Schede riassuntive di Analisi Numerica

Analisi dell'errore

Se $\tilde{x}$ è un'approssimazione di x definiamo $\tilde{x} - x$ come l'*errore assoluto*, $\varepsilon = (\tilde{x} - x)/x$ come l'*errore relativo* (rispetto a x) e $\eta = (x - \tilde{x})/\tilde{x}$ come l'*errore relativo* (rispetto a $\tilde{x}$). In particolare vale che $\tilde{x} = x(1 + \varepsilon) = x/(1 + \eta)$.

Rappresentazione in base

Sia $B \geq 2$ un numero intero, vale allora il seguente risultato:

Teorema (di rappresentazione in base)**.** Per ogni numero reale $x \neq 0$ esistono e sono unici un intero p ed una successione $\{d_i\}_{i\geq 1}$ con le seguenti proprietà:

(1.) $0 \leq d_i \leq B - 1$

(2.) $d_1 \neq 0$

(3.) per ogni $k > 0$ esiste un $j \geq k$ tale che $d_j \neq B - 1$ (ossia $\{d_i\}_{i\geq 1}$ è frequentemente diversa da $B - 1$)

(4.) x si scrive come:

$$x = \operatorname{sgn}(x) B^p \sum_{i\geq 1} d_i B^{-i}$$

L'intero B è detto *base della rappresentazione*, gli interi d_i sono dette *cifre della rappresentazione* ed il numero $\sum_{i\geq 1} d_i B^{-i}$ viene chiamato *mantissa*. Si scrive $\mathtt{p}(x)$ per indicare l'esponente p relativo a x.

La condizione (2.) è detta di *normalizzazione* e serve a garantire l'unicità della rappresentazione e a memorizzare il numero in maniera più efficiente, mentre la condizione 3 esclude rappresentazioni con cifre uguali a $B - 1$ da un certo punto in poi (per esempio la rappresentazione di 1 come $0.\overline{9}$ è esclusa).

Numeri *floating point*

Dati gli interi $B \geq 2$, $t \geq 1$ ed M, $m > 0$, si definisce l'insieme $\mathcal{F}(t, B, m, M)$ dei **numeri di macchina** o **dei numeri in virgola mobile** o ancora dei **numeri *floating point*** come l'insieme:

$$\{0\} \cup \{\pm B^p \sum_{i}^{t} d_i B^{-i} | d_1 \neq 0, 0 \leq d_i \leq B - 1, -m \leq p \leq M\},$$

in particolare B rappresenta la base della rappresentazione, t la lunghezza della mantissa, B^{-m} il minimo numero che moltiplica la mantissa e B^M il massimo.

Troncamento su $\mathbb{R}$ e $\mathbb{R}^n$ e numero di macchina u

Sia x un numero reale. Se $-m \leq \mathtt{p}(x) \leq M$, allora x viene ben rappresentato in $\mathcal{F}$ dal numero $\tilde{x} = \mathtt{fl}(x)$, dove:

$$x = \operatorname{sgn}(x) B^p \sum_{i\geq 1} d_i B^{-i} \implies \tilde{x} = \operatorname{sgn}(x) B^p \sum_{i=1}^{t} d_i B^{-i},$$

ossia $\tilde{x}$ è ottenuto da x troncandone la mantissa al t-esimo termine (**troncamento**). Facendo così si ottiene un errore relativo di rappresentazione tale per cui:

$$\left|\frac{\tilde{x} - x}{x}\right| < B^{1-t}, \quad \left|\frac{\tilde{x} - x}{\tilde{x}}\right| < B^{1-t},$$

e si definisce pertanto $u := B^{1-t}$ come la **precisione di macchina**. In generale, se y è un'approssimazione di $x \in \mathbb{R}$ tale per cui $|(y - x)/x|$ è minore di B^{1-c} con c intero, si dice che y ha c *cifre significative*.

Se invece $\mathtt{p}(x) < -m$ o $\mathtt{p}(x) > M$ il numero non è rappresentabile e ci ritroviamo rispettivamente in una situazione di *underflow* o di *overflow*.

Dato $x \in \mathbb{R}^n$, si definisce la rappresentazione $\tilde{x}$ rispetto a x in $\mathcal{F}$ come:

$$\tilde{x} = (\tilde{x}_i) = (\mathtt{fl}(x_i) = (x_i(1 + \varepsilon_i)), \quad |\varepsilon_i| < u.$$

Precisione di un insieme $\mathcal{F}$

Dati due sistemi floating point in base B_1 e B_2 di t_1 e t_2 cifre, il primo è più preciso del secondo se e solo se vale:

$$B_1^{1-t_1} < B_2^{1-t_2} \iff t_1 > (t_2 - 1)\frac{\log B_2}{\log B_1} + 1$$

Si può definire in modo analogo l'approssimazione per arrotondamento, osservando che in questo caso l'errore relativo di rappresentazione è limitato da $\frac{1}{2}B^{1-t} = \frac{1}{2}u$.

Aritmetica di macchina

Siano a, $b \in \mathcal{F}(t, B, m, M)$ e sia $\mathtt{op}$ una delle quattro operazioni aritmetiche ($+$, $-$, $\cdot$ e $/$). Si consideri $c = a \,\mathtt{op}\, b$. Allora la macchina calcola c con l'approssimazione $\hat{c} := \mathtt{fl}(c)$. Vale pertanto che:

$$\hat{c} = \mathtt{fl}(a \,\mathtt{op}\, b) = c(1 + \delta) = c/(1 + \eta) \text{ con } |\delta|, |\eta| < u.$$

Gli errori relativi δ e η sono detti **errori locali** generati dall'operazione $\mathtt{op}$. Un'operazione del tipo $\mathtt{fl} \circ \mathtt{op}$ è detta **flop**.

Errori nel calcolo di una funzione

Si utilizza il simbolo $\doteq$ per indicare l'uguaglianza di termini al *primo ordine*. Si scrive $\varepsilon(x)$ per valutare l'errore nella variabile x.

Errore inerente

L'errore relativo di $f(\tilde{x})$ rispetto $f(x)$ viene detto **errore inerente** e vale:

$$\varepsilon_{\text{in}} = \frac{f(\tilde{x}) - f(x)}{f(x)}.$$

Tale errore misura il *condizionamento* di un problema, ossia quanto varia il risultato perturbando l'input.

Nel caso di funzioni $f : \mathbb{R}^n \to \mathbb{R}$ sufficientemente regolari vale che:

$$\varepsilon_{\text{in}} \doteq \sum_{i=1}^{n} C_i \delta_i,$$

dove:

$$\delta_i := \varepsilon_{x_i} = \frac{\tilde{x}_i - x_i}{x_i}, \quad C_i = x_i \frac{\partial f/\partial x_i(x)}{f(x)}.$$

I termini C_i sono detti **coefficienti di amplificazione** rispetto alla variabile x_i.

Un problema si dice **ben condizionato** se una piccola variazione nelle condizioni iniziali produce una piccola variazione in output e si dice **mal condizionato** altrimenti. Il condizionamento di un problema varia a seconda della grandezza in modulo dei coefficienti di amplificazione (più sono grandi, più il problema è mal condizionato).

Per le 4 operazioni aritmetiche elementari $x \,\mathtt{op}\, y$ sono presentati i relativi coeff. di amplificazione:

Operazione ($\mathtt{op}$)	C_1	C_2
addizione ($+$)	$x/(x + y)$	$y/(x + y)$
sottrazione ($-$)	$x/(x - y)$	$-y/(x - y)$
moltiplicazione ($\cdot$)	1	1
divisione ($/$)	1	-1

Funzioni razionali ed errore algoritmico

Si ricorda che una funzione f si dice *razionale* se si può esprimere tramite una formula *finita* che comprende le quattro operazioni aritmetiche.

Se la funzione $f : \mathbb{R}^n \to \mathbb{R}$ è razionale, si può agevolmente calcolarne l'approssimazione $\varphi = \tilde{f}$ tramite troncamento. Generalmente esistono più modi di implementare la priorità delle operazioni, e una tale scelta di priorità φ è detta *algoritmo*.

Il discostamento relativo di $\varphi(\tilde{x})$ da $f(\tilde{x})$ (il valore effettivamente calcolato sull'input perturbato) viene detto **errore algoritmico**, che pertanto è così definito:

$$\varepsilon_{\text{alg}} = \frac{\varphi(\tilde{x}) - f(\tilde{x})}{f(\tilde{x})}.$$

Un problema si dice **numericamente stabile** (in avanti) se $|\varepsilon_{\text{alg}}| < ku$ per k costante, e **numericamente instabile** (in avanti) se non lo è.

Funzioni regolari non razionali ed errore analitico

Se $f : \mathbb{R}^n \to \mathbb{R}$ non è razionale, ma è comunque sufficientemente regolare, si può approssimare f tramite una funzione razionale g. Chiamiamo **errore analitico** il discostamento relativo di g da f sull'input non perturbato (ossia a priori dell'uso di aritmetica di macchina):

$$\varepsilon_{\text{an}} = \frac{g(x) - f(x)}{f(x)}.$$

Questo errore può essere studiato attraverso gli strumenti dell'analisi matematica e della teoria dell'approssimazione delle funzioni (e.g. polinomi di Taylor e resti di Peano, Lagrange, Cauchy o resto integrale).

Errore totale

Definiamo l'**errore totale** come il discostamento relativo dell'algoritmo φ valutato sull'input perturbato dal valore esatto $f(x)$:

$$\varepsilon_{\text{tot}} = \frac{\varphi(\tilde{x}) - f(x)}{f(x)}.$$

Proposizione. A meno di termini non lineari vale sempre che:

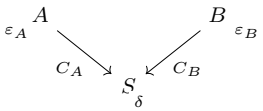
$$\varepsilon_{\text{tot}} \doteq \varepsilon_{\text{in}} + \varepsilon_{\text{alg}} + \varepsilon_{\text{an}}(\tilde{x}).$$

Pertanto, se la funzione f è razionale, $\varepsilon_{\text{tot}} \doteq \varepsilon_{\text{in}} + \varepsilon_{\text{alg}}$. Questa proposizione giustifica lo studio separato dei tre errori.

Calcolo dell'errore algoritmico e dell'errore totale (per funzioni razionali)

L'unica differenza tra lo studio dell'errore algoritmico e dell'errore totale (per funzioni razionali) secondo il metodo presentato sta nell'errore ε_{x_i} che viene dato alle variabili di input. Infatti, nel caso dell'errore algoritmico, le variabili sono già ben rappresentate come numeri di macchina e quindi non hanno errore, mentre nel caso dell'errore totale lo hanno. Pertanto, se una traccia richiede di calcolare un errore totale su dei numeri di macchina, ciò è equivalente a calcolarne l'errore algoritmico.

Per procedere allo studio dell'errore algoritmico conviene utilizzare dei grafi orientati con le op. aritmetiche segnando ogni flop come un nuovo nodo, a cui sono rivolte le frecce da tutte le variabili impiegate per portare a termine l'op. aritmetica. Il più semplice di questi grafi, che rappresenta $S = A \,\texttt{op}\, B$, è presentato di seguito:



In tal caso vale che:

$$\varepsilon_S = \delta + C_A\varepsilon_A + C_B\varepsilon_B.$$

Una volta ricavato l'errore, è sufficiente maggiorarlo in modulo con una certa funzione di u per determinare se l'algoritmo è stabile o meno.

Cancellazione numerica

In generale per la scelta di un algoritmo è consigliabile evitare di *sottrarre numeri dello stesso segno* o di *sommare numeri discordi*. Infatti, se la somma o la differenza di tali numeri è molto vicina a 0, i coefficienti di amplificazione hanno modulo molto maggiore di 1, e dunque il problema diviene *numericamente instabile*.

Analisi dell'errore all'indietro

Sia $f(x_1, \dots, x_n)$ una funzione razionale e si denoti con $\varphi(x_1, \dots, x_n)$ l'algoritmo definito su $\mathcal{F}^n$ i cui valori sono ottenuti calcolando $f(x_1, \dots, x_n)$ con l'aritmetica di macchina.

Nell'analisi *all'indietro* dell'errore si cercano delle perturbazioni δ_i tali che denotando $\hat{x}_i = x_i(1 + \delta_i)$ risulti:

$$\varphi(x_1, \dots, x_n) = f(\hat{x}_1, \dots, \hat{x}_n),$$

ovverosia valutare l'algoritmo φ su $(x_1, \dots, x_n)$ diviene equivalente a calcolare il valore esatto di $f(\hat{x}_1, \dots, \hat{x}_n)$.

Agendo in questo modo per calcolare l'errore algoritmico ε_{alg} è sufficiente calcolare in realtà l'errore inerente ε_{in} sulle perturbazioni δ_i.

Si riporta lo schema risolutivo di un tipico esercizio sull'analisi all'indietro:

- Si calcoli $f(\hat{x}_1, \dots, \hat{x}_n)$, dove $\hat{x}_i = x(1 + \delta_i)$ e i vari δ_i sono variabili libere.
- Si calcoli $\varphi(x_1, \dots, x_n)$ sostituendo a ogni j-esima flop $\texttt{fl}(\tilde{s_1} \,\texttt{op}\, \tilde{s_2})$ il valore esatto $(\tilde{s_1} \,\texttt{op}\, \tilde{s_2})(1 + \varepsilon_j)$ dove $|\varepsilon_j| < u$, procedendo poi a ritroso nelle flop di $\tilde{s_1}$ e di $\tilde{s_2}$.
- Si risolva il sistema $f(\hat{x}_1, \dots, \hat{x}_n) = \varphi(x_1, \dots, x_n)$ nei δ_i in funzione degli ε_j, maggiorandoli alla fine.

Un algoritmo si dice **numericamente stabile all'indietro** se è possibile effettuare un analisi all'indietro per cui $\varepsilon_{\text{alg}} < ku$ con k costante.

Algebra lineare numerica

Matrici di permutazione e matrici riducibili

Definizione. Si dice che $P \in \mathbb{C}^{n\times n}$ è una **matrice di permutazione** se esiste $\sigma \in S_n$ tale per cui $P^j = e_{\sigma(j)}$; e in tal caso si scrive $P_\sigma := P$ per rimarcare la permutazione a cui P è associata.

Si osserva in particolare che $\sigma \mapsto P_\sigma$ è un omomorfismo da S_n alle matrici di permutazioni, che dunque formano un gruppo moltiplicativo. Per le matrici di permutazioni valgono dunque le seguenti proprietà:

- $P_\sigma \in \mathbb{R}^{n\times n}$,
- $P_\sigma P_\tau = P_{\sigma\circ\tau}$,
- $P_\sigma^T P_\sigma = P_\sigma P_\sigma^T = I$ (P_σ è ortogonale e unitaria),
- $P_\sigma^{-1} = P_\sigma^T = P_{\sigma^{-1}}$,
- $\det(P_\sigma) = \text{sgn}(\sigma)$,
- P_σ rappresenta la matrice di cambio di base da quella canonica a quella canonica permutata secondo σ, ovverosia da $(e_1, \dots, e_n)$ a $\big(e_{\sigma(1)}, \dots, e_{\sigma(n)}\big)$,
- Se $A \in \mathbb{C}^{n\times n}$ e $B = P_\sigma^T A P_\sigma$, allora $b_{ij} = a_{\sigma(i)\sigma(j)}$.
- Analogamente, se $A \in \mathbb{C}^{n\times n}$ e $B = P_\sigma A P_\sigma^T = P_{\sigma^{-1}}^T A P_{\sigma^{-1}}$, allora $b_{ij} = a_{\sigma^{-1}(i)\sigma^{-1}(j)}$, e dunque $b_{\sigma(i)\sigma(j)} = a_{ij}$.

Definizione (Matrice riducibile)**.** Una matrice $A \in \mathbb{C}^{n\times n}$ si dice **riducibile** se esiste una matrice di permutazione P per cui:

$$PAP^\top = PAP^{-1} = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}$$

con A_{11} e A_{22} matrici quadrate, ovverosia se $PAP^\top$ è triangolare a blocchi.

Una matrice si dice **irriducibile** se non è riducibile.

Equivalentemente una matrice è riducibile se esiste un sottoinsieme proprio non vuoto I di $\{e_1, \dots, e_n\}$ tale per cui $A \cdot I \subseteq \text{Span}(I)$, ovverosia se I è A-invariante.

Grafo associato a una matrice

Data $A \in \mathbb{C}^{n\times n}$ si definisce **grafo associato** $G[A]$ ad A come il grafo orientato con le seguenti proprietà:

- $G[A]$ ha come nodi $\{1, \dots, n\}$,
- Esiste un arco orientato da i a j se e solo se $a_{ij} \neq 0$.

Un cammino è una sequenza di archi $i_1 \to i_2 \to \cdots i_m$. Un grafo si dice **fortemente connesso** se $\forall\, i, j$ esiste un cammino da i a j.

Se P è una matrice di permutazione, $B := PAP^\top$ ha lo stesso grafo di A a meno di rinominare i nodi secondo σ, ovverosia esiste un arco da $\sigma(i)$ a $\sigma(j)$ in $G[B]$ se e solo se esiste un arco da i a j in $G[A]$. Pertanto il grafo di A è fortemente connesso se e solo se quello di $PAP^\top$ lo è. Questo fatto risulta fondamentale nel dimostrare il seguente teorema:

Teorema. *A è irriducibile se e solo se il suo grafo associato è fortemente connesso. Equivalentemente, A è riducibile se e solo se il suo grafo associato non è fortemente connesso.*

Teoremi di Gershgorin e applicazioni

Sia $A \in \mathbb{C}^{n\times n}$. Allora si definisce l'$i$-esimo **cerchio di Gershgorin** come l'insieme $K_i \subset \mathbb{C}$ tale per cui:

$$K_i = \{z \in \mathbb{C} : |z - a_{ii}| \leq r_i\},$$

dove

$$r_i = \sum_{j=1,\, j\neq i}^{n} |a_{ij}|$$

è detto i-esimo **raggio di Gershgorin**. In particolare K_i è un cerchio in $\mathbb{C} \cong \mathbb{R}^2$ di centro a_{ii} e raggio r_i.

Primo teorema di Gershgorin

Teorema (Gershgorin I)**.** Data $A \in \mathbb{C}^{n\times n}$, gli autovalori di A appartengono all'insieme $\bigcup_{i=1}^{n} K_i$, ossia per ogni autovalore λ di A esiste un i tale per cui λ appartiene all'i-esimo cerchio K_i di Gershgorin.

Inoltre, se v è un autovettore relativo all'autovalore λ, $\lambda \in K_h$, dove $h = \operatorname{argmax} |v_i|$, ovverosia h è l'indice tra gli i per cui $|v_i|$ è massimo.

Dal momento che $A^\top$ è simile ad A, $A^\top$ e A condividono gli stessi autovalori. Si può dunque rafforzare la tesi del teorema cercando gli autovalori nell'intersezione delle unioni dei cerchi relativi ad A e a $A^\top$, ovverosia, se H_i rappresenta l'i-esimo cerchio di Gershgorin di $A^\top$, ogni autovalore di A appartiene a:

$$\left(\bigcup_{i=1}^{n} K_i\right) \cap \left(\bigcup_{i=1}^{n} H_i\right).$$

Per lo stesso motivo è utile coniugare A per similitudine, specie se si coniuga A utilizzando matrici diagonali, dacché il coniugio per matrice diagonale, oltre a non alterare gli autovalori, lascia invariati i centri dei cerchi e riscalare i raggi.

In particolare, se $D = \operatorname{diag}(d_1, \dots, d_n)$, allora $DAD^{-1} = (d_i a_{ij} d_j^{-1})$ e $D^{-1}AD = (d_i^{-1} a_{ij} d_j)$. Pertanto vale il seguente teorema:

Teorema. Sia $A \in \mathbb{C}^{n\times n}$. Allora gli autovalori di A appartengono all'insieme:

$$\bigcap_{d\in D} \bigcup_{i=1}^{n} \left\{ z \in \mathbb{C} : |z - a_{ii}| \le \frac{1}{|d_i|} \sum_{j=1,\, j\neq i}^{n} |a_{ij}|\,|d_j| \right\},$$

dove D è un sottoinsieme non vuoto di $\mathbb{C}^n$.

Matrici fortemente dominanti diagonali

Una matrice A si dice **fortemente dominante diagonale** se $|a_{ii}| > r_i = \sum_{j=1,\, j\neq i}^{n} |a_{ij}|$ $\forall i = 1, ..., n$. Per il primo teorema di Gershgorin, A non può essere singolare (0 non appartiene per ipotesi a nessun cerchio). Ogni sottomatrice principale di una matrice fortemente diagonale in senso stretto è allo stesso modo fortemente diagonale. In particolare, se A è fortemente dominante diagonale, allora $a_{ii} \neq 0$.

Si possono definire in modo analogo le matrici dominanti diagonali (in senso debole) cambiando $>$ in $\geq$, ma non è detto che tali matrici siano non singolari. Un controesempio è infatti la matrice:

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

Secondo teorema di Gershgorin

Teorema (Secondo teorema di Gershgorin)**.** Sia $A \in \mathbb{C}^{n\times n}$ e sia $K = \bigcup_{i=1}^{n} K_i$ l'unione dei cerchi K_i relativi ad A. Sia inoltre

$$K = M_1 \cup M_2 \text{ con } M_1 \cap M_2 = \varnothing,$$

dove M_1 è costituito da n_1 cerchi e M_2 è costituito da n_2 cerchi. Allora M_1 contiene n_1 autovalori e M_2 ne contiene n_2.

Nota. La dimostrazione di questo teorema sfrutta un argomento di continuità riguarda al segmento $A(t) = D + t(A - D)$ con $t \in [0, 1]$ e $D = \operatorname{diag}(A)$. Un simile argomento di continuità può risultare utile in altri contesti.

Quest'ultimo teorema risulta utile per dimostrare che una matrice $A \in \mathbb{R}^{n\times n}$ ha un autovalore reale. Se infatti K_i è un cerchio di A disgiunto dagli altri, allora K_i contiene un unico autovalore, che deve essere dunque necessariamente reale (altrimenti, siccome A è reale, vi sarebbe anche il suo coniugato, e quindi vi sarebbero almeno 2 autovalori, assurdo).

Terzo teorema di Gershgorin

Teorema (Terzo teorema di Gershgorin)**.** Sia λ un autovalore di A. Si ipotizzi che val le seguenti condizioni

(1.) A è irriducibile,
(2.) $\lambda \in K_i \implies \lambda \in \partial K_i$ (dove ∂K_i è la frontiera di K_i).

Allora $\lambda \in \bigcap \partial K_i$.

Matrici irriducibilmente dominanti diagonali (i.d.d.)

Una matrice $A \in \mathbb{C}^{n\times n}$ si dice **irriducibilmente dominante diagonale** se

(i) A è irriducibile,
(ii) A è dominante diagonale <u>in senso debole</u>,
(iii) $\exists\, h$ t.c. $|a_{hh}| > r_h = \sum_{j=1,\, j\neq h}^{n} a_{hj}$ (ossia con $0 \notin K_h$).

Per il terzo teorema di Gershgorin le matrici i.d.d. sono non singolari. Inoltre, se A è i.d.d., allora $a_{ii} \neq 0$ per ogni i.

Forma normale di Schur

Teorema. Data $A \in \mathbb{C}^{n\times n}$, esiste una matrice $U \in \mathbb{C}^{n\times n}$ unitaria, ovverosia tale per cui $U^H U = UU^H = I_n$, tale che:

$$U^H A U = T \in T_+(n, \mathbb{C}),$$

ovverosia con $U^H AU$ triangolare superiore. Si dice in tal caso che T è una **forma normale di Schur** di A.

Generalmente esistono più forme normali di Schur per una matrice, e possono essere ottenute riordinando gli autovalori e/o le basi scelte.

Teorema spettrale

Teorema (spettrale)**.** Se A è hermitiana, allora una sua forma normale di Schur è sempre una matrice diagonale con elementi reali, ovverosia A è ortogonalmente diagonalizzabile con autovalori reali.

Se invece A è anti-hermitiana ($A^H = -A$), allora una sua forma normale di Schur è una matrice diagonale con elementi immaginari puri, ovverosia A è ortogonalmente diagonalizzabile con autovalori immaginari puri.

Caratterizzazione delle matrici normali

Definizione. Una matrice $A \in \mathbb{C}^{n\times n}$ si dice **normale** se $AA^H = A^H A$.

Si enuncia una caratterizzazione delle matrici triangolari normali:

Proposizione. Una matrice triangolare T è normale se e solo se è diagonale.

Dalla precedente proposizione si ottiene facilmente la seguente fondamentale altra caratterizzazione:

Teorema. Una matrice $A \in \mathbb{C}^{n\times n}$ è normale se e solo se la sua forma normale di Schur è diagonale (i.e. è ortogonalmente diagonalizzabile).

Autovalori di una matrice unitaria

Gli autovalori di una matrice unitaria A hanno modulo 1. Se infatti v è un autovettore relativo all'autovalore λ, vale che:

$$Av = \lambda v \implies v^H A^H = \overline{\lambda} v^H,$$

da cui:

$$v^H v = v^H A^H A v = \overline{\lambda} v^H A v = \overline{\lambda}\lambda v^H v \implies |\lambda| = 1.$$

In particolare, se A è reale, $\lambda \in \{\pm 1\}$.

Forma normale di Schur reale

Definizione (Matrice quasi-triangolare superiore)**.** Una matrice $T \in \mathbb{R}^{n\times n}$ si dice **quasi-triangolare superiore** se si può scrivere nella forma:

$$\begin{pmatrix} T_{11} & \dots & T_{1n} \\ & \ddots & \vdots \\ & & T_{mm} \end{pmatrix},$$

dove i blocchi matriciali T_{ii} possono essere matrici 2×2 oppure matrici 1×1, ossia numeri reali.

Si definisce analogamente la nozione di matrice quasi-triangolare *inferiore*.

Gli autovalori di una matrice quasi-triangolare T sono gli autovalori delle sottomatrici T_{ii}. Il polinomio caratteristico di T è il prodotto dei polinomi dei blocchi T_{ii}. Il determinante di T è il prodotto dei determinanti dei blocchi T_{ii}, la traccia è la somma delle tracce dei blocchi T_{ii}.

Usando la nozione di matrice quasi-triangolare si può enunciare un teorema analogo a quello della forma normale di Schur, ma riadattato per le matrici reali:

Teorema (Forma normale di Schur reale)**.** Per ogni matrice $A \in \mathbb{R}^{n\times n}$ esistono $T \in \mathbb{R}^{n\times n}$ quasi-triangolare superiore e $Q \in O(n)$ tali per cui:

$$Q^t A Q = T.$$

Norme di vettori

Una **norma** (vettoriale) su $\mathbb{C}^n$ è un'applicazione

$$\|\cdot\| : \mathbb{C}^n \longrightarrow \mathbb{R}$$

tale per cui:

(1.) $\forall\, x \in \mathbb{C}^n$, $\|x\| \geq 0$ e $\|x\| = 0 \iff x = 0$ (definitezza positiva).

(2.) $\forall\, \alpha \in \mathbb{C}$, $\forall\, x \in \mathbb{C}^n$, $\|\alpha x\| = |\alpha|\, \|x\|$ (omogeneità).

(3.) $\forall\, x,\, y \in \mathbb{C}^n$, $\|x + y\| \leq \|x\| + \|y\|$ (disuguaglianza triangolare).

Ogni prodotto hermitiano $\langle \cdot, \cdot \rangle$ definito positivo di $\mathbb{C}^n$ induce una norma ponendo $\|v\| = \sqrt{\langle v, v \rangle}$.

Per $p \in [1, \infty)$ si dice **norma di Hölder** di ordine p la norma:

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}.$$

Alcuni esempi noti di norme di Hölder sono:

- $\|x\|_1 = \sum_{i=1}^n |x_i|$, la norma 1,
- $\|x\|_2 = \left(\sum_{i=1}^n x_i^2\right)^{\frac{1}{2}} = \sqrt{x^H x}$, detta anche *norma euclidea*, indotta dal prodotto hermitiano standard di $\mathbb{C}^n$,
- $\|x\|_\infty := \lim_{p \to \infty} \|x\|_p = \max_i |x_i|$.

Per una norma l'insieme $S = \{x \in \mathbb{C}^n \mid \|x\| \leq 1\}$ è un insieme convesso. Da ciò si deduce che per $0 < p < 1$ l'espressione (generalizzata) $\|x\|_p$ non è una norma.

Data una norma $\|\cdot\|$ su $\mathbb{C}^n$ e data $S \in \mathbb{C}^{n \times n}$ non singolare, allora esiste la norma (indotta da) S, definita in modo tale che:

$$\|\cdot\|_S : x \mapsto \|Sx\|.$$

Inoltre vale la seguente caratterizzazione:

Proposizione. Una norma $\|\cdot\|$ è indotta da un prodotto scalare se e solo se vale la *legge del parallelogramma*, ossia se e solo se:

$$\|x + y\|^2 + \|x - y\|^2 = 2(\|x\|^2 + \|y\|^2)$$

A partire da questo risultato, si deduce che le norme 1 e ∞ non sono indotte da nessun prodotto scalare.

Ogni norma è una **funzione uniformemente continua**, cioè vale che, $\forall\, \varepsilon > 0$, esiste $\delta > 0$ tale per cui:

$$\forall\, x, y \in \mathbb{C}^n, \|x - y\|_2 < \delta \implies |\|x\| - \|y\|| < \varepsilon,$$

espressione che segue dal fatto che una norma è Lipschitziana (deriva dalla *disuguaglianza triangolare*). Equivalentemente vale che per ogni $\varepsilon > 0$ esiste $\delta > 0$ tale per cui:

$$\forall\, x, y \in \mathbb{C}^n, |x_i - y_i| < \delta \implies |\|x\| - \|y\|| \leq \varepsilon.$$

A partire da questo risultato si può dimostrare il seguente fondamentale teorema:

Teorema (Equivalenza tra norme)**.** Per ogni coppia di norme $\|\cdot\|'$ e $\|\cdot\|''$ su $\mathbb{C}^n$, esistono due costanti positive α e β (dipendenti da n) tali per cui:

$$\alpha \|x\|' \leq \|x\|'' \leq \beta \|x\|' \quad \forall\, x \in \mathbb{C}^n.$$

In altre parole, le norme su $\mathbb{C}^n$ sono bi-lipschitziane tra loro.

Infatti $S = \{x \in \mathbb{C}^n \mid \|x\|_2 = 1\}$ è un compatto euclideo (essendo chiuso e limitato, per il teorema di Heine-Borel), e $\|\cdot\|$ è continua sulla topologia euclidea. Pertanto tutte le norme di $\mathbb{C}^n$ inducono la stessa topologia, ossia quella indotta da $\|\cdot\|_2$, e quindi coincidono tutti gli aperti e i chiusi.

In particolare per le norme 1, 2 e ∞ vale che:

- $\|x\|_\infty \leq \|x\|_1 \leq n \|x\|_\infty$,
- $\|x\|_2 \leq \|x\|_1 \leq \sqrt{n} \|x\|_2$,
- $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n} \|x\|_\infty$.

Se U è una matrice unitaria, allora U agisce lasciando invariata la norma 2, ovverosia $\|Ux\|_2 = \|x\|_2$.

Norme di matrici

Una **norma matriciale** è un'applicazione

$$\|\cdot\| : \mathbb{C}^{n \times n} \to \mathbb{R}$$

per la quale valgono le seguenti quattro proprietà:

(1.) $\forall\, A \in \mathbb{C}^{n \times n}$, $\|A\| \geq 0$ e $\|A\| = 0 \iff A = 0$ (definitezza positiva),

(2.) $\forall\, \lambda \in \mathbb{C}$, $\forall\, A \in \mathbb{C}^{n \times n}$, $\|\lambda A\| = |\lambda|\, \|A\|$ (omogeneità),

(3.) $\|A + B\| \leq \|A\| + \|B\|$ $\forall\, A, B \in \mathbb{C}^{n \times n}$ (disuguauglianza triangolare),

(4.) $\|AB\| \leq \|A\|\, \|B\|$ $\forall\, A,\, B \in \mathbb{C}^{n \times n}$ (proprietà submoltiplicativa).

Norma di Frobenius

Si definisce la **norma di Frobenius** l'applicazione che agisce su $A \in \mathbb{C}^{n \times n}$ in modo tale che:

$$\|A\|_F = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{\frac{1}{2}} = \sqrt{\mathrm{tr}(A^H A)}.$$

In particolare la norma di Frobenius è esattamente la norma euclidea dello spazio $\mathbb{C}^{n \times n}$ immerso in $\mathbb{C}^{n^2}$ a cui è naturalmente isomorfo associando ad A il vettore $(A_i^\top)_i$ ottenuto trasponendo le righe e sovrapponendole ordinatamente.

Norme matriciali indotte

Nota. Si ricorda che, data una norma $\|\cdot\|$, l'insieme $S = \{x \in \mathbb{C}^n \text{ t.c. } \|x\| = 1\}$ è chiuso e limitato (in tutte le topologie indotte da norme, essendo equivalenti). Allora, per il teorema di Heine-Borel, S è compatto, e dunque una funzione continua $f : S \to \mathbb{R}$ ammette massimo per il teorema di Weierstrass.

Nota. Una mappa lineare da $\mathbb{C}^n$ in sé è continua. Dunque è possibile applicare il teorema di Weierstrass alla funzione $\|\cdot\| \circ f_A$ ristretta su S, dove f_A è l'app. indotta da una matrice A.

Definizione. Data una norma vettoriale $\|\cdot\| : \mathbb{C}^n \to \mathbb{R}$, si definisce la **norma matriciale indotta** (o la corrispondente *norma operatore*), come la norma che agisce su $A \in \mathbb{C}^{n \times n}$ in modo tale che:

$$\|A\| := \max_{\|x\|=1} \|Ax\| = \max_{x \in S} \|Ax\| = \max_{x \in \mathbb{C}^n \setminus \{0\}} \frac{\|Ax\|}{\|x\|}.$$

Per ogni norma operatore valgono le due seguenti aggiuntive proprietà, *oltre quelle caratterizzanti una norma*:

(1.) $\|Ax\| \leq \|A\|\, \|x\|$, $\forall\, A \in \mathbb{C}^{n \times n}$, $\forall\, x \in \mathbb{C}^n$,

(2.) $\|I\| = 1$.

Poiché $\|I\|_F = \sqrt{n}$, la norma di Frobenius non è in generale indotta da alcuna norma vettoriale.

Per le norme 1, 2, ∞ valgono le seguenti identità:

- $\|A\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^{n} |a_{ij}| = \max_{j=1,\dots,n} \left\|A^j\right\|_1$,
- $\|A\|_\infty = \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}| = \max_{i=1,\dots,n} \|A_i\|_1$,
- $\|A\|_2 = \sqrt{\rho(A^H A)}$, dove $\rho(A^H A)$ è l'autovalore di valore assoluto maggiore di $A^H A$ (*raggio spettrale*).

Si osserva immediatamente che la norma matriciale 2 è computazionalmente più difficile da calcolare rispetto alle norme 1 e ∞. Inoltre, se A è simmetrica, le norme 1 e ∞ coincidono.

Poiché $\|A\|_F = \sqrt{\mathrm{tr}(A^H A)}$ è la radice della somma degli autovalori di $A^H A$, vale in particolare che $\|A\|_F \geq \|A\|_2$. Si osserva che $A^H A$ è semidefinita positiva, e dunque i suoi autovalori sono non negativi.

Se U è unitaria e $B = UA$, vale che:

$$B^H B = (UA)^H UA = A^H U^H UA = A^H A,$$

e quindi B e A condividono la stessa norma di Frobenius e la stessa norma 2. Analogamente si vede che AU e A condividono le stesse due norme. Equivalentemente, la moltiplicazione per matrice unitaria (sia a destra che a sinistra) lascia invariata sia la norma 2 che quella di Frobenius. In particolare, se U e V sono unitarie, $\|UAV\|_2 = \|A\|_2$ e $\|UAV\|_F = \|A\|_F$.

Se A è normale, allora gli autovalori di $A^H A$ sono i moduli quadrati degli autovalori di A. Pertanto, per A normale, $\|A\|_2 = \max_{\lambda \in \mathrm{sp}(A)} |\lambda| = \rho(A)$ e $\|A\|_F = \sqrt{\sum_{\lambda \in \mathrm{sp}(A)} |\lambda|^2}$. Si è usato che la forma normale di Schur di A è diagonale e che le trasformazioni unitarie non variano né $\|A\|_2$ né $\|A\|_F$.

Valgono inoltre le seguenti altre due disuguaglianze:

- $\|A\|_F \leq \sqrt{r}\, \|A\|_2 \leq \sqrt{n}\, \|A\|_2$, dove $r = \mathrm{rg}(A)$,
- $\|A\|_2^2 \leq \|A\|_1 \cdot \|A\|_\infty$.

Norma indotta da una matrice non singolare S

Sia $\|\cdot\|$ una norma vettoriale e $S \in \mathbb{C}^{n\times n}$ una matrice non singolare. Allora, data la norma vettoriale $\|x\|_S := \|Sx\|$, vale che:

$$\|A\|_S = \max_{\|x\|_S=1} \|Ax\|_S = \max_{\|x\|_S=1} \|SAx\| = \max_{\|Sx\|=1} \left\|SAS^{-1}Sx\right\|,$$

e quindi, sfruttando che S è non singolare:

$$\|A\|_S = \max_{\|y\|=1} \left\|SAS^{-1}y\right\|,$$

ossia vale che:

$$\|A\|_S = \left\|SAS^{-1}\right\|.$$

Norme e raggi spettrali

Definizione. Data $A \in \mathbb{C}^{n\times n}$, si definisce **raggio spettrale** $\rho(A)$ il modulo dell'autovalore massimo di A, ossia:

$$\rho(A) = \max\{|\lambda| \mid \lambda \text{ autovalore di } A\}.$$

Se x è un autovettore relativo a $\rho(A)$ di modulo unitario rispetto a una norma $\|\cdot\|$, allora vale che:

$$\|Ax\| = \|\rho(A)x\| = \rho(A)\|x\| = \rho(A),$$

e dunque $\rho(A) \leq \|A\|$ per ogni A. In particolare $\|A\| \geq |\lambda|$ per ogni autovalore λ.

Si enunciano inoltre i seguenti teoremi:

Teorema. Sia $A \in \mathbb{C}^{n\times n}$. Allora per ogni $\varepsilon > 0$ esiste una norma indotta $\|\cdot\|$ tale per cui:

$$\rho(A) \leq \|A\| \leq \rho(A) + \varepsilon.$$

Inoltre, se gli autovalori di modulo massimo di A appartengono solo a blocchi di Jordan di taglia 1, allora esiste una norma per cui $\|A\| = \rho(A)$.

Teorema. Sia $\|\cdot\|$ una norma matriciale. Allora vale la seguente identità:

$$\lim_{k\to+\infty} \left\|A^k\right\|^{\frac{1}{k}} = \rho(A).$$

Se $A \in \mathbb{C}^{n\times n}$ è tale per cui $\|A\| < 1$ dove $\|\cdot\|$ è una norma matriciale indotta, allora 1 non può essere autovalore di A, e dunque $I - A$ è invertibile. Inoltre vale che:

$$\left\|(I-A)^{-1}\right\| \leq \frac{1}{1-\|A\|}.$$

Condizionamento di un sistema lineare e numero di condizionamento

Data $A \in \mathbb{C}^{n\times n}$ non singolare e dato un vettore $b \in \mathbb{C}^n \setminus \{0\}$ si vuole studiare il condizionamento del problema $Ax = b$, ossia di un sistema lineare.

Consideriamo il problema $(A + \delta_A)y = b + \delta_b$, dove perturbiamo il sistema originale mediante dei parametri δ_A e δ_B di cui conosciamo i rapporti $\frac{\|\delta_b\|}{\|b\|}$ e $\frac{\|\delta_A\|}{\|A\|}$, cercando di ottenere informazioni riguardo $\frac{\|\delta_x\|}{\|x\|}$, sostituendo $y = x + \delta_x$.

Definizione. Si dice **numero di condizionamento** $\mu(A)$ di A nella norma $\|\cdot\|$ il valore:

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

Si scrive $\mu_p(A)$ per intendere $\|A\|_p \left\|A^{-1}\right\|_p$.

Studiamo in particolare la perturbazione di $Ax = b$ nel caso in cui $\delta_A = 0$.

Proposizione. Se $\delta_A = 0$, allora vale la seguente disuguaglianza:

$$\frac{\|\delta_x\|}{\|x\|} \leq \|A\| \left\|A^{-1}\right\| \frac{\|\delta_b\|}{\|b\|} = \mu(A)\frac{\|\delta_b\|}{\|b\|}.$$

In generale, per $\delta_A \neq 0$ e $A + \delta_A$ invertibile, vale che:

$$\varepsilon_x \leq \frac{\|A\|\left\|A^{-1}\right\|}{1 - \varepsilon_A \|A\| \left\|A^{-1}\right\|}(\varepsilon_A + \varepsilon_B),$$

dove $\varepsilon_t = \|\delta_t\| / \|t\|$.

Pertanto il sistema è **ben condizionato** se $\mu(A)$ è relativamente piccolo.

Per il numero di condizionamento valgono le seguenti proprietà:

- $\mu(A) \geq \|I\|$ per la *proprietà submoltiplicativa*,
- $\mu(A) \geq 1$ per $\|\cdot\|$ norma operatore ($\|I\| = 1$),
- $\mu_2(U) = 1$ per U unitaria.

Per A normale vale la seguente identità:

$$\mu_2(A) = \frac{\max_{\lambda\in \mathrm{sp}(A)} |\lambda|}{\min_{\lambda\in \mathrm{sp}(A)} |\lambda|}.$$

Inoltre vale la seguente disuguaglianza:

$$\mu_2(A) \leq \mu(A),$$

dove $\mu(A)$ è riferito a qualsiasi altra norma operatore.

Metodi diretti per sistemi lineari

Ci si propone di risolvere il sistema $Ax = b$ scrivendo A come prodotto di matrici "consone" e facilmente invertibili. Se infatti $A = PQ$, allora il sistema $PQx = b$ può essere risolto come:

$$\begin{cases} Py = b \\ Qx = y \end{cases}$$

risolvendo dunque prima $Py = b$ e poi $Qx = y$.

Risoluzione di $Ax = b$ per A triangolare o unitaria

Sono presentati di seguito tre tipi di matrice per le quali l'invertibilità è garantita e per cui il sistema $Ax = b$ è facilmente risolvibile.

Se A è **triangolare inferiore** con $a_{ii} \neq 0\ \forall i$, allora, per risolvere $Ax = b$, è possibile applicare il *metodo di sostituzione in avanti* ponendo:

$$x_1 = \frac{b_1}{a_1}, \quad x_i = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j\right) \text{ con } i \geq 2.$$

Se A è **triangolare superiore** con $a_{ii} \neq 0\ \forall i$, allora, per risolvere $Ax = b$, è possibile applicare il *metodo di sostituzione all'indietro* ponendo:

$$x_n = \frac{b_n}{a_n}, \quad x_i = \frac{1}{a_{ii}}\left(b_i - \sum_{j=i+1}^{n} a_{ij}x_j\right) \text{ con } i < n.$$

I due algoritmi presentati hanno un costo computazionale di n^2 flops e sono entrambi numericamente stabili all'indietro.

Se A è **unitaria**, $Ax = b \implies x = A^H b$, e dunque si verifica che:

$$x_j = \sum_{i=1}^{n} \overline{a_{ij}} b_i.$$

Questo algoritmo richiede un costo computazionale di $2n^2 - n$ flops ed è ancora numericamente stabile all'indietro.

Fattorizzazione classiche di matrici

In letteratura esistono 4 fattorizzazioni classiche:

1. $A = LU$ con L triangolare inferiore con tutti 1 sulla diagonale e U triangolare superiore (possibile solo per alcune classi di matrici);
2. $A = PLU$ con L, U come sopra e P matrice di permutazione (sempre possibile);
3. $A = P_1LUP_2$ con L, U come sopra e P_1, P_2 matrici di permutazione (sempre possibile);
4. $A = QR$ con Q unitaria e R triangolare superiore (sempre possibile).

Condizioni per l'esistenza e l'unicità di una fattorizzazione LU

Definizione. Si dicono **sottomatrici principali di testa** le sottomatrici di A di cui prendiamo le prime k righe e colonne. Quando si utilizza l'aggettivo *proprie*, si esclude A stessa.

Vale la seguente condizione sufficiente per l'esistenza e l'unicità di una fattorizzazione LU:

Proposizione. Se tutte le sottomatrici principali di testa proprie sono non singolari allora esiste ed è unica la fattorizzazione LU di A.

Se non sono verificate le ipotesi può comunque esistere una fattorizzazione LU. Per esempio:

$$\begin{pmatrix} 0 & 1 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}.$$

Se A è invertibile, allora la non singolarità delle sottomatrici principali di testa diventa una condizione necessaria oltre che sufficiente per l'esistenza e l'unicità della fattorizzazione LU.

Le seguenti classi di matrici ammettono sempre un'unica fattorizzazione LU:

- Matrici fortemente dominanti diagonali, le cui sottomatrici principali di testa sono fortemente dominanti diagonali e dunque invertibili,
- Matrici hermitiane definite positive (criterio di Sylvester),
- Matrici hermitiane definite positive su $\mathrm{Span}(e_1, \dots, e_{n-1})$ e semidefinite positive su $\mathbb{C}^n$ (metodo di Jacobi, criterio di Sylvester).

Matrici elementari

Dati $\sigma \in \mathbb{C}$, u, $v \in \mathbb{C}^n$ si dice **matrice elementare** una matrice M del tipo:

$$M = I - \sigma u v^H.$$

Si osserva che uv^H è della seguente forma:

$$uv^H = (u_i \overline{v_j})_{ij}.$$

Inoltre $\mathrm{rg}(uv^H) \leq \mathrm{rg}(u) \leq 1$. Se u o v sono nulli uv^H è necessariamente nullo. Non si deve confondere uv^H con $u^H v$, che invece è il prodotto hermitiano complesso computato su u e v.

Un vettore x è autovettore di M se:

- $x = u \implies Mu = (1 - \sigma(v^H u))u$, e dunque u è relativo all'autovalore $1 - \sigma(v^H x)$;
- x t.c. $v^H x = 0$ (x è ortogonale a v) $\implies Mx = x$, il cui relativo autovalore è 1.

Supponiamo che $\sigma(v^H u)$ sia diverso da 0. La traccia di M è:

$$\mathrm{tr}(M) = \mathrm{tr}(I) - \sigma \cdot \mathrm{tr}(uv^H)) = (n-1) + (1 - \sigma v^H u).$$

Dal momento che $v^\perp$ ha dimensione $n-1$ – il prodotto hermitiano è positivo definito –, allora $\mu_g(1) \geq n-1$.

Osservando allora che $\mathrm{tr}(M)$ è la somma degli autovalori di M e che $\mu_g(1 - \sigma(v^H u)) \geq 1$, si conclude che gli unici autovalori di M sono proprio 1, con molteplicità algebrica e geometrica $n-1$, e $1 - \sigma(v^H u)$, con molteplicità 1. In particolare M è sempre diagonalizzabile e vale la seguente proposizione:

Proposizione. Se $\sigma(v^H u) \neq 0$, gli unici autovettori di M sono i vettori ortogonali a v, che sono punti fissi, e i multipli di u.

M è non singolare se e solo se 0 non è autovalore, ossia se e solo se:

$$1 - \sigma(v^H u) \neq 0 \iff \sigma(v^H u) \neq 1.$$

Se M è non singolare, allora vale che:

$$M^{-1} = I - \tau u v^H \text{ con } \tau = \frac{-\sigma}{1 - \sigma v^H u},$$

e dunque anche M^{-1} è una matrice elementare.

Proposizione. Per ogni x e $y \in \mathbb{C}^n \setminus \{0\}$ esiste M matrice elementare con $\det M \neq 0$ tale che $Mx = y$. In particolare è sufficiente scegliere v non ortogonale sia a x che a y e porre $\sigma u = \frac{x-y}{v^H x}$.

Matrici elementari di Gauss e applicazione alla fattorizzazione LU

Definizione. Dato $x \in \mathbb{C}^n$ con $x_1 \neq 0$ si definisce la relativa **matrice elementare di Gauss** come la matrice elementare:

$$M = I - u e_1^T, \quad u^T = \frac{1}{x_1}(0, x_2, \dots, x_n).$$

Vale sempre $Mx = x_1\, e_1$. Inoltre una matrice elementare di Gauss è sempre invertibile, e $M^{-1} = I + u e_1^T$. Vale sempre $\det(M) = 1$. Una matrice elementare di Gauss corrisponde a uno step di annichilimento degli elementi sotto il *pivot* dell'algoritmo di Gauss.

In particolare:

$$M = \begin{pmatrix} 1 & 0 & \dots & 0 \\ -u_2 & 1 & 0 & \\ \vdots & 0 & \ddots & \\ -u_n & 0 & \dots & 1 \end{pmatrix} \text{ e } M^{-1} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ u_2 & 1 & 0 & \\ \vdots & 0 & \ddots & \\ u_n & 0 & \dots & 1 \end{pmatrix}.$$

Algoritmo (Fattorizzazione LU con le matrici di Gauss)**.** Sia $A \in \mathbb{C}^{n \times n}$ che soddisfa l'esistenza e unicità della fattorizzazione LU. Se M_1 è la matrice elementare di Gauss associata alla prima colonna di A, allora:

$$M_1 A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ 0 & & \\ \vdots & A_1 & \\ 0 & & \end{pmatrix}.$$

Se $A_1 = (a_{ij}^{(1)})_{i,j>1}$, allora $a_{ij}^{(1)} = a_{ij} - \frac{a_{i1}}{a_{11}} a_{1j}$. Se $\hat{M}_2$ è la matrice elementare di Gauss che annulla la prima colonna di A_1, si definisce M_2 in modo tale che:

$$M_2 = \begin{pmatrix} 1 & 0 \\ 0 & \hat{M}_2 \end{pmatrix}.$$

Iterando questo procedimento si ottiene il seguente prodotto:

$$M_{n-1} \dots M_1 A = U, \quad M_1^{-1} \dots M_{n-1}^{-1} = L,$$

dove L è triangolare inferiore con 1 sulla diagonale e U è triangolare superiore.

In particolare vale che $\mathrm{diag}(U) = (a_{11}, a_{22}^{(1)}, \dots, a_{nn}^{(n-1)})$ e la j-esima colonna di $L = (\ell_{ij})$ è la j-esima colonna di M_j cambiata di segno. Sono espresse di seguito le relazioni di ricorrenza:

$$\begin{cases} a_{ij}^{(k+1)} = a_{ij}^{(k)} + m_{ij}^{(k)} a_{jk}^{(k)} & i,j = k+1, \dots, n \\ m_{ik}^{(k)} = -\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} & i = k+1, \dots, n \\ \ell_{ij} = -m_{ij}^{(j)} & i \geq j \end{cases}$$

Nel caso della risoluzione di un sistema lineare $Ax = b$, l'algoritmo può essere esteso aggiornando b ad ogni step; in particolare $b^{(k+1)} = M_k b^{(k)}$ dove M_k è la k-esima matrice di Gauss. In tal caso vale che:

$$b_i^{(k+1)} = b_i^{(k)} + m_{ik}^{(k)} b_k^{(k)}, \quad i = k+1, \dots, n.$$

Il costo totale dell'algoritmo è di $\frac{2}{3}n^3 + O(n^2)$ operazioni aritmetiche. Se A è anche tridiagonale il costo è $O(n)$.

Riguardo al precedente algoritmo vale la seguente proposizione:

Proposizione. Sia $A \in \mathbb{C}^{n \times n}$ una matrice con sottomatrici principali proprie di testa non singolari per cui esiste ed è unica la fattorizzazione $A = LU$. Se $\tilde{L}$ e $\tilde{U}$ i valori effettivamente calcolati di L e U tramite il precedente algoritmo in aritmetica di macchina e sia $\Delta_A = A - \tilde{L}\tilde{U}$. Allora vale elemento per elemento di Δ_A la seguente disuguaglianza:

$$|\Delta_A| \leq 2nu \left(|A| + |\tilde{L}|\,|\tilde{U}| \right) + O(u^2),$$

dove $|T| := (|t_{ij}|)$.

Se $\tilde{y}$ è la soluzione del sistema $\tilde{L}y = b$ calcolato realmente in aritmetica di macchina mediante l'algoritmo di sostituzione in avanti e $\tilde{x}$ è il vettore effettivamente calcolato risolvendo $\tilde{U}x = \tilde{y}$ mediante sostituzione all'indietro, allora vale che:

$$(A + \hat{\Delta}_A)\tilde{x} = b,$$

con

$$|\hat{\Delta}_A| \leq 4nu \left(|A| + |\tilde{L}|\,|\tilde{U}| \right) + O(u^2),$$

dove $|T| := (|t_{ij}|)$.

In particolare questa proposizione mostra che i valori $\tilde{L}$ e $\tilde{U}$ formano una decomposizione LU di una perturbazione di A, e dunque è possibile effettuare un'analisi all'indietro dell'algoritmo di Gauss. Se i valori in modulo di A sono troppo grandi, ci si può aspettare un mal funzionamento in senso numerico del metodo di eliminazione gaussiana (anche per il calcolo della soluzione $\tilde{x}$).

Strategia di *pivoting* parziale per la decomposizione PLU

Permutando i pivot è sempre possibile fornire una fattorizzazione del tipo PLU. La **strategia del *pivoting parziale*** (*massimo pivot parziale*) consiste nel scegliere al passo k come pivot il termine $a_{hk}^{(k)}$ con $h \geq k$ tale che $|a_{hk}^{(k)}| \geq |a_{ik}^{(k)}|$ per $i = k, \dots, n$. In questo modo se $a_{hk}^{(k)} = 0$, allora tutta la parte di colonna è nulla e si può procedere allo step successivo dell'algoritmo; altrimenti si può scambiare la riga h-esima con la riga k-esima e applicare poi l'algoritmo di Gauss.

In questo modo si sta moltiplicando per una matrice di permutazione P relativa alla trasposizione $\tau = (h, k)$. Nel caso in cui $a_{hk}^{(k)} \neq 0$ vale allora che:

$$A_{k+1} = M_k(P_k A_k).$$

In particolare vale sempre $m_{ij}^k \leq 1$ per $i \geq j$.

Si osserva inoltre che:

$$M_k P_k = P_k \tilde{M}_k,$$

dove, se $M_k = I - ue_q^T$, allora $\tilde{M}_k = I - P_k u e_q^T$ (infatti $\tilde{M}_k = P_k^T M_k P_k$). Applicando questa strategia ad ogni passo si ottiene allora una fattorizzazione $PA = LU$ dove P è un'opportuna matrice di permutazione, e dunque si ottiene una fattorizzazione PLU.

Matrici elementari di Householder e applicazioni alla fattorizzazione QR

Definizione. Si definiscono **matrici elementari di Householder** le matrici elementari della forma:

$$M = I - \beta u u^H,$$

con $u \in \mathbb{C}^n$ e $\beta \in \mathbb{R}$.

Se $u \neq 0$ e $\beta = 0$ o $\beta = 2/(u^H u) = 2/\langle u, u \rangle$, allora una matrice di Householder è unitaria. Inoltre una matrice di Householder è sempre hermitiana.

Proposizione. Per ogni $x \in \mathbb{C}^n \setminus \{0\}$ esiste M matrice elementare di Householder tale per cui $Mx = \alpha e_1$ per un dato $\alpha = \theta \|x\|_2$, dove:

$$\theta = \begin{cases} \pm 1 & \text{se } x_1 = 0, \\ \pm \frac{x_1}{|x_1|} & \text{se } x_1 \neq 0. \end{cases}$$

Tale matrice M si ottiene ponendo
$u^t = x - \alpha e_1 = (x_1 - \alpha, x_2, \dots, x_n)$ e $\beta = \frac{2}{u^H u}$.

Per evitare le cancellazioni nell'implementazione del calcolo di una matrice di Householder, e dunque migliorare la stabilità numerica, è consigliato scegliere sempre $\theta = -\frac{x_1}{|x_1|}$ per $x_1 \neq 0$ (nel caso in cui $x_1 = 0$, la scelta è indiffernete).

Algoritmo. Applicando la stessa filosofia dell'algoritmo di Gauss si possono utilizzare le matrici di Householder per calcolare la fattorizzazione QR di una matrice A. Sia $u^{(k)}$ il vettore relativo alla k-esima matrice di Householder. Allora vale che:

$$u_i^{(k)} = \begin{cases} 0 & \text{se } i < k, \\ a_{kk}^{(k)} \left(1 + \frac{\sqrt{\sum_{i=k}^n \left| a_{ik}^{(k)} \right|^2}}{\left| a_{kk}^{(k)} \right|} \right) & \text{se } i = k, \\ a_{ik}^{(k)} & \text{altrimenti.} \end{cases}$$

mentre il parametro $\beta^{(k)}$ della stessa matrice è così dato:

$$\beta^{(k)} = 2 \Big/ \sum_{i=k}^n \left| u_i^{(k)} \right|^2$$

A partire da questi, si ottengono i termini di A_{k+1}:

$$a_{i,j}^{(k+1)} = \begin{cases} a_{ij}^{(k)} & \text{se } j < k, \text{se } i \leq k, \\ 0 & \text{se } j = k, \text{ e } i > k, \\ a_{ij}^{(k)} - \beta^{(k)} u_i^{(k)} \sum_{r=k}^n \overline{u}_r^{(k)} a_{rj}^{(k)} & \text{se } i \geq k, j > k. \end{cases}$$

Nel caso della risoluzione di un sistema lineare $Ax = b$, l'algoritmo può essere esteso aggiornando b ad ogni step; in particolare $b^{(k+1)} = M_k b^{(k)}$ dove M_k è la k-esima matrice di Householder. In tal caso vale che:

$$b_i^{(k+1)} = b_i^{(k)} - \beta^{(k)} u_i^{(k)} \sum_{r=k}^n \overline{u}_r^{(k)} b_r^{(k)}, \quad i = k, \dots, n.$$

Il costo totale è di $\frac{4}{3}n^3 + O(n^2)$ operazioni, il *doppio* di quello dell'algoritmo gaussiano.

Riguardo a quest'ultimo algoritmo vale la seguente proposizione:

Proposizione. Sia $A \in \mathbb{C}^{n \times n}$ e sia $\tilde{R}$ la matrice triangolare superiore ottenuta in aritmetica di macchina applicando il metodo di Householder con l'algoritmo dato in precedenza. Sia inoltre $\tilde{x}$ la soluzione ottenuta risolvendo in aritmetica di macchina il sistema $Ax = b$ con il metodo di Householder, aggiornando le entrate di b. Allora $\tilde{x}$ risolve il sistema $(A + \Delta_A)\tilde{x} = b + \delta_b$, dove:

$$\|\Delta_A\|_F \leq u(\gamma n^2 \|A\|_F + n \left\| \tilde{R} \right\|_F) + O(u^2),$$

$$\|\delta_b\|_2 \leq \gamma n^2 u \|b\|_2 + O(u^2),$$

dove γ è una costante positiva. Inoltre $\left\| \tilde{R} \right\|_F$ differisce da $\left\| \tilde{R} \right\|_F = \left\| \tilde{A} \right\|_F$ ($A = QR$) per un termine proporzionale a u, e dunque:

$$\|\Delta_A\|_F \leq \gamma u(n^2 + n) \|A\|_F + O(u^2).$$

In particolare, il metodo di Householder è numericamente stabile.

Metodi stazionari iterativi per sistemi lineari

Definizione. Dato il sistema lineare $Ax = b$ con $A \in \mathbb{C}^{n \times n}$ si definisce un generico **partizionamento additivo di** A una partizione della forma $A = M - N$ con $\det M \neq 0$.

A partire da ciò possiamo riscrivere equivalentemente il sistema come $Mx = Nx + b$, il quale riconduce alla seguente scrittura equivalente del sistema originale:

$$x = Px + q, \quad P = M^{-1}N,\ q = M^{-1}b.$$

Questa formulazione del sistema viene detta **problema di punto fisso**. Una volta fissato $x^{(0)} \in \mathbb{C}^n$ si può generare in modo naturale una successione di vettori:

$$x^{(k+1)} = Px^{(k)} + q.$$

Se tale successione ammette limite x^*, allora il punto limite x^* è un punto fisso di $Px + q$, ed è dunque l'unica soluzione di $Ax = b$. Questo metodo è detto **metodo iterativo stazionario**. La matrice P ed il vettore q sono indipendenti da k, e quindi la matrice P viene detta **matrice di iterazione** del metodo.

Si dice che il metodo iterativo è **convergente** se per ogni scelta del vettore $x^{(0)}$ la successione $x^{(k)}$ converge alla soluzione x^* del sistema.

Se A è invertibile, definiamo la successione $e^{(k)} = x^{(k)} - x^*$ dove x^* è soluzione del sistema. Allora $e^{(k+1)} = Px^{(k)} + Q - (Px^* + Q) = Pe^{(k)}$. A partire da questa osservazione si dimostra il seguente teorema:

Teorema. Se esiste una norma di matrice indotta $\|\cdot\|$ tale che $\|P\| < 1$, allora la matrice A è invertibile. In tal caso il metodo iterativo è convergente
($\lim_{k\to\infty} \left\|e^{(k)}\right\| \leq \lim_{k\to\infty} \|P\|^k \left\|e^{(0)}\right\| = 0$).

Teorema. Il metodo iterativo è convergente e $\det A \neq 0$ se e solo se $\rho(P) < 1$.

Il quoziente $\left\|e^{(k)}\right\| / \left\|e^{(k-1)}\right\|$ rappresenta la riduzione dell'errore al passo k-esimo del metodo iterativo rispetto alla norma scelta. Se si considera la media geometrica $\theta_k\left(e^{(0)}\right)$ delle prime k riduzioni, si ricava che:

$$\theta_k\left(e^{(0)}\right) = \left(\frac{\left\|e^{(1)}\right\|}{\left\|e^{(0)}\right\|} \cdots \frac{\left\|e^{(k)}\right\|}{\left\|e^{(k-1)}\right\|} \right)^{\frac{1}{k}} = \left(\frac{\left\|P^k e^{(0)}\right\|}{\left\|e^{(0)}\right\|} \right)^{\frac{1}{k}} \leq \left\|P^k\right\|^{\frac{1}{k}}.$$

Definizione. Si definisce **riduzione asintotica media per passo** di un metodo iterativo con errore iniziale $e^{(0)}$ il valore:

$$\theta(e^{(0)}) = \lim_{k\to\infty} \theta_k(e^{(0)}).$$

Tale riduzione rappresenta la velocità di convergenza del metodo: più è piccolo il valore, più veloce è il metodo.

In particolare vale che:

$$\theta(e^{(0)}) \leq \lim_{k\to\infty} \left\|P^k\right\|^{\frac{1}{k}} = \rho(P).$$

L'uguaglianza è raggiunta per $e^{(0)}$ pari ad un autovettore relativo al raggio spettrale di P. Se P è nilpotente, allora il metodo converge in un numero finito di passi.

Un esempio di metodo iterativo è dato dal **metodo di Richardson**, che pone $P = I - \alpha A$, $q = M^{-1}b = \alpha A$. In particolare tale metodo genera la successione $x^{(k+1)} = x^{(k)} - \alpha(Ax^{(k)} - b)$, dove $\alpha \neq 0$ è un opportuno parametro e $Ax_n - b$ è detto *errore residuo*.

Poiché il metodo converge se $\rho(P) < 1$, l'idea naturale è quella di scegliere α in modo tale che $\rho(P)$ sia minimo, ricordandosi che λ è autovalore di P se e solo se $\alpha(\lambda + 1)$ è autovalore di A. Infatti:

$$\det(\lambda I - P) = \det(\lambda I + I - \alpha A) = \frac{1}{\alpha^n}\det(\alpha(\lambda+1)I - A).$$

Se la matrice è hermitiana e definita positiva, allora $\alpha = 1/\left\|A\right\|$, dove $\|\cdot\|$ è una norma indotta, garantisce la convergenza del metodo.

I metodi di Jacobi e Gauss-Seidel

Sia A tale per cui $a_{ii} \neq 0$. Si può allora decomporre A nel seguente modo:

$$A = D - B - C,$$

dove:

- D è diagonale con $d_{ii} = a_{ii} \neq 0$ ($D = \text{diag}(A)$, D invertibile);
- B è strettamente triangolare inferiore con $b_{ij} = -a_{ij}$ con $i > j$ ($B = -\text{tril}(A)$);
- C è strettamente triangolare superiore con $c_{ij} = -a_{ij}$ con $i < j$ ($C = -\text{triu}(A)$).

Il metodo iterativo ottenuto con il partizionamento additivo $M = D$ e $N = B + C$ è detto **metodo di Jacobi**, mentre ponendo $M = D - B$ e $N = C$ si applica il **metodo di Gauss-Seidel**.

Le matrici di iterazione $P = M^{-1}N$ dei due metodi illustrati sono rispettivamente:

$$J = D^{-1}(B + C), \quad G = (D - B)^{-1}C.$$

Teorema. Se vale una delle seguenti condizioni allora $\rho(J)$ e $\rho(G) < 1$ (e quindi A è invertibile e il metodo iterativo è convergente):

1. A è fortemente dominante diagonale;
2. A^T è fortemente dominante diagonale;
3. A è irriducibilmente dominante diagonale (i.d.d.);
4. A^T è irriducibilmente dominante diagonale (i.d.d.).

L'iterazione del metodo di Jacobi si scrive come:

$$x^{(k+1)} = D^{-1}((B + C)x^{(k)} + b),$$

che diventa:

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

richiedendo $O(n^2)$ flops.

L'iterazione del metodo di Gauss-Seidel si scrive invece come

$$x^{(k+1)} = (D - B)^{-1}(Cx^{(k)} + b),$$

$$x^{(k+1)} = D^{-1}(Bx^{(k+1)} + Cx^{(k)} + b),$$

che diventa:

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

richiedendo sempre $O(n^2)$ flops.

Il metodo di Jacobi esegue le operazioni su $x^{(k)}$ mentre quello di Gauss-Siedel usa anche le componenti già aggiornate di $x^{(k+1)}$. In generale dunque, il metodo di Gauss-Siedel ha migliori proprietà di convergenza, anche se non è sempre così.

Teorema (di Stein-Rosenberg)**.** Se $a_{ii} \neq 0\ \forall\, i$ e J ha elementi non negativi allora vale una sola delle seguenti proprietà:

- $\rho(J) = \rho(G) = 0$;
- $0 < \rho(G) < \rho(J) < 1$;
- $\rho(J) = \rho(G) = 1$;
- $1 < \rho(J) < \rho(G)$.

Teorema. Se A è una matrice tridiagonale con elementi diagonali non nulli, allora per ogni autovalore λ di J esiste un autovalore μ di G tale che $\mu = \lambda^2$. Per ogni autovalore non nullo di G esiste un autovalore λ di J tale che $\mu = \lambda^2$. In particolare $\rho(G) = \rho(J)^2$.

Metodi a blocchi

Se $A \in M_{nm}(\mathbb{C})$ è partizionata in n^2 blocchi $m \times n$ $A = (A_{ij})$ possiamo considerare una decomposizione additiva $A = D - B - C$ dove D matrice diagonale a blocchi con blocchi diagonali uguali a A_{ii}, B la matrice triangolare inferiore a blocchi tale che $B_{ij} = -A_{ij}$ con $i > j$ e C la matrice triangolare superiore a blocchi tale che $C_{ij} = -A_{ij}$ con $i < j$.

In questo modo se i blocchi diagonali A_{ii} sono non singolari possiamo considerare i metodi iterativi con $M = D$ e $N = B + C$, e con $M = D - B$ e $N = C$. Il primo è detto **metodo di Jacobi a blocchi**, mentre il secondo è detto **metodo di Gauss-Siedel a blocchi**.

Teorema. Sia $A = (A_{ij})$ una matrice tridiagonale a blocchi, cioè tale che $A_{ij} = 0$ se $|i - j| \geq 2$ con blocchi diagonali A_{ii} non singolari. Siano J_B e G_B le matrici di iterazione dei metodi di Jacobi a blocchi e di Gauss-Siedel a blocchi. Allora per ogni autovalore λ di J_B esiste un autovalore μ di G_B tale che $\mu = \lambda^2$. Per ogni autovalore non nullo μ di G_B esiste un autovalore λ di J_B tale che $\mu = \lambda^2$. In particolare vale che $\rho(G_B) = \rho(J_B)^2$.

Calcolo di zeri di funzioni continue su $\mathbb{R}$

Sia $f : [a, b] \to \mathbb{R}$ continua tale per cui $f(a)f(b) < 0$ (i.e. $f(a)$ ed $f(b)$ sono discordi). Allora, per il teorema degli zeri (o di Bolzano), esiste un punto $\alpha \in [a, b]$ tale per cui $f(\alpha) = 0$, ossia esiste uno zero in (a, b). Lo scopo di questa sezione è illustrare i metodi che permettono di generare una successione $\{x_k\}$ che converga ad un tale α.

Metodo di bisezione (o *dicotomico*)

Uno dei metodi più semplici, e il più costruttivo per la dimostrazione del teorema degli zeri, è il metodo di **bisezione** (o *dicotomico*). Innanzitutto si presuppone $f(b) > 0$ (e dunque $f(a) < 0$); altrimenti è sufficiente applicare l'algoritmo al contrario a $-f(x)$.

Al passo $(k + 1)$-esimo si considera $c_k = (a_k + b_k)/2$, dove $a_0 = a$ e $b_0 = b$. Se $f(c_k) = 0$, l'algoritmo termina; altrimenti, se $f(c_k) > 0$, $b_{k+1} = c_k$ e $a_{k+1} = a_k$, e se $f(c_k) < 0$, $b_{k+1} = b_k$ e $a_{k+1} = c_k$.

Dal momento che gli intervalli $I_k = [a_k, b_k]$ hanno lunghezza $(b - a)/2^k$, per $k \to \infty$ la successione x_k ha limite α, che è tale per cui $f(\alpha) = 0$. Inoltre l'intervallo si dimezza a ogni step, e dunque il metodo converge esponenzialmente (non per questo è veloce, anzi: come vedremo, ci sono metodi estremamenti più veloci che convergono in modo *doppiamente esponenziale*).

Metodi del punto fisso

I metodi del punto fisso si ottengono trasformando il problema $f(x) = 0$ in un problema del tipo $g(x) = x$. Questa trasformazione si può ottenere in infiniti modi diversi. Ad esempio, data una qualsiasi funzione $h(x)$ (diversa da zero nei punti del dominio di $f(x)$), si può porre:

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

i cui punti fissi sono gli zeri di $f(x)$.

Se $g(x)$ è continua, la successione $x_{k+1} = g(x_k)$, con $x_0 \in \mathbb{R}$, se convergente, converge ad un punto fisso di $g(x)$ (e dunque ad uno zero di $f(x)$). Questo metodo è detto **metodo del punto fisso** (o *di iterazione funzionale*) associato a $g(x)$.

Teorema (del punto fisso)**.** Sia $\mathcal{I} = [\alpha - \rho, \alpha + \rho]$ e $g(x) \in C^1(\mathcal{I})$ dove $\alpha = g(\alpha)$ e $\rho > 0$. Si denoti con $\lambda = \max_{x\in\mathcal{I}}|g'(x)|$. Se $\lambda < 1$ allora per ogni $x_0 \in \mathcal{I}$, posto $x_{k+1} = g(x_k)$, vale che:

$$|x_k - \alpha| \leq \lambda^k \rho.$$

Pertanto $x_k \in \mathcal{I}$ e $\lim_k x_k = \alpha$. Inoltre α è l'unico punto fisso di $g(x)$ in $\mathcal{I}$.

Supponiamo di avere un intervallo $[a, b]$ in cui è presente un punto fisso α e che $|g'(x)| < 1$ su tale intervallo. Sotto queste ipotesi siamo certi che almeno con una delle due scelte $x_0 = a$ o $x_0 = b$ la successione è ben definita e converge al punto fisso α. Infatti basta prendere rispettivamente $\rho = \alpha - a$ o $\rho = b - \alpha$ ed applicare il teorema. Se la successione x_k cade fuori dall'intervallo, allora si arrestano le iterazioni e si riparte con x_0 uguale all'altro estremo.

In questo caso la convergenza vale in un intorno opportuno del punto fisso α. Denotiamo questo fatto con l'espressione **convergenza locale**. Si parla di **convergenza globale** qualora la convergenza vi sia per ogni scelta iniziale.

Nell'aritmetica a virgola mobile il teorema diventa:

Teorema. Nelle ipotesi del teorema del punto fisso sia:

$$\tilde{x}_{k+1} = g(\tilde{x}_k) + \delta_k,$$

dove $|\delta_k| \leq \delta$ è l'errore commesso nel calcolo di $g(\tilde{x}_k)$ e δ è una quantità nota. Posto $\sigma = \delta/(1-\lambda)$, se $\sigma < \rho$ allora:

$$|\tilde{x}_k - \alpha| \leq (\rho - \sigma)\lambda^k + \sigma.$$

Questo teorema ci dice che la distanza di $\tilde{x}_k$ da α è limitata dalla somma di due parti. La prima converge a zero in modo esponenziale su base λ. La seconda è costante e rappresenta l'intervallo di incertezza sotto il quale non è consentito andare.

Per la funzione $g(x) = x - f(x)$ l'intervallo di incertezza è:

$$I = \left[\alpha - \frac{\delta}{|f'(\alpha)|}, \alpha + \frac{\delta}{|f'(\alpha)|}\right].$$

Se $g'(x) > 0$, allora $I \subseteq [\alpha - \sigma, \alpha + \sigma]$.

Si ricorda che:

$$x_{k+1} - \alpha = g'(\xi_k)(x_k - \alpha),$$

dove $\xi_k \in (\alpha, x_k)$ viene dall'applicazione del teorema di Lagrange. Ciò implica che se $0 < g'(x) < 1$ con $x \in [\alpha - \rho, \alpha + \rho]$ allora $x_{k+1} - \alpha$ ha lo stesso segno di $x_k - \alpha$. Vale quindi che:

$$x_0 > \alpha \Rightarrow x_k > \alpha \ \forall k, \quad \alpha < x_{k+1} < x_k,$$

cioè la successione è decrescente. Analogamente se $x_0 < \alpha$ la successione è crescente.

Se invece $-1 < g'(x) < 0$, la differenza di un punto con α alterna segno, in particolare per $x_0 > \alpha$ la sottosuccessione $\{x_{2k}\}$ cresce mentre $\{x_{2k+1}\}$ decresce (entrambe al punto α).

Velocità di convergenza

Definizione. Sia $\{x_k\}$ una successione tale che $\lim_k x_k = \alpha$. Si ponga allora $e_k = x_k - \alpha$ come l'errore assoluto al passo k-esimo. Supponiamo esista il limite della riduzione dell'errore al passo k-esimo:

$$\gamma = \lim_k \left|\frac{x_{k+1} - \alpha}{x_k - \alpha}\right| = \lim_k \left|\frac{e_{k+1}}{e_k}\right|.$$

La convergenza di $\{x_k\}$ a α è detta allora:

- lineare (o geometrica) se $0 < \gamma < 1$;
- sublineare se $\gamma = 1$;
- superlineare se $\gamma = 0$.

Nel caso di convergenza superlineare, se $p > 1$ ed esiste il limite:

$$\lim_k \left|\frac{x_{k+1} - \alpha}{(x_k - \alpha)^p}\right| = \sigma, \quad 0 < \sigma < \infty$$

si dice che la successione **converge con ordine** p.

Osservazione. L'ordine di convergenza non è obbligatoriamente intero.

Teorema. Sia $g(x) \in C^1([a, b])$ e $\alpha \in (a, b)$ t.c. $g(\alpha) = \alpha$. Se esiste un $x_0 \in [a, b]$ tale che la successione $x_{k+1} = g(x_k)$ converge linearmente ad α con fattore γ allora:

$$|g'(\alpha)| = \gamma.$$

Viceversa, se $0 < |g'(\alpha)| < 1$ allora esiste un intorno I di α contenuto in $[a, b]$ tale che per ogni $x_0 \in I$ la successione $\{x_k\}$ converge ad α in modo lineare con fattore $\gamma = |g'(\alpha)|$.

Teorema. Sia $g(x) \in C^1([a, b])$ e $\alpha \in (a, b)$ t.c. $g(\alpha) = \alpha$. Se esiste un $x_0 \in [a, b]$ tale che la successione $x_{k+1} = g(x_k)$ converge sublinearmente ad α allora:

$$|g'(\alpha)| = 1.$$

Viceversa, se $|g'(\alpha)| = 1$ esiste un intorno I di α contenuto in $[a, b]$ tale che per ogni $x \in I$, $x \neq \alpha$ vale $|g'(x)| < 1$ e $g'(x)$ non cambia segno su I allora tutte le successioni $\{x_k\}$ con $x_0 \in I$ convergono ad α in modo sublineare.

Teorema. Sia $g(x) \in C^p([a, b])$ con $p > 1$ intero e $\alpha \in (a, b)$ t.c. $g(\alpha) = \alpha$. Se esiste un $x_0 \in [a, b]$ tale che la successione $x_{k+1} = g(x_k)$ converge superlinearmente ad α con ordine di convergenza p allora:

$$\left|g^{(k)}(\alpha)\right| = 0 \quad 1 \leq k \leq p-1, \quad g^{(p)}(\alpha) \neq 0.$$

Viceversa se $|g^{(k)}(\alpha)| = 0$ per $k = 1, \ldots, p-1$ e $g^{(p)}(\alpha) \neq 0$ allora esiste un intorno I di α tale che per ogni $x_0 \in I$ tutte le successioni $\{x_k\}$ convergono ad α in modo superlineare con ordine p.

Definizione. La successione $\{x_k\}$ converge ad α **con ordine almeno** p se esiste una costante β tale che

$$|x_{k+1} - \alpha| \leq \beta \, |x_k - \alpha|^p .$$

Se una successione converge con ordine $q \geq p$ allora converge anche con ordine almeno p.

Se una successione x_k converge ad α in modo che l'errore relativo ε_k al passo k-esimo è limitato nel seguente modo:

$$\varepsilon_k = |x_k - \alpha/\alpha| \leq \beta\gamma^{p^k}$$

allora il numero di cifre significative $(1 + \log_2(\varepsilon_k^{-1}))$ è tale che

$$1 + \log_2 \varepsilon_k^{-1} \geq 1 + \log_2 \beta^{-1} + p^k \log_2 \gamma^{-1}.$$

Confronto tra metodi

Siano dati due metodi iterativi del punto fisso definiti da due funzioni $g_1(x)$ e $g_2(x)$. Si supponga che siano entrambi siano o a convergenza lineare o a convergenza superlineare. Denotiamo con c_1, c_2 il numero di flops per passo dei due metodi. Se siamo nel caso di convergenza lineare, detti γ_1, γ_2 i due fattori di convergenza allora il primo metodo risulta più conveniente se

$$\frac{c_1}{c_2} < \frac{\log \gamma_1}{\log \gamma_2}.$$

Nel caso di convergenza superlineare, se p_1 e p_2 sono i due ordini di convergenza, il primo metodo è più efficiente del secondo se

$$\frac{c_1}{c_2} < \frac{\log p_1}{\log p_2}.$$

Alcuni metodi del punto fisso

Metodo delle secanti

Sia $f(x) \in C^1([a, b])$ e $\alpha \in [a, b]$ t.c. $f(\alpha) = 0$. Il metodo definito dalla funzione

$$g(x) = x - \frac{f(x)}{m}$$

dove m è un'opportuna costante è detto **metodo delle costanti**.

Il metodo consiste nel tracciare la retta passante per il punto $(x_k, f(x_k))$ di coefficiente angolare m e considerare come x_{k+1} l'ascissa dell'intersezione di tale retta con l'asse delle ascisse, reiterando.

Una condizione sufficiente di convergenza è che valga $|1 - f'(x)/m| < 1$ in un intorno di circolare di α. È sufficiente quindi scegliere m in modo che abbia lo stesso segno di $f'(x)$ e $|m| > \frac{1}{2}|f'(x)|$. Se $f'(\alpha)$ fosse nota, la scelta $m = f'(\alpha)$ darebbe una convergenza superlineare.

Metodo delle tangenti di Newton

La funzione $g(x)$ del **metodo di Newton** è definita nel modo seguente:

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

Per tale metodo vale:

$$g'(x) = 1 - \frac{f'(x)^2 - f''(x)f(x)}{f'(x)^2} = \frac{f''(x)f(x)}{f'(x)^2}.$$

Il metodo consiste nel tracciare la retta passante per il punto $(x_k, f(x_k))$ e tangente a esso e considerare come x_{k+1} l'ascissa dell'intersezione di tale retta con l'asse delle ascisse, reiterando.

Teorema. Sia $f(x) \in C^2([a,b])$ e sia $\alpha \in (a,b)$ t.c. $f(\alpha) = 0$. Se $f'(\alpha) \neq 0$, allora esiste un intorno $I = [\alpha - \rho, \alpha + \rho] \subset [a,b]$ tale che per ogni $x_0 \in I$ la successione generata dal metodo di Newton converge ad α.

Inoltre se $f''(\alpha) \neq 0$ la convergenza è superlineare di ordine 2, se $f''(\alpha) = 0$ la convergenza è di ordine almeno 2.

Teorema. Sia $f(x) \in C^p([a,b])$ con $p > 2$ e $\alpha \in (a,b)$ t.c. $f(\alpha) = 0$. Se $f'(\alpha) = \cdots = f^{(p-1)}(\alpha) = 0$, $f^{(p)}(\alpha) \neq 0$ e $f'(x) \neq 0$ per $x \neq \alpha$, allora esiste un intorno $I = [\alpha - \rho, \alpha + \rho] \subset [a,b]$ in cui $f'(x) \neq 0$ per $x \in I, x \neq \alpha$ e tale che per ogni $x_0 \in I$, la successione generata dal metodo di Newton converge ad α. Inoltre α è l'unico zero di $f(x)$ in I. La convergenza è lineare con fattore di convergenza $\gamma = 1 - 1/p$.

Teorema. Se la funzione $f(x)$ è di classe C^2 sull'intervallo $I = [\alpha, \alpha + \rho]$ ed è tale che $f'(x)f''(x) > 0$ per $x \in I$ allora per ogni $x_0 \in I$ la successione generata dal metodo di Newton applicato ad $f(x)$ converge decrescendo ad α con ordine 2. Un risultato analogo vale su intervalli del tipo $[\alpha - \rho, \alpha]$, su cui invece la successione cresce.

Esempio (Calcolo del reciproco di un numero a)**.** Per calcolare $1/a$ si può porre $f(x) = a - 1/x$ e poi applicare il metodo di Newton.

Esempio (Calcolo della radice p-esima di un numero a)**.** Per calcolare $\sqrt[p]{a}$ è sufficiente applicare il metodo di Newton con $f(x) = (x^p - a)x^{-q}$, dove $q \geq 0$.

Esempio (Metodo di Aberth)**.** Se si hanno delle approssimazione t_1, ..., t_n di zeri di $p(x)$, si può applicare il metodo di Newton su $f(x) = p(x) / \prod_{i=1}^{n}(x - t_i)$, in modo tale che $f(t_i)$ sia sempre più grande, diminuendo gli errori e migliorando le approssimazioni.

Interpolazione di funzioni

Siano $\varphi_0(x)$, ..., $\varphi_n(x) : [a,b] \to \mathbb{R}$ funzioni linearmente indipendenti. Siano (x_i, y_i) $n+1$ valori assegnati (*nodi*) tali che $x_i \in [a,b]$ e $x_i \neq x_j$ se $i \neq j$.

Il problema dell'interpolazione consiste nel determinare i coefficienti $a_i \in \mathbb{R}$ tali per cui:

$$f(x) = \sum_{i=0}^{n} a_i \, \varphi_i(x), \qquad f(x_i) = y_i \ \forall\, i.$$

Tali condizioni sono dette **condizioni di interpolazione**.

Interpolazione polinomiale

L'interpolazione polinomiale ricerca il polinomio p di grado $n+2$ che soddisfi le condizioni di interpolazione di $n+1$ nodi. Sia pertanto $\varphi_i(x) = x^i$. Siano (x_i, y_i) per $i = 0$, ..., $n+1$ i nodi dell'interpolazione.

Interpolazione monomiale

Definizione. Si dice **matrice di Vandermonde** nei nodi x_0, ..., x_n la matrice di taglia $n+2$:

$$V_n = \begin{pmatrix} 1 & x_0 & \dots & x_0^n \\ 1 & x_1 & \dots & x_1^n \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \dots & x_n^n \end{pmatrix}$$

Teorema. Vale $\det V_n = \prod_{j<i}(x_i - x_j)$.

Se $a = (a_i)$ è il vettore dei coefficienti di $p(x)$ e $y = (y_i)$ è il vettore delle immagini nei nodi, il problema dell'interpolazione polinomiale si riduce a risolvere il sistema lineare $V_n a = y$.

Il problema ha dunque un costo computazionale di $\frac{2}{3}(n+2)^3$ (ossia quello di risoluzione di un sistema lineare), ma risulta essere numericamente instabile.

Interpolazione di Lagrange

Definizione (polinomio di Lagrange)**.** Si definisce l'i-esimo polinomio di Lagrange $L_i(x)$ come:

$$L_i(x) = \frac{\prod_{j=0, j\neq i}^{n}(x - x_j)}{\prod_{j=0, j\neq i}^{n}(x_i - x_j)}.$$

Questo polinomio è tale per cui $L_i(x_i) = 1$ e $L_i(x_j) = 0$ per $j \neq i$.

Considerando dunque come base dell'interpolazione $\phi_i = L_i$, si ottiene che:

$$p(x) = \sum_{i=0}^{n} y_i L_i(x).$$

Infatti, dacché V_n è invertibile ($\det V_n \neq 0$), allora esiste un solo polinomio di grado al più $n+2$ che interpola i nodi (x_i, y_i), e $p(x)$ soddisfa le condizioni di interpolazione.

Inoltre, per $x \neq x_i$, vale la seguente riscrittura:

$$p(x) = \left[\prod_{i=0}^{n}(x - x_i)\right] \left[\sum_{i=0}^{n} \frac{y_i/\theta_i}{x - x_i}\right], \quad \theta_i = \prod_{\substack{j=0 \\ j\neq i}}^{n}(x_i - x_j).$$

Tramite questa riscrittura il calcolo del primo valore di $p(x)$ impiega $O(n^2)$ flops, ed il costo della valutazione in un nuovo valore costa solo $O(n)$ flops (dacché σ_i è già calcolato).

Resto dell'interpolazione polinomiale

Se $f(x)$ è sufficientemente regolare, allora definiamo il **resto dell'interpolazione** come:

$$r_n(x) = f(x) - p_n(x),$$

dove $p_n(x)$ è il polinomio di interpolazione di $f(x)$ relativo ai nodi $x_0 < \cdots < x_n$.

Teorema. Sia $f(x) \in C^{n+1}([a,b])$ e sia $p_n(x)$ il polinomio di interpolazione di $f(x)$ relativo ai nodi $a \leq x_0 < \cdots < x_n \leq b$. Allora per ogni $x \in [a,b]$ esiste $\xi \in (a,b)$ tale per cui:

$$r_n(x) = \left[\prod_{i=0}^{n}(x - x_i)\right] \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

Teorema. Sia $f(x) \in C^{n+1}[a,b]$ e $p(x)$ il polinomio di interpolazione di $f(x)$ relativo ai nodi $a \leq x_0 < \cdots < x_n \leq b$. Per ogni $x \in [a,b]$ esiste $\xi \in (a,b)$ tale che

$$r_n(x) = \prod_{i=0}^{n}(x - x_i) \frac{f^{(n+1)}(\xi)}{(n+1)!}$$

Osservazione. Avvicinandosi con tutti i nodi a x_i si ottiene un'espressione simile a quella del resto di Lagrange (ponendo infatti $x_i = x_j$ per ogni i, j, la formula è proprio quella del resto di Lagrange).

Interpolazione polinomiale nelle radici n-esime dell'unità

Definizione. Si dice **radice n-esima dell'unità** una radice di $x^n - 1$ su $\mathbb{C}$. Si dice **radice primitiva n-esima dell'unità** il numero complesso ω_n tale per cui:

$$\omega_n = \cos\frac{2\pi}{n} + i \sin\frac{2\pi}{n}.$$

Si verifica facilmente che le radici n-esime dell'unità formano un gruppo moltiplicativo generato da ω_n. Le radici dell'unità soddisfano la seguente proposizione:

Proposizione. Vale l'identità:

$$\sum_{i=0}^{n-1} \omega_n^{ki} = \begin{cases} n & \text{se } k \equiv 0 \pmod n, \\ 0 & \text{altrimenti} \end{cases}$$

Scegliamo come nodi dell'interpolazione le radici n-esime dell'unità, detti *nodi di Fourier*. Dato il vettore $y = (y_0, \dots, y_{n-1})$, il nostro obiettivo è dunque trovare i coefficienti $z = (z_0, \dots, z_{n-1})$ del polinomio $p(t) = \sum_{j=0}^{n-1} z_j t^j$ che soddisfa $p(\omega_n^i) = y_i \ \forall\, i$.

Definizione. Si definisce $z = \mathrm{DFT}_n(y) = \mathrm{DFT}(y)$ come la **trasformata discreta di Fourier**, e $y = \mathrm{IDFT}_n(z) = \mathrm{IDFT}(z)$ come la **trasformata discreta inversa di Fourier** (l'inverso è ben definito dal momento che le radici n-esime sono distinte).

Definizione. Si definisce **matrice di Fourier** di taglia $n-1$ la matrice di Vandermonde nei nodi di Fourier:

$$\Omega_n = (\omega_n^{ij}) = (\omega_n^{ij \bmod n}).$$

Ω_n soddisfa la relazione $\mathrm{DFT}(y) = \Omega_n^{-1} y$ e $\mathrm{IDFT}(z) = \Omega_n z$.

Proposizione. Valgono le seguenti proprietà per la matrice di Fourier:

- $\Omega_n = \Omega_n^T$;
- $\Omega_n^H \Omega_n = nI$ (pertanto $F_n := \Omega_n/\sqrt{n}$ è unitaria);
- $\Omega_n^2 = n\,\Pi_n$, dove Π_n è la matrice di permutazione corrispondente alla permutazione σ su $\{0,\dots,n-1\}$ tale per cui $\sigma(0)=0$ e $\sigma(j) = n-j$ per $j>0$, ovverosia:

$$\Pi_n = \begin{pmatrix} 1 & 0 \\ 0 & P_{n-1} \end{pmatrix},$$

  dove P_{n-1} è la matrice identità specchiata orizzontalmente (ossia con soli 1 sull'antidiagonale), che è tale per cui $P_{n-1}^2 = I$.

Come corollario della precedente proposizione, si ottiene che $\Omega_n^{-1} = \frac{1}{n}\Omega_n^H$ e che $\Omega_n^{-1} = \frac{1}{n}\Pi_n\Omega_n = \frac{1}{n}\Omega_n\Pi_n$.

Dunque il problema dell'interpolazione si risolve ponendo $z = \Omega_n^{-1} y$ e applicando una delle formule proposte precedentemente.

Osservazione. La matrice $F_n = \Omega_n/\sqrt{n}$ è unitaria, e dunque $\|F_n\|_2 = 1$. Ciò implica che $\|\Omega_n\|_2 = \sqrt{n}$, dunque il numero di condizionamento di Ω_n è 1, ovverosia il problema dell'interpolazione ai nodi di Fourier è ben condizionato.

Calcolo di (I)DFT: *Fast Fourier Transform* (FFT)

Consideriamo il caso del calcolo di $\mathrm{IDFT}(y)$, dove abbiamo i coefficienti z_0, ..., z_{n-1} e vogliamo trovare i valori y_i tali per cui $\Omega_n(z_i) = (y_i)$. Il calcolo di $\mathrm{DFT}(z)$ si può poi fare eseguendo $\mathrm{IDFT}(z)$ e applicando le relazioni introdotte nella proposizione precedente, aggiungendo n divisioni e permutando gli indici.

Consideriamo il caso in cui $n = 2^q$. Allora:

$$y_i = \sum_{j=0}^{n-1} \omega_n^{ij} z_j.$$

Ricordando che $\omega_n^2 = \omega_{\frac{n}{2}}$, allora, separando gli indici pari da quelli dispari, vale che:

$$y_i = \sum_{j=0}^{\frac{n}{2}-1} \omega_{\frac{n}{2}}^{ij} z_{2j} + \omega_n^i \sum_{j=0}^{\frac{n}{2}-1} \omega_{\frac{n}{2}}^{ij} z_{2j+1}.$$

Pertanto, detti $Y = \left(y_0,\dots,y_{\frac{n}{2}-1}\right)^\top$, $Y' = \left(y_{\frac{n}{2}},\dots,y_{n-1}\right)^\top$ e $D_n = \mathrm{diag}\left(1,\omega_n,\dots,\omega_n^{\frac{n}{2}-1}\right)$, valgono le seguenti due identità:

$$Y = \mathrm{IDFT}_{\frac{n}{2}}(z_{\mathrm{pari}}) + D_n\ \mathrm{IDFT}_{\frac{n}{2}}(z_{\mathrm{disp}}),$$

$$Y' = \mathrm{IDFT}_{\frac{n}{2}}(z_{\mathrm{pari}}) - D_n\ \mathrm{IDFT}_{\frac{n}{2}}(z_{\mathrm{disp}}).$$

In forma matriciale le due identità si scrivono infine come:

$$y = \begin{pmatrix} \Omega_{\frac{n}{2}} z_{\mathrm{pari}} \\ \Omega_{\frac{n}{2}} z_{\mathrm{pari}} \end{pmatrix} + \begin{pmatrix} D_n \Omega_{\frac{n}{2}} z_{\mathrm{disp}} \\ -D_n \Omega_{\frac{n}{2}} z_{\mathrm{disp}} \end{pmatrix}.$$

Poiché n è potenza di 2 possiamo ripetere questa strategia per calcolare le due trasformate di ordine $\frac{n}{2}$ mediante quattro trasformate di ordine $\frac{n}{4}$, ecc... fino ad avere trasformate di ordine 1 che non richiedono alcuna operazione.

Il costo $c(n)$ di flops di una IDFT su n nodi con questo metodo è ricorsivamente:

$$c(n) = 2\,c\Big(\frac{n}{2}\Big) + \frac{n}{2} + n = 2\,c\Big(\frac{n}{2}\Big) + \frac{3}{2}n,$$

dove $n/2$ sono le moltiplicazioni effettuate e n le addizioni.

Dacché $c(1) = 0$ e $n = 2^q$ vale che:

$$c(n) = \frac{3}{2} n \log_2 n = 3q2^{q-1}.$$

In generale, l'algoritmo ha dunque complessità $O(n\log_2(n))$. L'algoritmo per il calcolo della IDFT che si ottiene in questo modo è noto come **algoritmo di Cooley-Tukey**.

Applicazioni della FFT: calcolo del prodotto di polinomi

Dati $a(t)$, $b(t) \in \mathbb{C}[x]$ allora $c(t) = a(t)b(t)$ si può calcolare trovando i coefficienti $c_i = \sum_{j+k=i} a_j b_k$. Se $a(t)$, $b(t)$ hanno grado rispettivamente p, q, questo algoritmo ha complessità $O((p+q)^2)$.

Sia $n = 2^q \geq p+q$. Con la FFT posso valutare $a(t)$, $b(t)$ nelle radici n-esime dell'unità. Dopo aver moltiplicato i valori ottenuti, si può applicare una DFT, ricavando i coefficienti di $c(t)$. Questo algoritmo utilizza 2 IDFT, n moltiplicazioni ed una DFT; pertanto ha un costo di $O(n\log_2(n))$ flops.

Metodi dell'integrazione approssimata

In questa sezione si illustrano alcuni metodi per approssimare gli integrali su un intervallo.

Definizione. Data $f : [a,b] \to \mathbb{R}$ continua e $n+1$ nodi x_0, ..., x_n si definiscono le due quantità:

$$S[f] = \int_a^b f(x)\,dx, \quad S_{n+1}[f] = \sum_{i=0}^{n} w_i f(x_i),$$

dove i termini w_i sono positivi reali detti *pesi*, possibilmente variabili. Una scelta di pesi corrisponde a una **formula di integrazione approssimata**.

Definizione. Si dice **resto** il valore

$$r_{n+1} = S[f] - S_{n+1}[f].$$

Definizione. Si dice che una formula di integrazione approssimata ha **grado di precisione** (massimo) $k \geq 0$ se vale

$$r_{n+1} = 0 \iff f(x) = x^j, \quad 0 \leq j \leq k$$

e se

$$r_{n+1} \neq 0 \iff f(x) = x^{k+1}.$$

Formule di integrazione dell'interpolazione di Lagrange

Si può approssimare un integrale utilizzando l'interpolazione di Lagrange sui nodi, ponendo

$$S_{n+1}[f] = \int_a^b \tilde{f}(x)\,dx = \sum_{i=0}^{n} \left[f(x_i) \int_a^b L_i(x)\,dx \right],$$

dove $\tilde{f}(x)$ è il polinomio ottenuto applicando l'interpolazione di Lagrange su f dati i nodi $(x_0, f(x_0))$, ..., $(x_n, f(x_n))$ e i pesi sono tali per cui $w_i = \int_a^b L_i(x)\,dx$.

Le formule di integrazione interpolatorie hanno grado di precisione almeno n (infatti se $j \leq n$, x^j è il polinomio che interpola i nodi dati) ed hanno grado di precisione massimo $2n+1$. Se una formula di integrazione ha grado di precisione almeno n, allora è interpolatoria.

Se $f \in C^{(n+1)}([a,b])$ e $|f^{(n+1)}(x)| \leq M$ allora

$$|r_{n+1}| \leq \frac{M}{(n+1)!} \left| \int_a^b \prod_{j=0}^{n} (x - x_j)\,dx \right|.$$

Formule di Newton-Cotes (semplici)

Scegliendo nodi equispaziati, con $h = \frac{b-a}{n}$, si ottengono le formule di Newton-Cotes. nel caso $n = 1$ vale

$$w_0 = \int_a^b L_0(x)\,dx = \frac{h}{2} = w_1 \Rightarrow S_2[f] = \frac{h}{2}(f(x_0) + f(x_1)),$$

ovverosia corrisponde all'area del trapezio con di base $|x_0|$ e $|x_1|$ e con altezza h. Il resto r_2 corrisponde invece a

$$\int_a^b f(x)\,dx - S_2[f] = -\frac{1}{12}h^3 f''(\xi), \quad \xi \in (a,b).$$

Nel caso $n = 2$ invece

$$w_0 = w_2 = \frac{h}{3}, \quad w_1 = \frac{4}{3}h$$

ed il resto è dunque

$$\int_a^b f(x)\,dx - S_3[f] = -\frac{1}{90}h^5 f^{(4)}(\xi), \quad \xi \in (a,b)$$

Formule di Newton-Cotes composte (trapezi, Cavalieri-Simpson)

Si può rendere più precisa la formula di integrazione introducendo altri nodi equispaziati z_0, ..., z_N e risolvendo

$$\int_a^b f(x)\,dx = \sum_{i=0}^{N-1} \int_{z_i}^{z_{i+1}} f(x)\,dx,$$

approssimando ciascuno degli integrali con le formule di Newton-Cotes semplici per $n = 1$ o 2.

Scegliendo sempre $n = 1$ si ottiene la **formula dei trapezi**, per la quale:

$$J_2^{(n)}[f] = \sum_{i=0}^{n-1} S_2^{(i)}[f] = \frac{b-a}{2n}\left[f(z_0) + f(z_n) + 2\sum_{i=1}^{n-1} f(z_i)\right]$$

Inoltre vale che:

$$S[f] - J_2^{(nN)}[f] = -\frac{(b-a)h^2}{12} f''(\xi).$$

Scegliendo invece sempre $n = 2$ si ottiene la **formula di Cavalieri-Simpson**, per la quale:

$$J_3^{(n)}[f] = \frac{b-a}{6n}\left[f(z_0) + f(z_n) + 2\sum_{k=1}^{n-1} f(z_k) + \right.$$
$$\left. + 4\sum_{k=0}^{n-1} f\left(\frac{z_k + z_{k+1}}{2}\right)\right].$$

Risoluzione di problemi di Cauchy con metodi a un passo

Sia $I \subseteq \mathbb{R}$ un intervallo e consideriamo una funzione $f(t, y) : I \times \mathbb{R} \to \mathbb{R}$ continua. Dati i parametri iniziali $t_0 \in I$, $y_0 \in \mathbb{R}$ si associa a questi il seguente *problema di Cauchy*:

$$\begin{cases} y'(t) = f(t, y(t)), \\ y(t_0) = y_0, \end{cases}$$

su cui si ipotizza che l'incognita y è $C^1(I, \mathbb{R})$.

Se $f(t, y)$ è Lipschitziana rispetto a y, ovverosia esiste $L \in \mathbb{R}$ t.c. $|f(t, y_1) - f(t, y_2)| \leq L\,|y_1 - y_2|$ per ogni $t \in I$, y_1, $y_2 \in \mathbb{R}$, allora esiste ed è unica la soluzione y al problema di Cauchy proposto.

Notazione (Intervallo destro di t_0 e successione per passi)**.** Per intervallo destro di un $t_0 \in \mathbb{R}$ intendiamo un intervallo della forma $[t_0, t_0 + T]$ con $t > 0$.

Dato $h > 0$, definiamo $t_n := t_0 + nh$ per $n = 0$, ..., N_h, dove N_h è il massimo naturale p per cui $ph \leq T$ (ossia il massimo naturale per cui la successione definita sta dentro l'intervallo destro definito).

Notazione (Approssimazioni)**.** Scriveremo y_n per indicare $y(t_n)$ e u_n per indicare l'approssimazione ricavata di $y(t_n)$, mentre $f_n := f(t_n, u_n)$.

Definizione (Metodo a un passo)**.** Si dice **metodo a un passo** un metodo di risoluzione approssimato di un problema di Cauchy che imposti una sequenza della forma:

$$\begin{cases} u_{n+1} = u_n + h\phi(t_n, u_n, f_n, h) \quad 0 \leq n \leq N_h - 1, \\ u_0 = y_0. \end{cases}$$

Definizione. Sia $\varepsilon_{n+1} := y_{n+1} - (y_n + h\phi(t_n, y_n, f(t_n, y_n), h))$ la differenza tra il valore esatto $y_{n+1} = y(t_{n+1})$ e quello dato dal metodo applicato al valore esatto $y_{n+1} = y(t_n{+}1)$. Si definisce allora **errore locale di troncamento** al nodo $(n+1)$-esimo il valore:

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

Si definisce inoltre l'**errore globale di troncamento** come massimo del modulo degli errori locali:

$$\tau(h) = \max_{n=0,\ldots,N_h-1} |\tau_{n+1}(h)|\,.$$

Definizione (Consistenza del metodo)**.** Si dice che un metodo è consistente di ordine p per il problema se $\tau(h) = o(h)$ ($\lim_{h\to 0} \tau(h) = 0$) e se $\tau(h) = O(h^p)$ per $h \to 0$.

Definizione (Convergenza del metodo)**.** Posto $e_n = y_n - u_n$, detto *errore globale*, un metodo si dice **convergente** di ordine p se esiste $C(h) = o(h)$ ($\lim_{h\to 0} C(h) = 0$), $C(h) = O(h^p)$ per $h \to 0$ tale per cui $|y_n - u_n| \leq C(h)$ per ogni n della successione.

Metodo di Eulero

Si definisce **metodo di Eulero** il metodo a un passo che si ottiene imponendo $\phi(t_n, u_n, f_n, h) = f_n$, ovverosia:

$$\begin{cases} u_{n+1} = u_n + hf_n \quad 0 \leq n \leq N_h - 1, \\ u_0 = y_0. \end{cases}$$

L'idea del metodo di Eulero è quella di approssimare innanzitutto i reali con una successione t_n prendendo $h \ll 1$ (in questo modo $N_h \gg 1$ e si ottiene una sequenza arbitrariamente lunga di reali). In questo modo l'equazione

$$y'(t) = f(t, y(t))$$

può essere approssimata ponendo l'equazione vera sui t_n con $y'(t) \approx \frac{u_{n+1}-u_n}{h}$ e $f(t, y(t)) \approx f_n$. La seconda equazione $y(t_0) = y_0$ si riconduce invece facilmente a $u_0 = y_0$.

Per ottenere dunque un'approssimazione è dunque necessario risolvere la relazione di ricorrenza tra u_{n+1} e u_n, ponendo $u_0 = y_0$, o calcolare direttamente la sequenza.

Per il metodo di Eulero vale che:

$$\varepsilon_{n+1} = y_{n+1} - (y_n + hf(t_n, y_n)).$$

Il metodo di Eulero è sempre consistente di ordine 1 ed è sempre convergente, ancora di ordine 1.

Metodo di Eulero implicito

Si definisce **metodo di Eulero implicito** il metodo che si ottiene sostituendo a $\phi(t_n, u_n, f_n, h)$ il valore f_{n+1}, ovverosia:

$$\begin{cases} u_{n+1} = u_n + hf_{n+1} \quad 0 \leq n \leq N_h - 1, \\ u_0 = y_0. \end{cases}$$

Tale metodo è consistente di ordine 1 e convergente, ancora di ordine 1.

Esempio (Problema test di Dahlquist)**.** Si consideri il seguente problema di Cauchy:

$$\begin{cases} y' = \lambda y, \\ y(0) = 1, \end{cases}$$

con $I = [0, \infty)$. Applicando il metodo di Eulero implicito si deve risolvere il sistema:

$$\left\{u_{n+1} = u_n + \lambda u_{n+1}, u_0 = 1,\right.$$

che ha come soluzione $u_n = \left(\frac{1}{1-h\lambda}\right)^n$.

Opera originale di Mario Zito, modifiche e aggiornamenti a cura di Gabriel Antonio Videtta.
Reperibile su https://notes.hearot.it.