

Calcolo numerico

Indice

1. Lezione 01	3
1.1. Problema matematico, metodo numerico e condizionamento	3
1.2. Aritmetica floating point	4
2. Lezione 02	6
2.1. Vettori e matrici	6
3. Lezione 03	9
3.1. Determinante, inversa e rango di matrici	9
4. Lezione 04	11
4.1. Sistemi lineari	11
5. Lezione 05	12
5.1. Metodi diretti per sistemi lineari	12
5.1.1. Metodo delle sostituzioni in avanti	12
5.1.2. Metodo delle sostituzioni all'indietro	12
5.1.3. Metodo di eliminazione gaussiana (MEG)	12
5.1.4. Fattorizzazione LU	13
6. Lezione 06	14
6.1. Metodi diretti per sistemi lineari II	14
6.1.1. Fattorizzazione di Cholesky	14
7. Lezione 07	15
7.1. Metodi iterativi per sistemi lineari	15
7.1.1. Metodo di Jacobi	16
7.1.2. Metodo di Gauss-Seidel	16
7.1.3. Osservazioni	16
7.1.4. Verificare la convergenza	16
7.1.5. Test d'arresto	16
7.1.5.1. Test del residuo	17
7.1.5.2. Test dell'incremento	17

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 successione 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

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

oppure

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

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 $f : [a, b] \rightarrow \mathbb{R}$ 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

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

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

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

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 $0 \leq a_i \leq \beta - 1$;
- $m = (.a_1 a_2 \dots a_t)_\beta = \frac{a_1}{\beta} + \frac{a_2}{\beta^2} + \dots + \frac{a_t}{\beta^t}$ 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

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

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:

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

- arrotondamento:

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

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

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

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

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

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

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

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

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

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

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

oppure la norma 1 tale che

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

oppure la norma ∞ (norma del massimo) tale che

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

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 $A = (a_{ij}) \in \mathbb{R}^{m \times n}$ si chiama matrice trasposta la matrice $A^T = (a_{ij}^T) \in \mathbb{R}^{n \times m}$ 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 $a_{ij} = a_{ji} \forall i, j = 1, \dots, n$.

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

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

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 $0 = \begin{bmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{bmatrix}$ 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 $A = (a_{ij}) \in \mathbb{R}^{m \times n}$ una matrice e $\beta \in \mathbb{R}$ uno scalare. Si chiama prodotto matrice-scalare la matrice $C = (c_{ij}) \in \mathbb{R}^{m \times n}$ tale che

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

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 $A = (a_{ij}) \in \mathbb{R}^{m \times n}$ 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

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

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

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

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 $I = (i_{kj})$ 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 $\|A\| = 0$ 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

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

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

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

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 $\det(A) = a_{11}a_{22}\dots a_{nn}$;
- 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 ($\text{rk}(A)$) l'ordine massimo dei minori non nulli che si possono estrarre da A .

In altre parole, $\text{rk}(A) = r$ 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 $\text{rk}(A) \geq 1$. Inoltre, $\text{rk}(A) \leq \min(m, n)$.

4. Lezione 04

4.1. Sistemi lineari

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

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = b_1 \\ \dots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n = b_m \end{cases}.$$

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 $A = (a_{ij})$ 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 $L = (l_{ij})$ 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

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

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

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

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 $Ux = \bar{b}$ 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 sistema 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

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

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 $R^T R x = b$, che andiamo a risolvere in due step:

1. $R^T y = b$ sistema triangolare inferiore;
2. $R x = y$ 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 $\lambda_i(A)$, $i = 1, \dots, n$ gli autovalori della matrice $A \in \mathbb{R}^{n \times n}$, allora

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

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

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

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

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

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

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

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

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

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

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

7.1.3. Osservazioni

Questi due metodi se inseriamo la condizione $a_{ii} \neq 0$ assicuriamo che il metodo si possa costruire.

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

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

7.1.4. Verificare la convergenza

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

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

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 $\text{toll} \ll 1$ arrestiamo il metodo iterativo se

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

7.1.5.2. Test dell'incremento

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

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