Calcolo Numerico

M. Ciampa

Ingegneria Elettronica, a.a. 2011/2012

Indice

Capitolo 0

Algebra lineare: complementi

A Norme

Sia V uno spazio vettoriale su $\mathbb{R}$ [su $\mathbb{C}$].

0.1 Definizione
Una funzione $N : V \to \mathbb{R}$ tale che

(1) $N(v) \geq 0$ per ogni $v \in V$

(2) $N(v) = 0$ se e solo se $v = 0$

(3) $N(\alpha\, v) = |\alpha| N(v)$ per ogni $v \in V$ e $\alpha \in \mathbb{R}$ [$\alpha \in \mathbb{C}$]

(4) $N(u + v) \leq N(u) + N(v)$ per ogni $u, v \in V$

si dice *norma* in V. Uno spazio vettoriale si dice *normato* quando in esso è assegnata una norma.

0.2 Esempio

(1) $V = V_L$, spazio vettoriale su $\mathbb{R}$ dei vettori geometrici nel piano. La funzione N definita da $N(v) =$ lunghezza del segmento che rappresenta v, è una norma in V_L.

(2) $V = \mathbb{R}^n$ [$V = \mathbb{C}^n$]. Le funzioni

$$N_1 : v = (v_1, \ldots, v_n)^{\mathsf{T}} \to |v_1| + \cdots + |v_n|$$
$$N_2 : v = (v_1, \ldots, v_n)^{\mathsf{T}} \to \sqrt{|v_1|^2 + \cdots + |v_n|^2}$$
$$N_\infty : v = (v_1, \ldots, v_n)^{\mathsf{T}} \to \max\{|v_1|, \ldots, |v_n|\}$$

sono norme in V, ed i valori $N_1(v), N_2(v)$ e $N_\infty(v)$ si indicano anche con $||v||_1$, $||v||_2$ e $||v||_\infty$ rispettivamente.

(3) $V = \mathcal{C}([0,1], \mathbb{R})$.* La funzione N definita da $N(f) = \max\{\, |f(x)|, x \in [0,1] \,\}$ è una norma in V.

(4) $V = P_2(\mathbb{R})$. La funzione N definita da $N(a_0 + a_1X + a_2X^2) = |a_0| + |a_1| + |a_2|$ è una norma in V.

(5) $V = P_2(\mathbb{R})$. La funzione N definita da $N(p) = |p(0)|$ non è una norma in V.

0.3 Esercizio

Siano N una norma in V e $v, w \in V$. Allora

$$|N(v) - N(w)| \leq N(v + w)$$

Soluzione

Si ha $v = v+w-w$ e $N(v) \leq N(v+w)+N(-w)$ quindi $N(v+w) \geq N(v)-N(w)$. Analogamente, da $w = v + w - v$, si ricava $N(v + w) \geq N(w) - N(v)$.

Tenuto conto della definizione di norma si ottiene quindi la relazione (geometricamente evidente — vedere la Figura 1)

$$|N(v) - N(w)| \leq N(v + w) \leq N(v) + N(w)$$

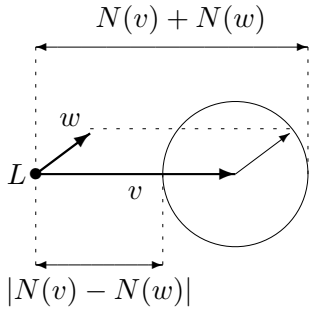

Figura 1. $V = V_L, N(v) =$ lunghezza ...

0.4 Definizione (intorno sferico)

Sia N una norma in V. Per ogni $v \in V$ ed ogni ϵ reale positivo, il sottoinsieme di V definito da

$$\mathcal{I}_N(v, \epsilon) = \{w \in V \text{ tali che } N(w - v) < \epsilon\}$$

*Per le notazioni, vedere la sezione apposita.

si dice *intorno sferico di v di raggio ϵ*.

0.5 Esempio

Sia $V = \mathbb{R}^2$. In Figura 2 è rappresentata la forma degli insiemi $\mathcal{I}_N(0,1)$ per $N = N_1, N_2, N_\infty$.

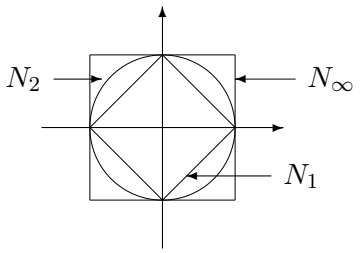

Figura 2. Intorni sferici in $\mathbb{R}^2$.

0.6 Definizione (limite, successione convergente)

Siano N una norma in V, $v_1, v_2, \ldots$ una successione di elementi di V, e $v \in V$.

Se

(1) per ogni $\mathcal{I}_N(v,\epsilon)$ esiste $n \in \mathbb{N}$ tale che $v_n, v_{n+1}, \ldots \in \mathcal{I}_N(v,\epsilon)$

allora

(2) dato comunque $w \in V$, con $w \neq v$, esistono $\mathcal{I}_N(w,\delta)$ e $m \in \mathbb{N}$ tali che $v_m, v_{m+1}, \ldots \notin \mathcal{I}_N(w,\delta)$

Infatti, posto $N(w-v) = 2\epsilon$ e $\delta = \epsilon$ si ha

$$\mathcal{I}_N(v,\epsilon) \cap \mathcal{I}_N(w,\delta) = \emptyset$$

(altrimenti esisterebbe $u \in V$ tale che $2\epsilon = N(w-u+u-v) \leq N(w-u) + N(u-v) < 2\epsilon$) quindi, scelto m tale che $v_m, v_{m+1}, \ldots \in \mathcal{I}_N(v,\epsilon)$, per $k \geq m$ si ha $v_k \notin \mathcal{I}_N(w,\delta)$.

Pertanto, esiste al più un elemento di V che verifica (1). Qualora esista, tale elemento si dice *limite* della successione *rispetto alla norma N*, si denota con $\lim_{k\to\infty} v_k$ e la successione si dice *convergente rispetto alla norma N*.

0.7 Osservazione

Siano N una norma in V, $v_1, v_2, \ldots$ una successione di elementi di V e $v \in V$. Si ha

$$\lim_{k\to\infty} v_k = v \text{ rispetto alla norma } N$$

se e solo se

$$\lim_{k\to\infty} N(v_k - v) = 0$$

0.8 Definizione (norme equivalenti)

Siano N, N^* norme in V. Le norme si dicono *equivalenti* se l'insieme delle successioni convergenti rispetto a N è uguale all'insieme delle successioni convergenti rispetto a N^*.

0.9 Osservazione

Le norme N ed N^* sono equivalenti se e solo se esistono α, β reali positivi tali che per ogni $v \in V$ si ha

$$\alpha N(v) \le N^*(v) \le \beta N(v)$$

Infatti, la condizione è ovviamente sufficiente. La necessità si prova per assurdo. Se per ogni $\beta > 0$ esiste v tale che $N^*(v) > \beta N(v)$, si consideri una successione $v_1, v_2, \ldots$ con v_k tale che $N^*(v_k) > k^2 N(v_k)$. Allora la successione definita da

$$w_k = \frac{1}{k}\frac{v_k}{N(v_k)}$$

è convergente (a 0) rispetto a N ($N(w_k) = \frac{1}{k} \to 0$) ma non è convergente rispetto a N^* :

$$N^*(w_k) = \frac{1}{k}\frac{N^*(v_k)}{N(v_k)} > k$$

0.10 Problema

Siano N, N^* norme equivalenti in V, e $v_1, v_2, \ldots$ una successione convergente a v rispetto ad N. Dimostrare che la successione converge a v *anche* rispetto a N^*. △

0.11 Teorema

Sia V uno spazio vettoriale di dimensione finita. Tutte le norme in V sono equivalenti.

In particolare, le norme N_1, N_2 e N_∞ in $\mathbb{R}^n$ [in $\mathbb{C}^n$] sono equivalenti.

Dimostrazione (solo caso particolare)

Basta dimostrare che N_1 ed N_2 sono equivalenti a N_∞ e poi usare la proprietà transitiva.

Per ogni v si ha

$$N_\infty(v) \le N_1(v) \le n\, N_\infty(v) \quad \text{e} \quad N_\infty(v) \le N_2(v) \le \sqrt{n}\, N_\infty(v)$$

che provano le equivalenze richieste. □

0.12 Problema

Siano N una norma in V, e $v_1, \ldots, v_m \in V$ linearmente indipendenti. Verificare che la funzione $n : \mathbb{R}^m \to \mathbb{R}$ definita da

$$n : (x_1, \ldots, x_m)^\mathsf{T} \to N(x_1 v_1 + \cdots + x_m v_m)$$

è una norma in $\mathbb{R}^m$. △

0.13 Problema

Sia V uno spazio vettoriale su $\mathbb{R}$ di dimensione m e $v_1, \ldots, v_m$ una sua base. Sia inoltre N una norma in $\mathbb{R}^m$. Verificare che la funzione $n : V \to \mathbb{R}$ definita da $n(v) = N(x)$, con $v = x_1 v_1 + \cdots + x_m v_m$ e $x = (x_1, \ldots, x_m)^\mathsf{T}$, è una norma in V.

(Applicazione: la convergenza di una successione in V è equivalente alla convergenza in $\mathbb{R}^m$ della successione delle coordinate.) △

0.14 Definizione (funzione continua)

Siano V, V^* spazi normati e N, N^* le rispettive norme.

Una funzione $f : \Omega \to V^*$, $\Omega \subset V$, si dice *continua in $v \in \Omega$ (rispetto alle norme N, N^*)* se per ogni successione $v_1, v_2, \ldots \in \Omega$ tale che

$$\lim_{k\to\infty} v_k = v \quad \text{rispetto ad } N$$

si ha

$$\lim_{k\to\infty} f(v_k) = f(v) \quad \text{rispetto ad } N^*$$

La funzione f si dice *continua in Ω (rispetto alle norme N, N^*)* se è continua in ogni $v \in \Omega$.

0.15 Problema

Siano V, V^* spazi normati, N, N^* le rispettive norme, e sia $\ell : \Omega \to V^*$, $\Omega \subset V$, un'applicazione lineare. Dimostrare che se ℓ è continua in 0 allora è continua. △

0.16 Problema

Sia N una norma in V. Dimostrare che la funzione $f : v \to N(v)$ è continua.

(Suggerimento: dimostrare, utilizzando l'Esercizio 0.3, che dati $v, w \in V$ si ha $|N(v) - N(w)| \le N(v - w)$ e poi usare la definizione.) △

B Norme di matrici

Sia N una norma in $\mathbb{R}^n$ [in $\mathbb{C}^n$].

0.17 Osservazione
Sia $A \in \mathbb{R}^{n\times n}$ $[A \in \mathbb{C}^{n\times n}]$. Si ha

$$\sup\left\{\frac{N(Av)}{N(v)}, v \neq 0\right\} < \infty$$

Infatti, dette $a_1, \ldots, a_n$ le colonne di A, per ogni $v = (v_1, \ldots, v_n)^{\mathsf{T}}$ si ha

$$\begin{aligned} N(Av) &= N(v_1a_1 + \cdots + v_na_n) \leq |v_1|N(a_1) + \cdots + |v_n|N(a_n) \leq \\ &\leq (N(a_1) + \cdots + N(a_n))N_\infty(v) \end{aligned}$$

e, per l'equivalenza di N e N_∞ (vedere Teorema 0.11), esiste $c > 0$ (indipendente da v) tale che $N_\infty(v) \leq c\,N(v)$. Quindi:

$$N(Av) \leq c\,(N(a_1) + \cdots + N(a_n))N(v)$$

da cui l'asserto.

0.18 Definizione (norma di una matrice)
Sia $A \in \mathbb{R}^{n\times n}$ $[A \in \mathbb{C}^{n\times n}]$. Il numero

$$||A||_N = \sup\left\{\frac{N(Av)}{N(v)}, v \neq 0\right\}$$

si dice *norma di A indotta da N*.

0.19 Problema
Siano N una norma in $\mathbb{C}^n$ e $I \in \mathbb{C}^{n\times n}$ la matrice identica. Provare che $||I||_N = 1$. △

0.20 Osservazione
Sia $A \in \mathbb{R}^{n\times n}$ $[A \in \mathbb{C}^{n\times n}]$. Per ogni $v \in \mathbb{R}^n$ $[v \in \mathbb{C}^n]$ si ha

$$N(Av) \leq ||A||_N N(v)$$

0.21 Problema
Siano N una norma in $\mathbb{R}^n$ e $A \in \mathbb{R}^{n\times n}$. Provare che l'applicazione lineare $v \to Av$ è continua. △

0.22 Osservazione
Sia $A \in \mathbb{R}^{n\times n}$ $[A \in \mathbb{C}^{n\times n}]$. Si ha

(1) $\left\{\frac{N(Av)}{N(v)}, v \neq 0\right\} = \{N(Av), v \text{ tale che } N(v) = 1\}$

(2) esistono $v_m, v_M \in \mathbb{R}^n$ $[v_m, v_M \in \mathbb{C}^n]$ tali che

$$N(Av_m) = \min\{N(Av), v \text{ tale che } N(v) = 1\}$$

e

$$N(Av_M) = \max\{N(Av), v \text{ tale che } N(v) = 1\}$$

(l'asserto si può dimostrare osservando che la funzione $v \to N(Av)$ è continua, che $\{\, v \text{ tali che } N(v) = 1 \,\}$ è un sottoinsieme compatto di $\mathbb{R}^n$ [di $\mathbb{C}^n$] e utilizzando il Teorema di Weierstrass.)

Allora

$$||A||_N = \max\{N(Av), v \text{ tale che } N(v) = 1\}$$

e, se A è invertibile:

$$||A^{-1}||_N = (\min\{N(Av), v \text{ tale che } N(v) = 1\})^{-1}$$

La prima relazione è immediata; per la seconda si ha

$$||A^{-1}||_N = \sup\left\{\frac{N(A^{-1}v)}{N(v)}, v \neq 0\right\}$$

e quindi, posto $A^{-1}v = w$:

$$\begin{aligned} ||A^{-1}||_N &= \sup\left\{\tfrac{N(w)}{N(Aw)}, w \ \neq 0\right\} = \left(\inf\left\{\tfrac{N(Aw)}{N(w)}, w \ \neq 0\right\}\right)^{-1} \\ &= (\min\{N(Av), v \text{ tale che } N(v) = 1\})^{-1} \end{aligned}$$

L'ultima uguaglianza segue da (1) e (2). Si osservi inoltre che, essendo A invertibile, $N(Av) = 0$ se e solo se $v = 0$.

0.23 Esempio

Siano $V = \mathbb{R}^n$ $[V = \mathbb{C}^n]$, e $A \in \mathbb{R}^{n\times n}$ $[A \in \mathbb{C}^{n\times n}]$. Siano $a_1, \ldots, a_n$ le colonne di A. La norma di A indotta dalle norme N_1, N_2 ed N_∞ si indica con $||A||_1, ||A||_2$ e $||A||_\infty$, rispettivamente, e vale[†]

$$||A||_1 = \max\{||a_1||_1, \ldots, ||a_n||_1\}$$

$$||A||_2 = \sqrt{\rho(A^{\mathsf{H}}A)}$$

$$||A||_\infty = ||A^{\mathsf{T}}||_1$$

Per la dimostrazione vedere [GLV] pagina 139.

0.24 Problema

Sia $A = (a_1, \ldots, a_n) \in \mathbb{C}^{n\times n}$. Dimostrare che

$$\max\{||a_1||_2, \ldots, ||a_n||_2\} \leq ||A||_2 \leq ||a_1||_2 + \cdots + ||a_n||_2$$

[†]Con il simbolo A^{H} si indica la matrice trasposta coniugata di $A \in \mathbb{C}^{n\times n}$ $[A \in \mathbb{R}^{n\times n}]$. Per la definizione di $\rho(A^{\mathsf{H}}A)$, vedere l'Appendice relativa a questo capitolo.

(Suggerimento: per la prima disuguaglianza usare l'Osservazione 0.20; per la seconda l'Osservazione 0.17 e la dimostrazione del Teorema 0.11.) $\triangle$

0.25 Osservazione

Siano $A, B \in \mathbb{R}^{n\times n}$ $[A, B \in \mathbb{C}^{n\times n}]$. Allora

$$||AB||_N \leq ||A||_N ||B||_N$$

Infatti, per il punto (2) dell'Osservazione 0.22, esiste v tale che $N(v) = 1$ e $N(ABv) = ||AB||_N$. Utilizzando l'Osservazione 0.20 si ottiene allora $||AB||_N = N(ABv) \leq ||A||_N N(Bv) \leq ||A||_N ||B||_N$ e l'asserto è provato.

0.26 Problema

Sia N una norma in $\mathbb{R}^n$ [in $\mathbb{C}^n$]. Verificare che la funzione $n : \mathbb{R}^{n\times n} \to \mathbb{R}$ $[n : \mathbb{C}^{n\times n} \to \mathbb{R}]$ definita da $n(A) = ||A||_N$ è una norma nello spazio vettoriale $\mathbb{R}^{n\times n}$ $[\mathbb{C}^{n\times n}]$. $\triangle$

0.27 Teorema

Sia N una norma in $\mathbb{C}^n$ ed $A \in \mathbb{C}^{n\times n}$. Allora $\rho(A) \leq ||A||_N$.

Dimostrazione

Sia $\lambda \in \sigma(A)^{\ddagger}$ e sia $v \in \mathbb{C}^n$ un autovettore di autovalore λ. Allora $N(Av) = |\lambda|\, N(v)$ e quindi $|\lambda| = \frac{N(Av)}{N(v)}$. L'asserto segue immediatamente dalle definizioni di $||A||_N$ e di $\rho(A)$. $\square$

0.28 Esercizio

In questo esercizio si utilizzano matrici a blocchi. Si veda l'Appendice al Capitolo 3.

Sia

$$L = \left[\begin{array}{rr|rrr} 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 1 & 3 & 0 & -3 \\ -1 & 1 & 0 & 2 & 0 \\ 5 & 0 & 0 & 0 & 7 \end{array}\right] \in \mathbb{C}^{5\times 5}$$

Determinare $||L||_1, ||L||_\infty, \sigma(L)$, $\rho(L)$ e verificare l'asserto del Teorema 0.27.

Soluzione

Si ha: $||L||_1 = \max\{7, 3, 3, 2, 10\} = 10$; $||L||_\infty = \max\{1, 1, 7, 4, 12\} = 12$.

Per lo spettro, si osservi che

(1) L ha elementi reali, allora $\lambda \in \sigma(L) \Rightarrow \overline{\lambda} \in \sigma(L)$;

[‡]Per la definizione di $\sigma(A)$, vedere l'Appendice relativa a questo capitolo.

(2) L è triangolare inferiore a blocchi:

$$L = \left[\begin{array}{c|c} L_{11} & 0 \\ \hline L_{21} & L_{22} \end{array}\right]$$

Allora si ha $\det L = \det L_{11} \det L_{22}$ — infatti:

$$L = \begin{bmatrix} L_{11} & 0 \\ 0 & I_3 \end{bmatrix} \begin{bmatrix} I_2 & 0 \\ L_{21} & L_{22} \end{bmatrix}$$

e

$$\det \begin{bmatrix} L_{11} & 0 \\ 0 & I_3 \end{bmatrix} = \det L_{11} \quad , \quad \det \begin{bmatrix} I_2 & 0 \\ L_{21} & L_{22} \end{bmatrix} = \det L_{22}$$

(queste uguaglianze si ottengono facendo lo sviluppo del determinante per righe o per colonne) —. Ne segue che il polinomio caratteristico di L è $P_L(\xi) = P_{L_{11}}(\xi)\, P_{L_{22}}(\xi)$ e quindi $\sigma(L) = \sigma(L_{11}) \cup \sigma(L_{22})$.

Si ha allora: $\sigma(L) = \{-i, i, 3, 2, 7\}$, $\rho(L) = 7$ e $\rho(L) = 7 \leq 10 = ||L||_1$, $\rho(L) = 7 \leq 12 = ||L||_\infty$.

C Prodotto hermitiano

0.29 Definizione

Sia V uno spazio vettoriale

su $\mathbb{R}$. Un'applicazione
$\bullet : V \times V \to \mathbb{R}$
tale che per ogni $u, v, w \in V$ ed $\alpha \in \mathbb{R}$

su $\mathbb{C}$. Un'applicazione
$\bullet : V \times V \to \mathbb{C}$
tale che per ogni $u, v, w \in V$ ed $\alpha \in \mathbb{C}$

1) $(u + v) \bullet w = u \bullet w + v \bullet w$
2) $\alpha u \bullet v = \alpha(u \bullet v)$ [è lineare a sinistra]

3) $u \bullet v = v \bullet u$ [è simmetrica]

(e quindi
$u \bullet (v + w) = u \bullet v + u \bullet w$
$u \bullet (\alpha v) = \alpha(u \bullet v)$
[è lineare a destra]
è un'applicazione bilineare)

3) $u \bullet v = \overline{v \bullet u}$ [è hermitiana]

(e quindi
$u \bullet (v + w) = u \bullet v + u \bullet w$
$u \bullet (\alpha v) = \overline{\alpha}(u \bullet v)$
[è antilineare a destra]
$u \bullet u \in \mathbb{R}$
è un'applicazione sesquilineare)

4) se $u \neq 0$ allora $u \bullet u$ è positivo [è definita positiva]

si dice *prodotto scalare* in V.
Quindi, un prodotto scalare in V è un'applicazione bilineare, simmetrica e definita positiva.

si dice *prodotto hermitiano* in V.
Quindi, un prodotto hermitiano in V è un'applicazione sesquilineare, hermitiana e definita positiva.

0.30 Esempio

(1) $V = \mathbb{R}^n$. Posto $u = (u_1, \ldots, u_n)^\mathsf{T}$ e $v = (v_1, \ldots, v_n)^\mathsf{T}$

   (a) l'applicazione $u \bullet v = \sum_{k=1}^n u_k v_k = u^\mathsf{T} v$ è un prodotto scalare (prodotto scalare canonico in $\mathbb{R}^n$);

   (b) dati $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ e posto $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, l'applicazione

$$u \bullet v = \sum_{k=1}^n \lambda_k u_k v_k = u^\mathsf{T} \Lambda\, v$$

   è un prodotto scalare se e solo se ogni λ_k è positivo.

(2) $V = \mathcal{C}(\mathbb{R}, \mathbb{R})$; l'applicazione $f \bullet g = \int_0^1 f(t)g(t)\,dt$ non è un prodotto scalare perché esistono $f \neq 0$ tali che $f \bullet f = 0$.

(3) $V = \mathcal{C}([0,1], \mathbb{R})$; l'applicazione $f \bullet g = \int_0^1 f(t)g(t)\,dt$ è un prodotto scalare.

(4) $V = P_{n-1}(\mathbb{R}); t_1, \ldots, t_n \in \mathbb{R}$. L'applicazione $f \bullet g = f(t_1)g(t_1) + \cdots + f(t_n)g(t_n)$ è un prodotto scalare se e solo se $t_1, \ldots, t_n$ distinti.

(5) $V = \mathbb{C}^n$. Posto $u = (u_1, \ldots, u_n)^\mathsf{T}$ e $v = (v_1, \ldots, v_n)^\mathsf{T}$

   (a) l'applicazione $u \bullet v = \sum_{k=1}^n u_k \overline{v_k} = u^\mathsf{T} \overline{v}$ è un prodotto hermitiano (prodotto hermitiano canonico in $\mathbb{C}^n$);

   (b) Dati $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ e posto $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, l'applicazione

$$u \bullet v = \sum_{k=1}^n \lambda_k u_k \overline{v_k} = u^\mathsf{T} \Lambda\, \overline{v}$$

   è un prodotto hermitiano se e solo se ogni λ_k è reale positivo.

(6) $V = \mathcal{C}([0, 2\pi], \mathbb{C})$. L'applicazione

$$f \bullet g = \int_0^{2\pi} f(t)\overline{g(t)}\,dt$$

è un prodotto hermitiano.

Sia V uno spazio vettoriale su $\mathbb{R}$ [su $\mathbb{C}$] con prodotto scalare [prodotto hermitiano]. Il vettore v si dice *ortogonale* al vettore w (notazione: $v \perp w$) se $v \bullet w = 0$.

Per ogni $v \in V$, indichiamo con $||v||$ la quantità $\sqrt{v \bullet v}$.

0.31 Esempio

Sia V come al punto (6) dell'Esempio 0.30. Per $k, m \in \mathbb{Z}$ si ha

$$e^{ikt} \bullet e^{imt} = \int_0^{2\pi} e^{i(k-m)t}\,dt = \begin{cases} 0 & \text{se } k \neq m \\ 2\pi & \text{se } k = m \end{cases}$$

I vettori e^{ikt} e e^{imt} sono quindi ortogonali per $k \neq m$, e $||e^{ikt}|| = \sqrt{2\pi}$.

0.32 Teorema (di scomposizione ortogonale, I)
Siano $u, v \in V$. Esistono unici $v', \omega \in V$ tali che (vedere la Figura 3)

(a) $v' \in \langle v \rangle$

(b) $\omega \perp v$

(c) $u = v' + \omega$

Dimostrazione
Se $v = 0$ allora $v' = 0$ e $\omega = u$ (scelte obbligate). Altrimenti

$$u - \alpha\, v \perp v \Leftrightarrow u \bullet v = \alpha\,(v \bullet v) \Leftrightarrow \alpha = \frac{u \bullet v}{v \bullet v}$$

e quindi

$$v' = \frac{u \bullet v}{||v||^2}\, v \quad \text{e} \quad \omega = u - v'$$

sono i vettori cercati.

Il vettore v' si dice la *proiezione ortogonale* di u su v; il vettore ω si dice il *complemento ortogonale* di u relativo a v. □

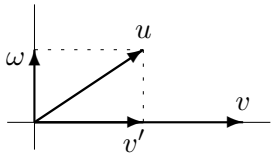

Figura 3.

0.33 Osservazione
Siano u, v, v', ω come nel Teorema 0.32. Si ha

(1) $\omega = 0$ se e solo se $u \in \langle v \rangle$

(2) $||u||^2 = ||v'||^2 + ||\omega||^2$

(3) Sia $F : \langle v \rangle \to \mathbb{R}$ definita da $F(w) = ||u - w||$. Allora

$$F(v') = \min\{\, F(w), w \in \langle v \rangle \,\}$$

Infatti, per ogni $w \in \langle v \rangle$ si ha:

$$u - w = u - v' + v' - w$$

Poiché $u - v' \perp v$ e $v' - w \in \langle v \rangle$ allora (vedere punto (2))

$$||u - w||^2 = ||u - v'||^2 + ||v' - w||^2 \geq ||u - v'||^2$$

e si ha uguaglianza se e solo se $w = v'$.

0.34 Problema

Sia $v \in V$. Provare che le applicazioni $P, Q : V \to V$ definite da

$$P(u) = \text{la proiezione ortogonale di } u \text{ su } v$$
$$Q(u) = \text{il complemento ortogonale di } u \text{ relativo a } v$$

sono lineari. △

0.35 Problema

(1) Siano V uno spazio vettoriale su $\mathbb{R}$ con prodotto scalare, $a, b \in V$. Provare che

$$||a + b||^2 = ||a||^2 + ||b||^2 \quad \text{se e solo se } a \perp b$$

(2) Siano V uno spazio vettoriale su $\mathbb{C}$ con prodotto hermitiano, $a, b \in V$. Provare che

$$||a + b||^2 = ||a||^2 + ||b||^2 \quad \text{se e solo se } \mathrm{re}(a \bullet b) = 0$$

△

0.36 Teorema (disuguaglianza di Schwarz)

Siano $u, v \in V$. Allora $|u \bullet v| \leq ||u||\,||v||$.

Dimostrazione

Se $v = 0$, l'asserto è verificato.

Siano $v \neq 0$ e v', ω come nel Teorema 0.32. Si ha

$$||u||^2 \geq ||v'||^2 = \frac{|u \bullet v|^2}{||v||^4} ||v||^2 = \frac{|u \bullet v|^2}{||v||^2}$$

da cui l'asserto.

Si osservi che l'uguaglianza si ottiene se e solo se $u \in \langle v \rangle$. □

0.37 Osservazione

(1) La funzione $v \to ||v||$ è una norma in V. Se $V = \mathbb{R}^n$ [$V = \mathbb{C}^n$] con prodotto scalare [prodotto hermitiano] canonico, si ottiene la funzione N_2 introdotta nell'Esempio 0.2.

(2) Non è vero che "ogni norma deriva da un prodotto scalare." Ad esempio, le norme N_1 e N_∞ introdotte nell'Esempio 0.2 non derivano da un prodotto scalare in $\mathbb{R}^n$.

Infatti, in caso contrario, per ogni v avremmo $||v||_N^2 = v \bullet v$ e quindi, per ogni u, v :

$$||u+v||_N^2 + ||u-v||_N^2 = 2\,||u||_N^2 + 2\,||v||_N^2$$

Ma, posto $u = e_1$ e $v = e_2, \ldots$

0.38 Problema

(1) Verificare che la funzione $N_F : \mathbb{C}^{n\times n} \to \mathbb{R}$ definita da

$$N_F(A) = \sqrt{\sum_{i,j=1}^{n} |a_{ij}|^2}$$

è una norma in $\mathbb{C}^{n\times n}$ (*norma di Frobenius*).

(Suggerimento: per verificare la proprietà (4) della definizione di norma, si utilizzi il fatto che un elemento di $\mathbb{C}^{n\times n}$ è "la stessa cosa" di un elemento di $\mathbb{C}^{n^2}$.)

(2) Determinare $N_F(I)$ e decidere (alla luce del Problema 0.19) se N_F sia una norma indotta per qualche norma in $\mathbb{C}^n$. △

0.39 Esercizio

Sia $v \in \mathbb{C}^n$. Si ha $||v||_1 \leq \sqrt{n}\,||v||_2$.

Soluzione

Siano $v_1, \ldots, v_n$ le componenti di v e sia w il vettore di componenti

$$w_k = \begin{cases} 1 & \text{se } v_k > 0 \\ -1 & \text{se } v_k < 0 \\ 0 & \text{se } v_k = 0 \end{cases}$$

Si ha: $||v||_1 = |v_1| + \cdots + |v_n| = v \bullet w$. Per la disuguaglianza di Schwarz: $|v \bullet w| \leq ||v||_2\,||w||_2$. Siccome

$$||w||_2 = \sqrt{\text{numero di componenti non nulle di } v} \leq \sqrt{n}$$

si ottiene l'asserto.

0.40 Problema

Sia $A \in \mathbb{C}^{n\times n}$. Dimostrare che per ogni $v \in \mathbb{C}^n$ non nullo si ha:

$$\frac{||Av||_2}{||v||_2} \leq \sqrt{n}||A||_\infty \quad \text{e} \quad \frac{||Av||_2}{||v||_2} \leq n||A||_1$$

e quindi

$$||A||_2 \leq \sqrt{n}||A||_\infty \quad \text{e} \quad ||A||_2 \leq n||A||_1$$

(Suggerimento: per dimostrare la prima relazione utilizzare la dimostrazione del Teorema 0.11; per la seconda l'Esercizio 0.39.) △

0.41 Problema

Siano $a, b \in \mathbb{C}^n$. Provare che $|b^{\mathsf{H}}a| \leq ||a||_1||b||_\infty$. △

D Proiezioni ortogonali su sottospazi e ortonormalizzazione di Gram–Schmidt

Sia V uno spazio vettoriale su $\mathbb{R}$ [su $\mathbb{C}$] con prodotto scalare [prodotto hermitiano].

Diremo che un insieme di vettori $v_1, \ldots, v_k$ è una *famiglia ortogonale* se $v_i \bullet v_j = 0$ per $i \neq j$, una *famiglia ortonormale* se, inoltre, $||v_j|| = 1$ per ogni j.

0.42 Osservazione

Se $v_1, \ldots, v_k$ sono una famiglia ortogonale di vettori non nulli, $v_1, \ldots, v_k$ sono linearmente indipendenti.

0.43 Teorema (di scomposizione ortogonale, II)

Sia $v_1, \ldots, v_k$ una famiglia ortogonale di vettori non nulli, sia $V' = \langle v_1, \ldots, v_k \rangle$ e sia $u \in V$.

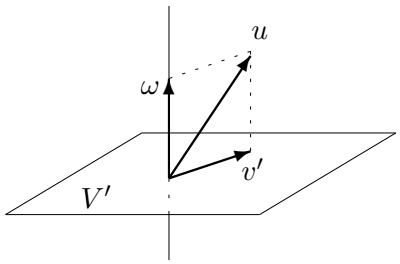

Figura 4.

Esistono unici $v', \omega \in V$ tali che (vedere la Figura 4)

(a) $v' \in V'$

(b) $\omega \perp V'$ (ossia $\omega \perp v$ per ogni $v \in V'$)[§]

(c) $u = v' + \omega$.

Dimostrazione

Analoga a quella del Teorema 0.32.

I vettori cercati sono

$$v' = \frac{u \bullet v_1}{||v_1||^2} v_1 + \cdots + \frac{u \bullet v_k}{||v_k||^2} v_k \quad \text{e} \quad \omega = u - v'$$

Il vettore v' si dice la *proiezione ortogonale* di u su V', il vettore ω si dice il *complemento ortogonale* di u relativo a V'. I numeri

$$\frac{u \bullet v_1}{||v_1||^2}, \ldots, \frac{u \bullet v_k}{||v_k||^2}$$

si chiamano *coefficienti di Fourier* di v relativi a $v_1, \ldots, v_k$. □

0.44 Osservazione

Siano V', u, v', ω come nel Teorema 0.43. Si ha

(1) $\omega = 0$ se e solo se $u \in V'$

(2) $||u||^2 = ||v'||^2 + ||\omega||^2$

(3) Sia $F : V' \to \mathbb{R}$ definita da $F(v) = ||u - v||$. Allora

$$F(v') = \min\{\, F(v), v \in V' \,\}$$

0.45 Osservazione

Siano $v_1, \ldots, v_k$ elementi linearmente indipendenti di V. Ci proponiamo di determinare una famiglia ortonormale $q_1, \ldots, q_k$ tale che

$$\begin{aligned} \langle v_1 \rangle &= \langle q_1 \rangle \\ \langle v_1, v_2 \rangle &= \langle q_1, q_2 \rangle \\ &\vdots \\ \langle v_1, \ldots, v_k \rangle &= \langle q_1, \ldots, q_k \rangle \end{aligned} \tag{0.1}$$

Il *procedimento di ortonormalizzazione di Gram–Schmidt* consente di risolvere il problema.

Il procedimento consiste nel determinare una famiglia ortogonale di vettori non nulli che verifica le condizioni (0.1) e poi normalizzarli.

[§]Questo accade se e solo se $\omega \perp v_j$ per $j = 1, \ldots, k$.

Descrizione del PROCEDIMENTO DI GRAM–SCHMIDT

`dati:` $v_1, \dots, v_k \in V$ linearmente indipendenti;

$\omega_1 = v_1$;
`per` $j = 2, \dots, k$ `ripeti`
`>` $\omega_j =$ complemento ortogonale di v_j relativo a $\langle \omega_1, \dots, \omega_{j-1} \rangle$;
`per` $j = 1, \dots, k$ `ripeti` $q_j = \dfrac{\omega_j}{||\omega_j||}$

`uscita:` $q_1, \dots, q_k$.

Si osservi che ad ogni passo si ha $\omega_j \neq 0$ e $\langle v_1, \dots, v_j \rangle = \langle \omega_1, \dots, \omega_j \rangle$.

0.46 Osservazione
Sia n la dimensione di V. Allora

(1) esiste una base ortonormale di V

(2) detta $v_1, \dots, v_n$ una base ortonormale di V, e $v \in V$ si ha

$$v = (v \bullet v_1)\, v_1 + \cdots + (v \bullet v_n)\, v_n$$

cioè: le componenti di v si ottengono tramite proiezioni ortogonali sugli elementi della base.

E Matrici hermitiane

Si consideri $\mathbb{R}^n$ con prodotto scalare canonico [$\mathbb{C}^n$ con prodotto hermitiano canonico].

0.47 Definizione
Una matrice $A \in \mathbb{C}^{n \times n}$ si dice *normale* se $A^{\mathsf{H}} A = A A^{\mathsf{H}}$, si dice *hermitiana* se $A^{\mathsf{H}} = A$, si dice *anti–hermitiana* se $A^{\mathsf{H}} = -A$.

0.48 Problema
Sia $A \in \mathbb{C}^{n \times n}$. Dimostrare che se A è hermitiana o se è anti–hermitiana, allora A è normale. $\triangle$

0.49 Problema
Dimostrare che se una matrice $T \in \mathbb{C}^{n\times n}$ è normale e triangolare, allora è diagonale. (Suggerimento: si uguaglino successivamente gli elementi diagonali di $T^{\mathsf{H}}T$ e TT^{H}.) $\triangle$

0.50 Osservazione

(1) Sia $A \in \mathbb{C}^{n\times n}$. Se A è hermitiana, tutte le radici del suo polinomio caratteristico sono reali. Se A è anti–hermitiana, tutte le radici del suo polinomio caratteristico sono immaginarie.

Infatti, sia $\lambda \in \mathbb{C}$ una radice del polinomio caratteristico. Allora λ è un autovalore ed esiste $v \in \mathbb{C}^n$ non nullo tale che $Av = \lambda v$. Si ha allora $Av \bullet v = \lambda v \bullet v$ e, poiché $v \bullet v$ è reale positivo:

$$\lambda = \frac{Av \bullet v}{v \bullet v}$$

Per il numeratore si ha

$$\overline{Av \bullet v} = v \bullet Av = v^{\mathsf{T}}\,\overline{A}\,\overline{v} = A^{\mathsf{H}}v \bullet v$$

e quindi, se A è hermitiana $Av \bullet v \in \mathbb{R}$, se A è anti–hermitiana $Av \bullet v$ è immaginario.

(2) Sia $A \in \mathbb{R}^{n\times n}$. Per il punto precedente si ha in particolare che se A è simmetrica, tutte le radici del suo polinomio caratteristico sono reali. Se A è anti–simmetrica, tutte le radici del suo polinomio caratteristico sono immaginarie.

0.51 Definizione
Una matrice $Q = (q_1, \ldots, q_n) \in \mathbb{R}^{n\times n}$ $[Q \in \mathbb{C}^{n\times n}]$ si dice *ortogonale* [*unitaria*] se ha le tre proprietà equivalenti

(1) $q_1, \ldots, q_n$ è una base ortonormale di $\mathbb{R}^n$ [di $\mathbb{C}^n$]

(2) $Q^{\mathsf{T}}Q = I \quad [Q^{\mathsf{H}}Q = I]$

(3) $Q^{-1} = Q^{\mathsf{T}} \quad [Q^{-1} = Q^{\mathsf{H}}]$

0.52 Problema
Sia $A \in \mathbb{C}^{n\times n}$. Dimostrare che se A è unitaria, allora A è normale. $\triangle$

0.53 Osservazione
Sia $Q \in \mathbb{R}^{n\times n}$ ortogonale.

(1) Per ogni $v \in \mathbb{R}^n$ si ha $||Qv|| = ||v||$. In particolare si ha $||Q|| = 1$.

(2) Se $\lambda \in \mathbb{R}$ è autovalore di Q si ha $\lambda \in \{-1, 1\}$.

Sia $Q \in \mathbb{C}^{n\times n}$ unitaria.

(1) Per ogni $v \in \mathbb{C}^n$ si ha $||Qv|| = ||v||$. In particolare si ha $||Q|| = 1$.

(2) Se $\lambda \in \mathbb{C}$ è autovalore di Q si ha $\lambda \in \Gamma$.[¶]

0.54 Teorema
Sia $A \in \mathbb{R}^{n\times n}$ simmetrica. Sussistono gli asserti equivalenti

(1) esiste una base ortonormale di $\mathbb{R}^n$ costituita da autovettori di A.

(2) A è diagonalizzabile tramite una matrice ortogonale. Cioè esistono $Q \in \mathbb{R}^{n\times n}$ ortogonale e $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ tali che

$$A = Q\,\mathrm{diag}(\alpha_1, \ldots, \alpha_n)Q^{\mathsf{T}}$$

Dimostrazione
Segue dall'Osservazione 0.50 punto (2) e dal Teorema 0.81 dell'Appendice.

0.55 Osservazione
Sia $A \in \mathbb{R}^{n\times n}$. Se A è diagonalizzabile tramite una matrice ortogonale, allora A è simmetrica.

0.56 Teorema
Sia $A \in \mathbb{C}^{n\times n}$ normale. Sussistono gli asserti equivalenti

(1) esiste una base ortonormale di $\mathbb{C}^n$ costituita da autovettori di A.

(2) A è diagonalizzabile tramite una matrice unitaria. Cioè esistono $Q \in \mathbb{C}^{n\times n}$ unitaria e $\alpha_1, \ldots, \alpha_n \in \mathbb{C}$ tali che

$$A = Q\,\mathrm{diag}(\alpha_1, \ldots, \alpha_n)Q^{\mathsf{H}}$$

Dimostrazione
Segue dal Teorema 0.81 dell'Appendice e dal Problema 0.49.

0.57 Osservazione
Sia $A \in \mathbb{C}^{n\times n}$. Se A è diagonalizzabile tramite una matrice unitaria, allora A è normale.

[¶] $\Gamma = \{z \in \mathbb{C} \mid z = e^{i\theta}, \theta \in \mathbb{R}\}$.

0.58 Osservazione
Sia $A \in \mathbb{C}^{n\times n}$, normale. Allora $\rho(A) = ||A||$.

0.59 Definizione (matrice definita positiva)
Una matrice $A \in \mathbb{R}^{n\times n}$ simmetrica [$A \in \mathbb{C}^{n\times n}$ hermitiana] si dice *definita positiva* se per ogni $v \in \mathbb{R}^n$ [$v \in \mathbb{C}^n$] non nullo si ha $Av \bullet v > 0$.

0.60 Osservazione
Sia $A \in \mathbb{R}^{n\times n}$ simmetrica [$A \in \mathbb{C}^{n\times n}$ hermitiana] definita positiva. Tutte le radici del suo polinomio caratteristico sono positive (vedere Osservazione 0.50).

0.61 Problema
Sia $A \in \mathbb{R}^{n\times n}$ simmetrica. Provare che se tutte le radici del polinomio caratteristico di A sono positive, allora A è definita positiva. △

0.62 Problema
Sia $A \in \mathbb{R}^{n\times m}$. Posto $B = A^{\mathsf{T}}A$, provare che B è simmetrica. Provare inoltre che B è definita positiva se e solo se le colonne di A sono linearmente indipendenti. △

0.63 Problema
Siano $\mathbb{C}^n$ con prodotto hermitiano canonico, $A \in \mathbb{C}^{n\times n}$ hermitiana definita positiva e $F : \mathbb{C}^n \to \mathbb{R}$ definita da $F(v) = Av \bullet v$.
Provare che

$$\min\{F(v) \mid v \in \mathbb{C}^n\} = 0$$

e che $v \neq 0 \Rightarrow F(v) \neq 0$. △

0.64 Osservazione
Sia $A \in \mathbb{C}^{n\times n}$ hermitiana definita positiva. Si ha

$$\max\{\, z^{\mathsf{H}}Az, ||z|| = 1 \,\} = \rho(A)$$

Dimostrazione
Tenuto conto dell'Osservazione 0.60, siano $\lambda_1 \geq \cdots \geq \lambda_n > 0$ le radici del polinomio caratteristico di A. Siano S unitaria e $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ tali che

$$A = S\Lambda\, S^{\mathsf{H}}$$

(vedere Teorema 0.56). Posto $w = S^{\mathsf{H}}z$ si ha

$$z^{\mathsf{H}}Az = z^{\mathsf{H}}S\Lambda\, S^{\mathsf{H}}z = w^{\mathsf{H}}\Lambda\, w = \lambda_1|y_1|^2 + \cdots + \lambda_n|y_n|^2 \leq \rho(A)||y||^2$$

e, per $v \in V(\lambda_1)$ con $||v|| = 1$ si ha $v^{\mathsf{H}}Av = \lambda_1$. L'asserto è provato osservando che $\lambda_1 = \rho(A)$.

F Localizzazione dello spettro

Il teorema che segue consente di determinare, in modo molto semplice, una regione del piano complesso che contiene tutti gli autovalori di una matrice.

0.65 Teorema (Gershgorin)

Sia $A \in \mathbb{C}^{n\times n}$. Per $k = 1, \ldots, n$, posto $\rho_k = \sum_{j\neq k} |a_{kj}|$, siano $\mathcal{C}_k = \{z \in \mathbb{C} \text{ tali che } |z - a_{kk}| \leq \rho_k\}$ i *cerchi di Gershgorin.*

Si ha

(1) $\sigma(A) \subset \cup_{k=1}^{n} \mathcal{C}_k$

(2) se per qualche intero m l'unione di m cerchi è *disgiunta* dall'unione dei rimanenti, allora l'unione degli m cerchi contiene m autovalori di A, ed i restanti $n - m$ appartengono all'unione degli $n - m$ cerchi rimanenti.

Dimostrazione (solo punto (1))

Siano λ un autovalore di A, v un autovettore di autovalore λ, e v_i una componente di v di modulo massimo. Allora: $(Av)_i = \lambda v_i \Rightarrow \sum_{k\neq i} a_{ik}v_k = (\lambda - a_{ii})v_i \Rightarrow |\lambda - a_{ii}||v_i| \leq \sum_{k\neq i} |a_{ik}||v_k| \leq \sum_{k\neq i} |a_{ik}||v_i| = \rho_i|v_i|$. Cioè $\lambda \in \mathcal{C}_i$. □

0.66 Esempio

Sia

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 0 & -i & 1 \\ i & 0 & 3 \end{bmatrix} \in \mathbb{C}^{3\times 3}$$

Applicando il Teorema di Gershgorin ad A e ad A^{T} si ottengono gli insiemi rappresentati in Figura 5. Gli autovalori di A, rappresentati dalle crocette, appartengono all'intersezione dei due insiemi.

Si osservi che, come asserito dal punto (2) del Teorema di Gershgorin, il cerchio $\mathcal{C}_4$ contiene *un solo* autovalore, $\mathcal{C}_5 \cup \mathcal{C}_6$ i restanti due.

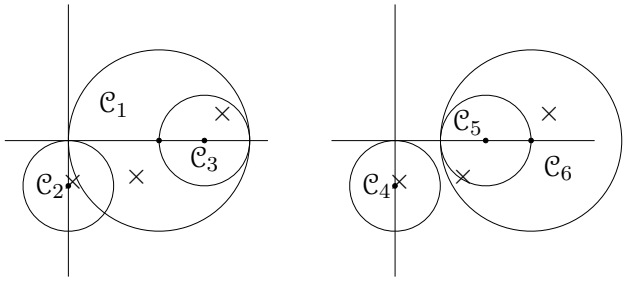

Figura 5. (a) Cerchi per A; (b) cerchi per A^{T}.

0.67 Esempio

Sia

$$A = \begin{bmatrix} 5 & 1 & 1 & -1 \\ 0 & 5 & 0 & 3 \\ 2 & 1 & 6 & 1 \\ 1 & 0 & 3 & 5 \end{bmatrix} \in \mathbb{C}^{4\times 4}$$

Siccome $0 \notin \cup_{k=1}^{4} \mathcal{C}_k$, A risulta invertibile.

G Matrici a predominanza diagonale forte

0.68 Definizione

Una matrice $A \in \mathbb{C}^{n\times n}$ si dice *a predominanza diagonale forte per righe* se per $k = 1, \ldots, n$ si ha

$$|a_{kk}| > \sum_{j\neq k} |a_{kj}|$$

ovvero *a predominanza diagonale forte per colonne* se per $j = 1, \ldots, n$ si ha

$$|a_{jj}| > \sum_{k\neq j} |a_{kj}|$$

Si osservi che in entrambi i casi si ha $a_{kk} \neq 0$ per $k = 1, \ldots, n$.

Il teorema seguente fornisce una caratterizzazione delle matrici a predominanza diagonale forte.

0.69 Teorema

Sia $A \in \mathbb{C}^{n\times n}$. Posto $D = \mathrm{diag}(a_{11}, \ldots, a_{nn})$ si ha:

(r) A è a predominanza diagonale forte per righe se e solo se D è invertibile e $||I - D^{-1}A||_\infty < 1$;

(c) A è a predominanza diagonale forte per colonne se e solo se D è invertibile e $||I - AD^{-1}||_1 < 1$.

Dimostrazione

Sia A a predominanza diagonale forte per righe. Allora D è invertibile e, posto $N = A - D$ si ha $||I - D^{-1}A||_\infty = || - D^{-1}N||_\infty < 1$. Viceversa ...

Questo prova l'asserto (r). L'altro si prova in modo analogo. □

Il Teorema di Gershgorin consente di provare i seguenti teoremi:

0.70 Teorema

Una matrice a predominanza diagonale forte (per righe o per colonne) è invertibile.

0.71 Teorema

Sia A una matrice hermitiana a predominanza diagonale forte. Se per ogni k si ha $a_{kk} > 0$, allora A è definita positiva.

Appendice: richiami di algebra lineare

0.72 Definizione

Sia $A \in \mathbb{R}^{n\times n}$ $[A \in \mathbb{C}^{n\times n}]$.

Un numero $\lambda \in \mathbb{R}$ $[\lambda \in \mathbb{C}]$ si dice *autovalore* di A se esiste $v \in \mathbb{R}^n$ $[v \in \mathbb{C}^n]$ non nullo tale che $Av = \lambda v$.

Un vettore $v \in \mathbb{R}^n$ $[v \in \mathbb{C}^n]$ non nullo si chiama *autovettore* di A se esiste $\lambda \in \mathbb{R}$ $[\lambda \in \mathbb{C}]$ tale che $Av = \lambda v$.

Sia λ un autovalore di A. Il sottospazio di $\mathbb{R}^n$ [di $\mathbb{C}^n$]

$$V(\lambda) = \{v \in \mathbb{R}^n \text{ tali che } Av = \lambda v\}$$
$$[\ V(\lambda) = \{v \in \mathbb{C}^n \text{ tali che } Av = \lambda v\}\ ]$$

si chiama *autospazio* di A relativo a λ.

0.73 Definizione (spettro, raggio spettrale)

Sia $A \in \mathbb{R}^{n\times n}$ $[A \in \mathbb{C}^{n\times n}]$. Il polinomio $P_A(\xi) = \det(A - \xi I)$ si chiama *polinomio caratteristico* di A. L'insieme

$$\sigma(A) = \{\lambda \in \mathbb{C} \text{ tali che } \lambda \text{ è radice del polinomio caratteristico di } A\}$$

si chiama *spettro* di A. Il numero

$$\rho(A) = \max\{|\lambda|, \lambda \in \sigma(A)\}$$

si chiama *raggio spettrale* di A.

0.74 Osservazione

(1) Siano $A \in \mathbb{R}^{n\times n}$ e $\lambda \in \mathbb{R}$.

λ è autovalore di A se e solo se $\lambda \in \sigma(A)$. In tal caso, per l'autospazio relativo a λ si ha

$$V(\lambda) = \ker(A - \lambda I) \subset \mathbb{R}^n$$

e $\dim V(\lambda) \geq 1$.

Se $\dim \ker(A - \lambda I) = 0$, λ non è autovalore di A.

(2) Siano $A \in \mathbb{C}^{n\times n}$ e $\lambda \in \mathbb{C}$.

λ è autovalore di A se e solo se $\lambda \in \sigma(A)$. In tal caso, per l'autospazio relativo a λ si ha

$$V(\lambda) = \ker(A - \lambda I) \subset \mathbb{C}^n$$

e $\dim V(\lambda) \geq 1$.

Se $\dim \ker(A - \lambda I) = 0$, λ non è autovalore di A.

(3) Sia $A \in \mathbb{C}^{n\times n}$. Dette $e_1, \ldots, e_n$ le colonne della matrice identica I_n, e $\lambda_1, \ldots, \lambda_n$ le radici del polinomio caratteristico di A, si ha:

$$\begin{aligned} P_A(\xi) = \det(A - \xi I) &= \det(a_1 - \xi e_1, \ldots, a_n - \xi e_n) \\ &= (-\xi)^n + \Sigma_1(-\xi)^{n-1} + \Sigma_2(-\xi)^{n-2} + \cdots + \Sigma_n \end{aligned}$$

dove Σ_k = somma dei determinanti dei minori principali di A di ordine k (ad esempio: $\Sigma_1 = a_{11} + \cdots + a_{nn}$, $\Sigma_n = \det A$), e

$$\begin{aligned} P_A(\xi) &= (-1)^n(\xi - \lambda_1)\cdots(\xi - \lambda_n) = (\lambda_1 - \xi)\cdots(\lambda_n - \xi) \\ &= (-\xi)^n + (-\xi)^{n-1}(\lambda_1 + \cdots + \lambda_n) + \cdots + (\lambda_1 \cdots \lambda_n) \end{aligned}$$

Confrontando le due espressioni si ottengono le uguaglianze

$$a_{11} + \cdots + a_{nn} = \lambda_1 + \cdots + \lambda_n \quad , \quad \det A = \lambda_1 \cdots \lambda_n$$

La quantità $a_{11} + \cdots + a_{nn}$ si dice *traccia* di A e si indica $\mathrm{tr}A$.

0.75 Definizione (matrici simili)

Due matrici $A, B \in \mathbb{R}^{n\times n}$ $[A, B \in \mathbb{C}^{n\times n}]$ si dicono *simili* se esiste $S \in \mathbb{R}^{n\times n}$ $[S \in \mathbb{C}^{n\times n}]$ invertibile tale che $A = SBS^{-1}$.

La similitudine è una relazione di equivalenza.

0.76 Teorema

Due matrici $A, B \in \mathbb{R}^{n\times n}$ $[A, B \in \mathbb{C}^{n\times n}]$ simili hanno lo stesso polinomio caratteristico (ed autospazi isomorfi).

Dimostrazione

Si ha: $P_B(\xi) = \det(B - \xi I) = \det(S^{-1}AS - \xi S^{-1}S) = \det S^{-1}(A - \xi I)S = \det(A - \xi I) = P_A(\xi)$. Inoltre, sia λ un autovalore. Allora: $v \in \ker(B - \lambda I) \Leftrightarrow (B - \lambda I)v = 0 \Leftrightarrow (S^{-1}AS - \lambda S^{-1}S)v = 0 \Leftrightarrow S^{-1}(A - \lambda I)Sv = 0 \Leftrightarrow Sv \in \ker(A - \lambda I)$.

0.77 Definizione (matrice diagonalizzabile)

Una matrice $A \in \mathbb{R}^{n\times n}$ $[A \in \mathbb{C}^{n\times n}]$ si dice *diagonalizzabile* se è simile ad una matrice diagonale: esistono $S \in \mathbb{R}^{n\times n}$ $[S \in \mathbb{C}^{n\times n}]$ invertibile e $\Lambda \in \mathbb{R}^{n\times n}$ $[\Lambda \in \mathbb{C}^{n\times n}]$ diagonale tali che $A = S\Lambda S^{-1}$. In tal caso, posto

$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ e $S = (s_1, \ldots, s_n)$ si ha: $\lambda_1, \ldots, \lambda_n$ sono gli autovalori di A e $s_k \in V(\lambda_k)$ per $k = 1, \ldots, n$.

0.78 Teorema (diagonalizzabilità)
Una matrice $A \in \mathbb{R}^{n\times n}$ $[A \in \mathbb{C}^{n\times n}]$ è diagonalizzabile se e solo se esiste una base di $\mathbb{R}^n$ [di $\mathbb{C}^n$] costituita da autovettori di A.

0.79 Problema
Dimostrare che la matrice

$$A = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \in \mathbb{C}^{2\times 2}$$

non è diagonalizzabile. △

0.80 Definizione (matrice triangolabile)
Una matrice $A \in \mathbb{R}^{n\times n}$ $[A \in \mathbb{C}^{n\times n}]$ si dice *triangolabile* se è simile ad una matrice triangolare: esistono $S \in \mathbb{R}^{n\times n}$ $[S \in \mathbb{C}^{n\times n}]$ invertibile e $T \in \mathbb{R}^{n\times n}$ $[T \in \mathbb{C}^{n\times n}]$ triangolare tali che $A = STS^{-1}$.

0.81 Teorema (triangolabilità, Schur)
Una matrice $A \in \mathbb{C}^{n\times n}$ è triangolabile tramite $S \in \mathbb{C}^{n\times n}$ unitaria.
Una matrice $A \in \mathbb{R}^{n\times n}$ è triangolabile se e solo se tutte le radici del polinomio caratteristico di A sono reali. In tal caso, A è triangolabile tramite $S \in \mathbb{R}^{n\times n}$ ortogonale.
(Per la dimostrazione, si veda [GLV] pagina 104.)

0.82 Esercizio
Sia $A \in \mathbb{C}^{n\times n}$ e $\lambda_1, \ldots, \lambda_n$ i suoi autovalori.

(1) Per $k \in \mathbb{Z}$, gli autovalori di A^k sono $\lambda_1^k, \ldots, \lambda_n^k$.

(2) Sia $q \in \mathbb{C}$. Gli autovalori di $B = A + qI$ sono $\lambda_1 + q, \ldots, \lambda_n + q$.

Soluzione
(1) Siano S unitaria e T triangolare tali che $A = STS^{-1}$. Gli autovalori di A sono gli elementi $t_{11}, \ldots, t_{nn}$ della diagonale principale di T; inoltre $A^k = ST^kS^{-1}$ e gli elementi della diagonale principale di T^k sono $t_{11}^k, \ldots, t_{nn}^k$. L'asserto segue dal Teorema 0.76. Si osservi che se $k < 0$, l'asserto ha senso solo se A è invertibile.
(2) Siano S unitaria e T triangolare tali che $A = STS^{-1}$. Allora $B = STS^{-1} + qSS^{-1} = S(T + qI)S^{-1}$ e quindi ...

0.83 Problema
Siano $A \in \mathbb{C}^{n\times n}$, λ autovalore di A, $V(\lambda)$ l'autospazio relativo a λ e $q \in \mathbb{C}$. Si dimostri che

(1) detto W l'autospazio relativo all'autovalore λ^k di A^k, si ha $V(\lambda) \subset W$;

(2) l'autospazio relativo all'autovalore $\lambda + q$ di $A + qI$ è $V(\lambda)$. $\triangle$

Capitolo 1

Funzionalità matematiche del calcolatore e teoria degli errori

In questo capitolo si descrivono le funzionalità matematiche di un calcolatore *ideale* (per le funzionalità matematiche di un calcolatore reale — troppo complesse per essere descritte qui — vedere [IEEE1], [IEEE2], [IEC]) e si studiano gli errori che l'uso di un calcolatore con tali funzionalità comporta nel calcolo del valore di una funzione.

A Numeri di macchina

1.1 Osservazione

Siano β intero ≥ 2 e x reale non zero. Esistono unici $s \in \{0,1\}, b \in \mathbb{Z}$ e $g \in \mathbb{R}$ tali che

(1) $x = (-1)^s \beta^b \, g$

(2) $\beta^{-1} \leq g < 1$

Infatti, posto $s = 0$ se $x > 0, s = 1$ se $x < 0$, $b =$ l'unico intero tale che $\beta^{b-1} \leq |x| < \beta^b$ e $g = \beta^{-b}|x|$, si ha $\beta^{-1} \leq g < 1$.

L'intero b si chiama *esponente* di x, il reale g *frazione* di x.

1.2 Definizione (numeri di macchina)

Siano β intero $\geq 2, m$ intero positivo e

$$F(\beta, m) = \{0\} \cup \Big\{ x \in \mathbb{R} \mid x = (-1)^s \beta^b \, 0.\gamma_1 \cdots \gamma_m \\ \text{con: } s \in \{0,1\}; b \in \mathbb{Z}; \\ \gamma_1, \ldots, \gamma_m \in \{0, \ldots, \beta - 1\}, \gamma_1 \neq 0 \Big\}$$

L'insieme $F(\beta, m)$ dei numeri in *virgola mobile*, base β e *precisione* m, si chiama *insieme dei numeri di macchina.*

Un calcolatore è un dispositivo capace di operare esclusivamente con numeri di macchina.

1.3 Esempio

Sia $M = F(10, 3)$. Si ha

$0 \in M$;
$76.3 \in M : 76.3 = 10^2\, 0.763$;
$10^2\, 0.071 \in M : 10^2\, 0.071 = 10^1\, 0.710$;
$10^{-1}\, 0.1005 \notin M$: la frazione non è compatibile con la precisione.

1.4 Osservazione

Siano β, m come nella Definizione 1.2.

(1) L'insieme $F(\beta, m)$ è un sottoinsieme (numerabile) di $\mathbb{Q}$ e lo si considera *ordinato* come tale.

(2) L'insieme $F(\beta, m)$ è *simmetrico* rispetto a 0.

(3) 0 è un punto di accumulazione di $F(\beta, m)$, e

$$\sup F(\beta, m) = +\infty \quad , \quad \inf F(\beta, m) = -\infty$$

1.5 Definizione (funzioni successore e predecessore)

Sia $M = F(\beta, m)$.

La funzione $\sigma : M \backslash \{0\} \to M \backslash \{0\}$ definita da $\sigma(\xi) = \min\{\eta \in M \mid \xi < \eta\}$ si dice funzione *successore*; la funzione $\pi : M \setminus \{0\} \to M \setminus \{0\}$ definita da $\pi(\xi) = \max\{\eta \in M \mid \eta < \xi\}$ si dice funzione *predecessore.*

Si osservi che $\sigma^{-1} = \pi$.

1.6 Osservazione

Sia $\xi \in M$ positivo, e siano b l'esponente e g la frazione di ξ. Si ha

$$\sigma(\xi) = \beta^b(g + \beta^{-m}) = \xi + \beta^{b-m} \tag{1.1}$$

$$\pi(\xi) = \begin{cases} \beta^b(g - \beta^{-m}) = \xi - \beta^{b-m} & \text{se } g > \beta^{-1} \\ \beta^{b-1}(1 - \beta^{-m}) = \xi - \beta^{b-m}\beta^{-1} & \text{se } g = \beta^{-1} \end{cases}$$

Sia $\xi \in M$ negativo. Si ha

$$\sigma(\xi) = -\pi(-\xi) \quad , \quad \pi(\xi) = -\sigma(-\xi)$$

1.7 Teorema (densità dei numeri di macchina)

Siano $M = F(\beta, m)$ e $\xi \in M$ positivo. Detti b l'esponente e g la frazione di ξ, si ha

$$\text{(a)} \qquad \frac{\sigma(\xi) - \xi}{\beta^b} = \beta^{-m};$$

$$\text{(b)} \quad \frac{\beta^{-m}}{1 - \beta^{-m}} \leq \frac{\sigma(\xi) - \xi}{\xi} \leq \beta^{1-m}$$

Dimostrazione

L'asserto (a) segue dalla (1.1). Inoltre, $\frac{\sigma(\xi)-\xi}{\xi} = \frac{\beta^{b-m}}{\beta^b g}$ da cui, essendo $\beta^{-1} \leq g \leq 1 - \beta^{-m}$, si ha (b). □

1.8 Osservazione

La principale differenza tra le funzionalità matematiche di un calcolatore ideale (come descritto qui) e quelle di un calcolatore reale (come descritto in [IEEE1], [IEEE2], [IEC]) è nella definizione dei numeri di macchina. Precisamente, in un calcolatore reale l'insieme dei numeri di macchina è *finito.*

B Arrotondamento

1.9 Definizione

Siano $M = F(\beta, m)$, $x \in \mathbb{R}$ e

$$\Theta(x) = \{\xi \in M \mid \forall \eta \in M \text{ si ha } |x - \xi| \leq |x - \eta|\}$$

L'insieme $\Theta(x)$ — degli elementi di M che hanno distanza minima da x — ha almeno un elemento, e ne ha due solo se x si trova al centro di un intervallo che ha per estremi due elementi consecutivi di M.

La funzione $\mathrm{rd} : \mathbb{R} \to M$ definita da

$$\mathrm{rd}(x) = \begin{cases} \max \Theta(x) & \text{se } x \geq 0 \\ \min \Theta(x) & \text{se } x < 0 \end{cases}$$

si dice funzione *arrotondamento.*

1.10 Osservazione

La funzione arrotondamento è *dispari* e *non decrescente.*

1.11 Osservazione

Sia $x \in \mathbb{R}$. Se $\mathrm{rd}(x) = 0$ allora $x = 0$.

1.12 Osservazione

Siano x reale positivo, b l'esponente di x, $M = F(\beta, m)$ e $\mathrm{rd} : \mathbb{R} \to M$.

(1) Se $x \in M$ allora $\mathrm{rd}(x) = x$.

(2) Si ha
$$\frac{\pi(\mathrm{rd}(x)) + \mathrm{rd}(x)}{2} \leq x < \frac{\mathrm{rd}(x) + \sigma(\mathrm{rd}(x))}{2}$$

(3) Se $x < \mathrm{rd}(x)$ e la frazione di $\mathrm{rd}(x)$ è β^{-1} allora l'esponente di $\mathrm{rd}(x)$ è $b+1$; altrimenti l'esponente di $\mathrm{rd}(x)$ è b.

(4) Si ha
$$|x - \mathrm{rd}(x)| \leq \frac{1}{2}\,\beta^{b-m}$$
(La relazione segue dai punti (2) e (3) di questa Osservazione, tenuto conto del punto (a) del Teorema 1.7.)

1.13 Definizione (funzioni δ, ϵ, η)
Sia $\mathrm{rd} : \mathbb{R} \to F(\beta, m)$.
La funzione $\delta : \mathbb{R} \to \mathbb{R}$ definita da
$$\delta(x) = \mathrm{rd}(x) - x$$
si dice funzione *errore assoluto.*
La funzione $\epsilon : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ $[\eta : \mathbb{R} \setminus \{0\} \to \mathbb{R}]$ definita da
$$\epsilon(x) = \frac{\mathrm{rd}(x) - x}{x} \qquad \left[\eta(x) = \frac{\mathrm{rd}(x) - x}{\mathrm{rd}(x)}\right]$$
si dice funzione *errore relativo.*
La funzione errore assoluto è *dispari,* le funzioni errore relativo sono *pari.*

1.14 Teorema
Sia x reale positivo, e b l'esponente di x. Si ha
$$|\delta(x)| \leq \frac{1}{2}\,\beta^{b-m} \quad , \quad |\epsilon(x)| \leq \frac{1}{2}\,\beta^{1-m} \qquad \left[|\eta(x)| \leq \frac{1}{2}\,\beta^{1-m}\right]$$

Dimostrazione
La prima relazione segue dal punto (4) dell'Osservazione 1.12, la seconda segue dalla prima e dalla considerazione che, detta g la frazione di x, si ha $\beta^{-1} \leq g < 1$. Dunque $\beta^{b-1} \leq x$ e
$$|\epsilon(x)| \leq \frac{1}{2}\,\frac{\beta^{b-m}}{x} \leq \frac{1}{2}\,\beta^{1-m}$$

Per l'ultima relazione, siccome $\beta^{b-1} \leq x$ e $\beta^{b-1} \in M$, si ha $\beta^{b-1} \leq \mathrm{rd}(x)$. Dunque

$$|\eta(x)| = \frac{|\delta(x)|}{\mathrm{rd}(x)} \leq \frac{1}{2}\,\frac{\beta^{b-m}}{\mathrm{rd}(x)} \leq \frac{1}{2}\,\beta^{1-m}$$

e l'asserto è provato. □

1.15 Definizione (precisione di macchina)

Chiameremo *precisione di macchina* la quantità

$$u = \frac{1}{2}\,\beta^{1-m}$$

In termini di precisione di macchina, la tesi del Teorema 1.14 riguardante la funzione errore assoluto si riformula:

$$|\delta(x)| \leq \beta^{b-1}\,u$$

ovvero, in modo più significativo:

$$|\delta(x)| \leq u\,|x|$$

e quella riguardante la funzione errore relativo:

$$|\epsilon(x)| \leq u \qquad [|\eta(x)| \leq u]$$

1.16 Problema

Siano $\beta = 2$, m intero positivo ed $M = F(\beta, m)$. Dimostrare che per ogni numero di macchina positivo ξ si ha

$$u < \frac{\sigma(\xi) - \xi}{\xi} \leq 2u$$

△

1.17 Osservazione

Dalle definizioni segue che

$$\mathrm{rd}(x) = \begin{cases} x + \delta(x) \\ x\,(1 + \epsilon(x)) \end{cases} \qquad \text{e} \quad x = \mathrm{rd}(x)\,(1 + \eta(x))$$

1.18 Esempio

(1) In $F(10, 4)$ si ha $\mathrm{rd}(\sqrt{2}) = 10^1\,0.1414$.

Un intervallo, di estremi razionali, che contiene $\sqrt{2}$ si può ottenere considerando che l'esponente di $\sqrt{2}$ è 1 e quindi $|\delta(\sqrt{2})| \leq 5 \cdot 10^{-4}$. Allora si ha $\sqrt{2} = \mathrm{rd}(\sqrt{2}) - \delta(\sqrt{2}) \in [1.4135\,;\, 1.4145]$.

(2) In $F(10,4)$ si ha $\mathrm{rd}(x) = 10^2\, 0.1000 = 10$.

Un intervallo, di estremi razionali, che contiene x si può ottenere considerando il successore ed il predecessore di $\mathrm{rd}(x)$. Si ottiene $x \in [9.9995\,;\, 10.005)$ (vedere la Figura 6.)

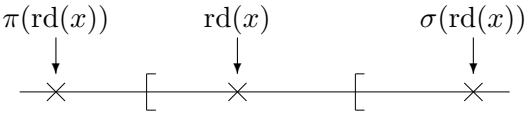

Figura 6.

C Funzioni predefinite ed algoritmi

Sia $M = F(\beta, m)$ e $\mathrm{rd} : \mathbb{R} \to M$. Per ogni intero positivo n, definiamo

$$M^n = \{\, (\xi_1, \ldots, \xi_n)^{\mathsf{T}} \text{ con } \xi_1, \ldots, \xi_n \in M \,\}$$

Siano $\Phi = \{\phi : A \to M^\ell$ tali che $A \subset M^n$ non vuoto e n, ℓ interi positivi$\}$ e $\mathcal{F}$ un sottoinsieme finito di Φ. Chiameremo *funzioni predefinite* gli elementi di $\mathcal{F}$.

1.19 Osservazione

L'insieme M eredita da $\mathbb{Q}$ i *confronti:* se $\diamond \in \{=, \neq, >, \geq, <, \leq\}$ allora per ogni $\xi_1, \xi_2 \in M$ si ha $\xi_1 \diamond \xi_2$ in M se e solo se $\xi_1 \diamond \xi_2$ in $\mathbb{Q}$.

Chiameremo *eleborazione elementare* il calcolo di una funzione predefinita oppure il confronto di due elementi di M.

Un calcolatore è un dispositivo capace di eseguire sequenze *finite* di eleborazioni elementari.

Le funzionalità matematiche di un calcolatore sono quindi definite dall'insieme dei numeri di macchina, dall'insieme delle funzioni predefinite e dai confronti.

1.20 Definizione (pseudo–operazioni aritmetiche)

Le funzioni definite da

$$\begin{aligned}
&\oplus : M \times M \to M \quad , \quad \xi_1 \oplus \xi_2 = \mathrm{rd}(\xi_1 + \xi_2) \\
&\ominus : M \times M \to M \quad , \quad \xi_1 \ominus \xi_2 = \mathrm{rd}(\xi_1 - \xi_2) \\
&\otimes : M \times M \to M \quad , \quad \xi_1 \otimes \xi_2 = \mathrm{rd}(\xi_1 \xi_2) \\
&\oslash : M \times M \setminus \{0\} \to M \quad , \quad \xi_1 \oslash \xi_2 = \mathrm{rd}(\xi_1 / \xi_2)
\end{aligned}$$

si chiamano *pseudo–operazioni aritmetiche.* In ogni calcolatore, le pseudo–operazioni aritmetiche sono funzioni predefinite.

1.21 Osservazione

Si consideri una pseudo–operazione $\circledast$ e siano $\xi_1, \xi_2 \in M$. Per la definizione di rd, il valore $\xi_1 \circledast \xi_2$ è uno dei numeri di macchina che rendono *minima* la distanza da $\xi_1 * \xi_2$.

In questo senso la definizione delle pseudo–operazioni è la migliore possibile.

1.22 Esempio

Sia $M = F(10, 2)$.

(A.1) $\oplus$ è simmetrica

(A.2) $\oplus$ non è associativa (es: $\xi_1 = 10^2\, 0.10, \xi_2 = \xi_3 = 10^0\, 0.38; (\xi_1 \oplus \xi_2) \oplus \xi_3 \neq \xi_1 \oplus (\xi_2 \oplus \xi_3)$)

(A.3) dati $\xi_1, \xi_2, \alpha \in M$, se $\xi_1 > \xi_2$ allora $\xi_1 \oplus \alpha \geq \xi_2 \oplus \alpha$ (monotonia: segue dalla corrispondente proprietà di rd — Osservazione 1.10)

(A.4) per ogni $\xi \in M$ si ha $\xi \oplus 0 = \xi$, ma "lo zero non è unico" (es: $\xi = 10^2\, 0.67; \xi \oplus 10^{-2}\, 0.11 = \xi$)

(A.5) per ogni $\xi \in M$ si ha $\xi \oplus (-\xi) = 0$, e "l'opposto è unico" (vedere l'Osservazione 1.11)

(M.1) $\otimes$ è simmetrica

(M.2) $\otimes$ non è associativa (es: $\xi_1 = 10^0\, 0.20, \xi_2 = 10^1\, 0.51, \xi_3 = 10^1\, 0.76; (\xi_1 \otimes \xi_2) \otimes \xi_3 \neq \xi_1 \otimes (\xi_2 \otimes \xi_3)$)

(M.3) dati $\xi_1, \xi_2, \alpha \in M$ con $\alpha > 0$. Se $\xi_1 > \xi_2$ allora $\xi_1 \otimes \alpha \geq \xi_2 \otimes \alpha$ (monotonia: segue dalla corrispondente proprietà di rd — Osservazione 1.10)

(M.4) per ogni $\xi \in M$ si ha $\xi \otimes 1 = \xi$, ma "l'unità non è unica" (es: $\xi = 10^0\, 0.49; \xi \otimes 10^0\, 0.99 = \xi$)

(M.5) sia $\xi \in M$ non zero: l'insieme degli inversi di ξ

$$\{\, \theta \in M \text{ tali che } \xi \otimes \theta = 1\}$$

può essere vuoto o avere più di un elemento: "l'inverso può non esistere o non essere unico" (es: $\xi = 10^0\, 0.20, \xi \otimes 10^1\, 0.50 = 1$ e $\xi \otimes 10^1\, 0.51 = 1$; $\xi = 10^1\, 0.89, \xi \otimes 10^0\, 0.11 = 10^0\, 0.98 < 1$ e $\xi \otimes 10^0\, 0.12 = 10^1\, 0.11 > 1$ e quindi, per la monotonia di $\otimes$ — (M.3) —, ξ non ha inverso)

Sia $\phi : A \to M^\ell$ un elemento di Φ. È possibile utilizzare un calcolatore per calcolare ϕ se per ogni $\xi \in A$ l'elemento $\phi(\xi)$ è ottenibile con un numero *finito* di eleborazioni elementari. Chiameremo *algoritmo* (che calcola ϕ) una descrizione di ϕ in termini di sequenze finite di elaborazioni elementari.

D Errori nel calcolo di una funzione

Siano $\Omega \subset \mathbb{R}^n$, $f : \Omega \to \mathbb{R}$ e $\phi : \Omega \cap M^n \to M$ un elemento di Φ.

Dati $x \in \Omega$ e $\xi \in \Omega \cap M^n$, si utilizza il valore $\phi(\xi)$ per approssimare il valore $f(x)$.

1.23 Definizione (errore totale)

Si chiama *errore totale*

(a) $\delta_t = \phi(\xi) - f(x)$ (errore assoluto)

(b) $\epsilon_t = \dfrac{\phi(\xi) - f(x)}{f(x)} \quad \left[\eta_t = \dfrac{\phi(\xi) - f(x)}{\phi(\xi)}\right]$ (errore relativo)

L'errore relativo è definito solo se $f(x) \neq 0$ $[\phi(\xi) \neq 0]$.

1.24 Definizione (errore trasmesso dai dati)

Sia $\hat{x} \in \Omega$. Si chiama *errore trasmesso dai dati* (nel calcolo del valore di f in x):

(a) $\delta_d = f(\hat{x}) - f(x)$ (errore assoluto)

(b) $\epsilon_d = \dfrac{f(\hat{x}) - f(x)}{f(x)} \quad \left[\eta_d = \dfrac{f(\hat{x}) - f(x)}{f(\hat{x})}\right]$ (errore relativo)

L'errore relativo è definito solo se $f(x) \neq 0$ $[f(\hat{x}) \neq 0]$.

1.25 Osservazione

L'errore trasmesso dai dati è indipendente da ϕ.

1.26 Definizione (errore algoritmico)

Si chiama *errore algoritmico* (nell'uso di ϕ per approssimare f in ξ):

(a) $\delta_a = \phi(\xi) - f(\xi)$ (errore assoluto)

(b) $\epsilon_a = \dfrac{\phi(\xi) - f(\xi)}{f(\xi)} \quad \left[\eta_a = \dfrac{\phi(\xi) - f(\xi)}{\phi(\xi)}\right]$ (errore relativo)

L'errore relativo è definito solo se $f(\xi) \neq 0$ $[\phi(\xi) \neq 0]$.*

1.27 Osservazione

L'errore algoritmico non dipende da x.

Utilizzando le definizioni date (con $\hat{x} = \xi$) si ha

*Si osservi che l'errore algoritmico dipende da ϕ, *non* dall'algoritmo che calcola ϕ.

(a) $\delta_t = \delta_a + \delta_d$

(b) $\epsilon_t = \epsilon_a + \epsilon_d + \epsilon_a\epsilon_d \quad [\eta_t = \eta_a + \eta_d + \eta_a\eta_d]$

La relazione (a) è immediata; per la relazione (b) si ha

$$\begin{aligned}\tfrac{\phi(\xi)-f(x)}{f(x)} &= \tfrac{\phi(\xi)-f(\xi)}{f(\xi)}\tfrac{f(\xi)}{f(x)} + \tfrac{f(\xi)-f(x)}{f(x)} \\ &= \tfrac{\phi(\xi)-f(\xi)}{f(\xi)}\left[1 + \tfrac{f(\xi)}{f(x)} - 1\right] + \tfrac{f(\xi)-f(x)}{f(x)} \\ &= \epsilon_a[1+\epsilon_d] + \epsilon_d\end{aligned}$$

ed analogamente per η_t.

Studio dell'errore trasmesso dai dati ("condizionamento")

1.28 Definizione (errore sui dati)
Sia $\hat{x} \in \Omega$. Per $k = 1, \ldots, n$, si chiama *errore sul dato* k–esimo

(a) $\delta_k = \hat{x}_k - x_k$ (errore assoluto)

(b) $\epsilon_k = \dfrac{\hat{x}_k - x_k}{x_k} \quad \left[\eta_k = \dfrac{\hat{x}_k - x_k}{\hat{x}_k}\right]$ (errore relativo, $x_k \neq 0$ $[\hat{x}_k \neq 0]$)

1.29 Esempio

(1) Sia $f(x_1, x_2) = x_1 + x_2$; allora

(a) $\delta_d = \delta_1 + \delta_2$

(b) $\epsilon_d = \dfrac{x_1}{x_1+x_2}\,\epsilon_1 + \dfrac{x_2}{x_1+x_2}\,\epsilon_2 \quad \left[\eta_d = \dfrac{\hat{x}_1}{\hat{x}_1+\hat{x}_2}\,\eta_1 + \dfrac{\hat{x}_2}{\hat{x}_1+\hat{x}_2}\,\eta_2\right]$

(2) Sia $f(x_1, x_2) = x_1 x_2$; allora

(a) $\delta_d = x_2\delta_1 + x_1\delta_2 + \delta_1\delta_2$

(b) $\epsilon_d = \epsilon_1 + \epsilon_2 + \epsilon_1\epsilon_2 \quad [\eta_d = \eta_1 + \eta_2 - \eta_1\eta_2]$

(3) Sia $f(x_1, x_2) = x_1/x_2$; allora

(a) $\delta_d = \dfrac{x_2\delta_1 - x_1\delta_2}{x_2(x_2+\delta_2)}$

(b) $\epsilon_d = \dfrac{\epsilon_1 - \epsilon_2}{1+\epsilon_2} \quad \left[\eta_d = \dfrac{\eta_1 - \eta_2}{1-\eta_2}\right]$

Le relazioni si ottengono immediatamente sostituendo $\hat{x}_k = x_k + \delta_k$ nel caso di errore assoluto e $\hat{x}_k = x_k\,(1+\epsilon_k)$ $[\hat{x}_k = x_k\,(1-\eta_k)]$ nel caso di errore relativo.

1.30 Esempio

Sia $f(x) = \sqrt{x}$. Allora, per ogni $x, \hat{x} > 0$ si ha

$$\text{(a)} \quad \delta_d = \frac{\delta_1}{\sqrt{x+\delta_1}+\sqrt{x}}$$

$$\text{(b)} \quad \epsilon_d = \frac{\epsilon_1}{1+\sqrt{1+\epsilon_1}} \qquad \left[\eta_d = \frac{\eta_1}{1+\sqrt{1-\eta_1}}\right]$$

Per l'errore assoluto si ha infatti

$$\delta_d = \sqrt{\hat{x}} - \sqrt{x} = \frac{\hat{x}-x}{\sqrt{\hat{x}}+\sqrt{x}} = \frac{\delta_1}{\sqrt{x+\delta_1}+\sqrt{x}}$$

Per l'errore relativo

$$\epsilon_d = \frac{\sqrt{\hat{x}}-\sqrt{x}}{\sqrt{x}} = \frac{\hat{x}-x}{\sqrt{x}}\,\frac{1}{\sqrt{\hat{x}}+\sqrt{x}} = \frac{\hat{x}-x}{x}\,\frac{\sqrt{x}}{\sqrt{\hat{x}}+\sqrt{x}} = \frac{\epsilon_1}{1+\sqrt{1+\epsilon_1}}$$

e analogamente per η_d.

1.31 Esempio

Siano $n = 1$, Ω un intervallo non vuoto, $f : \Omega \to \mathbb{R}$ derivabile nei punti interni di Ω e $x, \hat{x} \in \Omega$.

Per l'errore trasmesso dai dati δ_d si ha, utilizzando il Teorema del valor medio per le derivate (vedere [A], volume 1, pagina 227):

$$\delta_d = f'(\theta(x, \delta_1))\,\delta_1$$

con $\delta_1 = \hat{x} - x$ e $\theta(x, \delta_1)$ tra $\hat{x}$ e x.

Per l'errore trasmesso dai dati ϵ_d $[\eta_d]$ (se $x \neq 0$ e $f(x) \neq 0$ $[\hat{x} \neq 0$ e $f(\hat{x}) \neq 0]$) si ha

$$\epsilon_d = f'(\theta(x, \epsilon_1))\frac{x}{f(x)}\,\epsilon_1 \qquad \left[\eta_d = f'(\theta(x, \eta_1))\frac{\hat{x}}{f(\hat{x})}\,\eta_1\right]$$

con $\epsilon_1 = \frac{\hat{x}-x}{x}$, $\eta_1 = \frac{\hat{x}-x}{\hat{x}}$ e $\theta(x, \epsilon_1)$ $[\theta(x, \eta_1)]$ tra $\hat{x}$ e x.

Nel caso di una funzione $f \in \mathcal{C}^1(\Omega, \mathbb{R})$, $\Omega \subset \mathbb{R}^n$ aperto convesso, si hanno le espressioni analoghe

$$\delta_d = \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(\theta(x, \delta_1, \ldots, \delta_n))\,\delta_k$$

e

$$\epsilon_d = \textstyle\sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(\theta(x, \epsilon_1, \ldots, \epsilon_n))\frac{x_k}{f(x)}\,\epsilon_k$$

$$\left[\eta_d = \textstyle\sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(\theta(x, \eta_1, \ldots, \eta_n))\frac{\hat{x}_k}{f(\hat{x})}\,\eta_k\right]$$

con $\theta(x, \delta_1, \ldots, \delta_n)$, $\theta(x, \epsilon_1, \ldots, \epsilon_n)$, $\theta(x, \eta_1, \ldots, \eta_n)$ nel segmento congiungente $\hat{x}$ ed x.

Chiameremo *funzione di condizionamento* per il problema del calcolo di f in $x \in \Omega$, una funzione che esprime l'errore trasmesso dai dati in termini di x e dell'errore sui dati.

Studio dell'errore algoritmico ("stabilità")

1.32 Osservazione

Siano $f : \Omega \to \mathbb{R}$ e $\phi : M^n \cap \Omega \to M$ la funzione definita da $\phi(\xi) = \mathrm{rd}(f(\xi))$. Se $f(\xi) \neq 0$, detto b l'esponente di $f(\xi)$, per l'errore algoritmico si ha

$$\text{(a)} \quad |\delta_a| = |\phi(\xi) - f(\xi)| = |\mathrm{rd}(f(\xi)) - f(\xi)| \leq \tfrac{1}{2}\,\beta^{b-m}$$

$$\text{(b)} \quad |\epsilon_a| = \left|\frac{\phi(\xi) - f(\xi)}{f(\xi)}\right| = \left|\frac{\mathrm{rd}(f(\xi)) - f(\xi)}{f(\xi)}\right| \leq \frac{1}{2}\,\beta^{1-m} = u$$

$$\left[|\eta_a| = \left|\frac{\phi(\xi) - f(\xi)}{\phi(\xi)}\right| = \left|\frac{\mathrm{rd}(f(\xi)) - f(\xi)}{\mathrm{rd}(f(\xi))}\right| \leq \frac{1}{2}\,\beta^{1-m} = u\right]$$

Le limitazioni si ottengono applicando il Teorema 1.14.

1.33 Esempio

Siano $M = F(10, 12), \mathtt{SQRT}$ la funzione definita, per ogni numero di macchina non negativo ξ, da $\mathtt{SQRT}(\xi) = \mathrm{rd}(\sqrt{\xi})$ e sia $\mathtt{SQRT} \in \mathcal{F}$.

Per approssimare $f(x) = \sqrt{x} - \sqrt{x-1}$, definita in $[1, \infty)$, si considerano le funzioni

$$\phi_1(\xi) = \mathtt{SQRT}(\xi) \ominus \mathtt{SQRT}(\xi \ominus 1)$$
$$\phi_2(\xi) = 1 \oslash (\mathtt{SQRT}(\xi) \oplus \mathtt{SQRT}(\xi \ominus 1))$$

Per $x = 2365 \in M$ si ottiene $\phi_1(2365) = 0.0102825386$ e $\phi_2(2365) = 10^{-2}\,1.02825385784$.

I due risultati sono diversi. Si vuole capire quale è *più affidabile.*

Per ϕ_1 si consideri il diagramma di Figura 7, in cui r_k sono i risultati delle operazioni su $\mathbb{R}$, ρ_k sono i risultati dei singoli passi elementari di elaborazione, $\epsilon_{k,t}$ gli errori (relativi) totali. Ad esempio: $\rho_2 = \mathtt{SQRT}(\rho_1), r_2 = \sqrt{r_1}$;

$$\epsilon_{2,t} = \frac{\rho_2 - r_2}{r_2}$$

Si ha

$$\epsilon_{1,t} \begin{cases} |\epsilon_{1,a}| \leq u \\ \epsilon_{1,d} = 0 \quad \text{perché } \xi, 1 \in M \end{cases}$$

$$\epsilon_{2,t} \begin{cases} |\epsilon_{2,a}| \leq u \\ \epsilon_{2,d} = \frac{\epsilon_{1,t}}{1+\sqrt{1+\epsilon_{1,t}}} \end{cases}$$

$$\epsilon_{3,t} \begin{cases} |\epsilon_{3,a}| \leq u \\ \epsilon_{3,d} = 0 \quad \text{perché } \xi \in M \end{cases}$$

$$\epsilon_{4,t} \begin{cases} |\epsilon_{4,a}| \leq u \\ \epsilon_{4,d} = \frac{r_3}{r_3 - r_2}\,\epsilon_{3,t} - \frac{r_2}{r_3 - r_2}\,\epsilon_{2,t} \end{cases}$$

Usando i valori numerici si ottiene $\rho_1 = 2364$, $\epsilon_{1,t} \approx u$; $\rho_2 = 48.620\ldots$, $\epsilon_{2,t} \approx u$; $\rho_3 = 48.631\ldots$, $\epsilon_{3,t} \approx u$. Siccome

$$\frac{r_3}{r_3 - r_2} > 4700$$

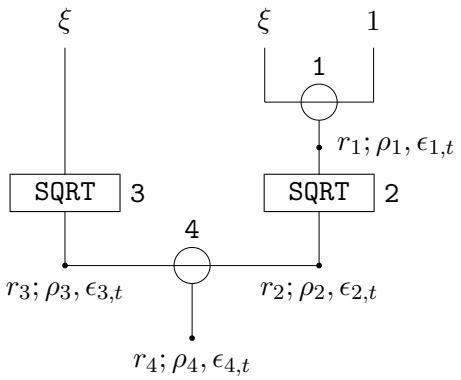

Figura 7.

potrebbe risultare $\epsilon_{4,t} \approx 4700\,u$. Il risultato è quindi "poco affidabile."

Per ϕ_2 si consideri il diagramma di Figura 8.

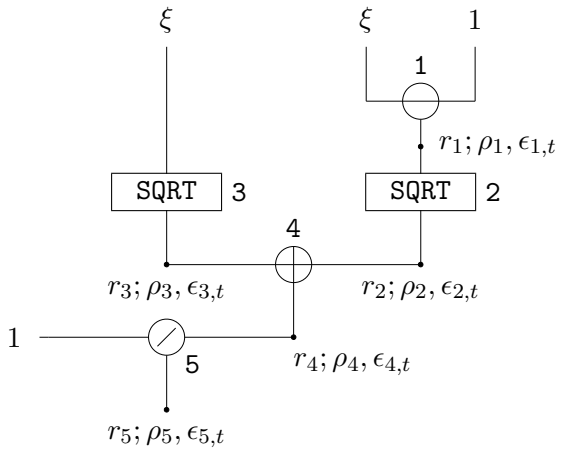

Figura 8.

Per $\epsilon_{1,t}, \epsilon_{2,t}$ e $\epsilon_{3,t}$ si hanno gli stessi risultati del caso di ϕ_1. Per $\epsilon_{4,t}$ ed $\epsilon_{5,t}$ si ha

$$\epsilon_{4,t} \begin{cases} |\epsilon_{4,a}| \leq u \\ \epsilon_{4,d} = \frac{r_3}{r_3+r_2}\,\epsilon_{3,t} + \frac{r_2}{r_3+r_2}\,\epsilon_{2,t} \end{cases}$$

$$\epsilon_{5,t} \begin{cases} |\epsilon_{5,a}| \leq u \\ \epsilon_{5,d} = -\frac{\epsilon_{4,t}}{1+\epsilon_{4,t}} \end{cases}$$

Usando i valori numerici si ottiene $\rho_1 = 2364$, $\epsilon_{1,t} \approx u$; $\rho_2 = 48.620\ldots$,

$\epsilon_{2,t} \approx u$; $\rho_3 = 48.631\ldots$, $\epsilon_{3,t} \approx u$; $\rho_4 = 97.252\ldots$. Poiché

$$\frac{r_3}{r_3 + r_2} < 1 \quad \text{e} \quad \frac{r_2}{r_3 + r_2} < 1$$

risulta $\epsilon_{4,t} \approx u$. Infine, $\epsilon_{5,t} \approx u$, e il risultato è più affidabile del precedente.

1.34 Osservazione (meccanismo di propagazione dell'errore algoritmico)

(a) Sia $\Omega \subset \mathbb{R}^n$ e siano $\phi_1 : \Omega \cap M^n \to M^\ell$ e $\phi_2 : M^\ell \to M$ gli elementi di Φ utilizzati per approssimare, rispettivamente, $f_1 : \Omega \to \mathbb{R}^\ell$ e $f_2 : \mathbb{R}^\ell \to \mathbb{R}$.

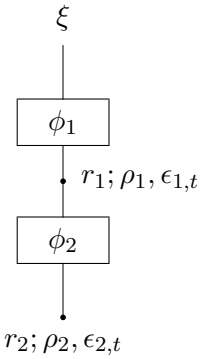

Figura 9.

Si considerino $f : \Omega \to \mathbb{R}$ definita da $f(x) = f_2(f_1(x))$ e $\phi : \Omega \cap M^n \to M$ definita da $\phi(\xi) = \phi_2(\phi_1(\xi))$. Se si utilizza ϕ per approssimare f, con riferimento alla Figura 9 in cui i simboli hanno il significato definito nell'Esempio 1.33, si ha

$$\epsilon_{2,t} = \frac{\phi_2(\rho_1) - f_2(r_1)}{f_2(r_1)} = \epsilon_{2,a}(\epsilon_{2,d} + 1) + \epsilon_{2,d}$$

con

$$\epsilon_{2,a} = \frac{\phi_2(\rho_1) - f_2(\rho_1)}{f_2(\rho_1)} \quad , \quad \epsilon_{2,d} = \frac{f_2(\rho_1) - f_2(r_1)}{f_2(r_1)}$$

L'errore algoritmico del primo blocco è l'errore sul dato r_1 e, come tale, *può* essere amplificato da f_2 *come errore trasmesso dai dati.*

(b) Siano $\alpha \in (0,1), \Omega = (1-\alpha, 1+\alpha)$, $f(x) = \ln x$ e $\phi : \Omega \cap M \to M$ definita da $\phi(\xi) = \xi \ominus 1$.

Si vuole studiare l'errore algoritmico

$$\epsilon_a = \frac{\phi(\xi) - f(\xi)}{f(\xi)}$$

Sia $g : \Omega \to \mathbb{R}$ definita da $g(x) = x - 1$. Allora

$$\epsilon_a = \epsilon'_a(1 + \epsilon'') + \epsilon''$$

con

$$\epsilon'_a = \frac{\phi(\xi) - g(\xi)}{g(\xi)} \quad , \quad \epsilon'' = \frac{g(\xi) - f(\xi)}{f(\xi)}$$

Per ϵ'_a si ha (se $g(\xi) \neq 0$): $|\epsilon'_a| \leq u$.

Per stimare ϵ'' si consideri che, utilizzando il Teorema del valor medio per le derivate, si ha

$$\ln x - \ln 1 = \frac{1}{\theta}(x - 1) \quad , \quad \theta \text{ tra } x \text{ e } 1$$

e quindi, per $x \in \Omega \setminus \{1\}$, $\theta = \frac{x-1}{\ln x}$ e

$$1 - \alpha \leq \frac{x - 1}{\ln x} \leq 1 + \alpha$$

Allora, per $\xi \neq 1$:

$$|\epsilon''| = |\tfrac{\xi - 1}{\ln \xi} - 1| \leq \alpha$$

Infine

$$|\epsilon_a| \leq |\epsilon'_a| + |\epsilon''| + |\epsilon'_a|\,|\epsilon''|$$

In questo caso, l'errore algoritmico può risultare grande perché *g non è* una buona approssimazione di f.

Sia $f : \Omega \to \mathbb{R}$ e sia $\phi : \Omega \cap M^n \to M$ la funzione utilizzata per approssimare f. Chiameremo *funzione di stabilità* dell'algoritmo ϕ (in realtà dovremmo dire *della funzione* ϕ) una funzione che esprime l'errore algoritmico

$$\epsilon_a = \frac{\phi(\xi) - f(\xi)}{f(\xi)} \quad \text{oppure} \quad \delta_a = \phi(\xi) - f(\xi)$$

in termini di ξ e dell'errore algoritmico delle singole funzioni predefinite.

Quando l'errore algoritmico delle singole funzioni predefinite si propaga in modo soddisfacente (cioè: "abbastanza poco"), l'algoritmo (in realtà dovremmo dire *la funzione calcolata dall'algoritmo*) si dice *stabile.*

E Il caso complesso

In questo paragrafo si descrive una comune estensione delle funzionalità matematiche del calcolatore: la possibilità di operare con numeri complessi.

1.35 Definizione (numeri complessi di macchina e pseudo–operazioni aritmetiche)

Sia $M = F(\beta, m)$. L'insieme

$$M_c = \{\xi + i\eta \quad \text{con} \quad \xi, \eta \in M\}$$

si chiama *insieme dei numeri complessi di macchina.*

Le pseudo–operazioni aritmetiche in M_c, sono definite in termini di pseudo–operazioni in M. Per la somma, la sottrazione ed il prodotto si ha:[†]

$$(\xi_1 + i\eta_1) \oplus_c (\xi_2 + i\eta_2) = (\xi_1 \oplus \xi_2) + i(\eta_1 \oplus \eta_2)$$
$$(\xi_1 + i\eta_1) \ominus_c (\xi_2 + i\eta_2) = (\xi_1 \ominus \xi_2) + i(\eta_1 \ominus \eta_2)$$
$$(\xi_1 + i\eta_1) \otimes_c (\xi_2 + i\eta_2) = [(\xi_1 \otimes \xi_2) \ominus (\eta_1 \otimes \eta_2)] + i[(\xi_1 \otimes \eta_2) \oplus (\eta_1 \otimes \xi_2)]$$

Infine, per la divisione si ha:

$$(\xi_1 + i\eta_1) \oslash_c (\xi_2 + i\eta_2) = [(\xi_1 \otimes \xi_2) \oplus (\eta_1 \otimes \eta_2)] \oslash [(\xi_2 \otimes \xi_2) \oplus (\eta_2 \otimes \eta_2)]$$
$$+i[(\eta_1 \otimes \xi_2) \ominus (\xi_1 \otimes \eta_2)] \oslash [(\xi_2 \otimes \xi_2) \oplus (\eta_2 \otimes \eta_2)]$$

1.36 Osservazione (errore trasmesso dai dati, errore algoritmico)

Siano $\Omega \subset \mathbb{C}^n$, $f : \Omega \to \mathbb{C}$ e $\phi : \Omega \cap M_c^n \to M_c$. Le definizioni di errore totale, trasmesso dai dati, algoritmico e sui dati sono identiche a quelle già date in 1.23, 1.24, 1.26 e 1.28.

Per l'errore trasmesso dai dati relativo alle funzioni somma, prodotto e divisione in $\mathbb{C}$, valgono risultati identici a quelli dati in 1.29.

Se $u \leq \frac{1}{20}$ allora per l'errore algoritmico della pseudo–operazione aritmetica $\circledast$ si ha:

$$\text{per ogni } \zeta_1, \zeta_2 \text{ esiste } \epsilon_a \in \mathbb{C} \text{ tale che } \zeta_1 \circledast \zeta_2 = (\zeta_1 * \zeta_2)(1 + \epsilon_a)$$

con

$$|\epsilon_a| \leq u \quad \text{per } \circledast \in \{\oplus, \ominus\}$$
$$|\epsilon_a| \leq 3u \quad \text{per } \circledast = \otimes$$
$$|\epsilon_a| \leq 7u \quad \text{per } \circledast = \oslash$$

(per la dimostrazione, e per stime più precise, si veda la Sezione G nell'Appendice relativa a questo Capitolo.)

[†]In questa Definizione, e solo qui, si adottano i simboli $\oplus_c, \ominus_c, \otimes_c$ e $\oslash_c$ per le pseudo–operazioni in M_c ed i consueti $\oplus, \ominus, \otimes$ e $\oslash$ per quelle in M. Nel resto di questi appunti, si useranno per entrambe i simboli $\oplus, \ominus, \otimes$ e $\oslash$ lasciando al contesto il compito di chiarire a quali pseudo–operazioni ci si riferisce.

Appendice

F Numeri in virgola fissa

In questo paragrafo sono descritte le funzionalità matematiche di un calcolatore *ideale* che opera, anzichè con numeri in virgola mobile, in un altro ambiente numerico: quello dei numeri in virgola fissa. Questo tipo di calcolatore è di uso frequente nel settore dell'elaborazione numerica dei segnali (si veda, ad esempio, il Capitolo 7 di [PM]).

1.37 Definizione (numeri di macchina)
Siano β intero $\geq 2, m$ intero non negativo e

$$Q(\beta, m) = \left\{ x \in \mathbb{R} \mid x = \beta^{-m}\alpha \quad , \quad \alpha \in \mathbb{Z} \right\}$$

$Q(\beta, m)$ è l'insieme dei numeri in *virgola fissa*, base β e *precisione* m.

1.38 Osservazione
Siano β, m come nella Definizione 1.37.

(1) L'insieme $Q(\beta, m)$ è un sottoinsieme (numerabile) di $\mathbb{Q}$ e lo si considera *ordinato* come tale.

(2) L'insieme $Q(\beta, m)$ è *simmetrico* rispetto a 0.

(3) $\sup Q(\beta, m) = +\infty, \inf Q(\beta, m) = -\infty$.

1.39 Osservazione (funzioni successore e predecessore)
Sia $M^* = Q(\beta, m)$.
Dette σ^*, π^* le funzioni successore e predecessore in M^* (definite in modo ovvio) si ha

$$\sigma^*(\xi) = \sigma^*(\beta^{-m}\alpha) = \beta^{-m}(\alpha + 1) = \xi + \beta^{-m}$$
$$\pi^*(\xi) = \pi^*(\beta^{-m}\alpha) = \beta^{-m}(\alpha - 1) = \xi - \beta^{-m}$$

1.40 Teorema

Sia $\xi = \beta^{-m}\alpha \in Q(\beta, m)$ non zero. Si ha

$$\text{(a)} \quad \sigma^*(\xi) - \xi = \beta^{-m}$$

$$\text{(b)} \quad \frac{\sigma^*(\xi) - \xi}{\xi} = \frac{1}{\alpha}$$

Si confronti l'asserto con quello del Teorema 1.7

1.41 Osservazione (funzione arrotondamento)

Sia $M^* = Q(\beta, m)$.

Detta rd^* la funzione arrotondamento in M^* (definita in modo ovvio e con le stesse proprietà di simmetria della corrispondente funzione in M) si ha

$$\mathrm{rd}^*(x) = 0 \text{ se e solo se } |x| < \tfrac{1}{2}\,\beta^{-m}$$

ed inoltre

$$|x - \mathrm{rd}^*(x)| \leq \tfrac{1}{2}\,\beta^{-m}$$

Si osservi che la stima è *indipendente da* x, e si confronti l'asserto con quello del punto (4) dell'Osservazione 1.12.

1.42 Osservazione (funzioni $\delta^*, \epsilon^*, \eta^*$)

Dette $\delta^*, \epsilon^*, \eta^*$ le funzioni errore assoluto ed errore relativo in M^* (definite in modo ovvio e con le stesse proprietà di simmetria delle corrispondenti funzioni in M) si ha

$$|\delta^*(x)| \leq \tfrac{1}{2}\,\beta^{-m}$$

$$|\epsilon^*(x)| \begin{cases} = 1 & \text{per } |x| \leq \frac{1}{2}\beta^{-m} \\ \leq \frac{1}{2}\,\frac{\beta^{-m}}{|x|} & \text{altrimenti} \end{cases}$$

$$|\eta^*(x)| \leq \tfrac{1}{2\alpha} \quad (\text{dove } \mathrm{rd}^*(x) = \beta^{-m}\alpha)$$

(La dimostrazione è immediata. Si osservi che la funzione η^* è definita per $|x| \geq \frac{1}{2}\,\beta^{-m}$.)

Chiameremo *precisione di macchina* la quantità

$$u^* = \tfrac{1}{2}\,\beta^{-m}$$

1.43 Osservazione (pseudo–operazioni aritmetiche)

Le pseudo–operazioni aritmetiche in M^* sono definite in termini delle corrispondenti operazioni in $\mathbb{R}$ e della funzione rd^* analogamente a quanto fatto nel caso di M (Definizione 1.20).

Le proprietà, invece, non sono identiche. Sia $M^* = Q(10, 2)$.[‡]

[‡]In questa Osservazione, $\oplus$ e $\otimes$ indicano le pseudo–operazioni in M^*.

(A.1*) Per ogni $\xi_1, \xi_2 \in M^*$ si ha

$$\xi_1 \oplus \xi_2 = \xi_1 + \xi_2$$

(M.1*) $\otimes$ è simmetrica

(M.2*) $\otimes$ non è associativa (ad esempio: $\xi_1 = 10^{-2}\,20, \xi_2 = 10^{-2}\,512, \xi_3 = 10^{-2}\,761; (\xi_1 \otimes \xi_2) \otimes \xi_3 \neq \xi_1 \otimes (\xi_2 \otimes \xi_3)$)

(M.3*) dati $\xi_1, \xi_2, \alpha \in M^*$ con $\alpha > 0$. Se $\xi_1 > \xi_2$ allora $\xi_1 \otimes \alpha \geq \xi_2 \otimes \alpha$ (monotonia: segue dalla corrispondente proprietà di rd* — Osservazione 1.41)

(M.4*) per ogni $\xi \in M^*$ si ha $\xi \otimes 1 = \xi$, ma "l'unità non è unica" (ad esempio: $\xi = 10^{-2}\,49; \xi \otimes 10^{-2}\,99 = \xi$)

(M.5*) sia $\xi \in M^*$ non zero: l'insieme degli inversi di ξ

$$\{\, \theta \in M^* \text{ tali che } \xi \otimes \theta = 1\}$$

può essere vuoto o avere più di un elemento: "l'inverso può non esistere o non essere unico" (es: $\xi = 10^{-2}\,20, \xi \otimes 10^{-2}\,500 = 1$ e $\xi \otimes 10^{-2}\,501 = 1$; $\xi = 10^{-2}\,890, \xi \otimes 10^{-2}\,11 = 10^{-2}\,98 < 1$ e $\xi \otimes 10^{-2}\,12 = 10^{-2}\,107 > 1$ e quindi, per la monotonia di $\otimes$ — (M.3*) — ξ non ha inverso)

1.44 Osservazione (errore algoritmico)

Siano $\Omega \subset \mathbb{R}^n$, $f : \Omega \to \mathbb{R}$ e $\phi : \Omega \cap (M^*)^n \to M^*$. La definizione di errore algoritmico è identica a quella già data in 1.26.

Per l'errore algoritmico assoluto relativo alla pseudo–operazione aritmetica $\circledast$ si ha:

$$|\delta_a^*| \leq \tfrac{1}{2}\beta^{-m} = u^*$$

(Per la dimostrazione di questa stima, come per stime di ϵ_a^* e η_a^*, si veda l'Osservazione 1.42.)

G Dimostrazioni

In questa Sezione si dimostrano le stime enunciate nella Sezione E.

Sia $\zeta_k = \sigma_k + i\,\omega_k$ per $k = 1, 2$.

Si consideri la pseudo–operazione $\oplus$ (per $\ominus$ si procede in modo analogo). Posto $\zeta_1 + \zeta_2 = z$, $\sigma_1 + \sigma_2 = s$, $\omega_1 + \omega_2 = w$ si ha:

$$\zeta_1 \oplus \zeta_2 = \begin{cases} (1 + \epsilon_a) z \\ (\sigma_1 \overset{1}{\oplus} \sigma_2) + i\,(\omega_1 \overset{2}{\oplus} \omega_2) \end{cases}$$

Indicando con ϵ_1 ed ϵ_2 gli errori algoritmici relativi alle pseudo–addizioni in M, si ha

$$\epsilon_a z = \epsilon_1 s + i\,\epsilon_2 w$$

Allora, tenuto conto che $|\epsilon_k| \leq u$:

$$|\epsilon_a|^2\,|z|^2 = \epsilon_1^2\,s^2 + \epsilon_2^2\,w^2 \leq |z|^2 u^2$$

da cui il primo asserto.

Per la pseudo–operazione $\otimes$, posto $\zeta_1\zeta_2 = z$, si ha:

$$\zeta_1 \otimes \zeta_2 = \begin{cases} (1+\epsilon_a)z \\ [(\sigma_1 \overset{1}{\otimes} \sigma_2) \overset{3}{\ominus} (\omega_1 \overset{2}{\otimes} \omega_2)] + i\,[(\sigma_1 \overset{4}{\otimes} \omega_2) \overset{6}{\oplus} (\omega_1 \overset{5}{\otimes} \sigma_2)] \end{cases}$$

Indicando con ϵ_j l'errore algoritmico relativo alla j–esima pseudo–operazione in M, e posto $e_{k\ell} = \epsilon_k + \epsilon_\ell + \epsilon_k\epsilon_\ell$ si ha

$$\epsilon_a z = (e_{13}\,\sigma_1\sigma_2 - e_{23}\,\omega_1\omega_2) + i\,(e_{46}\,\sigma_1\omega_2 + e_{56}\,\sigma_2\omega_1)$$

Tenuto conto che $|\epsilon_k| \leq u$, si ha

$$|e_{k\ell}| \leq u\,(2+u) \tag{1.2}$$

e quindi:

$$|\epsilon_a|^2|z|^2 \leq u^2\,(2+u)^2(\sigma_1^2\sigma_2^2+\omega_1^2\omega_2^2+\sigma_1^2\omega_2^2+\sigma_2^2\omega_1^2)+2\sigma_1\sigma_2\omega_1\omega_2(e_{46}e_{56}-e_{13}e_{23})$$

Ma:

$$\sigma_1^2\sigma_2^2 + \omega_1^2\omega_2^2 + \sigma_1^2\omega_2^2 + \sigma_2^2\omega_1^2 = |z|^2 \quad , \quad |e_{46}e_{56} - e_{13}e_{23}| \leq 2u^2(2+u)^2$$

e quindi

$$|\epsilon_a|^2 \leq u^2(2+u)^2\left(1 + 4\frac{|\sigma_1\sigma_2\omega_1\omega_2|}{|z|^2}\right)$$

Infine, essendo

$$\frac{|\sigma_1\sigma_2\omega_1\omega_2|}{\sigma_1^2\sigma_2^2 + \omega_1^2\omega_2^2 + \sigma_1^2\omega_2^2 + \sigma_2^2\omega_1^2} = \frac{1}{|\frac{\sigma_1\sigma_2}{\omega_1\omega_2}| + |\frac{\omega_1\omega_2}{\sigma_1\sigma_2}| + |\frac{\sigma_1\omega_2}{\sigma_2\omega_1}| + |\frac{\sigma_2\omega_1}{\sigma_1\omega_2}|} \leq \frac{1}{4}$$

— infatti: $\inf\{x + \frac{1}{x}, x > 0\} = 2$ — si ha

$$|\epsilon_a|^2 \leq 2u^2(2+u)^2$$

e quindi

$$|\epsilon_a| \leq \sqrt{2}\,u\,(2+u) \tag{1.3}$$

Il secondo asserto è una immediata conseguenza di questa stima.

Per la pseudo–operazione $\oslash$, posto $\frac{\zeta_1}{\zeta_2} = z$, si ha:

$$\zeta_1 \oslash \zeta_2 = \begin{cases} (1+\epsilon_a)z \\ (\zeta_1 \overset{1}{\otimes} \overline{\zeta}_2) \overset{5}{\oslash} [(\sigma_2 \overset{2}{\otimes} \sigma_2) \overset{4}{\oplus} (\omega_2 \overset{3}{\otimes} \omega_2)] \end{cases}$$

Si osservi che la quinta pseudo–operazione è una pseudo–divisione di un elemento di M_c per un elemento di M. In tal caso si ha:

$$(\sigma + i\,\omega) \oslash \xi = (\sigma \oslash \xi) + i\,(\omega \oslash \xi)$$

e, operando come già fatto per la pseudo–somma, si ottiene

$$(\sigma + i\,\omega) \oslash \xi = \frac{\sigma + i\,\omega}{\xi}(1+e)$$

con $|e| \leq u$.

Con i simboli già introdotti si ha

$$(\sigma_2 \overset{2}{\otimes} \sigma_2) \overset{4}{\oplus} (\omega_2 \overset{3}{\otimes} \omega_2) = (\sigma_2^2 + \omega_2^2)(1 + e_{234})$$

con

$$e_{234} = \frac{\sigma_2^2\, e_{24} + \omega_2^2\, e_{34}}{\sigma_2^2 + \omega_2^2}$$

e, per la (1.2):

$$|e_{234}| \leq u\,(2+u) \tag{1.4}$$

Allora si ottiene

$$\epsilon_a = \frac{(1+\epsilon_1)(1+e)}{1+e_{234}} - 1 = \frac{\epsilon_1 + e + \epsilon_1 e - e_{234}}{1+e_{234}}$$

da cui, tenuto conto di (1.3) e (1.4):

$$|\epsilon_a| \leq \frac{\sqrt{2}\,u\,(2+u) + u + \sqrt{2}\,u^2\,(2+u) + u\,(2+u)}{1 - u\,(2+u)}$$

cioé

$$|\epsilon_a| \leq \frac{(3+2\sqrt{2})u + (1+3\sqrt{2})u^2 + \sqrt{2}u^3}{1 - u\,(2+u)}$$

Il terzo asserto è una immediata conseguenza di questa stima.

Capitolo 2

Zeri di funzioni reali

PROBLEMA: data $f : \Omega \to \mathbb{R}$ continua, $\Omega \subset \mathbb{R}$, determinare gli zeri di f.

Un numero reale $\alpha \in \Omega$ si dice *zero* di f se $f(\alpha) = 0$.

A Metodo di bisezione

Idea del metodo:

costruire una successione di intervalli sempre più piccoli, contenenti uno zero di f, utilizzando il *Teorema di esistenza degli zeri* (Siano $a < b \in \mathbb{R}$ ed $f : [a, b] \to \mathbb{R}$ continua tali che $f(a)f(b) < 0$. Allora esiste uno zero di f in (a, b) — vedere [A], volume 1, Teorema 3.6, pagina 178, oppure [C], Teorema 4.17, pagina 214).

Descrizione del METODO DI BISEZIONE

`dati:` $f : [a, b] \to \mathbb{R}$ continua tale che $f(a)f(b) < 0$;

$a_0 = a, b_0 = b,\ I_0 = [a_0, b_0],\ x_0 = (a_0 + b_0)/2$;
`per` $k = 1, \ldots$ `ripeti`
  `se` $f(x_{k-1}) = 0$ `allora` STOP (x_{k-1} è uno zero di f) `altrimenti`
    `>` `se` $f(x_{k-1})f(b_{k-1}) < 0$ `allora` $a_k = x_{k-1}, b_k = b_{k-1}$;
    `>` `se` $f(a_{k-1})f(x_{k-1}) < 0$ `allora` $a_k = a_{k-1}, b_k = x_{k-1}$;
    `>` $I_k = [a_k, b_k],\ x_k = (a_k + b_k)/2$

`uscita:` quando un opportuno criterio di arresto è verificato: I_k, x_k.

2.1 Esempio

Con riferimento alla Figura 10, si ha $I_0 = [a, b],\ x_0 = \frac{a+b}{2};\ f(a)f(x_0) < 0$ allora $I_1 = [a, x_0],\ x_1 = \frac{a+x_0}{2}$ etc.

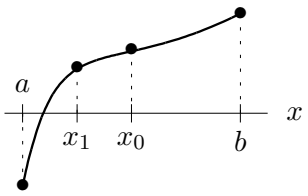

Figura 10.

Discussione del metodo.

2.2 Osservazione

(1) Se per un k si ha $f(x_k) = 0$, il metodo si arresta. Altrimenti si ha: I_k contiene uno zero di f, $I_k \subset I_{k-1}$, $\operatorname{mis} I_k = \frac{1}{2} \operatorname{mis} I_{k-1}$ e quindi $\lim_{k\to\infty} \operatorname{mis} I_k = 0$ e, infine, $\lim_{k\to\infty} x_k$ esiste ed è uno zero di f.

(2) (Criterio di arresto) Sia $\epsilon > 0$.

Un criterio di arresto che utilizza una stima dell'errore assoluto è:

$$\operatorname{mis} I_k < 2\epsilon$$

In tal caso si ha, detto α uno zero di f in I_k :

$$|x_k - \alpha| \leq \frac{\operatorname{mis} I_k}{2} < \epsilon$$

Se $0 \notin I_k$, posto $m_k = \min\{|a_k|, |b_k|\}$, un criterio di arresto che utilizza una stima dell'errore relativo è:

$$\frac{\operatorname{mis} I_k}{m_k} < 2\epsilon$$

In tal caso si ha, detto α uno zero di f in I_k :

$$\frac{|x_k - \alpha|}{|\alpha|} \leq \frac{\operatorname{mis} I_k}{2\, m_k} < \epsilon$$

(3) Utilizzando il criterio d'arresto "assoluto," la rapidità di convergenza del metodo di bisezione dipende solo dall'ampiezza dell'intervallo iniziale:

$$\operatorname{mis} I_k = \frac{\operatorname{mis} I_0}{2^k}$$

Utilizzando il criterio d'arresto "relativo," la rapidità di convergenza del metodo di bisezione dipende dall'ampiezza dell'intervallo iniziale *e* dalla posizione dello zero.

2.3 Esempio

Sia $f(x) = x^2 - 2$. Individuare il numero di zeri di f, *separarli*[*] e stimare un numero di passi sufficiente per conoscere ciascuno zero con errore assoluto $\leq 10^{-6}$.

Soluzione

La funzione ha due zeri: $\alpha_1 = \sqrt{2} \in [1, 2]$ e $\alpha_2 = -\sqrt{2} \in [-2, -1]$. Consideriamo α_1. Si ha: $f(1)f(2) < 0$ e, posto $I_0 = [1, 2]$, si ottiene $|x_k - \alpha_1| \leq 10^{-6}$ se $\operatorname{mis} I_k \leq 2 \cdot 10^{-6}$. Si ha:

$$\operatorname{mis} I_k = \frac{\operatorname{mis} I_0}{2^k} \leq 2 \cdot 10^{-6} \Rightarrow 2^{k+1} \geq 10^6$$

Il primo intero che verifica è $k = 19$, e x_{19} verifica la condizione richiesta.

2.4 Osservazione

Operando in $M = F(\beta, m)$, il metodo di bisezione produce una successione di intervalli $[a_k, b_k]$ ad estremi in M, ed una successione ξ_k di elementi di M definita da:

$$\xi_k = (a_k \oplus b_k) \oslash 2$$

In questo caso il criterio di arresto assoluto è:

$$b_k \ominus a_k < \operatorname{rd}(\epsilon)$$

che, quando verificato, comporta:

$$b_k - a_k < \epsilon$$

Dunque se la procedura si arresta con criterio di arresto verificato, si ha:

$$|\xi_k - \alpha| \leqslant b_k - a_k < \epsilon$$

B Metodi ad un punto

Idea del metodo:

trasformare il problema della determinazione degli zeri di f nel problema della determinazione dei punti uniti[†] di una funzione opportuna.

2.5 Esempio

(a) Siano $f : \mathbb{R} \to \mathbb{R}$ continua, $g : \mathbb{R} \to \mathbb{R} \setminus \{0\}$ continua. Posto $h(x) = x - f(x)g(x)$ si ha: α zero di f se e solo se α punto unito di h.

[*]Separare gli zeri di una funzione significa individuare intervalli *disgiunti* contenenti, ciascuno, *uno* zero della funzione.

[†]Un numero reale $\alpha \in \Omega$ si dice *punto unito* della funzione $h : \Omega \to \mathbb{R}$ se $\alpha = h(\alpha)$.

(b) Sia $f(x) = x + \log x$. Per le funzioni $h_1(x) = -\log x$, $h_2(x) = e^{-x}$, $h_3(x) = (x + e^{-x})/2$ i punti fissi coincidono con gli zeri di f.

Descrizione di un METODO AD UN PUNTO

`dati:` $h : \Omega \to \mathbb{R}$ continua, $\Omega \subset \mathbb{R}$, $\beta \in \Omega$;

$x_0 = \beta$;
`per` $k = 1, \ldots$ `ripeti`
`>` `se` $x_{k-1} \notin \Omega$ `allora` STOP `altrimenti` $x_k = h(x_{k-1})$

`uscita:` quando un opportuno criterio di arresto è verificato: x_k.

Discussione del metodo.

2.6 Teorema (di convergenza locale)
Siano $h \in \mathcal{C}^1(\Omega)$, $[a, b] \subset \Omega$, ed $x_0 \in [a, b]$ tali che

(1) $[a, b]$ contiene almeno un punto unito di h;

(2) esiste $L \in [0, 1)$ tale che $|h'(x)| \leq L$ per ogni $x \in [a, b]$;

(3) posto $x_k = h(x_{k-1}), k = 1, 2, \ldots$, si ha $x_k \in [a, b]$ per ogni k.

Allora: in $[a, b]$ vi è un solo punto unito di h e, detto α tale punto unito, si ha $\lim_{k\to\infty} x_k = \alpha$.

Dimostrazione
Proviamo l'unicità del punto unito per assurdo. Siano α e β punti uniti di h in $[a, b]$, distinti. Allora:

$$|\alpha - \beta| = |h(\alpha) - h(\beta)| = |h'(\xi)||\alpha - \beta|$$

con ξ tra α e β.[‡] Da (2) segue $|\alpha - \beta| < |\alpha - \beta|$, assurdo. Si osservi che questo risultato non dipende dall'ipotesi (3).

Per la successione si ha:

$$\begin{aligned} |x_k - \alpha| &= |h(x_{k-1}) - h(\alpha)| = |h'(\xi_k)||x_{k-1} - \alpha| \text{ (con } \xi_k \text{ tra } x_{k-1} \text{ e } \alpha) \\ &\leq L|x_{k-1} - \alpha| \end{aligned}$$

Ripetendo i passaggi:

$$|x_k - \alpha| \leq L^2|x_{k-2} - \alpha| \leq \cdots \leq L^k|x_0 - \alpha|$$

[‡]Si è applicato ad h il *Teorema del valor medio per le derivate*: Sia f continua su $[a, b]$ e derivabile su (a, b); allora esiste $c \in (a, b)$ tale che $f(b) - f(a) = f'(c)(b - a)$ (vedere [A], volume 1, Teorema 4.5, pagina 227, oppure [C], Teorema 4.29, pagina 227).

Siccome $L \in [0,1)$, si ha $\lim_{k\to\infty} |x_k - \alpha| = 0$. □

2.7 Esempio

Sia $h(x) = \frac{1}{2}\cos x$. L'intervallo $[0, \frac{\pi}{2}]$ contiene almeno un punto unito di h (vedere la Figura 11).

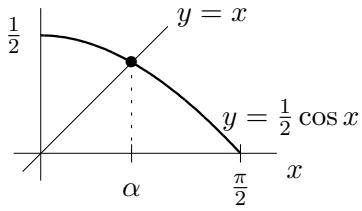

Figura 11.

Inoltre, essendo $h'(x) = -\frac{1}{2}\operatorname{sen} x$, si ha $|h'(x)| \leq \frac{1}{2} = L$ per ogni $x \in [0, \frac{\pi}{2}]$. Infine: $x \in [0, \frac{\pi}{2}] \Rightarrow h(x) \in [0, \frac{1}{2}] \subset [0, \frac{\pi}{2}]$. (Si osservi che se $h([a,b]) \subset [a,b]$ allora $x_0 \in [a,b] \Rightarrow x_k \in [a,b]$ per $k = 1, 2, \ldots$)

Allora, *ogni* $x_0 \in [a,b]$ genera una successione che verifica l'ipotesi (3) del Teorema 2.6. Si può concludere che: h ha un unico punto unito in $[0, \frac{\pi}{2}]$ e per *ogni* $x_0 \in [0, \frac{\pi}{2}]$ la successione generata dal metodo converge a tale punto unito.

2.8 Osservazione

Siano h ed $[a,b]$ come nelle ipotesi (1) e (2) del Teorema di convergenza locale.

Non è detto che un qualsiasi $x_0 \in [a,b]$ verifichi l'ipotesi (3) del Teorema. In particolare, non è detto che per ogni $x_0 \in [a,b]$ si abbia $\lim_{k\to\infty} x_k = \alpha$.

Si consideri, ad esempio, la funzione di Figura 12 (non è di classe $\mathcal{C}^1$, però ...)

2.9 Osservazione (scelta del punto iniziale)

Siano h ed $[a,b]$ come nelle ipotesi (1) e (2) del Teorema 2.6, ed α punto unito di h in $[a,b]$. Allora, posto $d = \min\{|\alpha - a|, |b - \alpha|\}$ e $I = \{x : |x - \alpha| \leq d\} \subset [a,b]$, si ha $h(I) \subset I$. Quindi, ogni $x_0 \in I$ verifica l'ipotesi (3) del Teorema 2.6.

In particolare, l'estremo di $[a,b]$ più vicino ad α verifica tale ipotesi.

(*Dim.* Infatti, per $x \in I$, si ha $|h(x) - \alpha| = |h(x) - h(\alpha)| < |x - \alpha| \leq d$ e quindi $h(x) \in I$.)

2.10 Esempio

Sia $f(x) = x + \log x$.

(1) Determinare il numero di zeri di f e separarli.

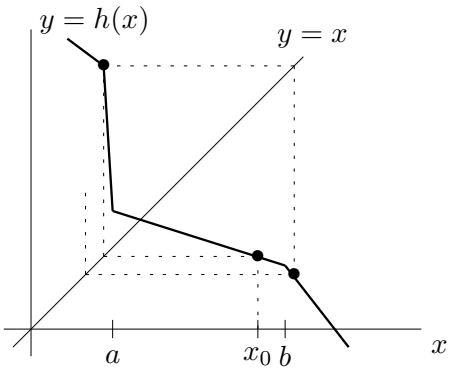

Figura 12.

(2) Discutere i metodi iterativi definiti dalle funzioni $h_1(x) = -\log x$, $h_2(x) = e^{-x}$, $h_3(x) = (x + e^{-x})/2$.

Soluzione

(1) La funzione, definita per $x > 0$, ha un solo zero (infatti è monotona crescente e ...) e, detto α tale zero, si ha $\alpha \in (0, 1)$ — vedere la Figura 13.

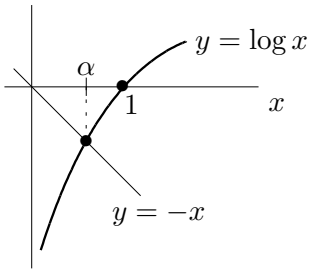

Figura 13.

(2) Si osserva che i punti fissi delle tre funzioni coincidono con gli zeri di f (vedere il punto (b) dell'Esempio 2.5).

Per la funzione h_1 si ha: $h_1'(x) = -1/x$ e per ogni $x \in (0, 1)$ risulta $|h_1'(x)| > 1$. Quindi non è possibile trovare un intervallo che contiene α e che verifica l'ipotesi (2) del Teorema di convergenza locale.

Per la funzione h_2 si ha: $h_2'(x) = -e^{-x}$. Allora, per ogni $x \in (0, 1)$ risulta $|h_2'(x)| < 1$, ma $|h_2'(0)| = 1$. Quindi l'ipotesi (2) del Teorema di convergenza locale non è verificata da $[0, 1]$. Però, restringendo l'intervallo che separa la radice a $[\frac{1}{2}, 1]$ — con un passo di bisezione — si ottiene $|h_2'(x)| \leq e^{-\frac{1}{2}} = L_2 < 1$ (vedere la Figura 14). Quindi, in base all'Osservazione 2.9, una

successione convergente ad α si ottiene ponendo $x_0 = \frac{1}{2}$ oppure $x_0 = 1$ a seconda ...

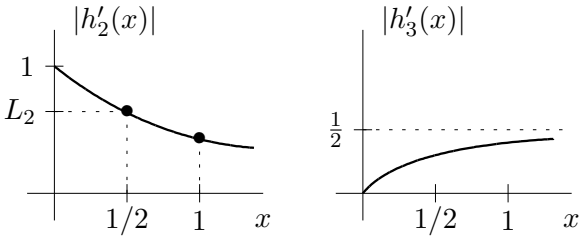

Figura 14.

Per la funzione h_3 si ha: $h_3'(x) = (1 - e^{-x})/2$. Per ogni $x \in [\frac{1}{2}, 1]$ si ha $|h_3'(x)| \leq (1 - e^{-1})/2 = L_3 < 1$ (vedere la Figura 14), e quindi anche in questo caso ...

2.11 Osservazione (numero di zeri)

Sia $n \in \mathbb{N}$ e $f \in \mathcal{C}^n(a, b)$, continua in $[a, b]$. Se $f^{(n)} \neq 0$ in (a, b), allora f ha *al più* n zeri in $[a, b]$.

(*Dim.* Se f ha $n + 1$ zeri in $[a, b]$, applicando il Teorema del valor medio per la derivata ad $f, f^{(1)}, \ldots, f^{(n-1)}$ si dimostra l'esistenza di un valore $\theta \in (a, b)$ tale che $f^{(n)}(\theta) = 0$.)

2.12 Osservazione

Si consideri una successione x_k generata dal metodo ad un punto definito dalla funzione h, e sia α un punto unito di h. Se $|h'(\alpha)| > 1$ allora: o la successione è definitivamente[§] uguale ad α, oppure $\lim_{k\to\infty} x_k \neq \alpha$.

(*Dim.* Supponiamo che per ogni k sia $x_k \neq \alpha$ e che $\lim_{k\to\infty} x_k = \alpha$ cioé: per ogni $\epsilon > 0$ esiste N tale che $k \geq N \Rightarrow |x_k - \alpha| < \epsilon$. Scelto ϵ tale che $|h'(x)| \geq M > 1$ per $|x - \alpha| < \epsilon$, si ha $|x_{N+n} - \alpha| = |h(x_{N+n-1}) - h(\alpha)| = |h'(\xi)||x_{N+n-1} - \alpha|$ con ξ tra x_{N+n-1} ed α e, siccome $|x_{N+n-1} - \alpha| < \epsilon$, risulta $|h'(\xi)| \geq M$ e quindi: $|x_{N+n} - \alpha| \geq M|x_{N+n-1} - \alpha|$. Ripetendo i passaggi: $|x_{N+n} - \alpha| \geq M^n|x_N - \alpha|$. Poichè $M > 1$, scelto n sufficientemente elevato, si ha $|x_{N+n} - \alpha| > \epsilon$, assurdo.)

In base a questa osservazione, la funzione h_1 dell'Esempio 2.10 non è utilizzabile per approssimare lo zero di f.

2.13 Osservazione (ordine di convergenza)

Si consideri il metodo ad un punto definito dalla funzione $h : \Omega \to \mathbb{R}$. Sia α un punto unito di h ed x_k una successione generata dal metodo e convergente (ma non definitivamente uguale) ad α.

[§]Una successione x_k ha *definitivamente* una proprietà P se esiste $n \in \mathbb{N}$ tale che per ogni $k \geq n, x_k$ ha la proprietà P.

(1) Se esiste $(a,b) \subset \Omega$ tale che

$$h \in \mathcal{C}^1(a,b) \quad , \quad \alpha \in (a,b) \quad , \quad 0 < |h'(\alpha)| < 1 \tag{2.1}$$

allora esistono un intervallo chiuso $I \subset (a,b)$ contenente α, L_1 ed L_2 tali che

$$0 < L_1 \leq |h'(x)| \leq L_2 < 1$$

per ogni $x \in I$. Sia N tale che $x_k \in I$ per $k \geq N$. Utilizzando il Teorema del valor medio per le derivate si ottiene

$$L_1|x_{N+j-1} - \alpha| \leq |x_{N+j} - \alpha| \leq L_2|x_{N+j-1} - \alpha|$$

per $j = 1, 2, \ldots$ e quindi

$$L_1^j|x_N - \alpha| \leq |x_{N+j} - \alpha| \leq L_2^j|x_N - \alpha|$$

Cioè: la successione x_{N+j} tende ad α *almeno* rapidamente come L_2^j, ma *non più* rapidamente di L_1^j.

Se esiste $(a,b) \subset \Omega$ che verifica le condizioni (2.1), si dice che il metodo definito dalla funzione h ha *ordine di convergenza ad* α pari ad 1.

(2) Se esiste $(a,b) \subset \Omega$ tale che

$$h \in \mathcal{C}^2(a,b) \quad , \quad \alpha \in (a,b) \quad , \quad h'(\alpha) = 0 \quad , \quad h^{(2)}(\alpha) \neq 0 \tag{2.2}$$

allora esistono un intervallo chiuso $I \subset (a,b)$ contenente α, M_1 ed M_2 tali che

$$0 < M_1 \leq |h^{(2)}(x)| \leq M_2$$

per ogni $x \in I$. Sia N tale che $x_k \in I$ per $k \geq N$ e $\frac{M_2}{2}|x_N - \alpha| < 1$. Procedendo come al punto (1) si ottiene

$$\frac{M_1}{2}|x_{N+j-1} - \alpha|^2 \leq |x_{N+j} - \alpha| \leq \frac{M_2}{2}|x_{N+j-1} - \alpha|^2$$

per $j = 1, 2, \ldots$ e quindi, posto $c_r = \frac{M_r}{2}|x_N - \alpha|, r = 1, 2$, si ha $0 < c_1 \leq c_2 < 1$ e

$$c_1^{2^j-1}|x_N - \alpha| \leq |x_{N+j} - \alpha| \leq c_2^{2^j-1}|x_N - \alpha|$$

Cioè: la successione x_{N+j} tende ad α almeno rapidamente come $c_2^{2^j-1}$, ma non più rapidamente di $c_1^{2^j-1}$.

Se esiste $(a,b) \subset \Omega$ che verifica le condizioni (2.2), si dice che il metodo definito dalla funzione h ha ordine di convergenza ad α pari a 2.

In modo analogo si definisce l'ordine di convergenza ad α maggiore di 2.

2.14 Esempio (continua)

Sia $f(x) = x + \log x$ ed α lo zero di f in $[\frac{1}{2}, 1]$. Approssimare, con errore assoluto $\leq 10^{-3}$, il valore di α.

Soluzione

I metodi definiti dalle funzioni h_2 e h_3 risultano entrambi avere ordine di convergenza pari ad 1. Inoltre, per ogni $x \in [\frac{1}{2}, 1]$, risulta

$$e^{-1} \leq |h_2'(x)| \leq e^{-\frac{1}{2}} \quad , \quad \frac{1 - e^{-\frac{1}{2}}}{2} \leq |h_3'(x)| \leq \frac{1 - e^{-1}}{2}$$

Poiché $\frac{1-e^{-1}}{2} < e^{-1}$, il metodo definito da h_3 garantisce, per quanto detto nell'Osservazione 2.13, una rapidità di convergenza superiore a quello definito da h_2. Si utilizza pertanto il metodo definito dalla funzione h_3 :

$$x_{k+1} = \frac{x_k + e^{-x_k}}{2}$$

Siccome $f(\frac{3}{4}) > 0$, si sceglie come punto iniziale $x_0 = \frac{1}{2}$. Si ha, inoltre: $h_3'(x) > 0$ per $x \in [\frac{1}{2}, 1]$, dunque x_k ed x_{k-1} sono "dalla stessa parte" rispetto ad α.

Per determinare un intervallo di ampiezza $\leq 10^{-3}$ contenente α, si utilizza la procedura descritta in Figura 15 supponendo:

(a) di determinare x_k' operando in $M = F(10, 3)$ secondo la regola

$$x_k' = \begin{cases} \mathrm{rd}(x_k) & \text{se } \mathrm{rd}(x_k) \geq x_k \\ \sigma(\mathrm{rd}(x_k)) & \text{altrimenti} \end{cases}$$

(b) di saper calcolare i valori di f con errore *relativo* < 1 (cioè, di saper calcolare esattamente il *segno* di f)

(c) di saper calcolare i valori di h_3 con errore sufficientemente piccolo.

Si osservi che, poiché $x_k \in [\frac{1}{2}, \frac{3}{4})$, si ha

$$x_k' - x_k < 10^{-3}$$

(vedere Capitolo 1, Paragrafi A,B).

Si ha: $x_0' = 0.5$ e, siccome $f(x_0') < 0$, risulta $\alpha > x_0'$. Procedendo si ha $x_1 = 0,5532\ldots$ e $x_1' = 0.554$ e, essendo $f(x_1') < 0$, risulta ancora $\alpha > x_1'$.

Si procede allo stesso modo fino a: $x_4 = h_3(x_3') = 0.5671\ldots$ e $x_4' = 0.568$. Essendo $f(x_4') > 0$, risulta $\alpha \in [x_4, x_4']$ e $0 < x_4' - \alpha < 10^{-3}$.

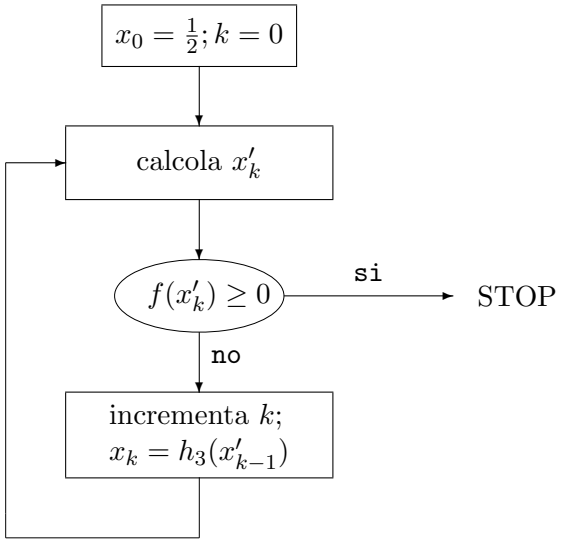

Figura 15. Procedura utilizzata nell'Esempio 2.14

C Metodo di Newton

Sia $f : \Omega \to \mathbb{R}$ derivabile tale che $f'(x) \neq 0$ per ogni $x \in \Omega$. Il metodo di Newton è il metodo ad un punto che si ottiene ponendo $h(x) = x - \frac{f(x)}{f'(x)}$. Si osservi (vedere il punto (a) dell'Esempio 2.5) che α è zero di f se e solo se è punto unito di h.

Descrizione del METODO DI NEWTON

`dati:` $f : \Omega \to \mathbb{R}$ derivabile, $\beta \in \Omega$;

$x_0 = \beta$;
`per` $k = 1, \ldots$ `ripeti`
`>` `se` $x_{k-1} \notin \Omega$ o $f'(x_{k-1}) = 0$ `allora` STOP
`altrimenti` $x_k = x_{k-1} - (f'(x_{k-1}))^{-1} f(x_{k-1})$

`uscita:` quando un opportuno criterio di arresto è verificato: x_k.

Discussione del metodo.

Siano $f : \Omega \to \mathbb{R}$ ed $(a, b) \subset \Omega$ tali che $f \in \mathcal{C}^2(a, b)$, $f' \neq 0$ in (a, b) ed esiste $\alpha \in (a, b)$ zero di f. Posto

$$h(x) = x - \frac{f(x)}{f'(x)}$$

si ha:

$$h'(\alpha) = 0$$

Allora:

(a) esistono un intervallo chiuso $I \ni \alpha$ ed $L \in [0, 1)$ tali che $|h'(x)| \leq L$ per ogni $x \in I$ — cioè che verificano le condizioni (1) e (2) del Teorema di convergenza locale;

(b) l'ordine di convergenza ad α del metodo è *almeno* 2 (si veda il punto (2) dell'Osservazione 2.13).

2.15 Osservazione

(a) Il metodo di Newton è detto *metodo delle tangenti* per l'interpretazione geometrica suggerita dalla Figura 16. Il valore x_k è tale che $f'(x_{k-1})(x_k - x_{k-1}) + f(x_{k-1}) = 0$.

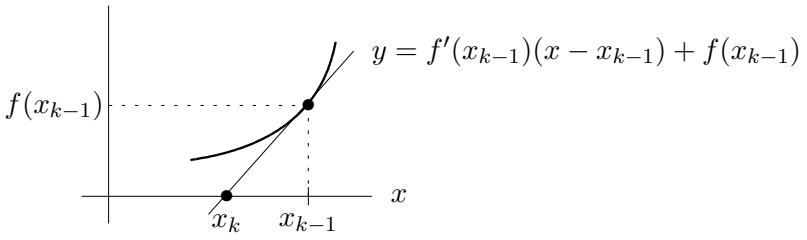

Figura 16. Interpretazione geometrica del metodo di Newton

(b) (Scelta di x_0) Siano $f \in \mathcal{C}^2(\Omega)$ ed $[a, b] \subset \Omega$ tali che $f' \neq 0$ e $f^{(2)} \neq 0$ in $[a, b]$ ed esiste $\alpha \in [a, b]$ zero di f. Posto $x_0 =$ l'estremo di $[a, b]$ in cui $f(x_0)f^{(2)}(x_0) > 0$, la successione x_k risulta *monotona* e convergente ad α (vedere l'esempio in Figura 17).

2.16 Esempio

Sia $f(x) = x^2 - 3$. Approssimare lo zero positivo di f ($\alpha = \sqrt{3}$) con errore assoluto $\leq 10^{-6}$, utilizzando il metodo di Newton.

Soluzione

Si ha: $\alpha \in [1, 2] = I$. Inoltre, $f' > 0$, $f^{(2)} > 0$ in I. Poi: $f(1)f^{(2)}(1) < 0, f(2)f^{(2)}(2) > 0$ e quindi possiamo scegliere $x_0 = 2$.

L'iterazione risulta definita da

$$x_k = h(x_{k-1}) = \frac{1}{2}\left(x_{k-1} + \frac{3}{x_{k-1}}\right)$$

Per determinare un intervallo di ampiezza $\leq 10^{-6}$ contenente α si opera come descritto dalla procedura seguente

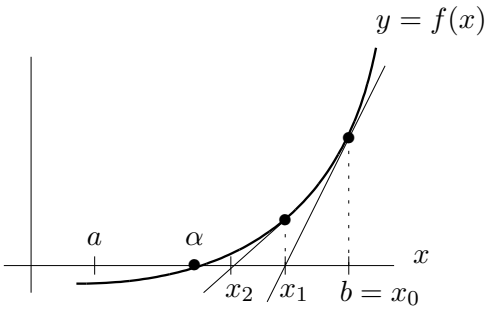

Figura 17. Scelta di x_0 nel metodo di Newton. In questo caso $f' > 0$ e $f^{(2)} > 0$ in $[a, b]$

$x_0 = 2, k = 0$;
`per` $k = 0, 1, \ldots$ `ripeti`
`>` calcola x'_k;
`> se` $f(x'_k) \leq 0$ `allora` STOP
`altrimenti` $x_{k+1} = h(x'_k)$

---

supponendo:

(a) di determinare x'_k operando in $M = F(10, 7)$ secondo la regola

$$x'_k = \begin{cases} \mathrm{rd}(x_k) & \text{se } \mathrm{rd}(x_k) \leq x_k \\ \pi(\mathrm{rd}(x_k)) & \text{altrimenti} \end{cases}$$

(b) di saper calcolare i valori di f con errore *relativo* < 1

(c) di saper calcolare i valori di h con errore sufficientemente piccolo.

Si ottiene:

$x_0 = 2$	$x'_0 = 2$	$f(x'_0) > 0$
$x_1 = 1.75$	$x'_1 = 1.75$	$f(x'_1) > 0$
$x_2 = 1.7321428\ldots$	$x'_2 = 1.732142$	$f(x'_2) > 0$
$x_3 = 1.7320508\ldots$	$x'_3 = 1.732050$	$f(x'_3) < 0$

dunque: $\alpha \in [\, x'_3 \,, x_3 \,]$ e $0 < \alpha - x'_3 < 10^{-6}$.

2.17 Osservazione

Con il metodo di bisezione, partendo da $I = [1, 2]$, occorrono 20 passi per ottenere un intervallo di ampiezza $\leq 10^{-6}$.

D Condizionamento del problema

Siano $f \in \mathcal{C}^1(\Omega)$, $[a,b] \subset \Omega$ e $\delta > 0$ tali che

(1) $f'(x) \neq 0$ per $x \in [a,b]$;

(2) $f(a)f(b) < 0,\ |f(a)| > \delta, |f(b)| > \delta$.

Da (1) e (2) segue l'esistenza di un solo zero di f in $[a,b]$. Sia α tale zero.
Sia infine g continua tale che

(3) $|f(x) - g(x)| \leq \delta$ per ogni $x \in [a,b]$.

Allora

(i) esiste $\beta \in [a,b]$, zero di g;

(ii) posto $m = \min_{[a,b]} |f'|$, si ha $|\alpha - \beta| \leq \frac{\delta}{m}$

(*Dim.* Si ha $f(\alpha) - f(\beta) = g(\beta) - f(\beta)$ ma anche $f(\alpha) - f(\beta) = |f'(\xi)|(\alpha - \beta)$, con ξ tra α e β. Allora $|f'(\xi)|\,|\alpha - \beta| = |g(\beta) - f(\beta)|$, etc.)

Un esempio di problema mal condizionato è quello di Figura 18. In tal caso è $|\alpha - \beta| \gg \delta$.

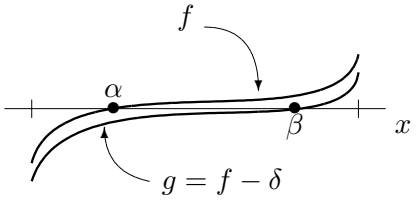

Figura 18. Esempio di problema mal condizionato

2.18 Osservazione

La disuguaglianza del punto (ii) può essere così riletta: l'errore assoluto trasmesso dai dati ($|\alpha - \beta|$) non supera un multiplo ($\frac{1}{m}$) dell'errore assoluto sul dato (δ). Si osservi che siccome

$$m \leq |f'(\alpha)|$$

la quantità $\frac{1}{|f'(\alpha)|}$ è una stima *per difetto* del coefficiente $\frac{1}{m}$.

E Errore algoritmico

Siano $h : \Omega \to \mathbb{R}$, $\phi : \Omega \cap M \to M$, $[a, b] \subset \Omega$, $\xi_0 \in [a, b] \cap M$ tali che

(i) $h, [a, b], \xi_0$ verificano le ipotesi del Teorema di convergenza locale con $L \in [0, 1)$;

(ii) $|h(\xi) - \phi(\xi)| \leq \delta$ per ogni $\xi \in [a, b] \cap M$;

(iii) posto $\xi_k = \phi(\xi_{k-1})$ per $k = 1, 2, \ldots$ si ha $\xi_k \in [a, b]$ per ogni k.

Allora, detto α l'unico punto unito di h in $[a, b]$ si ha:

(A) per ogni $\xi \in [a, b] \cap M$:

$$\text{se } |\xi - \alpha| > \frac{\delta}{1 - L} \text{ allora } |\phi(\xi) - \alpha| < |\xi - \alpha|$$

(B) per ogni k si ha:

$$|\xi_k - x_k| \leq \frac{1 - L^k}{1 - L}\delta$$

(C) per ogni k si ha:

$$|\xi_k - \alpha| \leq \frac{1 - L^k}{1 - L}\,\delta + L^k|\xi_0 - \alpha| = \frac{\delta}{1 - L} + L^k\left(|\xi_0 - \alpha| - \frac{\delta}{1 - L}\right)$$

Infatti:

$$\frac{|\phi(\xi) - \alpha|}{|\xi - \alpha|} \leq \frac{|\phi(\xi) - h(\xi)| + |h(\xi) - h(\alpha)|}{|\xi - \alpha|} \leq \frac{\delta}{|\xi - \alpha|} + \frac{|h(\xi) - h(\alpha)|}{|\xi - \alpha|}$$

Applicando il Teorema del valor medio per le derivate si ottiene inoltre:

$$\frac{|h(\xi) - h(\alpha)|}{|\xi - \alpha|} = |h'(\theta)|$$

con θ tra ξ ed α, e quindi $\theta \in [a, b]$. Per l'ipotesi (i) si ha infine:

$$\frac{|\phi(\xi) - \alpha|}{|\xi - \alpha|} \leq \frac{\delta}{|\xi - \alpha|} + L$$

Siccome $|\xi - \alpha| > \frac{\delta}{1-L}$ equivale a $\frac{\delta}{|\xi - \alpha|} < 1 - L$, si ottiene l'asserto (A).

Per l'asserto (B) si ha, detta x_k la successione ottenuta con il metodo iterativo definito da h a partire da ξ_0 (dunque operando in $\mathbb{R}$):

$$|\xi_k - x_k| = |\phi(\xi_{k-1}) - h(x_{k-1})| \leq |\phi(\xi_{k-1}) - h(\xi_{k-1})| + |h(\xi_{k-1}) - h(x_{k-1})|$$

da cui, utilizzando l'ipotesi (ii) e, di nuovo, il Teorema del valor medio per le derivate:

$$|\xi_k - x_k| \leq \delta + L|\xi_{k-1} - x_{k-1}|$$

Iterando, e ricordando che $x_0 = \xi_0$:

$$|\xi_k - x_k| \leq \delta \left(1 + L + \cdots + L^k\right) = \frac{1 - L^k}{1 - L}\, \delta$$

L'asserto (C) si ottiene immediatamente dall'ipotesi (i) e dall'asserto (B):

$$|\xi_k - \alpha| \leq |\xi_k - x_k| + |x_k - \alpha| \leq \frac{1 - L^k}{1 - L}\, \delta + L^k |\xi_0 - \alpha|$$

2.19 Osservazione

L'asserto (A) garantisce che *finchè la successione ξ_k è sufficientemente lontana da α, la successione delle distanze $|\xi_k - \alpha|$ è decrescente.*

L'asserto (C) afferma che *la successione ξ_k può risultare non convergente ad α, ma in ogni caso* (se le ipotesi sono verificate!) *si "avvicina" all'insieme* $[\alpha - \frac{\delta}{1-L}, \alpha + \frac{\delta}{1-L}]$.

L'asserto (B) afferma che *le successioni ottenute da ξ_0 operando in M, ξ_k, ed operando in $\mathbb{R}$, x_k, non sono mai troppo lontane.*

F Criteri d'arresto

Siano h, ϕ, L e δ come nel paragrafo E. Un criterio d'arresto usato è

$$|\xi_k - \xi_{k-1}| \leq \epsilon$$

con $\epsilon > 0$ dato. Si ha infatti:

$$\xi_k - \xi_{k-1} = \Big(\phi(\xi_{k-1}) - h(\xi_{k-1})\Big) + \Big(h(\xi_{k-1}) - h(\alpha)\Big) + \Big(\alpha - \xi_{k-1}\Big)$$

Applicando il Teorema del valor medio per le derivate al secondo addendo si ottiene

$$\xi_k - \xi_{k-1} = \Big(\phi(\xi_{k-1}) - h(\xi_{k-1})\Big) + \Big(h'(\theta_{k-1}) - 1\Big)\Big(\xi_{k-1} - \alpha\Big)$$

con θ_{k-1} tra ξ_{k-1} ed α. Considerando che $h'(\theta_{k-1}) - 1 \neq 0$ si ottiene

$$\xi_{k-1} - \alpha = \frac{\Big(\phi(\xi_{k-1}) - h(\xi_{k-1})\Big) + \Big(\xi_k - \xi_{k-1}\Big)}{h'(\theta_{k-1}) - 1}$$

e quindi

$$|\xi_{k-1} - \alpha| \leq \frac{\delta + |\xi_k - \xi_{k-1}|}{1 - L}$$

Un diverso criterio d'arresto è basato sul valore di $\psi(\xi_k)$ — approssimazione del valore di $f(\xi_k)$ calcolata dall'algoritmo ψ :

$$|\psi(\xi_k)| \leq \epsilon$$

Infatti, se $f \in \mathcal{C}^1(\Omega), f' \neq 0$ in $[a,b]$ e $|f(\xi) - \psi(\xi)| < \gamma$ in $[a,b] \cap M$, si ha, detto α uno zero di f in $[a,b]$ e posto $m = \min_{[a,b]} |f'|$:

$$|\xi_k - \alpha| \leq \frac{|\psi(\xi_k)| + \gamma}{m}$$

(*Dim.* Infatti: $f(\xi_k) - f(\alpha) = f'(\tau)(\xi_k - \alpha)$, con τ tra ξ_k ed α, e quindi ...)

2.20 Esempio (continua)

Per $f(x) = x^2 - 3$ e $[a,b] = [1,2]$ si ha $m = 2$. Operando come nell'Esempio 1.33, con $\psi(\xi) = (\xi \otimes \xi) \ominus 3$ e $M = F(10,12)$, si ottiene

$$|\eta_t| \leq 8.1 \cdot 10^6 \, u(1+u) \approx 4 \cdot 10^{-5}$$

e quindi $\gamma \approx 1.5 \cdot 10^{-11}$. Essendo $\psi(x_3') = 10^{-6}\, 0.37263$ si ottiene

$$|x_3' - \alpha| \leq \frac{10^{-6}\, 0.37263 + 1.5 \cdot 10^{-11}}{2} \approx 2 \cdot 10^{-7}$$

Capitolo 3

Sistemi di Equazioni Lineari

PROBLEMA: dati $A \in \mathbb{C}^{n\times n}$ invertibile e $b \in \mathbb{C}^n$, determinare $z^* \in \mathbb{C}^n$ tale che $Az^* = b$.

Prima di affrontare il problema nella sua formulazione generale, si osservino i seguenti *casi semplici*:

(D) Se A è *diagonale* ($a_{ij} = 0$ per $i \neq j$), A è invertibile se e solo se $a_{kk} \neq 0$ per $k = 1, \ldots, n$. In tal caso, la soluzione del Problema è:

$$z_k^* = \frac{b_k}{a_{kk}} \quad , \quad k = 1, \ldots, n$$

(T) Se A è *triangolare superiore* ($a_{ij} = 0$ per $i > j$), A è invertibile se e solo se $a_{kk} \neq 0$ per $k = 1, \ldots, n$. In tal caso, la soluzione del Problema si ottiene con la seguente procedura *si* di SOSTITUZIONE ALL'INDIETRO:

$z = si(R, c)$

`dati:` $R \in \mathbb{C}^{n\times n}$ triangolare superiore, $c \in \mathbb{C}^n$;

`se` $r_{nn} \neq 0$ `allora` $z_n = \dfrac{c_n}{r_{nn}}$ `altrimenti` STOP

`per` $k = n-1, \ldots, 1$ `ripeti`

  `se` $r_{kk} \neq 0$ `allora` $z_k = \dfrac{c_k - \sum_{j=k+1}^{n} r_{kj} z_j^*}{r_{kk}}$ `altrimenti` STOP

`uscita:` z

Precisamente:

$$z^* = si(A, b)$$

Se A è *triangolare inferiore* ($a_{ij} = 0$ per $i < j$), si opera in modo analogo, utilizzando la procedura sa di SOSTITUZIONE IN AVANTI (Esercizio: descrivere la procedura!):

$$z^* = sa(A, b)$$

(U) Se A è unitaria, è certamente invertibile. In tal caso la soluzione del Problema è:

$$z^* = A^{\mathsf{H}} b$$

(P) Una matrice si dice *di permutazione* se si ottiene dalla matrice I cambiando l'ordine delle righe o delle colonne. Una matrice di permutazione è dunque unitaria e, se A è la matrice, per ogni $v \in \mathbb{C}^n$, le componenti di Av si ottengono permutando opportunamente quelle di v.

(Ad esempio: la matrice $P = (e_2, e_4, e_1, e_3) \in \mathbb{C}^{4\times 4}$ è di permutazione.)

Se A è una matrice *di permutazione*, è certamente invertibile. In tal caso la soluzione del Problema è:

$$z^* = A^{\mathsf{T}} b$$

e si ottiene permutando le componenti della colonna b.

Nel caso generale, si cerca di *fattorizzare* A con (scrivere A come prodotto di) fattori di uno dei quattro tipi suddetti ...

3.1 Esempio

(I) fattorizzazione LR: $S, D \in \mathbb{C}^{n\times n}$ tali che

(1) S triangolare inferiore con $s_{kk} = 1$ (certamente invertibile)

(2) D triangolare superiore

(3) $SD = A$

Si osservi che se S, D è una fattorizzazione LR di A, quest'ultima risulta invertibile se e solo se lo è il fattore D.

(II) fattorizzazione QR: $U, T \in \mathbb{C}^{n\times n}$ tali che

(1) U unitaria (certamente invertibile)

(2) T triangolare superiore

(3) $UT = A$

Si osservi che se U, T è una fattorizzazione QR di A, quest'ultima risulta invertibile se e solo se lo è il fattore T.

... poi (uso della fattorizzazione $A = MN$), riscritto il sistema nella forma equivalente:

$$MNz = b$$

si opera il *cambio di variabile* $Nz = c$ che trasforma il sistema in:

$$Mc = b$$

Risolvendo quest'ultimo sistema (caso semplice) si ricava c. Infine, risolvendo il sistema (caso semplice):

$$Nz = c$$

si ricava z^*.

3.2 Problema

Siano:

$$A = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 4 & 0 \\ 1 & 6 & 0 \end{bmatrix} \quad , \quad M = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix} \quad , \quad N = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 2 & -1 \\ 0 & 0 & 1 \end{bmatrix}$$

Decidere se M, N sia una fattorizzazione LR oppure QR – o nessuna delle due – di A, ed utilizzarla per risolvere il sistema:

$$Az = \begin{bmatrix} 4 \\ 5 \\ 7 \end{bmatrix}$$

Assegnata $A \in \mathbb{C}^{n \times n}$, i problemi che affronteremo adesso sono: *come determinare una fattorizzazione LR di A* e *come determinare una fattorizzazione QR di A.*

A-1 Fattorizzazione LR – Metodo di Gauss (aritmetica esatta)

La seguente procedura *eg* (ELIMINAZIONE DI GAUSS), che deriva dal *metodo di eliminazione di Gauss*, fornisce, se termina, una fattorizzazione LR di una assegnata matrice $A \in \mathbb{C}^{n \times n}$:

$(S, D) = eg(A)$

`dati:` $A \in \mathbb{C}^{n \times n}$

$A^{(1)} = A$;
`per` $k = 1, \ldots, n-1$ `ripeti`
    `>` `se` $a_{kk}^{(k)} \neq 0$ `allora`

> $s_{ik} = a_{ik}^{(k)}/a_{kk}^{(k)}, \quad i = k+1, \ldots, n;$
> $l_k = (0, \ldots, 0, s_{k+1,k}, \ldots, s_{nk})^{\mathsf{T}};$
> $H_k = I - l_k \hat{e}_k;$
> $A^{(k+1)} = H_k A^{(k)}$

`altrimenti` STOP

`uscita:` $S = H_1^{-1} \cdots H_{n-1}^{-1}; D = A^{(n)}$

Gli elementi s_{ik} si chiamano *moltiplicatori,* gli elementi $a_{kk}^{(k)}$ si chiamano *pivot.*

3.3 Osservazione

Sia $k \in \{1, \ldots, n-1\}$ tale che $a_{kk}^{(k)} \neq 0$. Allora:

(1) H_k è triangolare inferiore con 1 sulla diagonale;

(2) H_k è invertibile e $H_k^{-1} = I + l_k \hat{e}_k$;

(3) gli elementi $a_{k+1,k}^{(k+1)}, \ldots, a_{n,k}^{(k+1)}$ sono nulli e $\hat{a}_1^{(k+1)} = \hat{a}_1^{(k)}, \ldots, \hat{a}_k^{(k+1)} = \hat{a}_k^{(k)}$.

3.4 Osservazione

Il procedimento termina se e solo se $a_{kk}^{(k)} \neq 0$ per $k = 1, \ldots, n-1$. In tal caso, posto $S_k = H_k^{-1}$ e quindi $S = S_1 \cdots S_{n-1}$, si ha:

(1) $S = I + l_1 \hat{e}_1 + \cdots + l_{n-1} \hat{e}_{n-1}$, e gli elementi di S al di sotto della diagonale sono i moltiplicatori s_{ik};

(2) $D = H_{n-1} \cdots H_1 A$, con $d_{kk} = a_{kk}^{(k)}$.

(3) $A = S_1 \cdots S_{n-1} D = SD$.

3.5 Osservazione

Il prossimo esempio mostra che la procedura *eg* può non terminare (il dominio della funzione *eg non è* l'intero insieme $\mathbb{C}^{n \times n}$).

3.6 Esempio

Sia:

$$A = \begin{bmatrix} 1 & 1 & i \\ 1 & 1 & 0 \\ 2 & 0 & 2i \end{bmatrix}$$

Si ha:

$A^{(1)} = \ldots$;
$k = 1$:

> pivot: $a_{11}^{(1)} = 1 \neq 0 \Rightarrow s_{21} = 1, s_{31} = 2$;

> $l_1 = (0, 1, 2)^{\mathsf{T}}$;

> $H_1 = I - l_1 \hat{e}_1 = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix}$;

> $A^{(2)} = \begin{bmatrix} 1 & 1 & i \\ 0 & 0 & -i \\ 0 & -2 & 0 \end{bmatrix}$

$k = 2$:

> pivot: $a_{22}^{(2)} = 0 \Rightarrow$ STOP

Pur essendo A invertibile, la procedura eg non termina.

Il teorema seguente caratterizza tutte le matrici per le quali l'eliminazione termina (l'insieme di definizione della funzione eg).

Sia $A \in \mathbb{C}^{n \times n}$. Indichiamo con $A[k]$ il minore principale di A ottenuto dalle prime k righe e dalle prime k colonne di A.

3.7 Teorema (di terminazione: insieme di definizione di eg)

Sia $A \in \mathbb{C}^{n \times n}$. La procedura eg termina su A (ovvero $a_{kk}^{(k)} \neq 0$ per $k = 1, \ldots, n-1$) se e solo se $\det A[k] \neq 0$ per $k = 1, \ldots, n-1$.

Dimostrazione

Si ha $\det A[1] = a_{11}^{(1)}$.

Se $\det A[1] = 0$ la procedura si arresta; se $\det A[1] \neq 0$ il passo della procedura per $k = 1$ è possibile e porta a: $A = S_1 A^{(2)}$. Allora, essendo S_1 triangolare inferiore: $A[2] = S_1[2] A^{(2)}[2]$ (aiutarsi con un disegno). Quindi $\det A[2] = a_{11}^{(1)} a_{22}^{(2)}$.

Se $\det A[2] = 0$ allora $a_{22}^{(2)} = 0$ e la procedura si arresta; se $\det A[2] \neq 0$ allora $a_{22}^{(2)} \neq 0$ e il passo della procedura con $k = 2$ è possibile, etc.

In generale, se il passo $k-1$ è possibile, si ha

$$\det A[k] = a_{11}^{(1)} \cdots a_{kk}^{(k)} \tag{3.1}$$

Questo prova l'asserto. □

3.8 Osservazione

Sia $A \in \mathbb{C}^{n \times n}$.

(1) Siano S ed D una fattorizzazione LR di A. Se $d_{kk} \neq 0$ per $k = 1, \ldots, n-1$, allora S e D sono l'*unica* fattorizzazione LR di A.

In particolare, se eg termina su A, le matrici S e D ottenute costituiscono l'unica fattorizzazione LR di A.

(*Dim.* Siano $S^{(1)}$ e $D^{(1)}$ una diversa fattorizzazione LR di A. Allora, procedendo come suggerito dal metodo di Doolittle (vedere GGM pag. 60) si ottiene: (1) $\hat{d}_1^{(1)} = \hat{d}_1$; (2) siccome $d_{11}^{(1)} = d_{11} \neq 0$, da $d_{11}^{(1)} s_1^{(1)} = d_{11} s_1$ segue $s_1^{(1)} = s_1$; (3) siccome $s_{21}^{(1)} = s_{21}$ e $\hat{d}_1^{(1)} = \hat{d}_1$, da $s_{21}^{(1)} \hat{d}_1^{(1)} + \hat{d}_2^{(1)} = s_{21}\hat{d}_1 + \hat{d}_2$ segue $\hat{d}_2^{(1)} = \hat{d}_2$; etc. Dunque: $S^{(1)} = S, D^{(1)} = D$.)

(2) Se eg non termina su A ed A è invertibile, allora A non è fattorizzabile LR (esempio: $A = (e_2, e_1)$).

(*Dim.* Se S, D sono una fattorizzazione LR di A, allora per $k = 1, \ldots, n-1$ si ha $\det A[k] = d_{11} \cdots d_{kk}$ e quindi, essendo A — e dunque D — invertibile, $\det A[k] \neq 0$; in tal caso, per il teorema precedente, eg termina: assurdo.)

(3) Se eg non termina su A ed A non è invertibile, allora o A non è fattorizzabile LR o ammette infinite fattorizzazioni LR (ad esempio: $A = (e_2, e_2)$ non ammette fattorizzazione LR; $A = 0$ ne ammette infinite).

(*Dim.* Si dimostra che la fattorizzazione non è unica. Siano S, D una fattorizzazione LR di A, $s_1, \ldots, s_n$ le colonne di S, $\hat{d}_1, \ldots, \hat{d}_n$ le righe di D. Sia $k \in \{1, \ldots, n-1\}$ un indice tale che $d_{kk} = 0$ — un tale indice esiste perché il procedimento di eliminazione non termina su A. Per ogni $\theta \in \mathbb{R}$, le matrici $S + \theta s_{k+1}\hat{e}_k$ e $D - \theta e_{k+1}\hat{d}_k$ costituiscono fattorizzazioni LR di A. Infatti: (a) $S + \theta s_{k+1}\hat{e}_k$ è triangolare inferiore con 1 sulla diagonale; (b) $D - \theta e_{k+1}\hat{d}_k$ è triangolare superiore; (c) $(S + \theta s_{k+1}\hat{e}_k)(D - \theta e_{k+1}\hat{d}_k) = SD + \theta s_{k+1}\hat{e}_k D - \theta S e_{k+1}\hat{d}_k - \theta^2 s_{k+1}\hat{e}_k e_{k+1}\hat{d}_k = SD + \theta s_{k+1}\hat{d}_k - \theta s_{k+1}\hat{d}_k = A$.)

Graficamente la situazione è:

$\exists!$

$\exists!$

Una classe di matrici per le quali *eg* termina è quella delle matrici a predominanza diagonale forte (per righe o per colonne — vedere Esempio 0.67).

(*Dim.* Sia A a predominanza diagonale forte. Siccome, per $k = 1, \ldots, n$, $A[k]$ ha predominanza diagonale forte, per il Teorema di Gershgorin si ha $\det A[k] \neq 0$.)

Un'altra classe di matrici per le quali *eg* termina è quella delle hermitiane definite positive.

3.9 Teorema

Sia $A \in \mathbb{C}^{n\times n}$ hermitiana. A è definita positiva se e solo se $\det A[k] > 0$ per $k = 1, \ldots, n$.

Dimostrazione

($\Rightarrow$) Segue dai due asserti:
(1) Se $B \in \mathbb{C}^{n\times n}$ è hermitiana definita positiva, allora anche $B[k]$ lo è, per $k = 1, \ldots, n$. (Sia $w \in \mathbb{C}^k$ non nullo. Posto $v = (w_1, \ldots, w_k, 0, \ldots, 0)^\mathsf{T} \in \mathbb{C}^n$ si ha: $B[k]w \bullet w = w^H B[k]w = v^H Bv = Bv \bullet v > 0$.)
(2) Se $B \in \mathbb{C}^{n\times n}$ è hermitiana definita positiva, allora $\det B > 0$ ($\det B = \lambda_1 \cdots \lambda_n$ — vedere Osservazione 0.74, punto (3), nell'Appendice relativa al Capitolo 0).

($\Leftarrow$) Siccome $\det A[k] \neq 0$ per $k = 1, \ldots, n-1$, allora A ammette un'unica fattorizzazione LR: S, D. Posto $\Delta = \mathrm{diag}(d_{11}, \ldots, d_{nn})$ e $S_1 = \Delta^{-1}S$ (triangolare superiore con 1 sulla diagonale) si ha $A = S\Delta D_1$ e quindi $A^H = D_1^H \Delta S^H$. Siccome A è hermitiana, si ha: $D_1^H \Delta S^H = A$. Ma D_1^H e ΔS^H sono una fattorizzazione LR di A. Per l'unicità della fattorizzazione si ha $S^H = D_1$. Dalla (3.1) e dall'Osservazione 3.4 punto (2), si ha che $d_{11} > 0, \ldots, d_{nn} > 0$. Allora, per ogni $z \neq 0$, si ha: $Az \bullet z = S\Delta S^H z \bullet z = \Delta(S^H z) \bullet (S^H z) > 0$ perché S invertibile. □

3.10 Osservazione

Sia $A \in \mathbb{C}^{n\times n}$ hermitiana. A è definita positiva se e solo se la procedura *eg* termina e $d_{kk} > 0$ per $k = 1, \ldots, n$.

(*Dim.* Segue dal teorema precedente, dalla (3.1) e dal punto (2) dell'Osservazione 3.4.)

Come rilevato nell'Osservazione 3.5, la procedura *eg* *può* non terminare anche se A è invertibile. Questa eventualità rende la procedura non soddisfacente per la soluzione del problema iniziale.

La procedura *egp* (ELIMINAZIONE DI GAUSS CON PIVOTING) descritta di seguito, variante della procedura *eg*, ha la proprietà di terminare se applicata ad una matrice invertibile.

3.11 Esempio

Si consideri la matrice di permutazione $P_1 = (e_2, e_4, e_1, e_3) \in \mathbb{C}^{4\times 4}$.

Dette $a_1, \ldots, a_4$ le colonne di A, le colonne di AP_1 sono, nell'ordine, a_2, a_4, a_1, a_3; dette $\hat{a}_1, \ldots, \hat{a}_4$ le righe di A, le righe di $P_1^{\mathsf{T}} A$ sono, nell'ordine, $\hat{a}_2, \hat{a}_4, \hat{a}_1, \hat{a}_3$.

3.12 Problema

Determinare la matrice $P \in \mathbb{C}^{4\times 4}$ tale che, dette $\hat{a}_1, \ldots, \hat{a}_4$ le righe di $A \in \mathbb{C}^{4\times 4}$, le righe di PA risultano, nell'ordine, $\hat{a}_4, \hat{a}_2, \hat{a}_1, \hat{a}_3$. △

$(P, S, D) = egp(A)$

`dati:` A;

$A^{(1)} = A$;
`per` $k = 1, \ldots, n-1$ `ripeti`
> `se` $a_{kk}^{(k)} = 0$ `allora`
`se` $\exists\, i > k$ tale che $a_{ik}^{(k)} \neq 0$ `allora`
P_k di permutazione, che scambia la riga k con la i
`altrimenti` STOP
`altrimenti` $P_k = I$
> $T^{(k)} = P_k A^{(k)}$; (le ultime due operazioni costituiscono il "pivoting")
... prosegui come nell'eliminazione senza pivoting,
operando su $T^{(k)}$ ...
> $A^{(k+1)} = H_k T^{(k)}$

`uscita:` $P = P_{n-1} \cdots P_1; D = A^{(n)}; S = P(H_{n-1}P_{n-1} \cdots H_1 P_1)^{-1}$

3.13 Osservazione

Se il procedimento termina si ha:

(1) $D = H_{n-1}P_{n-1} \cdots H_1 P_1 A$ con $d_{kk} = a_{kk}^{(k)}$;

(2) S è triangolare inferiore con 1 sulla diagonale;*

(3) le matrici S ed D costituiscono una fattorizzazione LR di PA.

3.14 Osservazione

Anche al procedura *egp* può non terminare.

*Infatti, per $j > k$ si ha: $H_k^{-1} P_j^{\mathsf{T}} = (I + l_k \hat{e}_k) P_j^{\mathsf{T}} = P_j^{\mathsf{T}} + l_k(\hat{e}_k P_j^{\mathsf{T}}) = P_j^{\mathsf{T}} + l_k \hat{e}_k = P_j^{\mathsf{T}}(I + (P_j l_k)\hat{e}_k) = P_j^{\mathsf{T}}(I + l_k' \hat{e}_k) = P_j^{\mathsf{T}} S_k'$; quindi $P_1^{\mathsf{T}} H_1^{-1} \cdots P_{n-1}^{\mathsf{T}} H_{n-1}^{-1} = P_1^{\mathsf{T}} \cdots P_{n-1}^{\mathsf{T}} S_1^{(n-2)} \cdots S_{n-2}^{(1)} H_{n-1}^{-1}$. Infine: $S = I + l_1^{(n-2)} \hat{e}_1 + \cdots + l_{n-2}^{(1)} \hat{e}_{n-2} + l_{n-1} \hat{e}_{n-1} = I + P_{n-1} \cdots P_2 l_1 \hat{e}_1 + \cdots + P_{n-1} l_{n-2} \hat{e}_{n-2} + l_{n-1} \hat{e}_{n-1}$ che risulta triangolare inferiore con 1 sulla diagonale.

3.15 Esempio

Sia

$$A = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

Applicando la procedura con pivoting si ha:

$A^{(1)} = \dots$

$k = 1$:

> $a_{11}^{(1)} = 0, a_{21}^{(1)} \neq 0 \Rightarrow P_1 = (\hat{e}_2^{\mathsf{T}}, \hat{e}_1^{\mathsf{T}}, \hat{e}_3^{\mathsf{T}})^{\mathsf{T}}$;

> $T^{(1)} = P_1 A^{(1)}$;

> pivot: $t_{11}^{(1)} = 1 \neq 0 \Rightarrow l_{21} = 0, l_{31} = 1$;

> $A^{(2)} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$

$k = 2$:

> $a_{22}^{(2)} = 0, \forall\ i > 2, a_{i2}^{(2)} = 0 \Rightarrow$ STOP

La procedura *egp* non termina, ma A non è invertibile.

Il teorema seguente caratterizza tutte le matrici per le quali l'eliminazione con pivoting termina (l'insieme di definizione della procedura *egp*).

3.16 Teorema (di terminazione della procedura *egp*)

Sia $A = (a_1, \dots, a_n) \in \mathbb{C}^{n \times n}$. La procedura *egp* termina su A se e solo se le colonne $a_1, \dots, a_{n-1}$ sono linearmente indipendenti.

Dimostrazione

Si ricordi che: se $M \in \mathbb{C}^{n \times n}$ è invertibile e $v_1, \dots, v_m \in \mathbb{C}^n$ sono linearmente indipendenti, allora anche $Mv_1, \dots, Mv_m$ sono linearmente indipendenti (vedere [L], pag. 90).

($\Rightarrow$) Se l'eliminazione con pivoting termina, si hanno P, S ed D tali che $PA = SD$ e $d_1, \dots, d_{n-1}$ linearmente indipendenti. Allora, essendo S invertibile, $Pa_1, \dots, Pa_{n-1}$ sono linearmente indipendenti. Anche P è invertibile, quindi $a_1 = P^{\mathsf{T}}(Pa_1), \dots, a_{n-1} = P^{\mathsf{T}}(Pa_{n-1})$ sono linearmente indipendenti.

($\Leftarrow$) Se al passo $k \in \{1, \dots, n-1\}$ la procedura *egp* si arresta, si ha: $A^{(k)} = M_{k-1}A$ con $a_{kk}^{(k)} = 0$, $a_1^{(k)} = M_{k-1}a_1, \dots, a_k^{(k)} = M_{k-1}a_k$ linearmente dipendenti e M_{k-1} invertibile. Allora $a_1, \dots, a_k$ sono linearmente dipendenti. □

3.17 Osservazione

Un procedimento per determinare la soluzione del sistema $Az = b$ è il seguente:

(1) $(P, S, D) = egp(A)$

(2) $c = sa(S, Pb)$

(3) $z^* = si(D, c)$

I sistemi da risolvere nei passi (2) e (3) sono *semplici*.

Il procedimento è soddisfacente nel senso che se A è invertibile trova la soluzione, altrimenti si arresta.

A-2 Fattorizzazione QR (aritmetica esatta)

Un metodo per calcolare *una* fattorizzazione QR di una matrice A si ottiene applicando il procedimento di ortonormalizzazione di Gram–Schmidt alle colonne di A, come suggerito dal seguente esempio.

3.18 Esempio

Sia $A = (a_1, a_2, a_3) \in \mathbb{C}^{3\times 3}$, e supponiamo (per semplicità) linearmente indipendenti le colonne di A. Il procedimento di ortonormalizzazione di Gram–Schmidt fornisce

$$\omega_1 = a_1 \quad , \quad \omega_2 = a_2 - d_{21}\omega_1 \quad , \quad \omega_3 = a_3 - d_{31}\omega_1 - d_{32}\omega_2$$

con

$$d_{21} = \frac{a_2 \bullet \omega_1}{||\omega_1||^2} \quad , \quad d_{31} = \frac{a_3 \bullet \omega_1}{||\omega_1||^2} \quad , \quad d_{32} = \frac{a_3 \bullet \omega_2}{||\omega_2||^2}$$

ovvero

$$a_1 = \omega_1 \quad , \quad a_2 = \omega_2 + d_{21}\omega_1 \quad , \quad a_3 = \omega_3 + d_{31}\omega_1 + d_{32}\omega_2$$

che si riscrivono

$$(a_1, a_2, a_3) = (\omega_1, \omega_2, \omega_3) \begin{bmatrix} 1 & d_{21} & d_{31} \\ 0 & 1 & d_{32} \\ 0 & 0 & 1 \end{bmatrix}$$

La fattorizzazione di A così ottenuta non è quella richiesta perché le colonne $\omega_1, \omega_2, \omega_3$ sono ortogonali ma non di norma unitaria. Per ottenere quanto si vuole si pone $N = \mathrm{diag}(||\omega_1||, ||\omega_2||, ||\omega_3||)$ e

$$U = (\omega_1, \omega_2, \omega_3)N^{-1} \quad , \quad T = N \begin{bmatrix} 1 & d_{21} & d_{31} \\ 0 & 1 & d_{32} \\ 0 & 0 & 1 \end{bmatrix}$$

Sussiste il seguente teorema (del quale l'Esempio precedente suggerisce una dimostrazione nel caso di matrice invertibile):

3.19 Teorema (fattorizzazione QR, esistenza)
Sia $A \in \mathbb{C}^{n\times n}$. Esiste una fattorizzazione QR di A.

Siano $A \in \mathbb{C}^{n\times n}$ e $b \in \mathbb{C}^n$. Detta qr una procedura che determina una fattorizzazione QR di di A, il procedimento seguente può essere utilizzato per risolvere il sistema $Az = b$:

(1) $(U, T) = qr(A)$

(2) $c = U^{\mathsf{T}} b$

(3) $z^* = si(T, c)$

Anche questo procedimento è soddisfacente.

A-3 Condizionamento

3.20 Esempio
Si vogliono determinare le tensioni di nodo nel circuito di Figura 19.

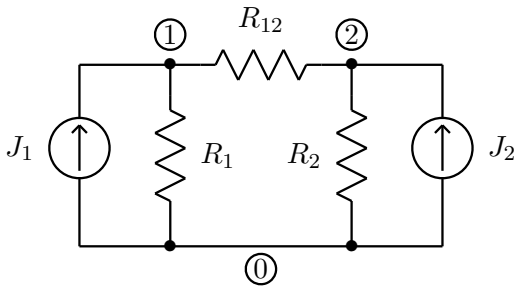

Figura 19.

Posto $G_1 = 1/R_1, \ldots$ le equazioni da risolvere (legge di Kirchhoff delle correnti) sono:

$$\begin{bmatrix} G_1 + G_{12} & -G_{12} \\ -G_{12} & G_2 + G_{12} \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} J_1 \\ J_2 \end{bmatrix} \tag{3.2}$$

con V_k tensione del nodo k rispetto al nodo 0 (di riferimento).

Siano $G_1 = G_2 = 10^{-2}$ S, $G_{12} = 10^2$ S; $J_1 = -1$ A, $J_2 = 1$ A. Riscrivendo $GV = J$ il sistema (3.2), la matrice G risulta a predominanza diagonale forte, dunque invertibile, e si ha:

$$V = \frac{1}{200.01} \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Se il vettore delle correnti J viene "*perturbato*" in $J + \delta J$ con

$$\delta J = \epsilon \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad (\epsilon > 0 \text{ assegnato})$$

la soluzione diviene $V + \delta V$ con

$$\delta V = 100\epsilon \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Una perturbazione di "misura" ϵ (ad esempio: $||\delta J||_\infty = \epsilon$) genera una variazione della soluzione di "misura" 100ϵ (ad esempio $||\delta V||_\infty = 100\epsilon$).

Siano N_n una norma in $\mathbb{R}^n$, N_m una norma in $\mathbf{R}^m$, $\Omega \subset \mathbb{R}^n, x, y \in \Omega$ e $f : \Omega \to \mathbb{R}^m$.

3.21 Definizione (errore trasmesso dai dati)
Si chiama *errore trasmesso dai dati* (nel calcolo del valore di f in x):

(1) $\delta_d = N_m(f(y) - f(x))$ (errore *assoluto*)

(2) $\epsilon_d = \dfrac{N_m(f(y) - f(x))}{N_m(f(x))}$ (errore *relativo*, $f(x) \neq 0$)

3.22 Definizione (errore sui dati)
Si chiama *errore sui dati*:

(1) $\delta_* = N_n(y - x)$ (errore *assoluto*)

(2) $\epsilon_* = N_n(y - x)/N_n(x)$ (errore *relativo*, $x \neq 0$)

3.23 Osservazione
Siano $A \in \mathbb{C}^{n\times n}$ invertibile, $b \in \mathbb{C}^n$. Per valutare il condizionamento del problema del calcolo della soluzione del sistema lineare $Az = b$, occorre considerare l'errore trasmesso dai dati per la funzione

$$f : \mathbb{C}^{n\times n} \times \mathbb{C}^n \to \mathbb{C}^n \quad \text{definita da} \quad f : A, b \to A^{-1}b$$

cioè (supponendo $A + \delta A$ invertibile, $b \neq 0$):

$$\epsilon_d = \frac{||f(A + \delta A, b + \delta b) - f(A, b)||}{||f(A, b)||}$$

in cui $\delta A, \delta b$ sono la perturbazione sui dati del problema.

Vediamo due casi particolari (per il caso generale, vedere GLV pag. 204).

Primo caso: $\delta A = 0$.

Si ha:

$$\epsilon_d = \frac{||A^{-1}(b+\delta b) - A^{-1}b||}{||A^{-1}b||} = \frac{||A^{-1}\delta b||}{||A^{-1}b||}$$

e, definito $\epsilon_* = \dfrac{||\delta b||}{||b||}$:

$$\epsilon_d = \frac{||b||}{||A^{-1}b||}\,\frac{||A^{-1}\delta b||}{||\delta b||}\,\epsilon_*$$

Dalle proprietà della norma di un operatore lineare esposte nell'Osservazione 0.22 segue che

$$\max_{\delta b \neq 0} \frac{||A^{-1}\delta b||}{||\delta b||} = ||A^{-1}|| \quad \text{e} \quad \max_{b \neq 0} \frac{||b||}{||A^{-1}b||} = ||A||$$

Si ha allora

$$\max\left\{ \frac{\epsilon_d}{\epsilon_*}, b \neq 0, \delta b \neq 0 \right\} = ||A||\,||A^{-1}|| \tag{3.3}$$

Definito $\mu(A) = ||A||\,||A^{-1}||$ *numero di condizionamento di* A si ottiene:

$$\epsilon_d \leq \mu(A)\frac{||\delta b||}{||b||}$$

Si osservi che la stima è *ottima* nel senso che vale la (3.3).

Secondo caso: $\delta b = 0$.

Posto $f(A+\delta A, b) = \hat{z}$ (cioè $\hat{z}$ tale che $(A+\delta A)\hat{z} = b$) e $f(A,b) = z^*$ si ha:

$$A(\hat{z} - z^*) = b - \delta A\,\hat{z} - b \Rightarrow \hat{z} - z^* = -A^{-1}\delta A\,\hat{z}$$

Allora:

$$\frac{||\hat{z} - z^*||}{||\hat{z}||} = \frac{||A^{-1}\delta A\,\hat{z}||}{||\hat{z}||}$$

e, definito $\epsilon_* = \dfrac{||\delta A||}{||A||}$:

$$\frac{||\hat{z} - z^*||}{||\hat{z}||} = \frac{||A^{-1}\delta A\,\hat{z}||}{||\hat{z}||}\,\frac{||A||}{||\delta A||}\,\epsilon_*$$

Definito $\Delta = \{\, \delta A \in \mathbb{C}^n \mid A + \delta A \text{ non singolare e } \delta A \neq 0 \,\}$, e tenuto conto che $b = (A+\delta A)\hat{z}$ con $A + \delta A$ non singolare, si ha

$$\begin{aligned}
&\max\left\{ \frac{1}{\epsilon_*}\frac{||\hat{z} - z^*||}{||\hat{z}||}, \delta A \in \Delta, b \neq 0 \right\} = \\
&= \max\left\{ \frac{||A^{-1}\delta A\,\hat{z}||}{||\hat{z}||}\frac{||A||}{||\delta A||}, \delta A \in \Delta, \hat{z} \neq 0 \right\} = \\
&= \max\left\{ \max\left\{ \frac{||A^{-1}\delta A\,\hat{z}||}{||\hat{z}||}\,\frac{||A||}{||\delta A||}, \hat{z} \neq 0 \right\}, \delta A \in \Delta \right\}
\end{aligned}$$

Inoltre:

$$\max\left\{ \frac{||A^{-1}\delta A\hat{z}||}{||\hat{z}||}\,\frac{||A||}{||\delta A||}, \hat{z}\neq 0 \right\} = ||A^{-1}\delta A||\,\frac{||A||}{||\delta A||}$$

e

$$\max\left\{ ||A^{-1}\delta A||\,\frac{||A||}{||\delta A||}, \delta A\in\Delta \right\} = ||A^{-1}||\,||A||$$

— infatti, per l'Osservazione 0.25, si ha

$$||A^{-1}\delta A||\,\frac{||A||}{||\delta A||} \leq ||A^{-1}||\,||A||$$

e, per α sufficientemente piccolo è

$$\alpha\, I\in\Delta \quad \text{e} \quad ||A^{-1}\alpha\, I||\,\frac{||A||}{||\alpha\, I||} = ||A^{-1}||\,||A||$$

Si ha dunque:

$$\max\left\{ \frac{1}{\epsilon_*}\,\frac{||\hat{z}-z^*||}{||\hat{z}||}, \delta A\in\Delta, b\neq 0 \right\} = ||A||\,||A^{-1}|| \tag{3.4}$$

e si ottiene la stima (*ottima* nel senso che vale (3.4))

$$\frac{||\hat{z}-z^*||}{||\hat{z}||} \leq \mu(A)\,\frac{||\delta A||}{||A||}$$

3.24 Esempio (continua)

Nel caso dell'esempio precedente si ha:

$$||G||_\infty = 200.01, \quad ||G^{-1}||_\infty = 100 \quad \Rightarrow \quad \mu_\infty(G) = 20{\cdot}001;$$
$$||J||_\infty = 1, \quad ||\delta J||_\infty = \epsilon; \quad ||V||_\infty = \frac{1}{200.01} \quad , \quad ||\delta V||_\infty = 100\epsilon$$

e allora:

$$\mu_\infty(G)\frac{||\delta J||_\infty}{||J||_\infty} = 20{\cdot}001\epsilon; \quad \frac{||\delta V||_\infty}{||V||_\infty} = 20{\cdot}001\epsilon$$

Siccome il numero di condizionamento è $\gg 1$, la perturbazione sui dati *può* dar luogo ad un errore trasmesso molto più grande. Per il dato e per la perturbazione in esame ($J=(-1,1)^\mathsf{T}, \delta J=\epsilon(1,1)^\mathsf{T}$) questo accade.

Se assumiamo

$$J = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad , \quad \delta J = \epsilon \begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

si ottiene:

$$V = 100 \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad , \quad \delta V = \epsilon \begin{bmatrix} 100 - \frac{1}{200.01} \\ 100 + \frac{1}{200.01} \end{bmatrix}$$

In questo caso, pur essendo ancora $\mu_\infty(G) \gg 1$, risulta

$$\frac{||\delta J||_\infty}{||J||_\infty} = 2\epsilon \quad , \quad \frac{||\delta V||_\infty}{||V||_\infty} \simeq \epsilon$$

Si può dimostrare che per ogni δJ con $||\delta J||_\infty = \gamma$ si ha $||\delta V||_\infty / ||V||_\infty \leq \gamma$, e quindi che per il dato considerato ($J = (1,1)^{\mathsf{T}}$) l'errore trasmesso è, per ogni perturbazione, dell'ordine dell'errore sui dati.

Allora: un numero di condizionamento elevato significa che *per qualche dato* esiste una perturbazione che genera un errore trasmesso molto più grande dell'errore sui dati; non significa, invece, che *per ogni* dato si possa trovare una perturbazione con tale proprietà.

3.25 Osservazione

Sia $A \in \mathbb{C}^{n\times n}$. Si ha:

(1) $\mu(A) \geq 1$;

(2) dal Teorema 0.27 segue che

$$\mu(A) \geq \rho(A)\rho(A^{-1})$$

(3) dall'Osservazione 0.22 si ottiene

$$\mu(A) = \frac{\max_{||v||=1} ||Av||}{\min_{||w||=1} ||Aw||}$$

3.26 Problema

Sia $A \in \mathbb{C}^{n\times n}$ hermitiana invertibile e $\sigma(A) = \{\lambda_1, \ldots, \lambda_n\}$. Dimostrare che

$$\mu_2(A) = \frac{\max\{\,|\lambda_1|, \ldots, |\lambda_n|\,\}}{\min\{\,|\lambda_1|, \ldots, |\lambda_n|\,\}} = \rho(A)\,\rho(A^{-1})$$

△

A-4 Propagazione dell'errore algoritmico (stabilità)

L'eliminazione di Gauss e la risoluzione all'indietro sono sequenze di operazioni aritmetiche (ed eventualmente di scambi di righe). Algoritmi per approssimare la soluzione di un sistema di equazioni lineari si possono ottenere *sostituendo* alle operazioni aritmetiche le corrispondenti pseudo–operazioni *e* specificandone l'ordine.

Per studiare la propagazione dell'errore algoritmico, dovremmo operare come nell'Esempio 1.33. Nel caso degli algoritmi per la soluzione dei sistemi lineari, tale studio non risulta semplice e, perciò, non sarà affrontato. Ci

limitiamo a giustificare, con un esempio, l'introduzione delle *tecniche di pivoting.*

3.27 Esempio

Siano

$$A = \begin{bmatrix} 10^{-3} & 1 \\ 1 & 2 \end{bmatrix} \quad , \quad b \in \mathbb{R}^2$$

Si ha

$$A^{-1} = \begin{bmatrix} -\frac{1000}{499} & \frac{500}{499} \\ \frac{500}{499} & -\frac{1}{998} \end{bmatrix}$$

ed il numero di condizionamento $\mu_\infty(A)$ vale $3\,\frac{1500}{499} \approx 9$.

Utilizzando il procedimento di ELIMINAZIONE si ottiene

$$R = \begin{bmatrix} 10^{-3} & 1 \\ 0 & -998 \end{bmatrix} \quad , \quad c = \dots$$

Si ha, inoltre: $\sigma(R) = \{10^{-3}, -998\}$ e quindi (punto (2) dell'Osservazione 3.25) $\mu_\infty(R) \geq 998\,10^3$.

Il procedimento di eliminazione di Gauss (senza pivoting: non è necessario alcuno scambio) ha generato un sistema *equivalente* a quello originario, *ma* la matrice del nuovo sistema ha un numero di condizionamento molto più elevato di quella iniziale.

Supponiamo adesso che per approssimare la soluzione del sistema si utilizzi un algoritmo che procede in due fasi: prima determina un'approssimazione di R e c, poi opera una risoluzione all'indietro del sistema trovato.

Ricordando il meccanismo di propagazione dell'errore algoritmico (Osservazione 1.34) si nota che l'elevato numero di condizionamento di R, può dar luogo ad una forte amplificazione dell'errore causato dall'algoritmo di eliminazione. Il procedimento di eliminazione di Gauss genera quindi algoritmi che *possono essere instabili.*

Per arginare il problema, si usa la tecnica del *pivoting parziale* (vedere [GGM] pagina 56.)

L'eliminazione con pivoting parziale produce

$$R' = \begin{bmatrix} 1 & 2 \\ 0 & \frac{998}{1000} \end{bmatrix}$$

con $\mu_\infty(R') = 3\,\frac{2998}{998} \approx 9$, senza un apprezzabile deterioramento del numero di condizionamento rispetto alla matrice iniziale e, quindi, senza pericolo di forte amplificazione dell'errore causato dall'algoritmo di eliminazione con pivoting.

A-5 Costo degli algoritmi

Il costo ("aritmetico") di un algoritmo è misurato dal numero di pseudo-operazioni elementari necessario per portare a termine il calcolo.

3.28 Esempio

Si consideri l'algoritmo $\phi_1 : M^n \times M^n \to M$, utilizzato per approssimare il prodotto scalare canonico di due vettori di $\mathbb{R}^n$, definito da[†]

$$\phi_1(a, b) = (a_1 \otimes b_1) \oplus \cdots \oplus (a_n \otimes b_n)$$

Il costo di ϕ_1 risulta

$$n\,\mathrm{P} + (n-1)\,\mathrm{S}$$

intendendo con questa scrittura che per calcolare $\phi_1(a, b)$ sono necessari n pseudo–prodotti (P) ed $n-1$ pseudo–somme (S).

Si consideri l'algoritmo $\phi_2 : M^{n\times n} \times M^n \to M^n$, utilizzato per approssimare il prodotto di una matrice per una colonna, definito, indicando con $\hat{a}_1, \ldots, \hat{a}_n$ le righe della matrice A, da

$$\phi_2(A, b) = \Big(\phi_1(\hat{a}_1, b), \ldots, \phi_1(\hat{a}_n, b)\Big)^{\mathsf{T}}$$

Il costo di ϕ_2 risulta n · costo di ϕ_1, cioè

$$n^2\,\mathrm{P} + n(n-1)\,\mathrm{S}$$

Se si applica l'algoritmo ϕ_2 ad una matrice triangolare superiore, tenuto conto che per ogni $\xi \in M$ si ha $\xi \otimes 0 = 0$ e $\xi \oplus 0 = \xi$, e che queste operazioni hanno "costo zero," il costo scende a

$$\sum_{i=1}^{n} i\,\mathrm{P} + \sum_{i=1}^{n-1} i\,\mathrm{S} = \frac{n(n+1)}{2}\,\mathrm{P} + \frac{n(n-1)}{2}\,\mathrm{S}$$

Si consideri l'algoritmo $\phi_3 : M^{n\times n} \times M^n \to M^n$, utilizzato per approssimare la soluzione di un sistema di equazioni lineari con matrice triangolare superiore, con il procedimento di risoluzione all'indietro. Per $T \in M^{n\times n}$ triangolare superiore e $b \in M^n$, gli elementi del vettore $\xi = \phi_3(T, b)$ sono definiti da

$$\xi_i = [b_i \ominus \underbrace{(t_{i,i+1} \otimes \xi_{i+1}) \ominus \cdots \ominus (t_{in} \otimes \xi_n)}_{n-i}] \oslash t_{ii}$$

per $i = n, n-1, \ldots, 1$. Il costo per il calcolo di ξ_i è: $1\,\mathrm{D} + (n-i)\,\mathrm{P} + (n-i)\,\mathrm{S}$, e il costo complessivo di ϕ_3 risulta

$$n\,\mathrm{D} + \sum_{i=1}^{n-1} i\,\mathrm{P} + \sum_{i=1}^{n-1} i\,\mathrm{S} = n\,\mathrm{D} + \frac{n(n-1)}{2}\,\mathrm{P} + \frac{n(n-1)}{2}\,\mathrm{S}$$

[†]Negli algoritmi di questo Esempio, dovremmo specificare meglio l'ordine delle pseudo–operazioni, ma per determinare il costo non è necessario farlo.

Sia $\phi_4 : M^{n\times n} \to M^{n\times n} \times M^{n\times n}$ un algoritmo che calcola una approssimazione della fattorizzazione LR di una matrice, utilizzando il procedimento di eliminazione di Gauss.

Il costo di ϕ_4 risulta

$$\frac{n^2-n}{2}\,\mathrm{D} + \frac{2n^3-3n^2+n}{6}\,(\mathrm{P}+\mathrm{S})$$

3.29 Osservazione

Un algoritmo per il calcolo dell'approssimazione della soluzione di un sistema di equazioni lineari si ottiene utilizzando gli algoritmi ϕ_3 e ϕ_4 come indicato nella Figura 20.

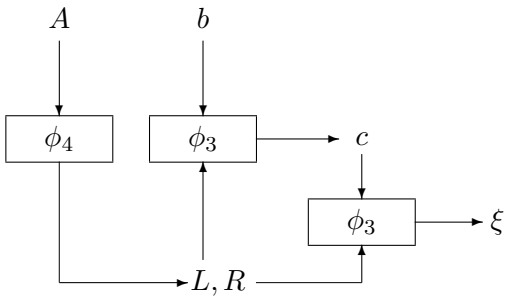

Figura 20. Algoritmo che calcola un'approssimazione della soluzione del sistema $Ax = b$.

Considerando soltanto i termini dominanti nell'espressione del costo dei singoli algoritmi, per il termine dominante nel costo dell'algoritmo complessivo si ottiene

$$\frac{n^2}{2}\,\mathrm{D} + \frac{n^3}{3}\,(\mathrm{P}+\mathrm{S})$$

Se si utilizza, invece, un algoritmo di soluzione basato sul metodo di Cramer, il termine dominante nel costo sale a: $n(n+1)!$

B-1 Metodi iterativi

I metodi di Gauss (con o senza pivoting) sono metodi "diretti" di risoluzione di un sistema di equazioni lineari. In un metodo *diretto*, con un numero finito di operazioni elementari (in aritmetica esatta) si determina *la soluzione* del problema.

Una classe alternativa di metodi per l'approssimazione della soluzione di un sistema di equazioni lineari è quella dei *metodi iterativi lineari*.

In un metodo iterativo, con un numero finito di operazioni elementari (in aritmetica esatta) si determina *l'elemento k-esimo di una successione* che converge alla soluzione del problema.

Idea del metodo:

determinare $H \in \mathbb{C}^{n\times n}, c \in \mathbb{C}^n$ e un vettore iniziale $z(0) \in \mathbb{C}^n$ tali che la successione definita da $z(k+1) = Hz(k) + c,\ k = 0, 1, \ldots$ risulti *convergente* a z^*.

Un metodo iterativo lineare è definito da una matrice H e da un vettore c. La matrice H si dice *matrice di iterazione* del metodo.

La successione generata dal metodo iterativo lineare definito da H e c *a partire da* $z(0) \in \mathbb{C}^n$ è la successione $z(0), z(1) = Hz(0) + c, z(2) = Hz(1) + c = H^2 z(0) + Hc + c, \ldots$.

3.30 Osservazione

Siano $H \in \mathbb{C}^{n\times n}, c \in \mathbb{C}^n$ e

$$\mathcal{Z}_\infty(H, c) = \{v \in \mathbb{C}^n \text{ tali che } v = Hv + c\}$$

l'insieme dei *punti uniti* della funzione (continua) $z \to Hz + c$. Questo insieme è interessante perché

> se la successione $z(k)$ generata dal metodo definito da H e c a partire da $z(0)$ è convergente, allora
>
> $$\lim_{k\to\infty} z(k) \in \mathcal{Z}_\infty(H, c)$$

Si osservi che $\mathcal{Z}_\infty(H, c)$ è l'insieme delle soluzioni del sistema $(I - H)\, v = c$. Allora:

(1) se $\mathcal{Z}_\infty(H, c)$ non è vuoto, si ha

$$\mathcal{Z}_\infty(H, c) = w + \ker(I - H)$$

per qualche $w \in \mathbb{C}^n$

(2) $\mathcal{Z}_\infty(H, c)$ ha un solo elemento se e solo se $\ker(I - H) = \{0\}$, ovvero se e solo se $1 \notin \sigma(H)$.

3.31 Definizione (metodo consistente)

Il metodo iterativo definito da H e c è *consistente* con il sistema $Az = b$ se

$$\mathcal{Z}_\infty(H, c) = \{z \in \mathbb{C}^n :\ Az = b\} \tag{3.5}$$

La definizione significa che se una successione generata dal metodo è convergente, il limite è una soluzione del sistema, e che ciascuna soluzione del sistema è limite di qualche successione convergente generata dal metodo.

3.32 Osservazione

Sia $H \in \mathbb{C}^{n\times n}$. Indichiamo con $C(H)$ il sottospazio di $\mathbb{C}^n$ dei vettori a partire dai quali la successione generata dal metodo definito da H e dal vettore nullo $0 \in \mathbb{C}^n$ risulta convergente.

Siano, inoltre, $c \in \mathbb{C}^n$ e H tale che $\mathcal{Z}_\infty(H, c)$ ha un solo elemento (vedere il punto (2) dell'Osservazione 3.30). Detto z_* l'unico elemento di $\mathcal{Z}_\infty(H, c)$, si ha

> l'insieme dei vettori di $\mathbb{C}^n$ a partire dai quali la successione generata dal metodo definito da H e c risulta convergente è $z_* + C(H)$.

Infatti: la successione $z(k)$ generata dal metodo definito da H e c a partire da $z(0)$ converge (a z_*) se e solo se la successione $e(k) = z(k) - z_*$ converge (a 0) e quindi, siccome $e(k)$ è la successione generata dal metodo definito da H e dal vettore nullo a partire da $z(0) - z_*$, se e solo se $z(0) - z_* \in C(H)$.

3.33 Definizione (metodo convergente)

Il *metodo* iterativo definito da H e c si dice *convergente* se

(1) $\mathcal{Z}_\infty(H, c)$ ha un solo elemento

(2) $C(H) = \mathbb{C}^n$

ossia se *per ogni* $z(0) \in \mathbb{C}^n$ la successione generata dal metodo a partire da $z(0)$ è convergente, ed il limite *non dipende* da $z(0)$.

Si osservi che, per quanto detto al punto (2) dell'Osservazione 3.30, la convergenza del metodo dipende *solo* da H.

Le condizioni (1) e (2) della Definizione 3.33 sono indipendenti.

3.34 Esempio

Si consideri il metodo iterativo definito da

$$H = \begin{bmatrix} 1/2 & 0 \\ 0 & 1 \end{bmatrix} \quad , \quad c = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Tutte le successioni generate dal metodo iterativo lineare definito da H e c sono convergenti (si ha infatti $C(H) = \mathbb{C}^n$), ma *il metodo* non è convergente: $\mathcal{Z}_\infty(H, c)$ non ha un solo elemento.

Se il metodo definito da H e c *non è* convergente, allora o $\mathcal{Z}_\infty(H, c)$ non ha un solo elemento (dunque è vuoto oppure ha infiniti elementi) o $\dim C(H) < n$.

Se si verifica la prima eventualità, il metodo è certamente non consistente con il sistema $Az = b$ (si ricordi che A è invertibile); se $\mathcal{Z}_\infty(H, c)$ ha un solo elemento ma $\dim C(H) < n$, si vedrà che risulta *praticamente* impossibile individuare un vettore $z(0) \in \mathbb{C}^n$ a partire dal quale la successione generata dal metodo risulti convergente (vedere Esempio 3.36).

3.35 Teorema (di convergenza)
Il metodo definito da H e c è convergente se e solo se $\rho(H) < 1$.

Dimostrazione
($\Leftarrow$, caso particolare) Sia $\rho(H) < 1$. Allora $I - H$ è invertibile (infatti $0 \notin \sigma(I - H)$) e quindi l'insieme $\mathcal{Z}_\infty(H, c)$ ha un solo elemento. Inoltre, in tal caso (si veda il punto (2) dell'Osservazione 3.30) si ha $\mathcal{Z}_\infty(H, 0) = \{\, 0 \,\}$.

Per dimostrare che $C(H) = \mathbb{C}^n$, occorre dimostrare che per ogni $e(0) \in \mathbb{C}^n$, posto $e(k) = H\, e(k-1)$, si ha $\lim_{k\to\infty} e(k) = 0$.

Supponiamo H diagonalizzabile. Siano $\lambda_1, \ldots, \lambda_n$ gli autovettori di H e $v_1 \in V(\lambda_1), \ldots, v_n \in V(\lambda_n)$ una base di $\mathbb{C}^n$. Posto $e(0) = \alpha_1\, v_1 + \cdots + \alpha_n\, v_n$ si ha:

$$e(k) = \alpha_1\, \lambda_1^k\, v_1 + \cdots + \alpha_n\, \lambda_n^k\, v_n \tag{3.6}$$

Per l'ipotesi, ciascuna delle successioni di (3.6) converge a 0 e quindi ...

($\Rightarrow$) Sia il metodo convergente. In tal caso (si veda il punto (2) dell'Osservazione 3.30) si ha $\mathcal{Z}_\infty(H, 0) = \{\, 0 \,\}$. Allora per ogni $e(0) \in \mathbb{C}^n$, si ha $\lim_{k\to\infty} e(k) = 0$.

Sia λ_j un autovalore di H e v_j un autovettore relativo a λ_j. Posto $e(0) = v_j$ si ha $e(k) = \lambda_j^k\, v_j$. L'asserto si ottiene considerando che $\lim_{k\to\infty} \lambda_j^k\, v_j = 0$ se e solo se $|\lambda_j| < 1$. □

Il Teorema precedente evidenzia che il vettore c non influenza la convergenza del metodo definito da H e c. Questo giustifica l'uso di chiamare *convergente* una matrice H qualora $\rho(H) < 1$.

3.36 Esempio
Si consideri il metodo iterativo definito da

$$H = \begin{bmatrix} 1/2 & 1 \\ 0 & -1 \end{bmatrix} \quad , \quad c = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

Determinare l'insieme dei vettori a partire dai quali la successione generata dal metodo è convergente.

Soluzione
Si ha $\sigma(H) = \{1/2, -1\}$ e quindi:

(a) $\rho(H) = 1$, dunque il metodo risulta *non convergente*;

(b) $I - H$ è invertibile, dunque $\mathcal{Z}_\infty(H, c)$ ha un solo elemento che risulta $z_* = (2, 1)^\mathsf{T}$;

(c) la matrice di iterazione H ha autovalori distinti, dunque esiste una base v_1, v_2 di $\mathbb{C}^2$ sostituita da autovettori di H; una delle possibili scelte è

$$v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \in V(\tfrac{1}{2}) \quad , \quad v_2 = \begin{bmatrix} 2 \\ -3 \end{bmatrix} \in V(-1)$$

Posto $e(0) = a_1\, v_1 + a_2\, v_2$ si ha

$$e(k) = a_1\, \frac{1}{2^k}\, v_1 + a_2\, (-1)^k\, v_2$$

e quindi $C(H) = V(\frac{1}{2})$.

Infine, l'insieme dei vettori a partire dai quali la successione generata dal metodo risulta convergente è $z_* + V(\frac{1}{2})$.

Si osservi che, come accade in generale, per determinare quest'ultimo insieme è necessario avere informazioni su H onerose da ottenere.

3.37 Osservazione

Perché il metodo definito da H e c sia utilizzabile per approssimare la soluzione del sistema $Az = b$, occorre dunque che:

(1) il metodo sia *consistente* con il sistema;

(2) la matrice H sia *convergente*.

La convergenza della matrice di iterazione, per un metodo consistente, garantisce che per ogni vettore iniziale $z(0)$ la successione degli errori $e(k)$ converge al vettore nullo, ossia che $\lim N(e(k)) = 0$ per qualsiasi norma N. Non è detto, invece, che la successione $N(e(k))$ risulti *monotona*.

3.38 Esempio

Si consideri la successione generata dal metodo definito da

$$H = \begin{bmatrix} -\frac{1}{2} & 1 \\ 0 & \frac{1}{2} \end{bmatrix} \quad , \quad c = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

a partire dal vettore $e(0) = (2, 1)^{\mathsf{T}}$.

Poiché $\sigma(H) = \{ -\frac{1}{2}, \frac{1}{2} \}$, il metodo risulta convergente e, poiché la matrice di iterazione ha autovalori distinti, esiste una base v_1, v_2 di $\mathbb{C}^2$ sostituita da autovettori; una delle possibili scelte è

$$v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \in V(-\tfrac{1}{2}) \quad , \quad v_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \in V(\tfrac{1}{2})$$

Essendo $e(0) = v_1 + v_2$ si ha:

$$e(k) = (-\tfrac{1}{2})^k\, v_1 + (\tfrac{1}{2})^k\, v_2$$

La successione $||e(k)||_2$ risulta convergente a 0 ma *non monotona*.

3.39 Problema

Sia N una norma in $\mathbb{C}^n$, $H \in \mathbb{C}^{n\times n}$ tale che $||H||_N < 1$ e $c \in \mathbb{C}^n$. Provare che il metodo definito da H e c è convergente. $\triangle$

3.40 Osservazione

Siano $H \in \mathbb{C}^{n\times n}$, $c \in \mathbb{C}^n$ e N una norma in $\mathbb{C}^n$ per cui risulti $||H||_N = q < 1$. Detti z_* l'unico elemento di $\mathcal{Z}_\infty(H,c)$, $z(k)$ la successione generata dal metodo definito da H e c a partire da $z(0)$ e posto $e(k) = z(k) - z_*$ si ha:

(1) poiché $N(e(k)) \le q\, N(e(k-1))$, la successione $N(e(k))$, se non nulla, risulta *monotona*

(2) vale la stima

$$N(e(k)) \le \frac{q^k}{1-q}\, N(z(1) - z(0))$$

infatti: $N(e(k)) \le q^k\, N(e(0))$ e inoltre $e(0) = z(0) - z(1) + e(1) = z(0) - z(1) + H\, e(0)$.

3.41 Osservazione

Siano N una norma in $\mathbb{C}^n$ e $H \in \mathbb{C}^{n\times n}$. Se per qualche intero positivo k si ha $||H^k||_N < 1$, allora $\rho(H) < 1$.

(*Dim.* $\sigma(H) = \{\lambda_1, \ldots, \lambda_n\} \Rightarrow \sigma(H^k) = \{\lambda_1^k, \ldots, \lambda_n^k\}$ e quindi $\rho(H^k) = (\rho(H))^k$. Ma $\rho(H^k) \le ||H^k||_N < 1$, allora ...)

3.42 Esempio

Sia:

$$H = \begin{bmatrix} 0 & 2 \\ 1/10 & 0 \end{bmatrix}$$

Si ha: $\sigma(H) = \{1/\sqrt{5}, -1/\sqrt{5}\}$, $||H||_\infty = 2$, $||H^2||_\infty = 1/5$.

B-2 Metodo di Jacobi

Si consideri il sistema $Az = b$ con $A \in \mathbb{C}^{n\times n}$ invertibile e tale che $a_{kk} \neq 0$ per $k = 1, \ldots, n$. Posto $D = \mathrm{diag}(a_{11}, \ldots, a_{nn})$, il *metodo di Jacobi* è il metodo iterativo lineare definito da $H_J = D^{-1}(D - A) = I - D^{-1}A$ e $c_J = D^{-1}b$.

3.43 Osservazione

Il metodo di Jacobi è consistente con il sistema $Az = b$. Infatti i sistemi $Az = b$ e $(I - H_J)z = c_J$ sono equivalenti.

3.44 Teorema

La matrice A è a predominanza diagonale forte *per righe* se e solo se $||H_J||_\infty < 1$.

In particolare, se A è a predominanza diagonale forte per righe il metodo di Jacobi è convergente.

Dimostrazione

Poiché $H_J = I - D^{-1}A$, il primo asserto è il punto (r) del Teorema 0.69. Il secondo asserto è conseguenza del Problema 3.39. □

3.45 Esempio

Sia

$$A = \begin{bmatrix} 2 & 0 & 1 \\ 1 & 2 & 0 \\ 0 & 2 & 4 \end{bmatrix} \in \mathbb{C}^{3\times 3}$$

Si determinino $H_J, \sigma(H_J), \rho(H_J)$ e si discuta la velocità di convergenza del metodo di Jacobi.

Soluzione

Si ha

$$H_J = \begin{bmatrix} 0 & 0 & -1/2 \\ -1/2 & 0 & 0 \\ 0 & -1/2 & 0 \end{bmatrix} \in \mathbb{C}^{3\times 3}$$

Il polinomio caratteristico è: $P(\lambda) = -\lambda^3 - 1/8$, perciò

$$\sigma(H_J) = \{-\frac{1}{2}, \frac{1}{4} + i\frac{\sqrt{3}}{4}, \frac{1}{4} - i\frac{\sqrt{3}}{4}\} \quad \text{e} \quad \rho(H_J) = \frac{1}{2}$$

Per la velocità di convergenza, essendo H_J diagonalizzabile (ha autovalori distinti) si ha:

$$\frac{||e^{(k)}||_\infty}{||e^{(0)}||_\infty} \leq \mu_\infty(S)\ (\rho(H_J))^k \tag{3.7}$$

Allora la quantità $||e^{(k)}||_\infty / ||e^{(0)}||_\infty$ tende a 0, per $k \to \infty$, almeno rapidamente quanto $(\rho(H_J))^k$. Quindi, più piccolo è $\rho(H_J)$ più rapidamente $||e^{(k)}||_\infty / ||e^{(0)}||_\infty$ tende a 0, *nel caso peggiore.*

3.46 Osservazione

Sia A a predominanza diagonale forte *per colonne.* Si ha:

(1) H_J è convergente;

(2) non è detto che $||H_J||_1 < 1$ oppure $||H_J||_2 < 1$ oppure $||H_J||_\infty < 1$.

Infatti: (1) Poiché A è a predominanza diagonale forte per colonne, A^T è a predominanza diagonale forte per righe e quindi, posto $\hat{H}_J = I - D^{-1}A^\mathsf{T}$, per il Teorema 3.44 si ha $\rho(\hat{H}_J) < 1$. Inoltre H_J e $\hat{H}_J{}^\mathsf{T}$ sono simili (si ha: $H_J D^{-1} = D^{-1}\hat{H}_J{}^\mathsf{T}$). Quindi anche $\rho(H_J) < 1$.

(2) Si consideri, ad esempio, la matrice

$$A = \begin{bmatrix} 5 & 2 \\ 4 & 3 \end{bmatrix} \in \mathbb{C}^{2\times 2}$$

3.47 Osservazione

Sia $B = A\,D^{-1}$. Se A è a predominanza diagonale forte per colonne, allora

(1) la matrice B è a predominanza diagonale forte per colonne;

(2) dette z_* la soluzione del sistema $Az = b$ e x_* la soluzione del sistema $Bx = b$, si ha: $z_* = D^{-1}x_*$;

(3) $\mathrm{diag}(b_{11}, \ldots, b_{nn}) = I$;

(4) la matrice di iterazione del metodo di Jacobi applicato al sistema $Bx = b$ è $I - B$ e, per la definizione di predominanza diagonale forte per colonne (vedere il punto (c) del Teorema 0.69), si ha $||I - B||_1 < 1$.

3.48 Esempio (costruzione alternativa)

Sia

$$A = \begin{bmatrix} 2 & 0 & 1 \\ 1 & 2 & 0 \\ 0 & 2 & 4 \end{bmatrix} \in \mathbb{C}^{3\times 3}$$

Posto $A = T - S$ con T triangolare superiore e S strettamente triangolare inferiore,[‡] si consideri un metodo iterativo lineare di matrice di iterazione $H = T^{-1}S$.

(1) Determinare $\mathcal{Z}_\infty(H, c)$; (2) decidere se il metodo è convergente ed eventualmente confrontare la rapidità di convergenza con quella del metodo di Jacobi.

Soluzione

(1) La funzione $z \to Hz + c$ ha come (unico) punto unito la soluzione del sistema $Az = b$ se $c = T^{-1}b$. Infatti: $z = T^{-1}Sz + c$ è equivalente a $T(I - T^{-1}S)z = Tc$ e quindi ...

(2) Si ha:

$$H = T^{-1}S = \left[\begin{array}{cc|c} 0 & 1/4 & 0 \\ -1/2 & 0 & 0 \\ \hline 0 & -1/2 & 0 \end{array}\right]$$

da cui: $\sigma(H) = \{0, i/2\sqrt{2}, -i/2\sqrt{2}\}$ e $\rho(H) = 1/2\sqrt{2}$. Il metodo risulta dunque convergente e, nel caso peggiore, tenuto conto della (3.7), converge più rapidamente del metodo di Jacobi.

[‡]Una matrice *strettamente* triangolare inferiore è una matrice triangolare inferiore con tutti gli elementi sulla diagonale principale uguali a 0.

3.49 Osservazione

In pratica l'elemento $z^{(k+1)}$ in metodi come quello dell'Esempio precedente, viene calcolato ad ogni passo risolvendo il sistema $Tz^{(k+1)} = Sz^{(k)}+b$, cioè senza calcolare T^{-1}.

3.50 Problema

Per $n \geq 2$ siano $u = (1, \ldots, 1)^{\mathsf{T}} \in \mathbb{R}^{n-1}$ e

$$A = \begin{bmatrix} nI & u \\ u^{\mathsf{T}} & 1 \end{bmatrix} \in \mathbb{R}^{n \times n}$$

(1) Calcolare $\det A$ (A risulta definita positiva per qualche valore di n ?)

(2) Calcolare la soluzione del sistema $Ax = b$ con

$$b = \begin{bmatrix} u \\ 2 \end{bmatrix} \in \mathbb{R}^n$$

(3) Calcolare H_J e decidere per quali valori di n il metodo di Jacobi è convergente.

Soluzione

(1) La matrice A ammette la fattorizzazione LR a blocchi:

$$A = \begin{bmatrix} I & 0 \\ \frac{1}{n}u^{\mathsf{T}} & 1 \end{bmatrix} \begin{bmatrix} nI & u \\ 0 & \frac{1}{n} \end{bmatrix}$$

quindi $\det A = n^{n-2}$ ed essendo $\det A[k] > 0$ per $k = 1, \ldots, n$ la matrice A risulta definita positiva per tutti i valori di n.

(2) Risolvendo i sistemi $Ly = b$ e $Rx = y$ si ottiene

$$y = \begin{bmatrix} u \\ 1 + \frac{1}{n} \end{bmatrix} \quad , \quad x = \begin{bmatrix} -u \\ 1 + n \end{bmatrix}$$

(3) Si ha

$$H_J = \begin{bmatrix} 0 & -\frac{1}{n}u \\ -u^{\mathsf{T}} & 0 \end{bmatrix}$$

Il polinomio caratteristico è

$$P_J(\lambda) = (-\lambda)^n + (-\lambda)^{n-2}\left(-\frac{n-1}{n}\right) = (-\lambda)^{n-2}\left(\lambda^2 - \frac{n-1}{n}\right)$$

e quindi

$$\sigma(H_J) = \Big\{ \underbrace{0, \ldots, 0}_{n-2}, \sqrt{\frac{n-1}{n}}, -\sqrt{\frac{n-1}{n}} \Big\}$$

ed il metodo risulta convergente per tutti i valori di n.

B-3 Costo

Il costo del calcolo dell'elemento $z^{(k)}$ della successione (a partire da $z^{(0)}$) è dato da

$$k \cdot (\text{costo di un passo})$$

Il costo di un passo è quello del procedimento che, dato $z^{(k)}$, consente di calcolare $z^{(k+1)}$. Normalmente quest'ultimo procedimento è la soluzione di un sistema lineare.

3.51 Esempio

Si consideri il metodo di Jacobi. Ad ogni passo si risolve il sistema

$$Dz^{(k+1)} = Mz^{(k)} + b$$

ed il termine dominante del costo è $n^2\,\mathrm{P} + n^2\,\mathrm{S}$.

Si consideri il metodo alternativo proposto nell'Esempio 3.48. Ad ogni passo si risolve il sistema

$$Tz^{(k+1)} = Sz^{(k)} + b$$

con T triangolare superiore e S strettamente triangolare inferiore. Tenendo conto della struttura delle matrici, il termine dominante del costo risulta ancora $n^2\,\mathrm{P} + n^2\,\mathrm{S}$.

In entrambi i casi esaminati, se il numero di passi necessario per ottenere l'approssimazione richiesta è $\leq \frac{n}{3}$, il costo del metodo iterativo risulta inferiore al costo del metodo di Gauss.

B-4 Criteri d'arresto

Sia $z^{(k)}$ la successione generata dal metodo iterativo definito da H e c a partire da $z^{(0)}$ (con aritmetica esatta), e sia $\xi^{(k)}$ la successione generata dall'algoritmo che realizza il metodo a partire da ξ_0 : $\xi^{(0)} = \xi_0$; $\xi^{(k+1)} = \Phi(\xi^{(k)}) = H\xi^{(k)} + c + \delta^{(k+1)}$.

Un criterio di arresto molto usato è:

$$||\xi^{(k)} - \xi^{(k-1)}|| < \epsilon$$

$\epsilon > 0$ dato. Si ha infatti, supponendo il metodo convergente e detto z_* il punto unito:

$$\xi^{(k)} - \xi^{(k-1)} = H\xi^{(k-1)} + c + \delta^{(k)} - \xi^{(k-1)} = (H - I)\xi^{(k-1)} + (I - H)z_* + \delta^{(k)}$$

da cui

$$||\xi^{(k-1)} - z_*|| \leq ||(I - H)^{-1}||\left(||\xi^{(k)} - \xi^{(k-1)}|| + ||\delta^{(k)}||\right)$$

Si osservi che, per la presenza del termine $\delta^{(k)}$, per $\epsilon \to 0$ non necessariamente si ha $\xi^{(k-1)} - z_* \to 0$.

3.52 Osservazione

Nel caso in cui sia $||H|| = q < 1$, poiché

$$I = (I - H)(I - H)^{-1} = (I - H)^{-1} - H(I - H)^{-1}$$

e quindi

$$(I - H)^{-1} = I + H(I - H)^{-1}$$

si ha

$$||(I - H)^{-1}|| \leq ||I|| + ||H||\,||(I - H)^{-1}||$$

da cui

$$||(I - H)^{-1}|| \leq \frac{1}{1 - q}$$

Sostituendo si ottiene

$$||\xi^{(k-1)} - z_*|| \leq \frac{||\xi^{(k)} - \xi^{(k-1)}|| + ||\delta^{(k)}||}{1 - q}$$

Si confronti questo risultato con quanto ricavato nel Paragrafo F, Capitolo 2.

3.53 Problema

Sia $c \neq 0$. Provare che

$$\frac{||\xi^{(k-1)} - z_*||}{||z_*||} \leq \mu(I - H)\,\frac{||\xi^{(k)} - \xi^{(k-1)}|| + ||\delta^{(k)}||}{||c||}$$

△

Un diverso criterio di arresto è basato sulla considerazione seguente. Sia $v \in \mathbb{C}^n$. Posto $r = Av - b$, si ha

$$||v - z_*|| \leq ||A^{-1}||\,||r|| \quad \text{e} \quad \frac{||v - z_*||}{||z_*||} \leq \mu(A)\frac{||r||}{||b||} \tag{3.8}$$

Posto $r^{(k)} = A\xi^{(k)} - b$, il criterio d'arresto è:

$$||r^{(k)}|| < \epsilon$$

$\epsilon > 0$ dato. Si osservi che

(a) questo secondo criterio di arresto risulta più costoso del precedente;

(b) le stime (3.8) sono utilizzabili per valutare la bontà della "soluzione approssimata" v *indipendentemente* dal metodo utilizzato per ottenerla.

Appendice: matrici a blocchi

Per n intero positivo si indicano con $e_1, \dots, e_n$ gli elementi della base canonica di $\mathbb{C}^n$, con I_n (con I se non vi è pericolo di confusione) la matrice $(e_1, \dots, e_n) \in \mathbb{C}^{n\times n}$, con J_n (con J se ...) la matrice $(e_n, \dots, e_1) \in \mathbb{C}^{n\times n}$, con u la colonna $(1, \dots, 1)^{\mathsf{T}} \in \mathbb{C}^n$ e con U la matrice $uu^{\mathsf{T}} \in \mathbb{C}^{n\times n}$. Per n, m interi positivi si indica con $0_{n\times m}$ (con 0 se ...) la matrice nulla in $\mathbb{C}^{n\times m}$.

C-1 Definizione e prime proprietà

3.54 Definizione (vettori e matrici a blocchi)

Siano $n_1, \dots, n_k$ interi positivi, e $n = n_1 + \cdots + n_k$.

Sia $\mathcal{V}(n_1, \dots, n_k)$ l'insieme delle colonne v con k elementi v_i tali che $v_i \in \mathbb{C}^{n_i}$, e per ogni $a, b \in \mathcal{V}(n_1, \dots, n_k)$ e ogni $\alpha \in \mathbb{C}$ definiamo $a + b$ e $\alpha\, a$ come gli elementi di $\mathcal{V}(n_1, \dots, n_k)$ tali che

$$(a + b)_i = a_i + b_i \quad \text{e} \quad (\alpha\, a)_i = \alpha\, a_i$$

La struttura che si ottiene è uno spazio vettoriale su $\mathbb{C}$ — una copia di $\mathbb{C}^n$.

Sia $\mathcal{M}(n_1, \dots, n_k)$ l'insieme delle matrici M di ordine $k \times k$ ad elementi m_{ij} tali che $m_{ij} \in \mathbb{C}^{n_i \times n_j}$, e per ogni $A, B \in \mathcal{M}(n_1, \dots, n_k)$ e ogni $\alpha \in \mathbb{C}$ definiamo $A + B$, AB e $\alpha\, A$ come gli elementi di $\mathcal{M}(n_1, \dots, n_k)$ tali che

$$(A + B)_{ij} = a_{ij} + b_{ij} \quad , \quad (AB)_{ij} = \sum_{\ell=1}^{k} a_{i\ell} b_{\ell j} \quad \text{e} \quad (\alpha\, A)_{ij} = \alpha\, a_{ij}$$

La struttura che si ottiene è un'algebra non commutativa su $\mathbb{C}$ — una copia di $\mathbb{C}^{n\times n}$.

Chiameremo *vettori a blocchi* gli elementi di $\mathcal{V}(n_1, \dots, n_k)$ e *matrici a blocchi* gli elementi di $M \in \mathcal{M}(n_1, \dots, n_k)$. Chiameremo *blocchi* gli elementi di un vettore a blocchi o di una matrice a blocchi.

3.55 Esempio

Siano $n_1 = 2, n_2 = 1$. In $\mathcal{V}(2, 1)$ si ha, ad esempio:

$$a = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \left[\begin{array}{c} i \\ 1 \\ \hline 2 \end{array}\right] \quad , \quad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \left[\begin{array}{c} 0 \\ -1 \\ \hline -i \end{array}\right] \tag{3.9}$$

e quindi

$$a+b=\left[\begin{array}{c} a_1+b_1 \\ a_2+b_2 \end{array}\right]=\left[\begin{array}{c} i \\ \hline 0 \\ \hline 2-i \end{array}\right] \quad , \quad 5\,a=\left[\begin{array}{c} 5a_1 \\ 5a_2 \end{array}\right]=\left[\begin{array}{c} 5i \\ \hline 5 \\ \hline 10 \end{array}\right]$$

Si osservi che a e b sono gli elementi di $\mathcal{V}(2,1)$ che si ottengono dai vettori

$$a'=\left[\begin{array}{c} i \\ 1 \\ 2 \end{array}\right] \quad \text{e} \quad b'\left[\begin{array}{c} 0 \\ -1 \\ -i \end{array}\right]$$

di $\mathbb{C}^3$ partizionandone gli elementi secondo lo schema di $\mathcal{V}(2,1)$ — e cioè come indicato in (3.9) — e che $a+b$ e $5\,a$ sono gli elementi di $\mathcal{V}(2,1)$ che si ottengono partizionando $a'+b'$ e $5\,a'$.

Questo spiega in che senso $\mathcal{V}(2,1)$ è una copia di $\mathbb{C}^3$.

Analogamente, in $\mathcal{M}(2,1)$ si ha:

$$A=\left[\begin{array}{cc} a_{11} & a_{12} \\ a_{21} & a_{22} \end{array}\right]=\left[\begin{array}{cc|c} 1 & i & 0 \\ 1 & 1 & 0 \\ \hline 1 & 1 & 2 \end{array}\right] \quad , \quad B=\left[\begin{array}{cc} b_{11} & b_{12} \\ b_{21} & b_{22} \end{array}\right]=\left[\begin{array}{cc|c} 2 & -i & 0 \\ 3 & 1 & 5 \\ \hline 1 & 2 & -4 \end{array}\right]$$

e quindi

$$A+B=\ldots \quad , \quad AB=\ldots \quad , \quad 7\,A=\ldots$$

Si osservi infine che A e B sono gli elementi di $\mathcal{M}(2,1)$ che si ottengono dalle matrici

$$A'=\left[\begin{array}{ccc} 1 & i & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 2 \end{array}\right] \quad \text{e} \quad B'=\left[\begin{array}{ccc} 2 & -i & 0 \\ 3 & 1 & 5 \\ 1 & 2 & -4 \end{array}\right]$$

in $\mathbb{C}^{3\times 3}$ partizionandone gli elementi secondo lo schema di $\mathcal{M}(2,1)$ — e cioè come indicato in (3.9) — e che $A+B$, AB e $7\,A$ sono gli elementi di $\mathcal{M}(2,1)$ che si ottengono partizionando $A'+B'$, $A'B'$ e $7\,A'$.

Questo spiega in che senso $\mathcal{M}(2,1)$ è una copia di $\mathbb{C}^{3\times 3}$.

3.56 Esercizio

Per n_1, n_2 interi positivi, sia

$$\mathcal{U}(n_1,n_2)=\{M\in\mathcal{M}(n_1,n_2) \text{ tali che } m_{21}=0_{n_2\times n_1}\}$$

l'insieme delle *matrici triangolari (superiori) in* $\mathcal{M}(n_1,n_2)$.

Verificare che

(1) $\mathcal{U}(n_1,n_2)$ è chiuso rispetto al prodotto;

(2) $A\in\mathcal{U}(n_1,n_2)$ ammette un inverso (in $\mathcal{M}(n_1,n_2)$) se e solo se $a_{11}\in\mathbb{C}^{n_1\times n_1}$ e $a_{22}\in\mathbb{C}^{n_2\times n_2}$ sono invertibili;

(3) se B è un inverso (in $\mathcal{M}(n_1,n_2)$) di $A\in\mathcal{U}(n_1,n_2)$ allora $B\in\mathcal{U}(n_1,n_2)$.

C-2 Sistemi di equazioni lineari

3.57 Definizione

Siano $n_1, \ldots, n_k$ interi positivi, e $n = n_1 + \cdots + n_k$.

Per $M \in \mathcal{M}(n_1, \ldots, n_k)$ e $v \in \mathcal{V}(n_1, \ldots, n_k)$ sia $Mv \in \mathcal{V}(n_1, \ldots, n_k)$ l'elemento definito da

$$(Mv)_i = \sum_{\ell=1}^{k} m_{i\ell} v_\ell$$

Si osservi che, se M si ottiene partizionando $M' \in \mathbb{C}^{n\times n}$ e v partizionando $v' \in \mathbb{C}^n$, Mv si ottiene partizionando $M'v'$.

3.58 Osservazione

Siano A l'elemento di $\mathcal{M}(n_1, \ldots, n_k)$ che si ottiene partizionando $A' \in \mathbb{C}^{n\times n}$ e v, b gli elementi di $\mathcal{V}(n_1, \ldots, n_k)$ che si ottengono partizionando $v', b' \in \mathbb{C}^n$. allora:

$$Av = b \quad \text{se e solo se} \quad A'v' = b'$$

Dunque: risolvere l'equazione $Ax = b$ in $\mathcal{V}(n_1, \ldots, n_k)$ è equivalente a risolvere il sistema $A'x = b'$ in $\mathbb{C}^n$.

C-3 Fattorizzazione LR a blocchi

Siano $n_1, \ldots, n_k$ interi positivi e $n = n_1 + \cdots + n_k$.

3.59 Definizione (fattorizzazione LR a blocchi)

Sia $A \in \mathcal{M}(n_1, \ldots, n_k)$. La coppia $S, D \in \mathcal{M}(n_1, \ldots, n_k)$ è una *fattorizzazione LR a blocchi* di A se

(1) S è triangolare inferiore con $s_{ii} = I_{n_i}$;

(2) D è triangolare superiore;

(3) $SD = A$.

3.60 Esempio

Siano

$$M = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \in \mathbb{C}^{2\times 2} \quad \text{e} \quad A = \begin{bmatrix} M & J \\ I & M^{\mathsf{T}} \end{bmatrix} \in \mathcal{M}(2,2)$$

Se esiste, una fattorizzazione LR a blocchi di A è costituita da due elementi

$$S = \begin{bmatrix} I & 0 \\ s_{21} & I \end{bmatrix} \quad , \quad D = \begin{bmatrix} d_{11} & d_{12} \\ 0 & d_{22} \end{bmatrix}$$

in $\mathcal{M}(2,2)$ tali che

$$SD = \begin{bmatrix} M & J \\ I & M^{\mathsf{T}} \end{bmatrix}$$

Il problema si riduce quindi a determinare blocchi s_{21}, d_{11}, d_{12} e d_{22} tali che

$$\begin{array}{ll} (1,1) & d_{11} = M \\ (1,2) & d_{12} = J \\ (2,1) & s_{21}d_{11} = I \\ (2,2) & s_{21}d_{12} + d_{22} = M^{\mathsf{T}} \end{array}$$

Le prime due relazioni forniscono immediatamente d_{11} e d_{12}. La terza relazione, *poiché* $d_{11} = M$ *è invertibile,* consente di ricavare $s_{21} = M^{-1}$. Dall'ultima si ottiene $d_{22} = M^{\mathsf{T}} - M^{-1}J$. Perciò una fattorizzazione LR a blocchi esiste ed è

$$S = \begin{bmatrix} I & 0 \\ M^{-1} & I \end{bmatrix} \quad , \quad D = \begin{bmatrix} M & J \\ 0 & M^{\mathsf{T}} - M^{-1}J \end{bmatrix}$$

Si osservi l'analogia del procedimento usato con il metodo di Doolittle utilizzato nello studio delle fattorizzazioni LR.

3.61 Osservazione

Siano M, S, D elementi di $\mathcal{M}(n_1, \ldots, n_k)$ e M', S', D' i corrispondenti elementi di $\mathbb{C}^{n\times n}$.

Se S, D è una fattorizzazione LR a blocchi di M, *non è detto* che S', D' sia una fattorizzazione LR di M'.

Analogamente, se S', D' è una fattorizzazione LR di M', *non è detto* che S, D sia una fattorizzazione LR a blocchi di M.

3.62 Esercizio

1) Per ciascuna delle seguenti matrici, determinare una fattorizzazione LR a blocchi:

$$\begin{bmatrix} I & I \\ J & 0 \end{bmatrix} \in \mathcal{M}(7,3) \qquad , \qquad \begin{bmatrix} 1 & 0 \\ 0 & J \end{bmatrix} \in \mathcal{M}(1,13)$$

2) Sia

$$M = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \in \mathbb{C}^{3\times 3}$$

Determinare una fattorizzazione LR di M e una fattorizzazione LR a blocchi dell'elemento di $\mathcal{M}(1,2)$ corrispondente ad M.

C-4 Uso della fattorizzazione LR a blocchi

Siano $n_1, \ldots, n_k$ interi positivi e $n = n_1 + \cdots + n_k$.

In questa Sezione si utilizzerà, senza pericolo di confusione, la stessa lettera per indicare un elemento di $\mathcal{M}(n_1, \ldots, n_k)$ ed il corrispondente elemento di $\mathbb{C}^{n\times n}$.

3.63 Osservazione

Sia

$$A = \begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{bmatrix} \in \mathcal{M}(n_1, n_2)$$

Si verifica facilmente che, posto

$$B = \begin{bmatrix} a_{11} & 0 \\ 0 & I \end{bmatrix} \in \mathcal{M}(n_1, n_2) \quad \text{e} \quad C = \begin{bmatrix} I & 0 \\ a_{21} & a_{22} \end{bmatrix} \in \mathcal{M}(n_1, n_2)$$

si ha $A = BC$. Per quanto detto nella Definizione 3.54, quest'ultima relazione sussiste se letta in $\mathbb{C}^{n\times n}$. Allora, per il Teorema di Binet, si ha: $\det A = \det B \det C$. Poiché $\det B = \det a_{11}$ e $\det C = \det a_{22}$ (verificare!), si ha infine che

$$\det A = \det a_{11} \det a_{22}$$

In generale, per $A \in \mathcal{M}(n_1, \ldots, n_k)$ triangolare, detti $a_{11}, \ldots, a_{kk}$ i blocchi sulla diagonale di A si ha

$$\det A = \det a_{11} \cdots \det a_{kk}$$

3.64 Esempio (uso della fattorizzazione LR a blocchi)

Sia

$$A = \left[\begin{array}{ccc|cc} 5 & 2 & 0 & 0 & -1 \\ 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ \hline 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \end{array}\right] = \begin{bmatrix} M & a_{12} \\ a_{21} & U \end{bmatrix} \in \mathcal{M}(3, 2)$$

Si vuole calcolare $\det A$ e, se possibile, A^{-1}.

Una fattorizzazione LR a blocchi di A è:

$$S = \begin{bmatrix} I & 0 \\ s_{21} & I \end{bmatrix} \quad \text{con} \quad s_{21} = \begin{bmatrix} 0 & 0 & 1 \\ -1 & 3 & 0 \end{bmatrix} \quad , \quad D = \begin{bmatrix} M & a_{12} \\ 0 & J \end{bmatrix}$$

Per l'Osservazione 3.63 si ha allora

$$\det A = \det M \det J = -1$$

Si osservi che il calcolo del determinante di una matrice 5×5 è ricondotto al calcolo del determinante di una matrice 3×3 e di una 2×2.

Si osservi inoltre che $A \in \mathbb{C}^{5\times 5}$ *non ammette* fattorizzazione LR (come mai?).

A risulta dunque invertibile e, poiché $A = SD$ si ha

$$A^{-1} = (SD)^{-1} = D^{-1}S^{-1}$$

Per calcolare D^{-1} si può operare in $\mathcal{M}(3,2)$: si cerca

$$X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} \in \mathcal{M}(3,2)$$

tale che

$$DX = \begin{bmatrix} I_3 & 0 \\ 0 & I_2 \end{bmatrix}$$

Il problema si riduce a determinare blocchi x_{11}, x_{21}, x_{21} e x_{22} tali che

$$\begin{array}{ll} (1,1) & Mx_{11} + a_{12}x_{21} = I_3 \\ (1,2) & Mx_{12} + a_{12}x_{22} = 0 \\ (2,1) & Jx_{21} = 0 \\ (2,2) & Jx_{22} = I_2 \end{array}$$

Dalle ultime due equazioni si ottiene $x_{21} = 0$ e $x_{22} = J^{-1}$. Dalla seconda si ha $x_{12} = -M^{-1}a_{12}J^{-1}$ e infine, dalla prima: $x_{11} = M^{-1}$. Si osservi che $D^{-1} \in \mathcal{M}(3,2)$ è triangolare superiore.

In modo analogo si calcola S^{-1}. Infine, eseguendo la moltiplicazione si ottiene l'inversa. Si osservi che il calcolo dell'inversa è ricondotto al calcolo dall'inversa di due matrici triangolari in $\mathcal{M}(3,2)$, e quindi al calcolo dell'inversa di una matrice 3×3 e di una 2×2.

Infine, posto

$$b = \left[\begin{array}{c} 1 \\ 0 \\ 0 \\ \hline 0 \\ 2 \end{array}\right] = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

si vogliono determinare le soluzioni in $\mathcal{V}(3,2)$ dell'equazione $Ax = b$.

Utilizzando la fattorizzazione LR a blocchi determinata sopra, il problema equivale a quello di risolvere in $\mathcal{M}(3,2)$, nell'ordine, le equzioni

$$[1] \quad Sc = b \qquad \text{e} \qquad [2] \quad Dx = c$$

Per la [1], il problema si riduce a determinare colonne c_1 e c_2 tali che

$$\begin{array}{ll} (1) & c_1 = b_1 \\ (2) & s_{21}c_1 + c_2 = b_2 \end{array}$$

Dalla prima si ottiene $c_1 = b_1$ e dalla seconda $c_2 = b_2 - s_{21}b_1$.

Per la [2], si cercano colonne x_1 e x_2 tali che

$$\begin{array}{ll}(1) & Mx_1 + a_{12}x_2 = b_1 \\ (2) & Jx_2 = b_2 - s_{21}b_1\end{array}$$

Dalla seconda si ottiene $x_2 = J^{-1}(b_2 - s_{21}b_1)$ e dalla prima $x_1 = \ldots$

Si osservi che si è ottenuto (come dovevamo aspettarci ...) un solo elemento di $\mathcal{V}(3,2)$ che risolve l'equazione $Ax = b$.

Capitolo 4

Interpolazione

In questo Capitolo considereremo il problema più classico dell'interpolazione, il *problema dell'interpolazione parabolica*, accenneremo ad una sua generalizzazione, il *problema lineare dell'interpolazione* ed infine discuteremo il caso più semplice di *campionamento e ricostruzione.*

A Interpolazione parabolica o polinomiale

Siano k un intero non negativo, $[a, b]$ un intervallo non degenere e si consideri $P_k(\mathbb{R})$ come sottospazio vettoriale di $\mathcal{C}^0([a, b], \mathbb{R})$, spazio vettoriale su $\mathbb{R}$ delle funzioni continue da $[a, b]$ in $\mathbb{R}$.

Assegnati $x_0, \ldots, x_k \in [a, b]$ *distinti* e $y_0, \ldots, y_k \in \mathbb{R}$, il *problema dell'interpolazione parabolica* consiste nel determinare gli elementi $p \in P_k(\mathbb{R})$ che verificano le condizioni

$$p(x_0) = y_0, \ldots, p(x_k) = y_k \tag{4.1}$$

("che interpolano i dati $(x_0, y_0), \ldots, (x_k, y_k)$").

Si osservi che $P_k(\mathbb{R})$ è uno spazio vettoriale di dimensione $k + 1$, e le condizioni (4.1) sono esattamente $k + 1$.

Sia $q_0, \ldots, q_k$ una base di $P_k(\mathbb{R})$. Un elemento $p = a_0 q_0 + \cdots + a_k q_k \in P_k(\mathbb{R})$ verifica le condizioni (4.1) se e solo se

$$\begin{cases} a_0 q_0(x_0) + \cdots + a_k q_k(x_0) &= y_0 \\ & \vdots \\ a_0 q_0(x_k) + \cdots + a_k q_k(x_k) &= y_k \end{cases}$$

ossia se e solo se le coordinate $(a_0, \ldots, a_k)^\mathsf{T}$ di p rispetto alla base verificano il sistema

$$\begin{bmatrix} q_0(x_0) & \cdots & q_k(x_0) \\ \vdots & & \vdots \\ q_0(x_k) & \cdots & q_k(x_k) \end{bmatrix} \begin{bmatrix} \xi_0 \\ \vdots \\ \xi_k \end{bmatrix} = \begin{bmatrix} y_0 \\ \vdots \\ y_k \end{bmatrix} \tag{4.2}$$

In particolare, il problema dell'interpolazione parabolica ha tante soluzioni quante ne ha il sistema (4.2).

4.1 Teorema

Esiste un solo elemento di $P_k(\mathbb{R})$ che verifica le condizioni (4.1).

Dimostrazione

Per $r = 0, \ldots, k$ sia

$$\ell_{k,r}(x) = \frac{(x - x_0) \cdots (x - x_{r-1})(x - x_{r+1}) \cdots (x - x_k)}{(x_r - x_0) \cdots (x_r - x_{r-1})(x_r - x_{r+1}) \cdots (x_r - x_k)}$$

Si ha

(a) $\ell_{k,0}, \ldots, \ell_{k,k} \in P_k(\mathbb{R})$

(b) siano $r, j \in \{\, 0, \ldots, k \,\}$; allora

$$\ell_{k,r}(x_j) = \begin{cases} 1 & \text{se} \quad j = r \\ 0 & \text{se} \quad j \neq r \end{cases}$$

Gli elementi $\ell_{k,0}, \ldots, \ell_{k,k}$ sono una base di $P_k(\mathbb{R})$. Infatti, sono $k + 1$ elementi linearmente indipendenti: se

$$\alpha_0 \ell_{k,0} + \cdots + \alpha_k \ell_{k,k} = 0$$

allora

$$\begin{cases} (\alpha_0 \ell_{k,0} + \cdots + \alpha_k \ell_{k,k})(x_0) &= 0 \\ & \vdots \\ (\alpha_0 \ell_{k,0} + \cdots + \alpha_k \ell_{k,k})(x_k) &= 0 \end{cases}$$

e quindi $\alpha_0 = 0, \ldots, \alpha_k = 0$.

Utilizzando la base trovata, la matrice del sistema (4.2) risulta la matrice identica di ordine $k + 1$ e quindi tale sistema ha la sola soluzione $(\xi_0, \ldots, \xi_k)^{\mathsf{T}} = (y_0, \ldots, y_k)^{\mathsf{T}}$.

Il polinomio

$$p_k = y_0 \ell_{k,0} + \cdots + y_k \ell_{k,k} \tag{4.3}$$

è l'elemento che verifica le condizioni, e prende il nome di *polinomio interpolante;* la scrittura (4.3) si chiama *forma di Lagrange* del polinomio interpolante. □

4.2 Osservazione

Utilizzando basi di $P_k(\mathbb{R})$ diverse da quella introdotta nella dimostrazione del Teorema 4.1, si ottengono *forme* diverse del polinomio interpolante.

(1) Utilizzando la base $1, x, \ldots, x^k$ il sistema (4.2) diventa $V\xi = y$ con $\xi = (\xi_0, \ldots, \xi_k)^\mathsf{T}$, $y = (y_0, \ldots, y_k)^\mathsf{T}$ e

$$V = \begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^k \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_k & x_k^2 & \cdots & x_k^k \end{bmatrix}$$

In virtù del Teorema 4.1, tale matrice (detta *matrice di Vandermonde* relativa ai punti $x_0, \ldots, x_k$) risulta non singolare; detta $(a_0, \ldots a_k)^\mathsf{T}$ la soluzione del sistema $V\xi = y$, il polinomio interpolante assume la forma

$$p_k(x) = a_0 + a_1 x + \cdots + a_k x^k$$

(2) Utilizzando la base $1, x - x_0, (x - x_0)(x - x_1), \ldots, (x - x_0) \cdots (x - x_{k-1})$ il sistema (4.2) diventa $T\xi = y$ con $\xi = (\xi_0, \ldots, \xi_k)^\mathsf{T}$, $y = (y_0, \ldots, y_k)^\mathsf{T}$ e

$$T = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & x_1 - x_0 & 0 & \cdots & 0 \\ 1 & x_2 - x_0 & (x_2 - x_0)(x_2 - x_1) & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_k - x_0 & (x_k - x_0)(x_k - x_1) & \cdots & (x_k - x_0) \cdots (x_k - x_{k-1}) \end{bmatrix}$$

Si osservi che la matrice del sistema risulta triangolare inferiore. Detta $(b_0, \ldots b_k)^\mathsf{T}$ la soluzione del sistema $T\xi = y$, il polinomio interpolante assume la forma

$$p_k(x) = b_0 + b_1(x - x_0) + \cdots + b_k(x - x_0) \cdots (x - x_{k-1})$$

chiamata *forma di Newton* del polinomio interpolante.

4.3 Esempio

Si considerino i dati

$$\begin{array}{c|c|c|c} x_j & 0 & 1 & -1 \\ \hline y_j & 1 & 3 & -1 \end{array} \quad , \quad j = 0, 1, 2$$

Studiare il problema dell'interpolazione polinomiale.

Soluzione

Il problema dell'interpolazione polinomiale, con $k = 2$, ha una soluzione unica, come garantito dal Teorema 4.1 (e, in questo caso, dall'intuito).

(a) Il polinomio interpolante in forma di Lagrange è

$$p_2(x) = \ell_{2,0}(x) + 3\ \ell_{2,1}(x) - \ell_{2,2}(x)$$

con

$$\ell_{2,0}(x) = \frac{(x-1)(x+1)}{(-1)\cdot 1},\ \ell_{2,1}(x) = \frac{x(x+1)}{1\cdot 2},\ \ell_{2,2}(x) = \frac{x(x-1)}{(-1)(-2)}$$

Quindi: $p_2(x) = 1 + 2x$.

(b) Il polinomio interpolante nella forma $a_0 + a_1x + a_2x^2$ si ottiene risolvendo il sistema

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} \xi_0 \\ \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ -1 \end{bmatrix}$$

da cui $p_2(x) = 1 + 2x$.

(c) Il polinomio interpolante in forma di Newton $b_0 + b_1x + b_2x(x-1)$ si ottiene risolvendo il sistema

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & -1 & 2 \end{bmatrix} \begin{bmatrix} \xi_0 \\ \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ -1 \end{bmatrix}$$

da cui $p_2(x) = 1 + 2x$.

4.4 Osservazione

Siano k un intero non negativo, $\Gamma \subset \mathbb{C}$ un aperto non vuoto, e si consideri $P_k(\mathbb{C})$ come sottospazio vettoriale di $\mathcal{C}(\Gamma, \mathbb{C})$.

Assegnati $z_0, \ldots, z_k \in \Gamma$ *distinti* e $w_0, \ldots, w_k \in \mathbb{C}$, il *problema dell'interpolazione parabolica in* $\mathbb{C}$ consiste nel determinare gli elementi $p \in P_k(\mathbb{C})$ che verificano le condizioni

$$p(z_0) = w_0, \ldots, p(z_k) = w_k$$

("che interpolano i dati $(z_0, w_0), \ldots, (z_k, w_k)$").

Le considerazioni precedenti il Teorema 4.1, il teorema stesso e l'Osservazione 4.2, restano validi per il problema dell'interpolazione parabolica in $\mathbb{C}$.

Un importante esempio di interpolazione parabolica in $\mathbb{C}$ è la Trasformata Discreta di Fourier.

L'esempio che segue introduce due possibili metodi di valutazione del condizionamento per il problema dell'interpolazione polinomiale.

4.5 Esempio

(1) La quota $q(t)$ di un corpo puntiforme pesante in caduta libera nel vuoto è data da

$$q(t) = -\frac{1}{2}\, g\, t^2 + v_0\, t + q_0$$

Il Teorema 4.1 assicura che tre misure di q ad istanti distinti t_0, t_1, t_2 consentono di determinare univocamente i valori g, v_0 e q_0 costruendo l'elemento di $P_2(\mathbb{R})$ che interpola i dati $(t_0, q(t_0)), (t_1, q(t_1)), (t_2, q(t_2))$.

Le misure, però, saranno affette da errore e quindi i valori ricavati per g, v_0 e q_0 (i valori "stimati") non sarano esattamente quelli desiderati.

Supponendo di effettuare una misuta ogni τ secondi, a partire da 0, si vuole studiare l'andamento dell'errore trasmesso ai coefficienti del polinomio interpolante dagli errori di misura, in funzione di τ.

In questo caso, la scelta naturale per la forma del polinomio interpolante è quella relativa alla base $1, t, t^2$ e, scegliendo di considerare gli errori relativi, il problema si riduce (vedere (3.3)) allo studio, per $\tau > 0$, del numero di condizionamento della matrice di Vandermonde relativa ai punti $0, \tau, 2\tau$:

$$V = \begin{bmatrix} 1 & 0 & 0 \\ 1 & \tau & \tau^2 \\ 1 & 2\tau & 4\tau^2 \end{bmatrix}$$

Utilizzando la norma infinito si ha

$$\mu_\infty(V) = \begin{cases} 1 + 2\tau + 4\tau^2 & \text{se } \tau > 4 \\ \frac{4}{\tau} + 8 + 16\tau & \text{se } \frac{1}{2} \leq \tau \leq 4 \\ \frac{2}{\tau^2} + \frac{4}{\tau} + 8 & \text{se } 0 \leq \tau < \frac{1}{2} \end{cases}$$

La situazione "meno rischiosa" si ottiene quando il numero di condizionamento è minimo. In questo caso il valore minimo si ottiene per $\tau = \frac{1}{2}$.

(2) Le tre misure di q agli istanti $0, \tau, 2\tau$ determinano la funzione $q(t)$ per ogni t. Detta q^* la funzione ricavata interpolando i dati $(t_0, q(t_0) + \delta_0), (t_1, q(t_1) + \delta_1), (t_2, q(t_2) + \delta_2)$, si vuole studiare l'andamento dell'errore trasmesso al valore di $q(t)$ dagli errori di misura δ_j.

In questo caso, scegliendo di considerare gli errori assoluti, misuriamo l'errore assoluto trasmesso dai dati con

$$\delta_d = \max\{\, |q^*(t) - q(t)|,\ t \in [0, 2\tau] \,\}$$

e l'errore assoluto sui dati con

$$\delta(q) = \max\{\, |\delta_j|,\ j = 0, 1, 2 \,\}$$

Si osservi che il polinomio $q^* - q$ interpola i dati $(0, \delta_0), (\tau, \delta_1), (2\tau, \delta_2)$. Si ha quindi

$$q^* - q = \delta_0\, \ell_{2,0} + \delta_1\, \ell_{2,1} + \delta_2\, \ell_{2,2}$$

e, tenuto conto che per $j = 0, 1, 2$ si ha $\max\{\, |\ell_{2,j}(t)|, t \in [0, 2\tau] \,\} = 1$, si ottiene (*indipendentemente dal valore di* τ)

$$\frac{\delta_d}{\delta(q)} \leq 3$$

B Il problema lineare dell'interpolazione

Siano k un intero non negativo, j un intero positivo, $[a,b]$ un intervallo non degenere e si consideri un sottospazio vettoriale $\mathcal{G}$ di $\mathcal{C}([a,b],\mathbb{R})$, di dimensione j.

Assegnati $L_0,\dots,L_k$ applicazioni lineari da $\mathcal{G}$ in $\mathbb{R}$ e $y_0,\dots,y_k \in \mathbb{R}$, il *problema lineare dell'interpolazione* consiste nel determinare gli elementi $g \in \mathcal{G}$ che verificano le $k+1$ condizioni

$$L_0(g) = y_0, \dots, L_k(g) = y_k \tag{4.4}$$

4.6 Esempio

(1) Il problema dell'interpolazione parabolica è il problema lineare di interpolazione con $\mathcal{G} = P_k(\mathbb{R}), L_0 : p \to p(x_0), \dots, L_k : p \to p(x_k)$.

(2) Determinare le $x \in \mathcal{C}^2(\mathbb{R},\mathbb{R})$ tali che

$$x'' - x' = 0 \quad , \quad x(0) = a \quad , \quad x'(0) = b$$

a, b reali assegnati.

Questo è il problema lineare di interpolazione con $\mathcal{G} = \{\, x \in \mathcal{C}^2(\mathbb{R},\mathbb{R})$ tali che $x'' - x' = 0\}$, $L_0 : x \to x(0), L_1 : x \to x'(0)$, $y_0 = a, y_1 = b$.

(3) Determinare le $x \in \mathcal{C}^2(\mathbb{R},\mathbb{R})$ tali che

$$x'' + x = 0 \quad , \quad x(0) = 0 \quad , \quad x(2\pi) = 0$$

Questo è il problema lineare di interpolazione con $\mathcal{G} = \{\, x \in \mathcal{C}^2(\mathbb{R},\mathbb{R})$ tali che $x'' + x = 0\}$, $L_0 : x \to x(0), L_1 : x \to x(2\pi)$, $y_0 = 0, y_1 = 0$.

4.7 Osservazione

Sia $g_1, ..., g_j$ una base di $\mathcal{G}$. Un elemento $g = a_1 g_1 + \cdots + a_j g_j \in \mathcal{G}$ verifica le condizioni (4.4) se e solo se

$$\begin{cases} a_1 L_0(g_1) + \cdots + a_j L_0(g_j) &= y_0 \\ & \vdots \\ a_1 L_k(g_1) + \cdots + a_j L_k(g_j) &= y_k \end{cases}$$

ossia se e solo se le coordinate $(a_1,\dots,a_j)^{\mathsf{T}}$ di g verificano il sistema di $k+1$ equazioni in j incognite

$$\begin{bmatrix} L_0(g_1) & \cdots & L_0(g_j) \\ \vdots & & \vdots \\ L_k(g_1) & \cdots & L_k(g_j) \end{bmatrix} \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_j \end{bmatrix} = \begin{bmatrix} y_0 \\ \vdots \\ y_k \end{bmatrix} \tag{4.5}$$

In particolare, il problema lineare dell'interpolazione ha tante soluzioni quante ne ha il sistema (4.5).

4.8 Problema

Determinare tutte le soluzioni dei problemi lineari di interpolazione dei punti (2) e (3) dell'Esempio 4.6. △

C Campionamento e ricostruzione

4.9 Definizione (campionamento, ricostruzione)

Siano $[a, b]$ un intervallo non degenere, k un intero non negativo, $t_0, ..., t_k$ reali distinti in $[a, b]$.

L'applicazione lineare $c : \mathcal{C}([a, b], \mathbb{R}) \to \mathbb{R}^{k+1}$ definita da

$$c(f) = (f(t_0), ..., f(t_k))^{\mathsf{T}}$$

si chiama *funzione di campionamento* (*agli istanti* $t_0, ..., t_k$); questi ultimi si chiamano *istanti di campionamento.*

Si osservi che c *non è invertibile.*

Un'applicazione lineare $r : \mathbb{R}^{k+1} \to \mathcal{C}([a, b], \mathbb{R})$ tale che

$$\text{per ogni } y \in \mathbb{R}^{k+1} \quad , \quad c(r(y)) = y$$

si chiama *funzione di ricostruzione* (*relativa a* c).

4.10 Esempio

Dati $t_0, ..., t_k$ distinti in $[a, b]$, sia c la funzione di campionamento ad essi relativa. Si indichino con $y_0, ..., y_k$ le componenti di $y \in \mathbb{R}^{k+1}$.

La funzione $r : \mathbb{R}^{k+1} \to \mathcal{C}([a, b], \mathbb{R})$ definita da

$$r : y \to \text{l'elemento di } P_k(\mathbb{R}) \text{ che interpola i dati } (t_0, y_0), ..., (t_k, y_k)$$

è una funzione di ricostruzione. Infatti: $c(r(y)) = y$ e, utilizzando la forma di Lagrange del polinomio interpolante, si verifica la linearità di r.

Con questa scelta di r si dice che *la ricostruzione è ottenuta mediante interpolazione parabolica.*

4.11 Definizione (errore di ricostruzione)

Siano $[a, b]$ un intervallo non degenere, k un intero non negativo, $t_0, ..., t_k$ reali distinti in $[a, b]$, c la funzione di campionamento ad essi relativa e r una funzione di ricostruzione relativa a c.

Data $f \in \mathcal{C}([a, b], \mathbb{R})$, la quantità

$$e(f) = \max\{\, |f(t) - r(c(f))(t)|, t \in [a, b] \,\}$$

si chiama *errore di ricostruzione* di f.

Il Teorema seguente consente di studiare l'errore nel caso di ricostruzione ottenuta mediante interpolazione parabolica.

4.12 Teorema

Siano $[a, b]$ un intervallo non degenere, k un intero non negativo, $t_0, ..., t_k$ reali distinti in $[a, b]$, $f \in \mathcal{C}^{k+1}([a, b], \mathbb{R})$.

Sia p_k l'elemento di $P_k(\mathbb{R})$ che interpola i dati $(t_j, f(t_j))$.

Per ogni $t \in [a,b]$ esiste $\xi \in [a,b]$ (dipendente da t) tale che

$$f(t) - p_k(t) = \frac{f^{(k+1)}(\xi)}{(k+1)!}(t-t_0)\cdots(t-t_k) \tag{4.6}$$

Dimostrazione

Se $t = t_j$ per qualche j, l'asserto è verificato per qualsiasi $\xi \in [a,b]$.

Sia $\tau \in [a,b]$ distinto da $t_0, ..., t_k$. Posto

$$\omega(t) = (t-t_0)\cdots(t-t_k) \quad , \quad \alpha = \frac{f(\tau) - p_k(\tau)}{\omega(\tau)}$$

si consideri la funzione

$$g(t) = f(t) - p_k(t) - \alpha\,\omega(t)$$

Si ha

(1) $\omega(t) = t^{k+1} + \cdots$

(2) $g \in \mathcal{C}^{k+1}([a,b], \mathbb{R})$

(3) $g(t_j) = 0$ per $j = 0, ..., k$ e $g(\tau) = 0$

Utilizzando ripetutamente il Teorema di Rolle,* si prova l'esistenza di un reale $\xi \in [a,b]$ tale che

$$g^{(k+1)}(\xi) = 0$$

Ma

$$g^{(k+1)}(t) = f^{(k+1)}(t) + \alpha\,(k+1)!$$

e quindi

$$\alpha = \frac{f^{(k+1)}(\xi)}{(k+1)!}$$

Uguagliando le due espressioni di α si prova quindi che l'asserto sussiste per ogni $t \in [a,b]$. □

4.13 Osservazione

Posto $M_j(f) = \max\{\, |f^{(j)}(t)|, t \in [a,b]\,\}$, per l'errore nel caso di ricostruzione ottenuta mediante interpolazione parabolica, dalla (4.6) si ottiene

$$e(f) \le \frac{M_{k+1}(f)}{(k+1)!} \max\{\, |t-t_0|\cdots|t-t_k|, t \in [a,b]\,\}$$

Si ha inoltre

*Vedere [A], volume 1, Teorema 4.4, pagina 226 oppure [C], Teorema 4.28, pagina 227.

(1) $\frac{(b-a)^{k+1}}{2^{2k+1}} \leq \max\{\, |t-t_0|\cdots|t-t_k|, t \in [a,b]\,\} < (b-a)^{k+1}$

(La seconda stima è immediata; per la prima si veda [A1], volume II, Teorema 15.12, pagina 599.)

(2) sia $f \in \mathcal{C}^\infty([a,b],\mathbb{R})$; se

$$\lim_{k\to\infty} \frac{M_{k+1}(f)}{(k+1)!}(b-a)^{k+1} = 0$$

allora l'errore di ricostruzione di f può essere reso piccolo quanto si vuole scegliendo k sufficientemente grande.

(3) per $f(t) = 1/t$ e $[a,b] = [\frac{1}{5}, 1]$ la successione $\frac{M_{k+1}(f)}{2^{2k+1}(k+1)!}(b-a)^{k+1}$ non tende a zero.

La condizione espressa al punto (2) è certamente verificata, ad esempio, da tutte le $f \in \mathcal{C}^\infty([a,b],\mathbb{R})$ per le quali $M_j(f) \leq K^j$ per qualche $K \in \mathbb{R}$. Il punto (3) fornisce un esempio di funzione che *non* la verifica.

Per estendere la classe di funzioni per le quali l'errore di ricostruzione possa essere reso piccolo quanto si vuole, modifichiamo la funzione di ricostruzione.

4.14 Definizione (funzioni continue lineari a tratti)

Siano $[a,b]$ un intervallo non degenere, k un intero non negativo e $t_0 = a < t_1 < \cdots < t_k = b$ reali in $[a,b]$. Per $j = 1, ..., k$ poniamo $I_j = [t_{j-1}, t_j]$.

Sia S l'insieme delle *funzioni continue lineari a tratti su* $I_1, ..., I_k$:

$f \in S$ se

(S1) $f \in \mathcal{C}([a,b],\mathbb{R})$

(S2) per $j = 1, ..., k$ esiste $p_j \in P_1(\mathbb{R})$ tale che $f = p_j$ su I_j.

Si ha:

(1) S è uno spazio vettoriale su $\mathbb{R}$ (ovvio)

(2) assegnati $y_0, ..., y_k \in \mathbb{R}$, esiste un solo elemento di S che interpola i dati $(t_0, y_0), ..., (t_k, y_k)$ (si osserva che, per ogni j, esiste un solo elemento di $P_1(\mathbb{R})$ che interpola i dati in $I_j \ldots$)

(3) gli elementi $s_0, ..., s_k \in S$ definiti da

$$s_j(t_i) = \begin{cases} 1 & \text{se } i = j \\ 0 & \text{altrimenti} \end{cases}$$

sono una base di S (infatti sono indipendenti e generano S)

4.15 Esempio

Si considerino i dati $(0,1),(1,-2),(4,2),(5,-1)$. Posto $t_0 = 0, t_1 = 1$, $t_2 = 4, t_3 = 5$ e

$$I_1 = [0,1] \quad , \quad I_2 = [1,4] \quad , \quad I_3 = [4,5]$$

l'unica funzione continua lineare a tratti su I_1, I_2, I_3 che interpola i dati è

$$s = s_0 - 2\,s_1 + 2\,s_2 - s_3$$

4.16 Osservazione (interpolazione lineare a tratti)

Siano $t_0, ..., t_k$ come nella Definizione 4.14 e si indichino con $y_0, ..., y_k$ le componenti di $y \in R^{k+1}$.

La funzione $r : \mathbb{R}^{k+1} \to \mathcal{C}([a,b],\mathbb{R})$ definita da

$$r : y \to \text{l'elemento di } S \text{ che interpola i dati } (t_0,y_0), ..., (t_k,y_k)$$

è una funzione di ricostruzione. Infatti: $c(r(y)) = y$ e, utilizzando la base $s_0, ..., s_k$, si verifica la linearità di r.

Con questa scelta di r si dice che *la ricostruzione è ottenuta mediante interpolazione lineare a tratti.*

Sia $f \in \mathcal{C}^2([a,b],\mathbb{R})$. Posto $M_j(f) = \max\{\, |f^{(j)}(t)|, t \in [a,b] \,\}$ si ha

(1) per $t \in I_j$, utilizzando il Teorema 4.12:

$$|f(t) - r(c(f))(t)| = |f(t) - p_j(t)| \leq$$
$$\leq \tfrac{M_2(f)}{2}|t - t_{j-1}|\,|t - t_j| \leq \tfrac{M_2(f)}{8}(t_j - t_{j-1})^2$$

(infatti, per $t \in I_j$, si ha $|t - t_{j-1}|\,|t - t_j| \leq \frac{(t_j - t_{j-1})^2}{4}$)

(2) posto $h = \max\{\, |t_j - t_{j-1}|, j = 1, ..., k \,\}$, si ha

$$e(f) \leq \frac{M_2(f)}{8}\,h^2$$

4.17 Esempio

Sia k intero positivo. Posto $h = \frac{b-a}{k}$ siano $t_j = a + hj$ per $j = 0, ..., k$.

Per $f \in \mathcal{C}^2([a,b],\mathbb{R})$ si ha

$$e(f) \leq \frac{M_2(f)}{8}\frac{(b-a)^2}{k^2}$$

e quindi l'errore di ricostruzione di f può essere reso piccolo quanto si vuole scegliendo k sufficientemente grande.

4.18 Osservazione

Siano $k, t_0, ..., t_k$ e f come nell'Esempio 4.17, $\delta_0, \ldots, \delta_k \in \mathbb{R}$. Posto $\delta = (\delta_0, \ldots, \delta_k)^\mathsf{T}$ e $r^* = r(c(f) + \delta)$, per $t \in [a, b]$ si ha

$$|f(t) - r^*(t)| \leq |f(t) - r(c(f))(t)| + |r(c(f))(t) - r^*(t)|$$

Poiché

$$|r(c(f))(t) - r^*(t)| = |r(\delta)| \leq ||\delta||_\infty$$

e $|f(t) - r(c(f))(t)| \leq e(f)$, dall'esempio precedente si ottiene

$$\max\{\, |f(t) - r^*(t)|, t \in [a, b] \,\} \leq \frac{M_2(f)}{8} \frac{(b-a)^2}{k^2} + ||\delta||_\infty$$

La presenza di δ pone quindi un limite inferiore all'errore di ricostruzione di f (un esempio di contesto in cui $\delta \neq 0$ è quello della *conversione analogico–digitale*).

La teoria relativa all'interpolazione lineare a tratti trova applicazione nella creazione di programmi per il tracciamento del grafico di una funzione di variabile reale e nel problema dell'approssimazione numerica dell'integrale di una funzione di variabile reale.

Capitolo 5

Approssimazione: minimi quadrati

In questo Capitolo tratteremo il problema dell'*approssimazione nel senso dei minimi quadrati* ed accenneremo ad alcune applicazioni: la *soluzione di un sistema di equazioni lineari nel senso dei minimi quadrati*, l'*approssimazione di dati* e l'*approssimazione di funzioni con polinomi trigonometrici.* Infine, introdurremo la *fattorizzazione QR* e ne discuteremo il legame con l'approssimazione nel senso dei minimi quadrati.

Siano V uno spazio vettoriale *con prodotto interno* e W un sottospazio vettoriale di V *di dimensione finita.*

5.1 Definizione

Dato $v \in V$, il vettore $w \in W$ è *una migliore approssimazione di v in W* se

$$\text{per ogni } w' \in W \quad , \quad ||v - w|| \leq ||v - w'||$$

dove la norma è quella definita dal prodotto interno.

Equivalentemente, $w \in W$ è una migliore approssimazione di v in W se

$$\text{per ogni } w' \in W \quad , \quad ||v - w||^2 \leq ||v - w'||^2$$

da cui il nome per w di migliore approssimazione di v in W *nel senso dei minimi quadrati.*

Il Teorema 0.43 e l'Osservazione 0.44 provano il Teorema seguente.

5.2 Teorema

Esiste una sola migliore approssimazione di v in W nel senso dei minimi quadrati: la proiezione ortogonale di v su W.

5.3 Osservazione

Siano $W = \langle w_1, \ldots, w_k \rangle$, $v \in V$ e $w \in W$.

Il vettore w è la proiezione ortogonale di v su W se e solo se

$$v - w \perp W$$

ovvero se e solo se

$$w \bullet w_j = v \bullet w_j \quad \text{per } j = 1, \ldots, k$$

Posto $w = a_1 w_1 + \cdots + a_k w_k$, queste ultime condizioni sono verificate se e solo se

$$a_1 w_1 \bullet w_j + \cdots + a_k w_k \bullet w_j = v \bullet w_j \quad \text{per } j = 1, \ldots, k$$

e quindi se e solo se la colonna $(a_1, \ldots, a_k)^{\mathsf{T}}$ è soluzione del *sistema delle equazioni normali*

$$\begin{bmatrix} w_1 \bullet w_1 & \cdots & w_k \bullet w_1 \\ \vdots & & \vdots \\ w_1 \bullet w_k & \cdots & w_k \bullet w_k \end{bmatrix} \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_k \end{bmatrix} = \begin{bmatrix} v \bullet w_1 \\ \vdots \\ v \bullet w_k \end{bmatrix} \tag{5.1}$$

Si osservi che la struttura della matrice del sistema delle equazioni normali dipende dai generatori $w_1, ..., w_k$. In ogni caso la matrice del sistema è simmetrica nel caso reale, hermitiana nel caso complesso.

5.4 Esempio

Siano $V = \mathcal{C}([0,1], \mathbb{R})$ con $f \bullet g = \int_0^1 f(t)g(t)\, dt$ e $W = P_2(\mathbb{R})$.

Determinare la migliore approssimazione di $v = e^t \in V$ in W nel senso dei minimi quadrati.

Soluzione

Posto $w_1(t) = 1, w_2(t) = t, w_3(t) = t^2$, si ha $W = \langle w_1, w_2, w_3 \rangle$. Inoltre

$$\begin{array}{lll} w_1 \bullet w_1 = 1 & w_2 \bullet w_1 = \int_0^1 t\, dt = \frac{1}{2} & w_3 \bullet w_1 = \int_0^1 t^2\, dt = \frac{1}{3} \\ & w_2 \bullet w_2 = \int_0^1 t^2\, dt = \frac{1}{3} & w_3 \bullet w_2 = \int_0^1 t^3\, dt = \frac{1}{4} \\ & & w_3 \bullet w_3 = \int_0^1 t^4\, dt = \frac{1}{5} \end{array}$$

e

$$\begin{aligned} v \bullet w_1 &= \int_0^1 e^t\, dt = e - 1 \\ v \bullet w_2 &= \int_0^1 t\, e^t\, dt = 1 \\ v \bullet w_3 &= \int_0^1 t^2 e^t\, dt = e - 2 \end{aligned}$$

Le equazioni normali sono

$$\begin{bmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{bmatrix} = \begin{bmatrix} e - 1 \\ 1 \\ e - 2 \end{bmatrix}$$

e, detta w la migliore approssimazione cercata, si ha

$$w = (39e - 105) - (216e - 588)t + (210e - 570)t^2.$$

5.5 Osservazione

Siano $W = \langle w_1, \ldots, w_k \rangle$ e $v \in V$ dato. Il Teorema 5.2 garantisce che esiste un solo elemento di W migliore approssimazione di v in W nel senso dei minimi quadrati.

Se $w_1, \ldots, w_k$ sono linearmente indipendenti, esiste *una sola* combinazione lineare di $w_1, \ldots, w_k$ che individua w. Quindi la matrice del sistema delle equazioni normali (5.1) risulta *non singolare.*

Se $w_1, \ldots, w_k$ sono linearmente dipendenti, esistono *infinite* combinazioni lineari di $w_1, \ldots, w_k$ che individuano w. Quindi la matrice del sistema delle equazioni normali (5.1) risulta *singolare.*

5.6 Esempio

Siano $V = \mathbb{R}^3$ con prodotto scalare canonico,

$$W = \langle \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 \\ 0 \\ 0 \end{bmatrix} \rangle \quad , \quad v = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \notin W$$

Il sistema delle equazioni normali è $F\xi = b$ con

$$F = \begin{bmatrix} w_1 \bullet w_1 & w_2 \bullet w_1 \\ w_1 \bullet w_2 & w_2 \bullet w_2 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 3 & 9 \end{bmatrix} \quad , \quad b = \begin{bmatrix} v \bullet w_1 \\ v \bullet w_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

F è singolare, e $\ker F = \langle (3, -1)^{\mathsf{T}} \rangle$. L'insieme delle soluzioni di $F\xi = b$ è

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} + \ker F = \left\{ \begin{bmatrix} 3\lambda + 1 \\ -\lambda \end{bmatrix}, \lambda \in \mathbb{R} \right\}$$

Quindi la migliore approssimazione... è

$$(3\lambda + 1) \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} - \lambda \begin{bmatrix} 3 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

indipendente da λ.

A Soluzione di un sistema di equazioni lineari nel senso dei minimi quadrati

Siano $\mathbb{R}^n$ con prodotto scalare canonico, $A = (a_1, \ldots, a_k) \in \mathbb{R}^{n\times k}$, $n \geq k$, e $b \in \mathbb{R}^n$. Gli elementi $\xi \in \mathbb{R}^k$ tali che

$$\text{per ogni } y \in \mathbb{R}^k \quad , \quad ||A\xi - b||^2 \leq ||Ay - b||^2$$

si chiamano *soluzioni del sistema $Ax = b$ nel senso dei minimi quadrati.* Tali elementi sono tutte e sole le soluzioni del sistema

$$A^{\mathsf{T}}Ax = A^{\mathsf{T}}b \tag{5.2}$$

Infatti, posto $V = \mathbb{R}^n$ con prodotto scalare canonico, $W = \langle a_1, \ldots, a_k\rangle$ e $v = b$, le soluzioni del sistema $Ax = b$ nel senso dei minimi quadrati sono *le coordinate* rispetto al sistema $a_1, \ldots, a_k$ della migliore approssimazione di v in W nel senso dei minimi quadrati, ed il sistema (5.2) è il sistema delle equazioni normali.

5.7 Osservazione

La matrice $A^{\mathsf{T}}A$ è simmetrica e per ogni $v \in \mathbb{R}^n$ si ha $A^{\mathsf{T}}Av \bullet v \geq 0$ (ossia $A^{\mathsf{T}}A$ è semidefinita positiva). Se le colonne di A sono linearmente indipendenti, $A^{\mathsf{T}}A$ è definita positiva — in particolare è invertibile (vedere il Problema 0.62).

5.8 Problema

Siano

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} \quad , \quad b = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

e

$$\widehat{A} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 2 \end{bmatrix} \quad , \quad \widehat{b} = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$$

Verificare (calcolandole) che la soluzione nel senso dei minimi quadrati di $Ax = b$ e quella di $\widehat{A}x = \widehat{b}$ sono diverse.

Si osservi che il secondo sistema si ottiene dal primo moltiplicando per 2 l'ultima equazione. △

5.9 Problema

Siano

$$A = \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \quad , \quad b = \begin{bmatrix} 2 \\ 1 \\ 0 \\ -1 \end{bmatrix}$$

Calcolare la soluzione $\hat{x}$ del sistema $Ax = b$ nel senso dei minimi quadrati e il *residuo quadratico* per $\hat{x}$, cioè: $||A\hat{x} - b||^2$. Come si spiega quest'ultimo risultato? △

5.10 Problema

Siano $A \in \mathbb{R}^{n\times k}$, $n \geq k$, e $b \in \mathbb{R}^n$. Provare che se A ha rango massimo, la soluzione del sistema $Ax = b$ nel senso dei minimi quadrati è

$$\hat{x} = (A^{\mathsf{T}}A)^{-1}\,A^{\mathsf{T}}b$$

e che, nel caso $n = k$, si ha

$$(A^{\mathsf{T}}A)^{-1}\,A^{\mathsf{T}} = A^{-1}$$

cioè: la soluzione nel senso dei minimi quadrati coincide con la soluzione del sistema. La matrice $(A^{\mathsf{T}}A)^{-1}\,A^{\mathsf{T}}$ si chiama *pseudoinversa* di A. △

B Approssimazione di dati nel senso dei minimi quadrati

Siano k un intero non negativo, $[a, b]$ un intervallo non degenere, e $\mathcal{G}$ un sottospazio vettoriale di $\mathcal{C}([a, b], \mathbb{R})$ di dimensione $j \leq k$.

Assegnati $x_0, \ldots, x_k \in [a, b]$, $y_0, \ldots, y_k \in \mathbb{R}$, il problema di *determinare gli elementi di $\mathcal{G}$ che meglio approssimano i dati $(x_0, y_0), \ldots, (x_k, y_k)$ nel senso dei minimi quadrati* consiste nel determinare le $g(x) \in \mathcal{G}$ che rendono *minima* la quantità

$$(g(x_0) - y_0)^2 + \cdots + (g(x_k) - y_k)^2$$

5.11 Esempio

Si considerino i dati

x_j	0	1	2	3
y_j	1	2	−1	0

Determinare gli elementi di $\langle 1, x\rangle$ che meglio approssimano i dati nel senso dei minimi quadrati.

Soluzione

Si cercano le $g(x)$ della forma $a_0 + a_1x$ tali che

$$\Big(g(0) - 1\Big)^2 + \Big(g(1) - 2\Big)^2 + \Big(g(2) + 1\Big)^2 + \Big(g(3)\Big)^2$$

risulti minimo. Si ha

$$\Big(g(0)-1\Big)^2+\Big(g(1)-2\Big)^2+\Big(g(2)+1\Big)^2+\Big(g(3)\Big)^2 = ||\begin{bmatrix} g(0)\\ g(1)\\ g(2)\\ g(3)\end{bmatrix} - \begin{bmatrix} 1\\ 2\\ -1\\ 0\end{bmatrix}||^2$$

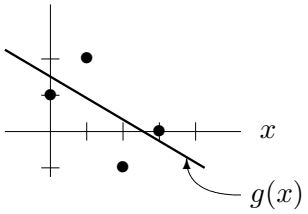

Figura 21.

(prodotto scalare canonico in $\mathbb{R}^4$) e

$$\begin{bmatrix} g(0) \\ g(1) \\ g(2) \\ g(3) \end{bmatrix} = a_0 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \ a_1 \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \end{bmatrix}$$

Dunque, posto

$$w_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad , \quad w_2 = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \end{bmatrix} \quad , \quad v = \begin{bmatrix} 1 \\ 2 \\ -1 \\ 0 \end{bmatrix}$$

si cercano $a_0, a_1 \in \mathbb{R}$ tali che $||a_0 w_1 + a_1 w_2 - v||^2$ risulti minimo.

Posto $V = \mathbb{R}^4$ con prodotto scalare canonico e $W = \langle w_1, w_2 \rangle$, si cerca la migliore approssimazione di v in W nel senso dei minimi quadrati.

Dalle equazioni normali

$$\begin{bmatrix} 4 & 6 \\ 6 & 14 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

si ottiene $a_1 = 7/5, a_2 = -3/5$ e l'elemento ... (vedere la Figura 21) è $g(x) = \frac{7}{5} - \frac{3}{5}x$.

5.12 Problema

Si considerino i dati dell'Esempio precedente. Si scriva il sistema di equazioni da studiare per determinare gli elementi di $\langle 1, x \rangle$ che interpolano i dati. Dopo aver verificato che il sistema non ammette soluzioni, determinare quelle nel senso dei minimi quadrati. Confrontare le soluzioni trovate con quella dell'Esempio precedente. $\triangle$

Sia $g_1, \ldots, g_j$ una base di $\mathcal{G}$. L'Esempio e il Problema precedenti mostrano che un elemento $g = a_1 g_1 + \cdots + a_j g_j$ è una migliore approssimazione in $\mathcal{G}$ dei dati $(x_0, y_0), \ldots, (x_k, y_k)$ nel senso dei minimi quadrati se e solo se le

coordinate $(a_1, \ldots, a_j)^{\mathsf{T}}$ di g rispetto alla base sono una soluzione nel senso dei minimi quadrati del sistema

$$\begin{bmatrix} g_1(x_0) & \cdots & g_j(x_0) \\ \vdots & & \vdots \\ g_1(x_k) & \cdots & g_j(x_k) \end{bmatrix} \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_j \end{bmatrix} = \begin{bmatrix} y_0 \\ \vdots \\ y_k \end{bmatrix}$$

Si osservi l'analogia con quanto detto per il problema lineare dell'interpolazione (si veda la Sezione B del Capitolo 4).

C Approssimazione con polinomi trigonometrici

Se $w_1, \ldots, w_k$ è una base ortonormale di W, la matrice del sistema delle equazioni normali è I, e la migliore approssimazione di v in W nel senso dei minimi quadrati è $(v \bullet w_1)w_1 + \cdots + (v \bullet w_k)w_k$. I numeri $v \bullet w_j, j = 1, \ldots, k$, si dicono in tal caso *coefficienti di Fourier* di v rispetto alla base ortonormale $w_1, \ldots, w_k$ (vedere la dimostrazione del Teorema 0.43).

5.13 Definizione

Sia $V = \mathcal{C}([0, 2\pi], \mathbb{C})$, con $f \bullet g = \frac{1}{2\pi} \int_0^{2\pi} f(t)\overline{g(t)}\, dt$. Il sottospazio vettoriale di V

$$T_k(\mathbb{C}) = \langle e^{i\ell t} \text{ con } \ell \in \mathbb{Z}, |\ell| \leq k \rangle_{\mathbf{C}}$$

si chiama insieme dei *polinomi trigonometrici di grado* $\leq k$.

$T_k(\mathbb{C})$ è un sottospazio di dimensione $2k + 1$, e i generatori $e^{i\ell t}$ sono ortonormali.

I coefficienti di Fourier di $v \in V$ sono

$$v \bullet e^{i\ell t} = \frac{1}{2\pi} \int_0^{2\pi} v(t) e^{-i\ell t} dt$$

e la migliore approssimazione di v in $T_k(\mathbb{C})$ è

$$\sum_{\ell=-k}^{k} (v \bullet e^{i\ell t})\, e^{i\ell t}$$

5.14 Esempio

Sia $v = t^2 \in V$; la migliore approssimazione in $T_1(\mathbb{C})$ nel senso dei minimi quadrati è $(v \bullet 1) + (v \bullet e^{it})\, e^{it} + (v \bullet e^{-it})\, e^{-it}$. Si ha

$$\begin{aligned} v \bullet 1 &= \tfrac{1}{2\pi} \textstyle\int_0^{2\pi} t^2 \, dt = \tfrac{4}{3}\pi^2 \\ v \bullet e^{it} &= \tfrac{1}{2\pi} \textstyle\int_0^{2\pi} t^2 e^{-it} \, dt = 2 + 2\pi i \\ v \bullet e^{-it} &= \tfrac{1}{2\pi} \textstyle\int_0^{2\pi} t^2 e^{it} \, dt = \overline{v \bullet w_2} = 2 - 2\pi i \end{aligned}$$

Quindi, la migliore approssimazione è

$$\frac{4}{3}\pi^2 + (2 + 2\pi i)\, e^{it} + (2 - 2\pi i)\, e^{-it} = \frac{4}{3}\pi^2 + 4\cos t - 4\pi \operatorname{sen} t.$$

5.15 Osservazione

Siano V come nella Definizione 5.13, k un intero non negativo e $v \in V$. Per $j = -k, \ldots, k$ sia

$$\widehat{c}_j = v \bullet e^{ijt}$$

Si verifica immediatamente che

(1) l'applicazione $v \to (\widehat{c}_{-k}, \ldots, \widehat{c}_k)^{\mathsf{T}}$ è lineare da V in $\mathbb{C}^{2k+1}$

(2) se $v(t) \in \mathbb{R}$ per ogni $t \in [0, 2\pi]$ allora $\widehat{c}_0 \in \mathbb{R}$ e, per $j \neq 0$, $\widehat{c}_{-j}$ è il coniugato di $\widehat{c}_j$ e quindi

$$\sum_{j=-k}^{k} \widehat{c}_j\, e^{ijt} \in \mathbb{R}$$

per ogni $t \in [0, 2\pi]$.

5.16 Osservazione

Sia V come nella Definizione 5.13. Poiché per $k = 0, 1, \ldots$ si ha $T_k(\mathbb{C}) \subset T_{k+1}(\mathbb{C})$, dato $v \in V$ la successione di reali non negativi

$$||v - \textstyle\sum_{\ell=-k}^{k} (v \bullet e^{i\ell t})\, e^{i\ell t}||$$

$k = 0, 1, \ldots$, è monotona decrescente e quindi *convergente.*

Si lascia al lettore il compito di approfondire lo studio del limite della successione (consultare un testo in cui si tratta di Serie di Fourier).

5.17 Osservazione

Lo spazio vettoriale V della Definizione 5.13 non ha dimensione finita e quindi l'ipotesi del Teorema 0.11 non è verificata. La convergenza di una successione in V rispetto alla norma dedotta dal prodotto interno *non* garantisce convergenza rispetto ad altre norme.

Si consideri, ad esempio, la norma in V definita da

$$||v||_\infty = \max\{\, |v(t)|, t \in [0, 2\pi] \,\}$$

Per la successione

$$s_k(t) = \begin{cases} 1 - kt & \text{per } t \in [0, 1/k] \\ 0 & \text{per } t \in [1/k, 2\pi] \end{cases}$$

si ha

$$\lim_{k\to\infty} ||s_k|| = 0$$

ma

$$\lim_{k\to\infty} ||s_k||_\infty = 1$$

Si lascia al lettore il compito di approfondire lo studio della successione

$$||v - \textstyle\sum_{\ell=-k}^{k} (v \bullet e^{i\ell t})\, e^{i\ell t}||_\infty$$

(consultare un testo in cui si tratta di Serie di Fourier).

D Minimi quadrati e fattorizzazione QR

Si consideri $\mathbb{R}^n$ con prodotto scalare canonico e siano $A = (a_1, \ldots, a_k), b$ come nella Sezione A, con $a_1, \ldots, a_k$ linearmente indipendenti. La fattorizzazione introdotta nella definizione seguente consente di determinare la soluzione del sistema $Ax = b$ nel senso dei minimi quadrati risolvendo un sistema, equivalente a quello delle equazioni normali, con matrice triangolare.

5.18 Definizione (fattorizzazione QR)
Sia $A = (a_1, \ldots, a_k) \in \mathbb{R}^{n\times k}, n \geq k$.
Una coppia $U \in \mathbb{R}^{n\times k}, T \in \mathbb{R}^{k\times k}$ si dice *fattorizzazione* QR di A se

(a) le colonne di U sono ortonormali (ossia U è ortogonale)

(b) T è triangolare superiore

(c) $A = UT$

Si osservi che se le colonne di A sono linearmente indipendenti, la matrice T è non singolare.*

Un metodo per calcolare una fattorizzazione QR di una matrice A si ottiene applicando il procedimento di ortonormalizzazione di Gram–Schmidt alle colonne di A.

5.19 Esempio
Sia $A = (a_1, a_2, a_3) \in \mathbb{R}^{4\times 3}$, e supponiamo (per semplicità) linearmente indipendenti le colonne di A. Il procedimento di ortonormalizzazione di Gram–Schmidt fornisce

$$\omega_1 = a_1 \quad , \quad \omega_2 = a_2 - d_{21}\omega_1 \quad , \quad \omega_3 = a_3 - d_{31}\omega_1 - d_{32}\omega_2$$

con

$$d_{21} = \frac{a_2 \bullet \omega_1}{||\omega_1||^2} \quad , \quad d_{31} = \frac{a_3 \bullet \omega_1}{||\omega_1||^2} \quad , \quad d_{32} = \frac{a_3 \bullet \omega_2}{||\omega_2||^2}$$

*Infatti: $Az = 0$ se e solo se $z = 0$ e quindi $UTz = 0$ se e solo se $z = 0$. Siccome $UTz = 0$ se e solo se $Tz = 0 \ldots$

ovvero

$$a_1 = \omega_1 \quad , \quad a_2 = \omega_2 + d_{21}\omega_1 \quad , \quad a_3 = \omega_3 + d_{31}\omega_1 + d_{32}\omega_2$$

che si riscrivono

$$(a_1, a_2, a_3) = (\omega_1, \omega_2, \omega_3) \begin{bmatrix} 1 & d_{21} & d_{31} \\ 0 & 1 & d_{32} \\ 0 & 0 & 1 \end{bmatrix}$$

La fattorizzazione di A così ottenuta non è quella richiesta perché le colonne $\omega_1, \omega_2, \omega_3$ sono ortogonali ma non di norma unitaria. Per ottenere quanto si vuole si pone $N = \mathrm{diag}(||\omega_1||, ||\omega_2||, ||\omega_3||)$ e

$$U = (\omega_1, \omega_2, \omega_3)N^{-1} \quad , \quad T = N \begin{bmatrix} 1 & d_{21} & d_{31} \\ 0 & 1 & d_{32} \\ 0 & 0 & 1 \end{bmatrix}$$

5.20 Problema

Sia

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix} \in \mathbb{R}^{4\times 3}$$

Calcolare una fattorizzazione QR di A. La fattorizzazione calcolata è l'unica fattorizzazione QR possibile per A?

5.21 Osservazione

Sia $A \in \mathbb{R}^{n\times k}, n \geq k$, a colonne linearmente indipendenti e $b \in \mathbb{R}^n$.

(a) Il procedimento di ortonormalizzazione di Gram–Schmidt prova che esiste almeno una fattorizzazione QR di A.

Se U, T è una fattorizzazione QR, per ogni scelta di $s_1, \ldots, s_k \in \{-1, 1\}$, posto $S = \mathrm{diag}(s_1, \ldots, s_k)$, anche US^{-1}, ST lo è.

(b) Sia U, T una fattorizzazione QR di A. Il sistema (con matrice triangolare)

$$Tx = U^{\mathsf{T}}b$$

è equivalente al sistema delle equazioni normali per $Ax = b$.

(Infatti: $A^{\mathsf{T}}A = T^{\mathsf{T}}U^{\mathsf{T}}UT = T^{\mathsf{T}}T, A^{\mathsf{T}}b = T^{\mathsf{T}}U^{\mathsf{T}}b$ e, essendo T non singolare ...)

(c) Sia U, T una fattorizzazione QR di A. Utilizzando la norma relativa al prodotto scalare canonico in $\mathbb{R}^n$, per il numero di condizionamento della matrice delle equazioni normali per $Ax = b$ si ha

$$\mu(A^{\mathsf{T}}A) = \mu(T^{\mathsf{T}}T) = \rho(T^{\mathsf{T}}T)\,\rho(\,(T^{\mathsf{T}}T)^{-1}\,)$$

(si veda il Problema 3.26: la matrice $T^{\mathsf{T}}T$ è simmetrica). Il numero di condizionamento della matrice del sistema $Tx = U^{\mathsf{T}}b$ (equivalente al sistema delle equazioni normali) è invece

$$\mu(T) = \sqrt{\rho(T^{\mathsf{T}}T)}\,\sqrt{\rho(\,(T^{\mathsf{T}}T)^{-1}\,)} = \sqrt{\mu(A^{\mathsf{T}}A)}$$

(infatti, tenuto conto che $\rho(\,(T\,T^{\mathsf{T}})^{-1}\,) = \rho(\,(T^{\mathsf{T}}T)^{-1}\,)$,[†] si ha $||T^{-1}|| = \sqrt{\rho((T^{\mathsf{T}}T)^{-1})}\,$).

Quindi, il numero di condizionamento di T è (quasi certamente) *più basso* di quello della matrice delle equazioni normali per $Ax = b$.

La fattorizzazione QR di una matrice A può essere utilizzata, analogamenta alla fattorizzazione LR, per la soluzione del sistema di equazioni $Az = b$.

5.22 Osservazione

Siano $A \in \mathbb{C}^{n\times n}$, invertibile, U, T una fattorizzazione QR di A e $b \in \mathbb{C}^n$. Allora:

(a) Il sistema $Az = b$ è equivalente al sistema (con matrice triangolare) $Tz = U^{\mathsf{H}}b$.

(b) Poiché $T = U^{\mathsf{H}}A$, allora

$$||T||_2 = ||U^{\mathsf{H}}A||_2 = ||A||_2$$

inoltre $T^{-1} = A^{-1}U$ e quindi, usando la definizione di norma 2

$$||T^{-1}||_2 = ||A^{-1}U||_2 = ||A^{-1}||_2$$

dunque

$$\mu_2(T) = \mu_2(A)$$

Si confronti questo risultato con quanto detto a proposito della fattorizzazione LR nel Paragrafo A-3 del Capitolo 3.

(c) Poiché $||U||_\infty \leqslant \sqrt{n}$ e $||U||_1 \leqslant \sqrt{n}$ (usare Esercizio 0.39) si ha

$$\mu_\infty(T) \leqslant n\,\mu_\infty(A) \qquad \text{e} \qquad \mu_1(T) \leqslant n\,\mu_1(A)$$

[†]Per $A, B \in \mathbb{R}^{n\times n}$ invertibili si ha infatti: $AB - \lambda I = A(BA - \lambda I)A^{-1}$ e quindi AB e BA hanno lo stesso polinomio caratteristico e dunque lo stesso raggio spettrale.

Riferimenti

[A] T.M. Apostol: *Calcolo*, Boringhieri, Torino, 1977.

[A1] T.M. Apostol: *Calculus*, Xerox Corp., Waltham, 1969.

[BBCM] R. Bevilacqua, D. Bini, M. Capovani, O. Menchi: *Introduzione alla Matematica Computazionale*, Zanichelli, Bologna, 1987.

[C] F. Conti: *Calcolo*, McGraw–Hill, Milano, 1993.

[GGM] P. Ghelardoni, G. Gheri, P. Marzulli: *Elementi di Calcolo Numerico*, Masson, Milano, 1995.

[GLV] S. Guerra, G. Lombardi, G. Vincenti: *Calcolo Numerico*, Felici, Pisa, 1996.

[GM] G. Ghelardoni, P. Marzulli: *Argomenti di Analisi Numerica*, ETS, Pisa, 1979.

[L] S. Lang: *Algebra Lineare*, Boringhieri, Torino, 1970.

[S] V.I. Smirnov: *Corso di Matematica Superiore*, Editori Riuniti, Roma, 1981.

[IEEE1] ANSI/IEEE Std 754–1985 — IEEE Standard for Binary Floating–Point Arithmetic.

[IEEE2] ANSI/IEEE Std 854–1987 — IEEE Standard for Radix–Independent Floating–Point Arithmetic.

[IEC] ISO/IEC 10967 Information Technology — Language Independent Arithmetic — Part 1, 2, 3.

[PM] J.G. Proakis, D.M. Manolakis: *Digital Signal Processing*, 3rd edition, Prentice Hall, New Jersey, 1996.

Notazioni

$\mathbf{N}$	l'insieme dei numeri naturali $(0, 1, \ldots)$
$\mathbf{Z}$	l'insieme dei numeri interi
$\mathbf{Q}$	l'insieme dei numeri razionali
$\mathbf{R}$	l'insieme dei numeri reali
$\mathbf{C}$	l'insieme dei numeri complessi
$\overline{z}$	il coniugato di $z \in \mathbb{C}$
re z	la parte reale di $z \in \mathbb{C}$
v^{T}, A^{T}	il vettore trasposto di v, la marice trasposta di A
v^{H}, A^{H}	il vettore trasposto coniugato di v, la matrice trasposta coniugata di A
$\langle v_1, \ldots, v_n \rangle$	il sottospazio vettoriale di V generato da $v_1, \ldots, v_n \in V$
$\langle v_1, \ldots, v_n \rangle_{\mathbb{K}}$	in uno spazio vettoriale V su $\mathbb{K}$ [$\mathbb{K} = \mathbb{R}$ o $\mathbb{K} = \mathbb{C}$], il sottospazio vettoriale generato da $v_1, \ldots, v_n \in V$
I_n	la matrice identica di ordine n
$\mathcal{C}(\Omega, \mathbb{R})$	con $\Omega \subset \mathbb{R}$, l'insieme delle funzioni continue da Ω in $\mathbb{R}$
$\mathcal{C}^k(\Omega, \mathbb{R})$	con $\Omega \subset \mathbb{R}$ [$\Omega \subset \mathbb{R}^n$] e $k \in \mathbb{N}$, l'insieme delle funzioni continue da Ω in $\mathbb{R}$ con derivata k-esima continua [derivate parziali k-esime continue] in Ω.
$\mathcal{C}^k(\Omega, \mathbb{R}^n)$	con $\Omega \subset \mathbb{R}^n$ e $k \in \mathbb{N}$, l'insieme delle funzioni da Ω in $\mathbb{R}^n$ con componenti in $\mathcal{C}^k(\Omega)$
$\mathcal{C}(\Omega, \mathbb{C})$	con $\Omega \subset \mathbb{R}$, l'insieme delle funzioni continue da Ω in $\mathbb{C}$
$\|\|v\|\|$	in uno spazio con prodotto scalare [con prodotto hermitiano], indica il numero $\sqrt{v \bullet v}$
$u \perp v$	in uno spazio con prodotto scalare [con prodotto hermitiano], è sinonimo di $u \bullet v = 0$
$\hat{a}_1, \ldots, \hat{a}_n$	le righe della matrice $A \in \mathbb{C}^{n \times n}$
mis I	con $I \subset \mathbb{R}$ intervallo, indica la misura dell'intervallo I
V_L	lo spazio vettoriale su $\mathbb{R}$ dei vettori geometrici nel piano
$P_n(\mathbb{R})[P_n(\mathbb{C})]$	lo spazio vettoriale dei polinomi a coefficienti reali [a coefficienti complessi] di grado $\leq n$.
$e_1, \ldots, e_n$	gli elementi della base canonica di $\mathbb{R}^n$ [di $\mathbb{C}^n$].