Calcolo numerico

Indice

1. Lezione 01

1.1. Problema matematico, metodo numerico e condizionamento

Un problema matematico in forma astratta è un problema che chiede di trovare u tale che

$$P(d, u) = 0,$$

con d insieme dei dati, u soluzione e P operatore che esprime la relazione funzionale tra u e d. Le due variabili possono essere numeri, vettori, funzioni, eccetera.

Un metodo numerico per la risoluzione approssimata di un problema matematico consiste nel costruire una successioni di problemi approssimati del tipo

$$P_n(d_n, u_n) = 0 \mid n \geq 1$$

oppure

$$P_h(d_h, u_h) = 0 \mid h > 0$$

che dipendono dai parametri n o h.

Un metodo numerico è convergente se

$$\lim_{n \to \infty} u_n = u$$

oppure

$$\lim_{h \to 0} u_h = u.$$

Il problema matematico $P(d, u) = 0$ è ben posto (o stabile) se, per un certo dato d, la soluzione u esiste ed è unica e dipende con continuità dai dati. Questa ultima proprietà indica che piccole perturbazioni (variazioni) dei dati d producono piccole perturbazioni nella soluzione u.

Per quantificare la dipendenza continua dai dati introduciamo il concetto di numero di condizionamento di un problema.

Consideriamo una funzione $f : [a, b] \longrightarrow \mathbb{R}$ in un punto x_0, ovvero

$$d := x_0 \quad u := f(x_0) \mid d, u \in \mathbb{R}.$$

Applichiamo lo sviluppo di Taylor di f in x_0, ovvero

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \ldots$$

Ma allora

$$\begin{aligned} f(x) - f(x_0) &\approx f'(x_0)(x - x_0) \\ \frac{f(x) - f(x_0)}{f(x_0)} &\approx \frac{x_0 f'(x_0)}{f(x_0)} \frac{x - x_0}{x_0} \\ \left| \frac{f(x) - f(x_0)}{f(x_0)} \right| &\approx \left| \frac{x_0 f'(x_0)}{f(x_0)} \right| \left| \frac{x - x_0}{x_0} \right| \end{aligned}$$

Osserviamo che

$$\Delta f(x_0) := \frac{f(x) - f(x_0)}{f(x_0)}$$

e

$$\Delta x_0 := \frac{x - x_0}{x_0}$$

sono le variazioni relative della soluzione $u := f(x_0)$ e del dato $d := x_0$.

Chiamiamo **numero di condizionamento del calcolo di una funzione f in x_0** la quantità

$$K_f(x_0) := \left| \frac{x_0 f'(x_0)}{f(x_0)} \right|.$$

Poiché vale

$$|\Delta f(x_0)| \approx K_f(x_0) |\Delta x_0|$$

diciamo che $K_f(x_0)$ esprime il rapporto tra la variazione relativa subita dalla soluzione e la variazione relativa introdotta nel dato.

Calcolare i numeri di condizionamento nei casi:
- $f(x) = 6$ e $x_0 = 4$;
- $f(x) = e^x$ e $x_0 = 4$;
- $f(x) = 6x - x^3$ e $x_0 = 4$.

Nell'approssimare numericamente un problema fisico si commettono errori di quattro tipi diversi:
1. errori sui dati, riducibili aumentando l'accuratezza nelle misurazioni dei dati;
2. errori dovuti al modello, controllabili nella fase modellistica matematica, quando si passa dal fisico al matematico;
3. errori di troncamento, dovuti al fatto che quando si passa al limite nel calcolatore questi passaggi vengono approssimati, essendo operazioni eseguite nel discreto;
4. errori di arrotondamento, dovuti alla rappresentazione finita dei calcolatori.

L'analisi numerica studia e controlla gli errori 3 e 4.

1.2. Aritmetica floating point

L'insieme dei numeri macchina è l'insieme

$$\mathcal{F}(\beta, t, L, U) = \left\{ \sigma(.a_1 a_2 ... a_t)_\beta \beta^e \right\} \cup \{0\}$$

e con il simbolo

$$\text{float}(x) \in \mathcal{F}(\beta, t, L, U)$$

il generico elemento dell'insieme, cioè il generico numero macchina.

Abbiamo:
- σ segno di $\text{float}(b)$;
- β base della rappresentazione;
- e esponente con $L \leq e \leq U$ con $L > 0$ e $U > 0$;
- t numero di cifre significative;
- $a_1 \neq 0$ e $0 \leq a_i \leq \beta - 1$;
- $m = (.a_1 a_2 ... a_t)_\beta = \frac{a_1}{\beta} + \frac{a_2}{\beta^2} + ... + \frac{a^t}{\beta^t}$ mantissa.

Facciamo un po' di osservazioni:
- $|\text{float}(x)| \in [\beta^{L-1}, (1 - \beta^{-t})\beta^U]$;
- in `MATLAB` si ha $\beta = 2$, $t = 53$, $L = -1021$ e $U = 1024$;
- il risultato di un'operazione fra numeri macchina non è necessariamente un numero macchina.

Preso il numero reale

$$x = \sigma\big(.a_1 a_2 ... a_t a_{t+1} a_{t+2}\big)_\beta \beta^e \in \mathbb{R}.$$

Distinguiamo i seguenti casi:

- $L \leq e \leq U, a_i = 0 \forall i > t$ allora si ha la rappresentazione esatta di x, ovvero $\text{float}(x) = x$;
- $e < L$ allora si ha underflow, ovvero $\text{float}(x) = 0$;
- $e > U$ allora si ha overflow, ovvero $\text{float}(x) = \infty$
- se $\exists i > t \mid a_i \neq 0$ allora:
  - troncamento:

    $$\text{float}(x) = \sigma(.a_1 a_2 ... a_t)_\beta \beta^e;$$

  - arrotondamento:

    $$\sigma \begin{cases} (.a_1 a_2 ... a_t)_\beta \beta^e \text{ se } 0 \leq a_{t+1} < \frac{\beta}{2} \\ (.a_1 a_2 ... a_t + 1)_\beta \beta^e \text{ se } \frac{\beta}{2} \geq a_{t+1} \leq \beta - 1 \end{cases}.$$

Si può dimostrare che l'errore commesso approssimando un numero reale x con la sua rappresentazione macchina $\text{float}(x)$ è maggiorato da

$$\left| \frac{\text{float}(x) - x}{x} \right| \leq k \beta^{1-t}$$

con $k = 1$ per troncamento e $k = \frac{1}{2}$ per arrotondamento.

La quantità

$$\text{eps} = k \beta^{1-t}$$

è detta precisione macchina nel fissato sistema floating point. La precisione si può caratterizzare come il più piccolo numero macchina per cui vale

$$\text{float}(1 + \text{eps}) > 1.$$

Esercizio: costruire $\mathcal{F}(\beta, t, L, U)$ con $\beta = 2, t = 3, L = -1, U = 2$.

2. Lezione 02

2.1. Vettori e matrici

Una tabella di $m \times n$ numeri reali disposti in m righe e n colonne del tipo

$$A = \begin{bmatrix} a_{11} & ... & a_1 n \\ \vdots & \ddots & \vdots \\ a_{m1} & ... & a_{mn} \end{bmatrix} = \left(a_{ij}\right) \mid i = 1, ..., m \quad j = 1, ..., n$$

si chiama matrice di m righe e n colonne. Ogni elemento a_{ij} ha un indice di riga i e un indice di colonna j che indicano riga e colonna di A in cui si trova quell'elemento.

Indichiamo con $\mathbb{R}^{m \times n}$ l'insieme delle matrici $m \times n$.

Chiamiamo **vettore colonna** di dimensione n una matrice $n \times 1$ formata da n righe e una sola colonna. Analogamente, il **vettore riga** è una matrice di dimensione $1 \times n$ formata da una sola riga e n colonne.

AGGIUNTI ESEMPI DI VETTORI COME PRIMA.

Con il termine vettore indicheremo un vettore colonna, e l'insieme dei vettori di dimensione n lo indichiamo con $\mathbb{R}^n$.

Usiamo vettori e matrici per rappresentare molte grandezze fisiche che non possono essere rappresentate come scalari, ma come vettori (tipo spostamento, velocità, accelerazione, eccetera).

Siano $a = (a_i), b = (b_i) \in \mathbb{R}^n$ due vettori, si chiama vettore somma il vettore $c = (c_i) \in \mathbb{R}^n$ tale che

$$c_i = a_i + b_i \forall i = 1...n.$$

Geometricamente parlando, il vettore somma è la diagonale del parallelogramma avente due lati coincidenti con a e b (regola del parallelogramma).

AGGIUNGI IMMAGINE CARINA.

La somma di vettori gode di alcune proprietà:

- **commutativa**: $\forall a, b \in \mathbb{R}^n \quad a + b = b + a$;
- **associativa**: $\forall a, b, c \in \mathbb{R}^n \quad (a + b) + c = a + (b + c)$;
- **esistenza del neutro**: il vettore $0 = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$ è l'elemento neutro della somma, cioè $\forall a \in \mathbb{R}^n \quad a + 0 = 0 + a = a$;
- **esistenza dell'opposto**: per ogni vettore $a \in \mathbb{R}^n$ esiste un altro vettore $b \in \mathbb{R}^n$ tale che $a + b = 0$; tale vettore b viene detto vettore opposto di a e si indica con $-a$.

Siano $a = (a_i) \in \mathbb{R}^n$ un vettore e $\beta \in \mathbb{R}$ uno scalare. Si chiama prodotto vettore-scalare il vettore $c = (c_i) \in \mathbb{R}^n$ tale che

$$c_i = \beta a_i \forall i = 1, ..., n.$$

Valgono le due proprietà distributive:

- $\forall \alpha \in \mathbb{R} \quad \forall a, b \in \mathbb{R}^n \quad \alpha(a + b) = \alpha a + \alpha b$;
- $\forall \alpha, \beta \in \mathbb{R} \quad \forall a \in \mathbb{R}^n \quad (\alpha + \beta)a = \alpha a + \beta a$.

Siano $a = (a_i), b = (b_i) \in \mathbb{R}^n$ due vettori; si chiama prodotto scalare lo scalare $c = a \cdot b \in \mathbb{R}$ tale che

$$c = a \cdot b = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + ... + a_n b_n.$$

Diciamo che l'applicazione

$$\|\cdot\| : \mathbb{R}^n \longrightarrow \mathbb{R}^+ \cup \{0\}$$

è una norma vettoriale se valgono le seguenti condizioni:

1. $\|x\| \geq 0 \forall x \in \mathbb{R}^n$ e $\|x\| = 0$ se e solo se $x = 0$;
2. $\|\alpha x\| = |\alpha| \|x\| \forall \alpha \in \mathbb{R} \quad \forall x \in \mathbb{R}^n$;
3. $\|x + y\| \leq \|x\| + \|y\| \forall x, y \in \mathbb{R}^n$.

Le norme più famose sono quella euclidea (detta norma 2) tale che

$$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}} \forall x \in \mathbb{R}^n$$

oppure la norma 1 tale che

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \forall x \in \mathbb{R}^n$$

oppure la norma ∞ (norma del massimo) tale che

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i| \forall x \in \mathbb{R}^n.$$

Una matrice si dice quadrata (di ordine n) se $m = n$. Una matrice quadrata è triangolare superiore (inferiore) se

$$a_{ij} = 0 \mid i > j (i < j),$$

cioè se sono nulli gli elementi al di sotto (sopra) della diagonale principale a_{ii}.

Se valgono entrambe le definizioni la matrice è detta diagonale.

Data la matrice $A = \left(a_{ij}\right) \in \mathbb{R}^{m \times n}$ si chiama matrice trasposta la matrice $A^T = \left(a_{ij}^T\right) \in \mathbb{R}^{n \times m}$ ottenuta dallo scambio delle righe e delle colonne di A, ovvero

$$a_{ij} = a_{ji}^T$$

Sia A una matrice quadrata di ordine n, essa si dice simmetrica se $A = A^T$, ovvero $a_{ij} = a_{ji} \forall i, j = 1, ..., n$.

Siano $A = \left(a_{ij}\right), B = \left(b_{ij}\right) \in \mathbb{R}^{m \times n}$ due matrici, si chiama matrice somma la matrice $C = \left(c_{ij}\right) \in \mathbb{R}^{m \times n}$ tale che

$$c_{ij} = a_{ij} + b_{ij} \forall i = 1, ..., m \forall j = 1, ..., n.$$

Anche la somma di matrici gode di alcune proprietà:

- **commutativa**: $\forall A, B \in \mathbb{R}^{m \times n} \quad A + B = B + A$;
- **associativa**: $\forall A, B, C \in \mathbb{R}^{m \times n} \quad (A + B) + C = A + (B + C)$;
- **esistenza del neutro**: la matrice $0 = \begin{bmatrix} 0 & ... & 0 \\ \vdots & \ddots & \vdots \\ 0 & ... & 0 \end{bmatrix}$ è l'elemento neutro della somma, cioè $\forall A \in \mathbb{R}^{m \times n} \quad A + 0 = 0 + A = A$;
- **esistenza dell'opposto**: per ogni matrice $A \in \mathbb{R}^n$ esiste un'altra matrice $B \in \mathbb{R}^{m \times n}$ tale che $A + B = 0$; tale matrice B viene detta matrice opposta di A e si indica con $-A$.

Siano $A = \left(a_{ij}\right) \in \mathbb{R}^{m \times n}$ una matrice e $\beta \in \mathbb{R}$ uno scalare. Si chiama prodotto matrice-scalare la matrice $C = \left(c_{ij}\right) \in \mathbb{R}^{m \times n}$ tale che

$$c_{ij} = \beta a_{ij} \forall i = 1, ..., m \forall j = 1, ..., n.$$

Valgono le due proprietà distributive:

- $\forall \alpha \in \mathbb{R} \quad \forall A, B \in \mathbb{R}^{m \times n} \quad \alpha(A + B) = \alpha A + \alpha B$;
- $\forall \alpha, \beta \in \mathbb{R} \quad \forall A \in \mathbb{R}^{m \times n} \quad (\alpha + \beta)A = \alpha A + \beta A$.

Sia $A = \left(a_{ij}\right) \in \mathbb{R}^{m \times n}$ una matrice e $b = (b_i) \in \mathbb{R}^n$ un vettore; si chiama prodotto matrice-vettore di A per b il vettore $c = (c_i) \in \mathbb{R}^m$ tale che

$$c_i = \sum_{j=1}^{n} a_{ij} b_j = a_{i1} b_1 + ... + a_{in} b_n \forall i = 1, ..., m.$$

Siano $A = \left(a_{ij}\right) \in \mathbb{R}^{m \times n}$ e $B = \left(b_{ij}\right) \in \mathbb{R}^{n \times k}$ due matrici; si chiama prodotto matrice-matrice di A per B la matrice $C = \left(c_{ij}\right) \in \mathbb{R}^{m \times k}$ tale che

$$c_{ij} = \sum_{t=1}^{n} a_{ik} b_{kj} = a_{i1} b_{1j} + ... + a_{in} b_{nj} \forall i = 1, ..., m \forall j = 1, ..., k.$$

Il prodotto di matrici in generale non è commutativo, cioè $A \cdot B \neq B \cdot A$.

Si chiama matrice identità di ordine n la matrice quadrata $I = \left(i_{kj}\right)$ di ordine n tale che

$$i_{kj} = \begin{cases} 1 \text{ se } k = j \\ 0 \text{ se } k \neq j \end{cases}.$$

Si può dimostrare che $A \cdot I = I \cdot A = A$.

L'applicazione

$$\|\cdot\| : \mathbb{R}^{n \times n} \longrightarrow \mathbb{R}^+ \cup \{0\}$$

è una norma matriciale se valgono le seguenti condizioni:

1. $\|A\| \geq 0 \forall A \in \mathbb{R}^{n \times n}$ e $\|A\|$ se e solo se $A = 0$;
2. $\|\alpha A\| = |\alpha| \|A\| \forall \alpha \in \mathbb{R} \forall A \in \mathbb{R}^{n \times n}$;
3. $\|A + B\| \leq \|A\| + \|B\| \forall A, B \in \mathbb{R}^{n \times n}$;
4. $\|A \cdot B\| \leq \|A\| \cdot \|B\| \forall A, B \in \mathbb{R}^{n \times n}$.

Definiamo la norma matriciale indotta dalla norma vettoriale come

$$\|A\| = \sup\left\{ \frac{\|A_x\|}{\|x\|}, \forall x \in \mathbb{R}^n / \{0\} \right\}.$$

Abbiamo alcuni casi particolari:

- norma 1 (calcolata colonna per colonna), calcolata come

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{n} |a_{ij}|;$$

- norma ∞ (calcolata per riga), calcolata come

$$\|A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|.$$

3. Lezione 03

3.1. Determinante, inversa e rango di matrici

Sia A una matrice quadrata di ordine due, ovvero

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

Si chiama determinante di A il numero reale

$$\det(A) := a_{11}a_{22} - a_{12}a_{21} \in \mathbb{R}.$$

Ora vediamo determinanti per matrici di ordine maggiore.

Siano A matrice quadrata di ordine n e a_{ij} il generico elemento; si chiama complemento algebrico di a_{ij} il numero reale

$$\mathrm{compl}(a_{ij}) := (-1)^{i+j} \det(A_{ij}),$$

dove la matrice A_{ij} è la matrice quadrata di ordine $n-1$ ottenuta da A eliminando la riga i e la colonna j.

Sia A una matrice quadrata di ordine n, fissata una qualunque riga o colonna di A, il determinante di A si ottiene sommando il prodotto di ogni elemento di tale riga o colonna per il suo complemento algebrico.

Il calcolo del determinante è indipendente dalla riga o colonna scelta, quindi conviene fissare la riga o colonna con il maggior numero di zeri.

Il determinante gode di alcune proprietà:

- se A è triangolare allora $\det(A) = a_{11}a_{22}...a_{nn}$;
- se A ha una riga o una colonna di soli zeri allora $\det(A) = 0$;
- se A ha due righe o colonne uguali allora $\det(A) = 0$;
- vale il Teorema di Binet, ovvero se A, B sono due matrici quadrate dello stesso ordine allora $\det(A \cdot B) = \det(A) \cdot \det(B)$.

Sia A una matrice quadrata di ordine n, si dice che A è invertibile se esiste una matrice A^{-1} detta matrice inversa di A quadrata di ordine n tale che $A \cdot A^{-1} = A^{-1} \cdot A = I_n$.

Teorema: sia A una matrice quadrata di ordine n, allora A è invertibile se e solo se $\det(A) \neq 0$.

Teorema: sia A una matrice quadrata di ordine due, cioè

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

e supponiamo $\det(A) \neq 0$, allora

$$A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}.$$

Sia A una matrice $m \times n$ e $k \in \mathbb{N}$ con $k \leq \min(m, n)$. Si chiama minore di ordine k estratto da A il determinante di una qualunque sottomatrice quadrata di ordine k di A, ottenuta prendendo gli elementi comuni a k righe di k colonne di A. Si chiama caratteristica o rango di A ($\mathrm{rk}(A)$) l'ordine massimo dei minori non nulli che si possono estrarre da A.

In altre parole, $\mathrm{rk}(A) = r$ se esiste un minore di ordine r diverso da zero e se tutti i minori di ordine $r+1$ sono nulli.

Sia A una matrice non nulla, allora $\mathrm{rk}(A) \geq 1$. Inoltre, $\mathrm{rk}(A) \leq \min(m, n)$.

4. Lezione 04

4.1. Sistemi lineari

Un sistema lineare di m equazioni in n incognite $x_1, x_2, ..., x_n$ è un sistema formato da m equazioni lineari in $x_1, x_2, ..., x_n$, ossia

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + ... + a_1nx_n = b_1 \\ ... \\ a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n = b_m \end{cases}.$$

Il vettore $x \in \mathbb{R}^n$ tale che $x = (x_i)$ si chiama vettore soluzione. La matrice $A \in \mathbb{R}^{m \times n}$ tale che $A = \big(a_{ij}\big)$ si chiama matrice dei coefficienti del sistema. Il vettore $b \in \mathbb{R}^m$ tale che $b = (b_i)$ si chiama vettore termine noto. La matrice $M \in \mathbb{R}^{m \times (n+1)}$ tale che $M = (A \mid b)$, ottenuta accostando alle colonne di A il vettore b, si chiama matrice completa del sistema.

In forma compatta, dati la matrice $A \in \mathbb{R}^{m \times n}$ e il vettore $b \in \mathbb{R}^m$, trovare il vettore $x \in \mathbb{R}^n$ tale che

$$Ax = b.$$

Abbiamo tre condizioni:

- sistema impossibile: il sistema non ammette soluzioni;
- sistema possibile determinato: il sistema ammette una e una sola soluzione;
- sistema possibile indeterminato: il sistema ammette infinite soluzioni.

Teorema di Cramer: siano A una matrice quadrata di ordine n e $b \in \mathbb{R}^n$, allora il sistema lineare $Ax = b$ ammette una e una sola soluzione se e solo se $\det(A) \neq 0$.

Se il determinante fosse zero potremmo avere sia sistema impossibile sia sistema determinato possibile.

Teorema di Rouché-Capelli: siano A una matrice $m \times n$ e $b \in \mathbb{R}^m$, allora il sistema lineare $Ax = b$ ammette soluzione se e solo se $\text{rk}(A) = \text{rk}(A \mid b)$.

Se $\text{rk}(A) = \text{rk}(A \mid b)$ possiamo avere $r = n$ e quindi una e una sola soluzione, altrimenti abbiamo infinite soluzioni.

5. Lezione 05

5.1. Metodi diretti per sistemi lineari

I metodi numerici per sistemi lineari si dividono in:

- metodi diretti: in assenza di errori di arrotondamento restituiscono la soluzione in un numero finito di passi;
- metodi iterativi: la soluzione è ottenuta come limite di una successione di vettori soluzione di sistemi lineari più semplici.

5.1.1. Metodo delle sostituzioni in avanti

Se vediamo che la matrice dei coefficienti è triangolare inferiore possiamo risolvere a cascata a partire dalla prima equazione, ovvero risolviamo per la prima variabile, poi sostituisco e faccio la seconda, e così via fino alla fine.

Sia $L = \left(l_{ij}\right)$ una matrice $n \times n$ triangolare inferiore e $b \in \mathbb{R}^n$, consideriamo il sistema lineare $Lx = b$. Il metodo delle sostituzioni in avanti consiste in

$$x_i = \frac{1}{l_{ii}} \left(b_i - \sum_{j=1}^{i} l_{ij} x_j \right) i = 1, ..., n.$$

Questo algoritmo ha complessità $O(n^2)$.

5.1.2. Metodo delle sostituzioni all'indietro

Se vediamo che la matrice dei coefficienti è triangolare superiore possiamo risolvere ad arrampicata a partire dall'ultima equazione, ovvero risolviamo per l'ultima variabile, poi sostituisco e faccio la penultima, e così via fino all'inizio.

Sia $U = \left(u_{ij}\right)$ una matrice $n \times n$ triangolare inferiore e $b \in \mathbb{R}^n$, consideriamo il sistema lineare $Ux = b$. Il metodo delle sostituzioni all'indietro consiste in

$$x_i = \frac{1}{u_{ii}} \left(b_i - \sum_{j=i}^{n} u_{ij} x_j \right) i = n, ..., 1.$$

Questo algoritmo ha complessità $O(n^2)$.

5.1.3. Metodo di eliminazione gaussiana (MEG)

Se non abbiamo triangolare superiore e inferiore usiamo MEG: trasformiamo il sistema $Ax = b$ in un sistema equivalente (con la stessa soluzione x) triangolare superiore $Ux = \overline{b}$ mediante combinazioni lineari di righe. Si risolve poi il sistema appena trovato con il metodo delle sostituzioni all'indietro.

L'algoritmo segue i seguenti passi:

1. pongo $A^{(0)} = A$ e $b^{(0)} = b$;
2. per costruire $A^{(t)}$ e $b^{(t)}$, con $1 \leq t \leq n$ a partire da $A^{(t-1)}$ e $b^{(t-1)}$ devo porre a zero gli elementi sulla colonna t a partire dalla riga $t + 1$ con:
    1. ricopio le prime t righe di $A^{(t-1)}$ nella prime t righe di $A^{(t)}$ e i primi t elementi di $b^{(t-1)}$ nei primi t elementi di $b^{(t)}$;
    2. per ogni riga successiva $i \geq t + 1$ calcolo il coefficiente $K_i = \frac{a_{it}^{(t-1)}}{a_{tt}^{(t-1)}}$;
    3. si modifica l'equazione i-esima modificando ogni coefficiente con se stesso meno coefficiente per valore della riga t-esima stessa colonna; modificare l'equazione vuol dire modificare ogni cella della riga i-esima della matrice ma anche il vettore dei termini noti;
3. mi fermo quando $A^{(t)}$ è triangolare superiore.

Il MEG costruisce anche una matrice triangolare inferiore L tale che $L \cdot U = A$.

5.1.4. Fattorizzazione LU

Una volta calcolata la fattorizzazione LU di A il sistemare lineare $Ax = b \iff LUx = b$ può essere risolto in due step:

- $Ly = b$ sistema triangolare inferiore;
- $Ux = y$ sistema triangolare superiore.

Come vantaggi offre quello di risolvere sistemi triangolari che costano meno del MEG, poiché questo applicato ogni volta può rallentare l'esecuzione.

Data $A \in \mathbb{R}^{n \times n}$, per applicare la fattorizzazione LU seguiamo i seguenti passi:

1. definiamo le matrici $U = A$ e $L = I_n$;
2. applichiamo MEG alla matrice U ma modificando al tempo stesso la matrice L: durante il calcolo del coefficiente K_i usando il valore $a_{it}^{(t-1)}$, mettiamo in l_{it} il coefficiente appena calcolato.

6. Lezione 06

6.1. Metodi diretti per sistemi lineari II

Matrici simmetriche definite positive

Una matrice simmetrica $A \in \mathbb{R}^{n \times n}$ si dice definita positiva se

$$Ax \cdot x \geq 0 \forall x \in \mathbb{R}^n$$

e

$$Ax \cdot x = 0 \iff x = 0.$$

Il criterio di Sylvester afferma che una matrice A simmetrica di ordine n è definita positiva se e solo se

$$\det(A_k) > 0, k = 1, ..., n$$

con A_k sottomatrice principale di ordine k formata dalle prime k righe e colonne.

6.1.1. Fattorizzazione di Cholesky

Teorema: sia $A \in \mathbb{R}^{n \times n}$ simmetrica definita positiva, allora esiste una matrice $R \in \mathbb{R}^{n \times n}$ triangolare superiore tale che

$$A = R^T \cdot R.$$

Tale fattorizzazione della matrice A è detta fattorizzazione di Cholesky.

Con questa trasformiamo il sistema $Ax = b$ nel sistema $R^T Rx = b$, che andiamo a risolvere in due step:

1. $R^T y = b$ sistema triangolare inferiore;
2. $Rx = y$ sistema triangolare superiore.

Cholesky aiuta nel risolvere sistemi triangolare più facili di applicare il MEG tutto insieme. Inoltre, il tempo di calcolo della fattorizzazione è $\approx \frac{1}{3}n^3$, che è la metà di quella LU ($\approx \frac{2}{3}n^3$).

7. Lezione 07

7.1. Metodi iterativi per sistemi lineari

Sia A una matrice quadrata di ordine n. Il numero $\lambda \in \mathbb{C}$ è detto autovalore di A se esiste un vettore $v \in \mathbb{C}^n \mid v \neq 0$ tale che

$$Av = \lambda v.$$

Il vettore è detto autovettore associato all'autovalore λ. L'insieme $\sigma(A)$ degli autovalori di A è detto spettro di A.

Proposizione: l'autovalore λ è soluzione dell'equazione caratteristica

$$p_A(\lambda) := \det(A - \lambda I) = 0,$$

dove $p_A(\lambda)$ è detto polinomio caratteristico.

Dal teorema fondamentale dell'algebra segue che una matrice di ordine n ha n autovalori.

Vediamo alcune proprietà:

- una matrice è singolare se e solo se ha almeno un autovalore nullo;
- A è simmetrica definita positiva allora gli autovalori di A sono tutti positivi;
- siano $\lambda_i(A), i = 1, ..., n$ gli autovalori della matrice $A \in \mathbb{R}^{n \times n}$, allora

$$\det(A) = \prod_{i=1}^{n} \lambda_i(A).$$

- $\mathrm{tr}(A) := \sum_{i=1}^{n} a_{ii} = \sum_{i=1}^{n} \lambda_i(A)$, con $\mathrm{tr}(A)$ traccia di A.

Sia A una matrice quadrata di ordine n, si chiama raggio spettrale di A ($\rho(A)$) il massimo valore assoluto degli autovalori di A, ovvero

$$\rho(A) := \max_{i=1,...,n} |\lambda_i(A)|.$$

Proposizione: sia A una matrice quadrata di ordine n, allora

$$\|A\|_2 = \sqrt{\rho(A^T A)}.$$

Siano A una matrice quadrata di ordine n non singolare e $\|\cdot\|$ una generica norma di matrice; si chiama numero di condizionamento della matrice A, e si indica con $K(A)$, la quantità scalare

$$K(A) = \|A\| \cdot \left\|A^{-1}\right\|.$$

Una matrice A si dice sparsa se ha un numero elevato di elementi $a_{ij} = 0$. Comunemente, una matrice quadrata di ordine n è ritenuta sparsa quando il numero di elementi diversi da zero è di ordine $O(n)$.

Può capitare che la fattorizzazione LU o la fattorizzazione di Cholesky di una matrice sparsa A generino due matrici piene. Questo fenomeno è detto fill in (riempimento). Questo è un problema se le matrici sono di grandi dimensioni, rendendo la risoluzione del sistema lineare inefficiente.

Per matrici sparse di grandi dimensioni i metodi iterativi possono essere più efficienti dei metodi diretti.

Un **metodo iterativo** per la risoluzione del sistema lineare $Ax = b$ consiste nel costruire una successione di vettori $x^{(k)} \in \mathbb{R}^n, k \geq 0$ con la speranza che

$$\lim_{k \to \infty} x^{(k)} = x,$$

a partire da un vettore iniziale $x^{(0)}$ dato.

In generale, un metodo iterativo per la risoluzione del sistema lineare $Ax = b$ ha la forma

$$x^{(k+1)} = Bx^{(k)} + g$$

con $B \in \mathbb{R}^{n \times n}$ è detta matrice di iterazione e $g \in \mathbb{R}^n$.

Teorema: un metodo iterativo nella forma descritta è convergente, cioè $\lim_{k \to \infty} x^{(k)} = x$, se e solo se

$$\rho(B) < 1,$$

dove $\rho(B)$ è il raggio spettrale della matrice B.

7.1.1. Metodo di Jacobi

Il metodo di Jacobi isola nell'i-esima equazione l'i-esima incognita e, a partire da un vettore $x^{(0)} \in \mathbb{R}^n$, genera i passi successivi

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1 \wedge j \neq i}^{n} a_{ij} x_j^{(k)} \right), i = 1, ..., n$$

per $k \geq 0$.

7.1.2. Metodo di Gauss-Seidel

Come prima, isoliamo l'i-esima incognita nell'i-esima equazione e partiamo da un vettore iniziale $x^{(0)}$. Il metodo di Gauss-Seidel genera tutte le soluzioni $x_i^{(k+1)}$ utilizzando come vettore di partenza non più quello formato dalle $x_i^{(k)}$ ma quello formato dai $x_j^{(k+1)}$ se $j < i$ e $x_t^{(k)}$ se $t \geq i$.

L'iterazione generica del metodo di Gauss-Seidel, dato il sistema lineare $Ax = b$ con $A \in \mathbb{R}^{n \times n}$ è

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i}^{n} a_{ij} x_j^{(k)} \right) i = 1, ..., n.$$

7.1.3. Osservazioni

Questi due metodi se inseriamo la condizione $a_{ii} \neq$ assicuriamo che il metodo si possa costruire.

Non è però garantita la convergenza, quindi non è sempre vero che

$$\lim_{k \to \infty} x^{(k)} = x.$$

7.1.4. Verificare la convergenza

Sia A una matrice quadrata di ordine n, allora essa è a dominanza diagonale stretta se

$$|a_{ii}| > \sum_{j=1 \wedge j \neq i}^{n} \left| a_{ij} \right| \forall i = 1, ..., n.$$

Teorema: sia $A \in \mathbb{R}^{n \times n}$ matrice a dominanza diagonale stretta per righe, allora i metodi di Jacobi e Gauss-Seidel applicati al sistema lineare $Ax = b$ sono convergenti.

Teorema: sia $A \in \mathbb{R}^{n \times n}$ una matrice simmetrica definita positiva, allora il metodo di Gauss-Seidel converge.

7.1.5. Test d'arresto

Vediamo qualche esempio. Notiamo che se il numero di condizionamento della matrice A è grande la convergenza è lenta.

7.1.5.1. Test del residuo

Fissata una tolleranza toll $\ll 1$ arrestiamo il metodo iterativo se

$$\frac{\left\|b - Ax^{(k)}\right\|}{\|b\|} < \text{toll} \ .$$

7.1.5.2. Test dell'incremento

Fissata una tolleranza toll $\ll 1$ arrestiamo il metodo iterativo se

$$\frac{\left\|x^{(k+1)} - x^{(k)}\right\|}{\left\|x^{(k)}\right\|} < \text{toll} \ .$$

8. Lezione 08

8.1. Metodi iterativi per sistemi lineari II

Vediamo un po' di matrici di iterazione.

8.1.1. Metodo di Jacobi

Dato il sistema $Ax = b$ creiamo le matrici D, E, F tali che:

- D è diagonale e contiene la diagonale di A;
- E è triangolare inferiore, contiene gli elementi triangolari inferiori di A cambiati di segno e ha 0 sulla diagonale;
- F è triangolare superiore, contiene gli elementi triangolari superiori di A cambiati di segno e ha 0 sulla diagonale;

Notiamo che $A = D - E - F$.

Chiamiamo matrice di iterazione di Jacobi la matrice

$$B_j := D^{-1}(E + F).$$

Si può verificare che questo metodo si scrive in forma compatta come

$$x^{(k+1)} = B_j x^{(k)} + D^{-1}b.$$

Grazie al teorema di convergenza, questo metodo converge se e solo se

$$\rho(B_j) < 1.$$

8.1.2. Metodo di Gauss-Seidel

Chiamiamo matrice di iterazione di Gauss-Seidel la matrice

$$B_{gs} := (D - E)^{-1}F.$$

Si può verificare che questo metodo si scrive in forma compatta come

$$x^{(k+1)} = B_{gs} x^{(k)} + (D - E)^{-1}b.$$

Grazie al teorema di convergenza, questo metodo converge se e solo se

$$\rho(B_{gs}) < 1.$$

8.1.3. Come calcolare gli autovalori di queste matrici

Si può dimostrare che:

- Jacobi: gli autovalori di B_j sono i λ tali che
$$\det(\lambda D - E - F) = 0;$$
- Gauss-Seidel: gli autovalori di B_{gs} sono i λ tali che
$$\det(\lambda(D - E) - F) = 0.$$

9. Lezione 09

9.1. Interpolazione polinomiale

Dati $N+1$ punti nel piano $(x_i, y_i) i = 0, ..., N$ (i valori y_i possono essere sia sperimentali che valutazioni di una funzione $f(\cdot)$ non nota in x_i), trovare il polinomio di grado N $P_N(x)$ tale che

$$P_N(x_i) = y_i \quad i = 0, ..., N$$

è il problema del polinomiale interpolatore.

Indichiamo con $\mathbb{P}_N$ l'insieme dei polinomi di grado N e i punti $x_i \quad i = 0, ..., N$ nodi di interpolazione.

Per fare sta cosa scriviamo il generico polinomio di grado N e imponiamo il passaggio per i punti, ottenendo un sistema lineare che sappiamo risolvere.

Teorema: dati $N+1$ punti distinti $x_0, ..., x_N$ e $N+1$ corrispondenti valori $y_0, ..., y_N$ allora esiste uno e un solo polinomio interpolatore $P_N(x)$ di grado N tale che $P_N(x_i) = y_i \quad \forall i = 0, ..., N$.

Dimostrazione: per assurdo esistano due polinomi $P_N(x)$ e $Q_N(x)$ in $\mathbb{P}_N$ tali che

$$P_N(x_i) = Q_N(x_i) = y_i \quad \forall i = 0, ..., N.$$

Ma allora $P_N(x) - Q_N(x) \in \mathbb{P}_N$ e $P_N(x_i) - Q_N(x_i) = 0 \quad \forall i = 0, ..., N$, cioè quel polinomio si annulla in $N+1$ punti distinti.

Questo implica che $P_N(x) - Q_N(x) = 0 \forall x \in \mathbb{R}$ perché per il teorema fondamentale dell'algebra, l'unico polinomio di grado N che si annulla in $N+1$ punti distinti è il polinomio banale identicamente nullo, quindi $P_N(x) = Q_N(x)$ unico.

Per dimostrare l'esistenza si procede in maniera costruttiva tramite metodo di Vandermonde o metodo di Lagrange.

9.1.1. Metodo di Vandermonde

Il generico polinomio è

$$P_N(x) = \sum_{j=0}^{N} c_j x^j = c_0 + c_1 x + ... + c_N x^N.$$

Se imponiamo il passaggio per i punti otteniamo un sistema lineare del tipo

$$\begin{cases} c_0 + c_1 x_0 + ... + c_N x_0^N = y_0 \\ ... \\ c_0 + c_1 x_N + ... + c_N x_N^N = y_N \end{cases}.$$

La matrice del sistema

$$V = \begin{bmatrix} 1 & x_0 & ... & x_0^N \\ ... & ... & ... & ... \\ 1 & x_N & ... & x_N^N \end{bmatrix}$$

è detta matrice di Vandermonde. Se i punti x_i sono distinti allora $\det(V) \neq 0$ e quindi la soluzione esiste ed è unica.

9.1.2. Metodo di Lagrange

Definiamo $N+1$ polinomi di Lagrange $L_i(x) i = 0, ..., N$ che soddisfano le proprietà di:

- $L_i(x) \in \mathbb{P}_N$;

- $L_i(x_j) = 0 \forall i, j = 0, ..., N \wedge i \neq j$;
- $L_i(x_j) = 1 \forall i = 0, ..., N$.

Ogni polinomio è quindi nella forma

$$L_i(x) = \prod_{j=0 \wedge j \neq i}^{N} \frac{x - x_j}{x_i - x_j} = \frac{(x - x_0) \cdot ... \cdot (x - x_{i-1}) \cdot (x - x_{i+1}) \cdot ... \cdot (x - x_N)}{(x_i - x_0) \cdot ... \cdot (x_i - x_{i-1}) \cdot (x_i - x_{i+1}) \cdot ... \cdot (x_i - x_N)}.$$

Il polinomio interpolatore è dato da

$$P_N(x) = \sum_{i=0}^{N} y_i L_i(x).$$

Infatti $\forall k = 0, ..., N$ vale

$$P_N(x_k) = \sum_{i=0}^{N} L_i(x_k) = y_0 L_0(x_k) + ... + y_k L_k(x_k) + ... + y_N L_N(x_k) = 0 + ... + y_k \cdot 1 + ... + 0 = y_k.$$

9.1.3. Errore di interpolazione

Consideriamo $f : \mathbb{R} \longrightarrow \mathbb{R}$ una funzione e $N + 1$ punti $(x_i, y_i) i = 0, ..., N$ tali che $y_i = f(x_i)$ e sia $P_N(x)$ il polinomio che interpola (x_i, y_i).

Dato $x \in \mathbb{R}$ chiamiamo errore di interpolazione nel punto x la quantità

$$|f(x) - P_N(x)|.$$

Teorema: siano $x_0, ..., x_N$ $N + 1$ nodi distinti, sia $x \neq x_i \forall i = 0, ..., N$ e sia $f \in C^{N+1}(I_x)$ dove I_x più piccolo intervallo chiuso e limitato contenente i nodi $x_0, ..., x_N, x$.

Allora l'errore di interpolazione nel punto x è dato da

$$f(x) - P_N(x) = \frac{\omega(x)}{(N + 1)!} f^{(N+1)}(\xi),$$

con $\xi \in I_x$ e

$$\omega(x) = (x - x_0) \cdot ... \cdot (x - x_N).$$

Corollario: nelle ipotesi del teorema precedente si ha

$$|f(x) - P_N(x)| \leq \frac{\max_{t \in I_x} |\omega(t)|}{(N + 1)!} \max_{t \in I_x} |f^{(N+1)}(t)|.$$

In generale non si può dedurre dal teorema e dal corollario che l'errore tende a 0 per $N \to \infty$. Infatti esistono funzioni per le quali l'errore può essere infinito, ossia

$$\lim_{n \to \infty} \max_{x \in I_x} |f(x) - P_N(x)| = +\infty.$$

Una funzione è il controesempio di Runge, ovvero interpoliamo $f(x) = \frac{1}{1+x^2}$ nell'intervallo $[-5, 5]$ su nodi equispaziati. Se $N \to \infty$ allora l'errore cresce.

Un altro rimedio è utilizzare i nodi di Chebishev, definiti:

- sull'intervallo $[-1, 1]$ da

$$x_i = \cos\left(\pi \frac{2i + 1}{2(N + 1)}\right) i = 0, ..., N;$$

- sul generico intervallo $[a, b]$ da

$$x_i = \frac{a+b}{2} + (b-a, 2)\cos\left(\pi\frac{2i+1}{2(N+1)}\right) i = 0, ..., N.$$

10. Lezione 10

11. Lezione 11

11.1. Minimi quadrati e spline lineari

Dati $N+1$ punti $(x_i, y_i) i = 0, ..., N$ dove eventualmente $y_i = f(x_i)$, vogliamo trovare la retta $R(x) = a_0 + a_1 x$ che renda minima la funzione

$$E(a_0, a_1) = \sum_{i=0}^{N} (y_i - R(x_i))^2 = \sum_{i=0}^{N} (y_i - (a_0 + a_1 x_1))^2$$

al variare dei coefficienti a_0, a_1.

Diciamo che $R(x)$ approssima l'insieme dei dati nel senso dei minimi quadrati e questa retta è la retta dei minimi quadrati o retta di regressione.

Il minimo della funzione $E(a_0, a_1)$ si ottiene imponendo le condizioni

$$\begin{cases} \frac{\partial E(a_0, a_1)}{\partial a_0} = 0 \\ \frac{\partial E(a_0, a_1)}{\partial a_1} = 0 \end{cases}.$$

Svolgendo i conti abbiamo

$$\begin{cases} \sum_{i=0}^{N} 2(y_i - a_0 - a_1 x_i)(-1) = 0 \\ \sum_{i=0}^{N} 2(y_i - a_0 - a_1 x_1)(-x_i) = 0 \end{cases}.$$

Dobbiamo quindi risolvere il sistema lineare

$$\begin{cases} (N+1)a_0 + \left(\sum_{i=0}^{N} x_i\right) a_1 = \sum_{i=0}^{N} y_i \\ \left(\sum_{i=0}^{N} x_i\right) a_0 + \left(\sum_{i=0}^{N} x_i^2\right) a_1 = \sum_{i=0}^{N} x_i y_i \end{cases}.$$

Tale sistema è detto sistema delle equazioni normali.

Dato un insieme di punti $(x_i, y_i) i = 0, ..., N$ con $a = x_0 < x_1 < ... < x_n = b$, una spline lineare interpolante è una funzione $S^1(x) : [a, b] \longrightarrow \mathbb{R}$ tale che:

- S^1 è un polinomio di grado 1 su ogni sotto-intervallo $[x_{i-1}, x_i] i = 1, ..., N$;
- S^1 è continua su $[a, b]$;
- $S^1(x_i) = y_i$ con $i = 0, ..., N$.

La possiamo vedere come una funzione a tratti formata da N funzioni lineari, ognuna delle quali passa per due punti consecutivi.

Consideriamo una funzione $f : \mathbb{R} \longrightarrow \mathbb{R}$ e $N+1$ punti $(x_i, y_i) i = 0, ..., N$ con $f(x_i) = y_i$. Sia $S^1(x)$ la spline lineare che interpola i punti (x_i, y_i).

Dato $x \in \mathbb{R}$ chiamiamo errore di interpolazione nel punto x la quantità

$$\left| f(x) - S^1(x) \right|.$$

Teorema: sia $f \in C^2([a, b])$, allora

$$\max_{x \in [a,b]} \left| f(x) - S^1(x) \right| \leq \frac{1}{8} h^2 \max_{x \in [a,b]} \left| f^{(2)}(x) \right|$$

con

$$h = \max_{0 \le i \le N-1} (x_{i+1} - x_i).$$

12. Lezione 12

13. Lezione 13

13.1. Integrazione numerica

Vogliamo calcolare l'integrale definito

$$I(f) = \int_a^b f(x)dx$$

data la funzione $f : [a, b] \longrightarrow \mathbb{R}$. In generale non possiamo calcolare $I(f)$ per via analitica, ma possiamo solo approssimarla numericamente tramite formule di quadratura.

Si chiama formula di quadratura una formula del tipo

$$I^{\tilde{f}} = \sum_{i=1}^{n} a_i f(x_i)$$

che approssima l'integrale $I(f) = \int_a^b f(x)dx$ mediante una combinazione lineare di valori della funzione in opportuni punti (x_i nodi di quadratura) moltiplicati per opportuni coefficienti (α_i pesi di quadratura).

Per costruire formule di quadratura approssimiamo con l'integrale di un polinomio $P(x)$ che interpola la funzione $f(x)$ in un determinato insieme di nodi nell'intervallo $[a, b]$, cioè

$$I(f) \approx I^{\sim}(f) \coloneqq I(P) = \int_a^b P(x)dx.$$

Al variare del numero di nodi di interpolazione e della loro posizione avremo diverse formule di quadratura, dette di tipo interpolatorio.

La formula del punto medio si ottiene scegliendo il polinomio di grado 0 che interpola $f(x)$ nel punto medio dell'intervallo $[a, b]$, cioè

$$I_{PM}^{\sim}(f) \coloneqq (b-a)f\left(\frac{a+b}{2}\right).$$

Il nodo di quadratura è $\frac{a+b}{2}$ mentre il peso di quadratura è $a_1 = b - a$. Si dimostra che l'errore di questa formula è

$$I(f) - I_{PM}^{\sim}(f) = \frac{(b-a)^3}{24} f^{(2)}(t) \quad t \in (a, b)$$

se $f \in C^2([a, b])$.

La formula del trapezio si ottiene scegliendo il polinomio di grado 1 che interpola $f(x)$ negli estremi dell'intervallo $[a, b]$, ovvero

$$I_T^{\sim}(f) \coloneqq \frac{b-a}{2}(f(a) + f(b)).$$

Abbiamo quindi due nodi di quadratura $x_1 = a, x_2 = b$ e due pesi di quadratura $\alpha_1 = \alpha_2 = \frac{b-a}{2}$. L'errore con questa formula di quadratura è

$$I(f) - I_T^{\sim}(f) = -\frac{(b-a^3)}{12} f^{(2)}(f) \quad t \in (a, b)$$

se $f \in C^2([a, b])$.

La formula di Cavalieri-Simpson si ottiene scegliendo il polinomio di grado 2 che interpola $f(x)$ negli estremi e nel punto medio dell'intervallo $[a,b]$, ovvero

$$I_{CS}^{\sim}(f) := \frac{b-a}{6}\left(f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right).$$

Abbiamo quindi tre nodi di quadratura $x_1 = a, x_2 = \frac{a+b}{2}, x_3 = b$ e due pesi di quadratura $\alpha_1 = \alpha_3 = \frac{b-a}{6}, \alpha_2 = \frac{2(b-a)}{3}$. L'errore con questa formula di quadratura è

$$I(f) - I_{CS}^{\sim}(f) = -\frac{(b-a)^5}{2880}f^{(4)}(t) \quad t \in (a,b)$$

se $f \in C^4([a,b])$.

Si chiama grado di precisione di una formula di quadratura il massimo intero $r \geq 0$ tale che $I^{\sim}(P) = I(P) \quad \forall P \in \mathbb{P}_r$.

Proposizione: una formula di quadratura ha grado di precisione r se e solo se

$$I^{\sim}\left(x^k\right) = I\left(x^k\right) \quad \forall k = 0, ..., r.$$

Per il punto medio, proviamo $k = 0$ quindi $f(x) = x^0 = 1$ e quindi

$$I(f) = I(1) = \int_a^b 1dx = [x]_a^b = b - a$$
$$I_{PM}^{\sim}(f) = I_{PM}^{\sim}(1) = (b-a)\cdot 1 = b - a$$

quindi

$$I(1) = I_{PM}^{\sim}(1).$$

Proviamo $k = 1$ quindi $f(x) = x$ e quindi

$$I(f) = I(x) = \int_a^b x dx = \left[\frac{x^2}{2}\right]_a^b = \frac{b^2-a^2}{2}$$
$$I_{PM}^{\sim}(f) = I_{PM}^{\sim}(x) = (b-a)\frac{a+b}{2} = \frac{b^2-a^2}{2}$$

quindi

$$I(x) = I_{PM}^{\sim}(x).$$

Se provassimo con $k = 2$ avremmo due risultati diversi, quindi PM ha grado di precisione 1.

Il trapezio ha grado di precisione 1, Cavalieri-Simpson ha grado di precisione 3.

Le formule di quadratura composite consistono in:

- introdurre una suddivisione dell'intervallo di integrazione $[a,b]$ in sotto-intervalli;
- utilizzando la proprietà additiva dell'integrale, scriverlo come una somma di integrali definiti su ciascun intervallo della suddivisione;
- approssimare tali integrali definiti mediante formule di quadratura semplici.

Sia M il numero di sotto-intervalli, $H = \frac{b-a}{M}$ ampiezza dei sotto-intervalli e $a_i = a + iH \quad i = 0, ..., M \quad a_0 = a \wedge a_M = b$ estremi dei sotto-intervalli.

La formula el punto medio composita approssima con

$$I_{PM}^{\tilde{C}}(f) = \sum_{i=1}^{M} Hf\left(\frac{a_{i-1} + a_i}{2}\right).$$

L'errore nella formula classica è

$$I(f) - I_{PM}^{\tilde{C}}(f) = \frac{b-a}{24} H^2 f^{(2)}(\eta) \quad \eta \in (a, b).$$

L'errore nella formula asintotica è

$$I(f) - I_{PM}^{\tilde{C}}(f) = \frac{H^2}{24}(f'(b) - f'(a)).$$

La formula del trapezio composita approssima con

$$I_{T}^{\tilde{C}}(f) = \sum_{i=1}^{M} \frac{H}{2}(f(a_{i-1}) + f(a_i)).$$

L'errore nella formula classica è

$$I(f) - I_{T}^{\tilde{C}} = -\frac{b-a}{12} H^2 f^{(2)}(\eta) \quad \eta \in (a, b).$$

L'errore nella formula asintotica è

$$I(f) - I_{T}^{\tilde{C}} = -\frac{H^2}{12}(f'(b) - f'(a)).$$

La formula di Cavalieri-Simpson composita approssima con

$$I_{CS}^{\tilde{C}} = \sum_{i=1}^{M} \frac{H}{6}\left(f(a_{i-1}) + 4f\left(\frac{a_{i-1} + a_i}{2}\right) + f(a_i)\right).$$

L'errore nella formula classica è

$$I(f) - I_{CS}^{\tilde{C}} = -\frac{b-a}{2880} H^4 f^{(4)}(\eta) \quad \eta \in (a, b).$$

L'errore nella formula asintotica è

$$I(f) - I_{CS}^{\tilde{C}} = -\frac{H^4}{2880}\big(f^{(3)}(b) - f^{(3)}(a)\big).$$

14. Lezione 14

15. Lezione 15

15.1. Zeri di funzione

Data una funzione $f : [a, b] \longrightarrow \mathbb{R}$ continua e tale che $f(a)f(b) < 0$ trovare $\alpha \in (a, b)$ tale che $f(\alpha) = 0$.

In generale α non riusciamo a calcolarlo per via analitica (a.e. eq non lineari), ma possiamo solo approssimarlo numericamente.

Teorema: sia $f : [a, b] \longrightarrow \mathbb{R}$ continua in $[a, b]$ e $f(a)f(b) < 0$ allora esiste $\alpha \in (a, b)$ tale che $f(\alpha) = 0$.

I metodi numerici per la ricerca degli zeri sono in generale iterativi, quindi costruiremo una serie di valori x_k con la speranza che

$$\lim_{k\to\infty} x_k = \alpha.$$

Nella pratica ci fermeremo ad un passo $\hat{k}$ tale che $x_{\hat{k}}$ sia vicino ad α.

Il **metodo di bisezione** parte con $a_0 = a$ e $b_0 = b$. Calcolo $x_1 = \frac{a_0+b_0}{2}$ con $|x_1 - \alpha| < \frac{b-a}{2}$. Se consideriamo la parte di intervallo e vale ancora il prodotto negativo allora restringo a questa parte la ricerca e ricomincio, altrimenti restringo sulla seconda parte di segmento. Al passo n-esimo ho

$$|x_n - \alpha| < \frac{b-a}{2^n}.$$

Vantaggi:

- robusto;
- converge sempre.

Svantaggi:

- convergenza lenta;
- buona approssimazione la si raggiunge lentamente.

Il **metodo di newton** parte da un x_0, calcola la retta tangente in x_0 e cerca di annullare questa funzione. Il nuovo punto sarà il punto di partenza per la nuova iterazione.

In generale, sia $f : [a, b] \longrightarrow \mathbb{R}$ derivabile tale che $f(a)f(b) < 0$ e $f'(x) \neq \quad \forall x \in [a, b]$. Sia $x_0 \in [a, b]$, allora per $k = 0, 1, 2, \ldots$ poniamo $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$.

Vantaggi:

- la convergenza è quadratica (veloce), infatti

$$\left|x_{k+1} - \alpha\right| \approx |x_k - \alpha|^2.$$

Svantaggi:

- la convergenza dipende dalla scelta di x_0, se non è sufficientemente vicino ad α il metodo può non convergere.

Teorema: supponiamo

- $f \in C^2([a, b])$ (regolarità);
- $f'(x) \neq 0 \quad \forall x \in [a, b]$ (monotonia);
- $f''(x) \neq 0 \quad \forall x \in [a, b]$ (non cambia convessità).

Chiamiamo estremo di Fourier x_0 l'unico punto tra a e b tale che

$$f(x_0)f''(x_0) > 0.$$

Allora il metodo di newton, innescato con dato iniziale x_0 estremo di Fourier, è convergente con convergenza quadratica.

Come test d'arresto abbiamo due possibilità:

- test del residuo: fissata una tolleranza toll $\ll 1$ arrestiamo il metodo iterativo se

$$\frac{|f(x_k)|}{|f(x_0)|} < \text{toll};$$

- test dell'incremento: fissata una tolleranza toll $\ll 1$ arrestiamo il metodi iterativo se

$$\frac{\left|x_{k+1} - x_k\right|}{|x_k|} < \text{toll} \ .$$

16. Lezione 16

17. Lezione 17

17.1. Metodi numerici per equazioni differenziali ordinarie

Consideriamo il problema di Cauchy

$$\begin{cases} \frac{dy(t)}{dt} = f(t, y(t)) \quad t \in (t_0, T) \\ y(t_0) = y_0 \end{cases}.$$

Supponiamo che $f = f(t, y) : (t_0) \times \mathbb{R} \longrightarrow \mathbb{R}$ sia Lipschitziana rispetto a y e uniformemente rispetto a t, cioè

$$\exists L > 0 \mid |f(t, y_1) - f(t, y_2)| \leq L|y_1 - y_2| \quad \forall t \in (t_0, T) \quad \forall y_1, y_2 \in \mathbb{R}.$$

Con queste ipotesi Cauchy ammette una e una sola soluzione.

Abbiamo $N + 1$ nodi di discretizzazione in $[t_0, T]$, ovvero

$$h > 0 \quad t_j = t_0 + jh \quad j = 0, ..., N \quad t_N \leq T.$$

Denotiamo con y_j la soluzione esatta $y(\cdot)$ valutata in t_j, ovvero $y_j \coloneqq y(t_j) \quad j = 0, ..., N$.

Un metodo numerico per l'approssimazione del problema di Cauchy è un algoritmo che costruisce $N + 1$ valori reali u_j che approssimano $y_j \quad \forall j = 0, ..., N$ i.e. $u_j \approx y_j$.

Un metodo numerico per l'approssimazione di Cauchy è detto metodo ad un passo se $\forall n \geq 0$ allora u_{n+1} dipende solo da u_n e non da u_j, per $j < n$. Altrimenti è detto multistep.

Un metodo numerico per l'approssimazione di Cauchy è detto esplicito se $\forall n \geq 0$ allora u_{n+1} si calcola come funzione dei passi precedenti u_j per $j \leq n$, altrimenti è detto implicito se $\forall n \geq 0$ allora u_{n+1} dipende implicitamente da se stesso attraverso la funzione f.

Eulero esplicito: posto $u_0 = y_0 \quad \forall n \geq 0$ faccio

$$u_{n+1} = u_n + hf(t_n, u_n).$$

Eulero implicito: posto $u_0 = y_0 \quad \forall n \geq 0$ faccio

$$u_{n+1} = u_n + ht(t_{n+1}, u_{n+1})$$

In questo caso, ad ogni passo dobbiamo risolvere un'equazione non lineare, ad esempio tramite Newton.

Metodo di Crank-Nicolson: posto $u_0 = y_0 \quad \forall n \geq 0$ faccio

$$u_{n+1} = u_n + \frac{h}{2}(f(t_n, u_n) + f(t_{n+1}, u_{n+1})).$$

Metodo di Heun: posto $u_0 = y_0 \quad \forall n \geq 0$ faccio

$$u^*_{n+1} = u_n + hf(t_n, u_n)$$
$$u_{n+1} = u_n + \frac{h}{2}(f(t_n, u_n) + f(t_{n+1}, u^*_{n+1})).$$

La forma generale di un metodo esplicito ad un passo è

$$u_{n+1} = u_n + h\phi(t_n, u_n, f(t_n, u_n), h),$$

dove ϕ è detta funzione incrementale.

Sia $y(\cdot)$ la soluzione esatta di Cauchy, poniamo

$$\varepsilon_{n+1} = y_{n+1} - y_n - h\phi(t_n, y_n, f(t_n, y_n), h) \quad 0 \leq n \leq N-1.$$

ε_{n+1} è l'errore che si commette pretendendo che la soluzione esatta soddisfi lo schema numerico.

Si chiama errore di troncamento locale la quantità

$$\tau_{n+1}(h) = \frac{\varepsilon_{n+1}}{h}.$$

Si chiama errore di troncamento globale la quantità

$$\tau(h) = \max_{0 \leq n \leq N-1} \tau_{n+1}(h).$$

Un metodo numerico è consistente se

$$\lim_{h \to 0} \tau(h) = 0.$$

Un metodo numerico è consistente di ordine p se

$$\tau(h) = O(h^p).$$

Un metodo numerico è detto zero-stabile se, in un dato intervallo limitato (t_0, T), piccole perturbazioni sui dati producono piccole perturbazioni sulla soluzione approssimata per $h \to 0$.

Un metodo numerico è detto convergente di ordine p se

$$\exists C > 0 \mid |u_n - y_n| \leq Ch^p \quad \forall n \mid 0 \leq n \leq N.$$

Teorema: un metodo numerico è convergente se e solo se è consistente e zero-stabile.

Consideriamo ora il problema modello

$$\begin{cases} \frac{dy(t)}{dt} = -\lambda y(t) \quad t \in (0, \infty) \quad \lambda > 0 \\ y(0) = 1 \end{cases}$$

la cui soluzione esatta è $y(t) = e^{-\lambda t}$.

Un metodo numerico è detto assolutamente stabile se, applicato al problema modello, allora

$$u_n \to 0 \quad t_n \to \infty.$$

Come sono i nostri metodi:

- Eulero esplicito se e solo se $h < \frac{2}{\lambda}$;
- Eulero implicito lo è incondizionatamente;
- Heun se e solo se $h < \frac{2}{\lambda}$;
- Crank-Nicolson lo è incondizionatamente.

18. Lezione 18

19. Lezione 19

19.1. Metodi numerici per sistemi di equazioni differenziali ordinarie

Consideriamo il problema di Cauchy

$$\begin{cases} \frac{\mathrm{d}y_1(t)}{\mathrm{d}t} = f_1(t, y_1(t), y_2(t)) \quad t \in (t_0, T) \\ \frac{\mathrm{d}y_2(t)}{\mathrm{d}t} = f_2(t, y_1(t), y_2(t)) \quad t \in (t_0, T) \\ y_1(t_0) = y_1^0 \\ y_2(t_0) = y_2^0 \end{cases}.$$

Sistemi di equazioni differenziali ordinarie.

Dati $f(\cdot, \cdot) : (t_0, T) \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ e $y_0 \in \mathbb{R}^n$ trovare $y(\cdot) : (t_0, T) \longrightarrow \mathbb{R}^n$ tale che

$$\begin{cases} \frac{\mathrm{d}y(t)}{\mathrm{d}t} = f(t, y(t)) \quad t \in (t_0, T) \\ y(t_0) = y^0 \end{cases}$$

dove

$$f(t, y(t)) = \begin{pmatrix} f_1(t, y(t)) \\ \dots \\ f_n(t, y(t)) \end{pmatrix} \quad | \quad y_0 = \begin{pmatrix} y_1^0 \\ \dots \\ y_n^0 \end{pmatrix} \quad | \quad y(t) = \begin{pmatrix} y_1(t) \\ \dots \\ y_n(t) \end{pmatrix}.$$

Abbiamo **due metodi di Eulero**:

- esplicito: posto $u_1^0 = y_1^0$ e $u_2^0 = y_2^0$ allora $\forall n \geq 0$ fare

$$\begin{cases} u_1^{n+1} = u_1^n + hf_1(t_n u_1^n, u_2^n) \\ u_2^{n+1} = u_2^n + hf_2(t_n, u_1^n, u_2^n) \end{cases};$$

- implicito: posto $u_1^0 = y_1^0$ e $u_2^0 = y_2^0$ allora $\forall n \geq 0$ fare

$$\begin{cases} u_1^{n+1} = u_1^n + hf_1\big(t_n u_1^{n+1}, u_2^{n+1}\big) \\ u_2^{n+1} = u_2^n + hf_2\big(t_n, u_1^{n+1}, u_2^{n+1}\big) \end{cases}.$$

Nel caso implicito, ad ogni passo temporale per trovare i valori di u_1^{n+1} e u_2^{n+1} bisogna risolvere in generale un sistema algebrico non lineare.