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

$p \in [L, U]$

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

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

$$0 \leq d_i \leq \beta - 1, \quad i = 1, 2, \dots$$
$$d_1 \neq 0, \quad \boxed{L \leq p \leq U}$$

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 $x = \pm.d_1 d_2 \dots d_t \beta^p$
- 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.
- $p \notin [L, U]$; il numero non può essere rappresentato esattamente. 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.

$x \notin \mathbb{F}$

Se una matrice A $n \times n$ ha un autovettore $\lambda = 0$, allora A e' singolare.

Il costo computazionale per la risoluzione di un sitema triangolare e' di :

$$O\left(\frac{n^2}{2}\right)$$

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 sis. 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

FATTORIZZAZIONE 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 $O(\frac{n^3}{3})$

FATTORIZZAZIONE LU CON PIVOT

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

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

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

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**
-

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 poliniomio di grado ≤ 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 e' un numero arbitrario grande di funzioni matematiche che interpolano un dato insieme di punti.

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

NUMERO DI CONDIZIONAMENTO

- In generale:
  - $K(A) = ||A^{-1}|| * ||A||$ (commutativa) ⟶ dipende solo dalla matrice
  - K(A) esiste solo per matrici quadrate non singolari
    - K (A) piccolo ~ n^p, p = 0, 1, 2, 3 ⟶ Problema ben condizionato.
    - K (A) grande ~ 10^n ⟶ Problema mal condizionato
      - Es: la matrice di Hilbert ⟶ $h_{i,j}$ = 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) ≥ 1
  - Si dimostra che 1/K(A) è la minima distanza tra A^{nxn} 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

NORME

- Le norme p sono tutte <u>equivalenti</u>, ovvero:
  - **∃ c_1, c_2 > 0 tali che: $c_1 * ||x||_p \le ||x||_q \le c_2 * ||x||_p$** con $1 < p, q < \infty$

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

$$\|x\|_p = \left(\sum_{i=1}^{m} |x_i|^p\right)^{1/p} , \quad 1 \le p < \infty$$

Altre norme importanti sono:

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

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

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

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

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

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 $\nabla \mathbf{f(x^*) = 0}$ 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);
  - $\nabla^2 f(x^*)$ è definita positiva.

  Allora x^* è un punto di minimo in senso stretto di f.
- Se f è convessa un punto di minimo locale è un punto di minimo globale. In particolare:
  - f convessa ⟶ ogni punto di minimo locale x^* è punto di minimo globale di f.
  - f strettamente convessa ⟶ esiste un unico punto di minimo globale.
    - **E OGNI STAZIONARIO È MINIMO GLOBALE**

DIREZIONI E METODI DI DISCESA

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

$$f(x + \alpha p) < f(x) \ \forall \alpha \in]0, \underline{\alpha}]$$

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

- Un metodo di discesa garantisce $f(x_k + 1) < f(x_k) \quad k = 0, 1, 2...$
- Nei metodi di discesa si calcola $\mathbf{x_{k+1} = x_k + a_k p_k}$
- 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)

MINIMI QUADRATI

Sia A una **matrice m × n, con m > n** e $rg(A) = k \le n$.
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 -> $\mathbf{A^T * Ax = A^T b}$
- Se k < n il problema ha infinite soluzioni ;
  - Tali soluzioni formano un sottospazio di R^n di dimensione n − k
  - Si risolve con scomposizione **SVD** (in valori singolari)
    - SVD SI PUÒ FARE SU QUALUNQUE MATRICE (anche per comprimerla)
    - Valori singolari $\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_k > \sigma_{k+1} = \sigma_{k+2} = \sigma_n = 0$ dove k = rg(A)
      "ha esattamente r (r = rg(A)) valori singolari > 0"