FORMULARIO CALCOLO NUMERICO

Libera Longo

2023-01-27

1 Floating Point

Si definisce insieme dei numeri macchina (floating-point) con t cifre significative, base β e range (L, U), l'insieme dei numeri reali definito nel modo seguente

$$\mathbb{F}(\beta, t, L, U) = \{0\} \cup \left\{ x \in \mathbb{R} = sign(x)\beta^p \sum_{i=1}^{t} d_i \beta^{-i} \right\}$$

ove t, β sono interi positivi con $\beta \geq 2$. Si ha inoltre

$$0 \leq d_i \leq \beta - 1, \qquad i = 1, 2, ...$$
$$d_i \neq 0, \qquad L \leq p \leq U \qquad p \in [L, U]$$

Usualmente U è positivo e L negativo.
I numeri dell'insieme $\mathbb{F}$ sono ugualmente spaziati tra le successive potenze di β, ma non su tutto l'intervallo.
Esempio $\beta = 2,\ t = 3,\ L = -1,\ U = 2$
$\mathbb{F} = \{0\} \cup \{0.100 \times 2^p,\ 0.101 \times 2^p,\ 0.110 \times 2^p,\ 0.111 \times 2^p,\ p = -1, 0, 1, 2\}$
dove 0.100 0.101 0.110 0.111 sono tutte le possibili mantisse e p il valore dell'esponente.

- In rappresentazione posizionale un numero macchina $x \neq 0$ viene denotato con $x = \pm .d_1 d_2 ... d_t \beta^p$
- La maggior parte dei calcolatori ha la possibilità di operare con lunghezze diverse di t, a cui corrispondono, ad esempio, la semplice e la doppia precisione.
- E' importante osservare che l'insieme $\mathbb{F}$ non è un insieme continuo e neppure infinito.

Come rappresentare un numero reale positivo x in un sistema di numeri macchina $\mathbb{F}(\beta, t, L, U)$?

- Il numero x è tale che $L \leq p \leq U$ e $d_i = 0$ per $i > t$; allora x è un numero macchina ed è rappresentato esattamente ($x \in \mathbb{F}$).
- $p \notin [L, U]$; il numero non può essere rappresentato esattamente ($x \notin \mathbb{F}$).
  Se $p < L$, si dice che si verifica un underflow; solitamente si assume come valore approssimato del numero x il numero zero.
  Se $p > U$ si verifica un overflow e solitamente non si effettua nessuna approssimazione, ma il sistema di calcolo dà un avvertimento più drastico, come ad esempio, l'arresto del calcolo.

Se una matrice A $n \times n$ ha un autovettore $\lambda = 0$, allora A è singolare.
Il costo computazionale per la risoluzione di un sistema triangolare è di: $O(\frac{n^2}{2})$

2 Condizionamento e Stabilità

- Un algoritmo è stabile se l'errore algoritmico è limitato
  - Può essere limitato da una costante c o da un'espressione

- Un sistema lineare è mal condizionato se l'errore relativo sul risultato è grande rispetto all'errore relativo sui **dati**
- Un sistema lineare è mal condizionato se il numero di condizione della matrice è grande
- Un problema è mal condizionato se ad una piccola perturbazione sui dati corrisponde una grande perturbazione sul risultato

$$K_2 = \frac{\rho}{\lambda_{min}}$$

dove ρ è il raggio spettrale e λ_{min} è il più piccolo degli autovalori

3 Fattorizzaizone LR o LU

- Non è sempre possibile
  - Ad esempio se un perno per cui dividere è 0
  - Oppure se A è singolare
- Potrebbe non essere esatta se si presentano errori di arrotondamento
- Costo computazionale di $O(\frac{n^3}{3})$

3.1 Fattorizzazione LU con pivot

Usando la fattorizzazione LU con pivoting ($PA = LU$) il sistema $Ax = b$ si può risolvere risolvendo i due sistemi triangolari:

$$\begin{cases} Ly = Pb \\ Ux = y \end{cases}$$

Ogni matrice A $n \times n$ non singolare è fattorizzabile $PA = LU$, con P matrice di permutazione, L matrice triangolare inferiore con tutti 1 sulla diagonale e U triangolare superiore non singolare.

4 Fattorizzazione di Cholesky

- Ogni matrice A simmetrica e definita positiva si può fattorizzare come prodotto di due matrici triangolari L e L' dove L' è la trasposta di L
- Costo computazionale di $O(\frac{n^3}{6})$ è minore della fattorizzazione LR

5 Interpolazione

- Interpolando punti equispaziati l'errore di interpol. aumenta all'aumentare dei punti.
- Per ogni insieme di coppie $\{x_i, y_i\}$ con $i = 0...n$ e i nodi x_i distinti tra loro,
  esiste un unico polinomio di *grado* $\leq n$, che chiamiamo polinomio interpolatore degli y_i negli x_i
- Esistono infiniti polinomi di grado n che interpolano n punti
  - ma solo uno che ne interpola $n + 1$

Vi è un numero arbitrario grande di funzioni matematiche che interpolano un dato insieme di punti.

6 Chebyshev

- NON si trovano per forza in $[-5, 5]$, ma **attenzione**
- Non sono equispaziati
- Scelta dei punti di Chebyshev come **ascisse** dei dati = interpolazione più stabile.

7 Numero di Condizionamento

In generale:

- $K(A) = \|A^{-1}\| * \|A\|$ (commutativa) $\rightarrow$ dipende solo dalla matrice
- $K(A)$ esiste solo per matrici quadrate non singolari
  - $K(A)$ piccolo - n^p, $p = 0, 1, 2, 3 \rightarrow$ Problema ben condizionato.
  - $K(A)$ grande - $10^n \rightarrow$ Problema mal condizionato
    - Es: la matrice di Hilbert $\rightarrow h_{i,j} = \frac{1}{i+j-1}$ con $i, j = 1...n$
- $K(A)$ dipende dalla norma usata ma l'ordine di grandezza è sempre lo stesso
- Si dimostra che per tutte le norme p, $K(A) \geq 1$
- Si dimostra che $\frac{1}{K(A)}$ è la minima distanza tra $A^{n\times n}$ e B, dove B è la più vicina matrice appartenente all'insieme delle matrici singolari
  - Questo significa che se $K(A)$ è alto, la matrice A si comporta quasi come una matrice singolare (il sistema non ha soluzioni) quindi, in questo caso, la soluzione è molto sensibile ai dati

8 Norme

- Le norme p sono tutte equivalenti, ovvero:
  - $\exists\ c_1,\ c_2$ tali che: $c_1 * \|x\|_p \leq \|x\|_q \leq c_2 * \|x\|_p$ con $1 < p, q < \infty$

  La classe più importante di norme vettoriali è costituita dalle norme p:

$$\|x\|_p = \left(\sum_{i=1}^{m} |x_i|^p \right)^{\frac{1}{p}} . \quad 1 \leq p < q$$

  altre norme importanti sono:

NORMA	DEFINIZIONE	ESEMPIO
Norma Euclidea $p = 2$	$\|x\|_2 = \sqrt{\sum_{i=1}^{m} \|x_i\|^2} = x^T x$	$x = (-1, 2, 3)$ $\|x\|_2 = \sqrt{1^2 + 2^2 + 3^2} = \sqrt{1 + 4 + 9} = \sqrt{14}$
Norma 1 $p = 1$	$\|x\|_1 = \sum_{i=1}^{m} \|x_i\|$	$x = (-1, 2, 3)$ $\|x\|_1 = \|1\| + \|2\| + \|3\| = 6$
Norma infinito	$\|x\|_\infty = \max_{1 \leq i \leq m} \|x_i\|$	$x = (-1, 2, 3)$ $\|x\|_\infty = \max(\|1\|, \|2\|, \|3\|) = 3$

$\|A\|_1$	$\max \sum_{i=1}^{m} \|a_{i,j}\|$ per $1 \leq j \leq n$
$\|A\|_\infty$	$\max \sum_{j=1}^{n} \|a_{i,j}\|$ per $1 \leq i \leq m$
Norma di Frobenius	$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} \|a_{i,j}\|^2}$
$\|A\|_2$	$\sqrt{\rho(A^T A)}$ Dove ρ è il raggio spettrale ovvero l'autovalore massimo in modulo

Se A è una matrice quadrata $n \times n$, allora:

$$\|A\|_2 = \sqrt{\max_{\lambda \in A^T A} \lambda} \qquad \|A\|_2 = \sqrt{\rho(A^T A)}$$

$$\|A\|_\infty = \max_i \sum_{j=1}^{n} |a_{i,j}|$$

9 Punti di massimo e minimo

- **Teorema** (Condizioni **necessarie** del primo ordine): se x^* è un punto di minimo locale e f è differenziabile con continuità in un intorno aperto di x^*, allora $\nabla f(x^*) = 0$. Un punto x^* tale che $\nabla f(x^*) = 0$ si chiama punto stazionario (minimo, massimo, sella).
- **Teorema** (Condizioni **necessarie** del secondo ordine): se x^* è un punto di minimo locale di f e f è due volte differenziabile con continuità in un intorno aperto di x^*, allora $\nabla f(x^*) = 0$ e $\nabla^2 f(x^*)$ è semidefinita positiva.
- **Teorema** (Condizioni **sufficienti** del secondo ordine): se:
  - f è due volte differenziabile con continuità in un intorno aperto di x^*;
  - $\nabla f(x^*) = 0$ (condizione di punto stazionario);
  - $\nabla f(x^*)$ è definita positiva.

  Allora x^* è un punto di minimo in senso streto di f.
- Se f è convessa, un punto di minimo locale è un punto di minimo globale. In particolare:
  - f convessa $\rightarrow$ ogni punto di minimo locale x^* è punto di minimo globale di f.
  - f strettamente convessa $\rightarrow$ esiste un unico punto di minimo globale.
    - **E OGNI PUNTO STAZIONARIO E' MINIMO GLOBALE**

10 direzione e metodi di discesa

Definizione: Il vettore p è una direzione di discesa in f se esiste un $m > 0$ tale che

$$f(x + \alpha p) < f(x) \forall \alpha \in]0, m]$$

Lemma: Sia $f \in C^1$, il vettore p è una direzione di discesa di f se $p^T \nabla f(x) < 0$

- Un metodo di discesa garantisce $f(x_{k+1}) < f(x_k),\ k = 0, 1, 2...$
- Nei metodi di discesa si calcola $x_{k+1} = x_k + a_k p_k$
- Nel metodo del gradiente la dir. di discesa di f in x_k è $-\nabla f(x_k)$
- $-\nabla f(x_k)$ ($\neq 0$) è sempre una direzione di discesa
- Un m. di discesa convergente converge al minimo locale (se str. convessa è globale)

11 minimi quadrati

Sia A una matrice $m \times n$, con $m > n$ e $rg(A) = k \leq n$. Allora il problema $min\|Ax - b\|_2{}^2$

- Ammette sempre almeno una soluzione;
- Se $k = n$ (rango massimo) il problema ha una ed una sola soluzione;
  - Si risolve con equazioni normali $\rightarrow$ $A^T * Ax = A^T b$
- Se $k < n$ il problema ha infinite soluzioni;
  - Tali soluzioni formano un sottospazio di $\mathbb{R}^n$ di dimensione $n - k$
  - Si risolve con scomposizione **SVD** (in valori singolari)
    - SVD SI PUO' FARE SU QUALUNQUE MATRICE (anche per decomprimerla)
    - Valori singolari $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_k > \sigma_k + 1 = ... = \sigma_n = 0$ dove $k = rg(A)$ "ha esattamente r ($r = rg(A)$) valori singolari > 0"