)-(1+\epsilon_y)}{1+\epsilon_y}\right| \lesssim 2\epsilon_m$$

La operazione critica è quindi la somma: in particolare, l'errore relativo dipende dal risultato e può essere molto grande se si sta eseguendo la differenza di due valori simili (come nell'esempio del rapporto incrementale). Tale situazione va sotto il nome di perdita di cifre significative per sottrazione.