IMPORTANTE: Premessa all'uso della dispensa

Questa dispensa è concepita per raccogliere in modo molto conciso solo gli argomenti oggetto del corso, che pur essendo sicuramente trattati in ogni libro "istituzionale" di Analisi Numerica, Calcolo Numerico, Introduzione al Calcolo Scientifico..., potrebbero esserlo con enfasi diversa da quella che viene loro assegnata all'interno di questa dispensa.

Un punto che deve essere chiaro a chi utilizza questa dispensa è che essa non è assolutamente sostitutiva di un buon libro di testo. Si tratta infatti di materiale privo di molti approfondimenti, presentato in modo molto sintetico, spesso in forma schematica, e sufficiente a preparare l'esame a patto di integrarlo con la frequenza alle lezioni. Inoltre, ovviamente, molti contenuti dell'Analisi Numerica non sono neppure trattati.

Utilizzare questo materiale didattico significa dunque assumersi la responsabilità di integrarlo con gli appunti delle lezioni, con colloqui con il docente per chiarimenti ed approfondimenti, ed eventualmente anche con altro materiale reperibile su links segnalati e con testi adeguati, scelti per esempio fra quelli indicati nella bibliografia del corso.

Introduzione al corso

Scopo del calcolo numerico.
"La matematica numerica è quel ramo della matematica che propone, sviluppa, analizza ed applica metodi per il calcolo scientifico nel contesto di vari campi, quali l'analisi, l'algebra lineare, la geometria, la teoria dell'approssimazione, la teoria delle equazioni funzionali, l'ottimizzazione, le equazioni differenziali. Anche altre discipline, come la fisica, le scienze naturali e biologiche, l'ingegneria, l'economia e la finanza, frequentemente originano problemi che richiedono di essere risolti ricorrendo al calcolo scientifico.

La matematica numerica è pertanto situata alla confluenza di diverse discipline di grande rilievo nelle moderne scienze applicate, e ne diventa strumento essenziale di indagine qualitativa e quantitativa. Tale ruolo decisivo è pure accentuato dallo sviluppo impetuoso ed inarrestabile di computer ed algoritmi, che rendono oggi possibile affrontare con il calcolo scientifico problemi di dimensioni tanto elevate da consentire la simulazione di fenomeni reali, fornendo risposte apprezzabili per accuratezza e tempo di calcolo. La corrispondente proliferazione di software numerico, se per un verso rappresenta una ricchezza, per l'altro pone spesso l'utilizzatore nella condizione di doversi orientare correttamente nella scelta del metodo (o dell'algoritmo) più efficace per affrontare il problema di suo interesse. È infatti evidente che non esistono metodi o algoritmi efficaci ed accurati per ogni tipo di problema".
[A. Quarteroni, R. Sacco e F. Saleri, *Matematica Numerica*, Springer-Verlag Italia, Milano 2000 - Prefazione]

Problema ben posto, numero di condizionamento di un problema, problema numerico.

$$\text{Problema fisico}$$
$$\Downarrow$$
$$\text{Problema matematico } P(d,u)=0$$
$$\Downarrow$$
$$\text{Problema numerico } P_n(d_n,u_n)=0$$

Si consideri il problema seguente: trovare u tale che

$$P(d,u)=0,$$

dove d è l'insieme dei dati da cui dipende la soluzione u, e P esprime la relazione funzionale tra u e d. A seconda del tipo di problema rappresentato, le variabili u e d potranno esprimere numeri reali, vettori o funzioni.

Un metodo numerico per la risoluzione approssimata del problema matematico consisterà, in generale, nel costruire una successione di problemi approssimati del tipo

$$P_n(d_n,u_n)=0,\ n\geq 1,$$

oppure,

$$P_h(d_h, u_h) = 0,\ h > 0,$$

dipendenti da un certo parametro n (risp. h), con la speranza che $u_n \to u$ per $n \to \infty$, (risp. $u_h \to u$ per $h \to 0$), ovvero la soluzione del problema numerico sia convergente alla soluzione del problema matematico.

Si dice che il problema matematico $P(d, u) = 0$ è ben posto se, per un certo dato d, la corrispondente soluzione u esiste ed è unica e dipende con continuità dai dati. Useremo il termine ben posto e stabile in maniera intercambiabile e ci occuperemo in questo corso solo di problemi ben posti.

La dipendenza continua dai dati significa che piccole perturbazioni (variazioni) sui dati d producono piccole perturbazioni nella soluzione u. Per quantificare la misura della dipendenza continua dai dati, introduciamo il concetto di numero di condizionamento di un problema. Per semplicità ci riferiamo al caso più semplice di problema matematico, che è il calcolo di una funzione $f : [a, b] \to \mathbb{R}$ in un punto x_0 fissato. Dunque:

$$d := x_0,\ u := f(x_0),\ d, u \in \mathbb{R}.$$

Consideriamo lo sviluppo di Taylor, arrestato al primo ordine, della funzione f in x con centro in x_0, ovviamente nelle opportune ipotesi di regolarità della funzione assegnata:

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + ...$$

$$\Big|\frac{f(x) - f(x_0)}{\boxed{f(x_0)}}\Big| \approx \Big|\frac{\boxed{x_0} f'(x_0)}{\boxed{f(x_0)}}\Big| \Big|\frac{x - x_0}{\boxed{x_0}}\Big|$$

Si osservi che le quantità

$$\Delta f(x_0) := \frac{f(x) - f(x_0)}{f(x_0)},$$

$$\Delta x_0 := \frac{x - x_0}{x_0}$$

rappresentano rispettivamente la variazione relativa della soluzione $u := f(x_0)$ e la variazione relativa del dato $d := x_0$. Dunque si può assumere come numero di condizionamento del calcolo di una funzione f nel punto x_0 la quantità

$$K_f(x_0) := \Big|\frac{x_0 f'(x_0)}{f(x_0)}\Big|,$$

che esprime appunto una misura di quanto la variazione relativa sul dato, (cioè Δx_0), viene amplificato nel calcolo di $f(x_0)$. In altre parole la quantità $K_f(x_0)$ esprime il rapporto tra la variazione relativa subita dalla soluzione, cioè $\Delta f(x_0)$, e la variazione relativa introdotta nel dato, cioè Δx_0.

In analogia a quanto detto per il problema matematico, affinchè il problema numerico sia a sua volta ben posto o stabile richiederemo che per ogni n (risp.

h) fissato la soluzione u_n (risp. u_h) in corrispondenza del dato d_n (risp. d_h) esista, sia unica e dipenda con continuità dai dati del problema numerico.

Esercizio.

Si consideri l'equazione di secondo grado

$$x^2 - 2p\,x + 1 = 0, \quad p \geq 1,$$

e la formula risolutiva

$$x = p \pm \sqrt{p^2 - 1}.$$

Consideriamo il condizionamento del problema del calcolo delle due radici al variare di $p \geq 1$:

$$x_+ = p + \sqrt{p^2 - 1}, \quad x_- = p - \sqrt{p^2 - 1}$$

definito da

$$K_+(p) = \left| \frac{p\,\frac{\mathrm{d}x_+}{\mathrm{d}p}}{x_+} \right|, \qquad K_-(p) = \left| \frac{p\,\frac{\mathrm{d}x_-}{\mathrm{d}p}}{x_-} \right|$$

con

$$\frac{\mathrm{d}x_+}{\mathrm{d}p} = 1 + \frac{p}{\sqrt{p^2 - 1}}, \qquad \frac{\mathrm{d}x_-}{\mathrm{d}p} = 1 - \frac{p}{\sqrt{p^2 - 1}}.$$

Si ottiene:

$$K_+(p) = \left| p\,\frac{\sqrt{p^2 - 1} + p}{\sqrt{p^2 - 1}}\,\frac{1}{p + \sqrt{p^2 - 1}} \right| = \frac{p}{\sqrt{p^2 - 1}};$$

$$K_-(p) = \left| p\,\frac{\sqrt{p^2 - 1} - p}{\sqrt{p^2 - 1}}\,\frac{1}{p - \sqrt{p^2 - 1}} \right| = \frac{p}{\sqrt{p^2 - 1}}.$$

Studiando la funzione $K(p) = K_+(p) = K_-(p)$ per $p > 1$, si ha:

$$\lim_{p \to 1^+} K(p) = +\infty, \qquad \lim_{p \to +\infty} K(p) = 1, \qquad K'(p) = \frac{-1}{\sqrt{p^2 - 1}} < 0, \ \forall p > 1.$$

Il problema del calcolo delle radici dell'equazione di secondo grado considerata risulta dunque tanto più malcondizionato quanto più p è vicino a 1, cioè quanto più le due radici sono vicine.
In generale si può dimostrare che il problema del calcolo delle radici di un'equazione di secondo grado qualsiasi è ben condizionato se le due radici sono "separate", ed è mal condizionato se le due radici sono "vicine".

Classificazione degli errori.
Nel passaggio dal problema fisico al problema numerico l'errore globale prodotto è espresso dalla differenza fra la soluzione effettivamente calcolata dal problema numerico e la soluzione del problema fisico di partenza. In quest'ottica

l'errore globale può essere visto come somma dell'errore del modello matematico (prodotto nel passaggio dal problema fisico al problema matematico) e dell'errore del modello numerico-computazionale (prodotto nel passaggio dal problema matematico al problema numerico).
In generale possiamo pensare di classificare gli errori in quattro insiemi:

1) Errori sui dati, riducibili aumentando l'accuratezza nelle misurazioni dei dati.

2) Errori dovuti al modello, controllabili nella fase propria della modellistica matematica, nel passaggio dal problema fisico al problema matematico.

3) Errori di troncamento, dovuti al fatto che nel modello numerico le operazioni di passaggio al limite vengono approssimate con operazioni che necessariamente richiedono un numero *finito* di operazioni (passaggio dal continuo al discreto).

4) Errori di arrotondamento, dovuti alla rappresentazione finita dei numeri reali sul calcolatore.

L'analisi numerica si occupa dello studio e del controllo degli errori 3 e 4.

Rappresentazione dei numeri sul calcolatore

- $x \in \mathbb{R} \to x \approx \text{float}(x) = \{\sigma(.a_1a_2...a_t)_\beta \beta^e\} \cup \{0\}$ (numero macchina)
  σ: segno di x.
  β: base della rappresentazione.
  e: esponente.
  t: numero di cifre significative.
  $a_1 \neq 0$; $0 \leq a_i \leq \beta - 1$.

$$m = \ .a_1a_2...a_t = \ \frac{a_1}{\beta} + \frac{a_2}{\beta^2} + ... + \frac{a_t}{\beta^t} \quad \text{(mantissa)}.$$

- $L \leq e \leq U, \quad L < 0,\ U > 0.$

- $|\text{float}(x)| \in [\beta^{-1}\beta^L, (1 - \beta^{-t})\beta^U]$.

- MATLAB: $\beta = 2;\ t = 53;\ L = -1021;\ U = 1024.$

Osservazione.
Il risultato di un'operazione fra numeri macchina non è necessariamente un numero macchina.

Rappresentazione di un numero reale.

$$x = \{\sigma(.a_1a_2...a_ta_{t+1}a_{t+2}..)_\beta \beta^e\} \ \in \mathbb{R}$$

1) $L \leq e \leq U,\ a_i = 0\ \forall i > t \Rightarrow$ rappresentazione esatta: $\text{float}(x) = x$.

2) $e < L \Rightarrow$ underflow; $e > U \Rightarrow$ overflow.

3) $\exists i > t : a_i \neq 0 \Rightarrow$

3.1) Troncamento

$$\text{float}(x) = \{\sigma(.a_1a_2...a_t)_\beta \beta^e\}$$

3.2) Arrotondamento

$$\text{float}(x) = \{\sigma(.a_1a_2...a_t)_\beta \beta^e\}, \ \text{se } 0 \leq a_{t+1} < \beta/2,$$

$$\text{float}(x) = \{\sigma(.a_1a_2... \underbrace{a_t + 1}_{\text{cifra } t-\text{esima}})_\beta \beta^e\}, \ \text{se } \beta/2 \leq a_{t+1} \leq \beta - 1.$$

Proprietà
Sia x un numero reale rappresentato in base β:

$$x = \{\sigma(.a_1 a_2 ... a_t a_{t+1} a_{t+2} ..)_\beta \beta^e\} \ \in \mathbb{R}, \quad a_1 \neq 0, \quad e \in [L, U].$$

Allora, se non si verifica una situazione di overflow, si ha la seguente maggiorazione:

$$\left| \frac{x - \text{float}(x)}{x} \right| \leq k\beta^{1-t},$$

dove $k = 1$ nel caso del troncamento e $k = 1/2$ nel caso dell'arrotondamento.

Definizione
La quantità

$$\boxed{\text{eps} = k\beta^{1-t}}$$

è detta precisione di macchina nel fissato sistema floating point. Si può caratterizzare eps come il più piccolo numero macchina per cui si verifica che:

$$\text{float}(1 + \text{eps}) > 1.$$

ESERCIZIO. Costruire l'insieme dei numeri macchina $F(\beta,\ t,\ L,\ U)$ con $\beta = 2,\ t = 3,\ L = -1,\ U = 2$.

SOLUZIONE. $a_i = 0$ oppure $a_i = 1$; $a_1 \neq 0 \ \rightarrow a_1 = 1$.

$$|x| = (.1a_2a_3)_2 \cdot 2^e = \left(\frac{1}{2} \ + \ \frac{a_2}{2^2} \ + \ \frac{a_3}{2^3}\right) \cdot 2^e$$

$\underline{e = -1}$

$(.100)_2 \cdot 2^{-1} = (\frac{1}{2} + \frac{0}{4} + \frac{0}{8}) \cdot \frac{1}{2} = \frac{1}{4} + \frac{0}{16} = \frac{4}{16}$

$(.101)_2 \cdot 2^{-1} = (\frac{1}{2} + \frac{0}{4} + \frac{1}{8}) \cdot \frac{1}{2} = \frac{1}{4} + \frac{1}{16} = \frac{5}{16}$

$(.110)_2 \cdot 2^{-1} = (\frac{1}{2} + \frac{1}{4} + \frac{0}{8}) \cdot \frac{1}{2} = \frac{1}{4} + \frac{2}{16} = \frac{6}{16}$

$(.111)_2 \cdot 2^{-1} = (\frac{1}{2} + \frac{1}{4} + \frac{1}{8}) \cdot \frac{1}{2} = \frac{1}{4} + \frac{3}{16} = \frac{7}{16}$

$\underline{e = 0}$

$(.100)_2 \cdot 2^0 = (\frac{1}{2} + \frac{0}{4} + \frac{0}{8}) \cdot 1 = \frac{1}{2} + \frac{0}{8} = \frac{4}{8}$

$(.101)_2 \cdot 2^0 = (\frac{1}{2} + \frac{0}{4} + \frac{1}{8}) \cdot 1 = \frac{1}{2} + \frac{1}{8} = \frac{5}{8}$

$(.110)_2 \cdot 2^0 = (\frac{1}{2} + \frac{1}{4} + \frac{0}{8}) \cdot 1 = \frac{1}{2} + \frac{2}{8} = \frac{6}{8}$

$(.111)_2 \cdot 2^0 = (\frac{1}{2} + \frac{1}{4} + \frac{1}{8}) \cdot 1 = \frac{1}{2} + \frac{3}{8} = \frac{7}{8}$

$\underline{e = 1}$

$(.100)_2 \cdot 2^1 = (\frac{1}{2} + \frac{0}{4} + \frac{0}{8}) \cdot 2 = 1 + \frac{0}{4} = \frac{4}{4}$

$(.101)_2 \cdot 2^1 = (\frac{1}{2} + \frac{0}{4} + \frac{1}{8}) \cdot 2 = 1 + \frac{1}{4} = \frac{5}{4}$

$(.110)_2 \cdot 2^1 = (\frac{1}{2} + \frac{1}{4} + \frac{0}{8}) \cdot 2 = 1 + \frac{2}{4} = \frac{6}{4}$

$(.111)_2 \cdot 2^1 = (\frac{1}{2} + \frac{1}{4} + \frac{1}{8}) \cdot 2 = 1 + \frac{3}{4} = \frac{7}{4}$

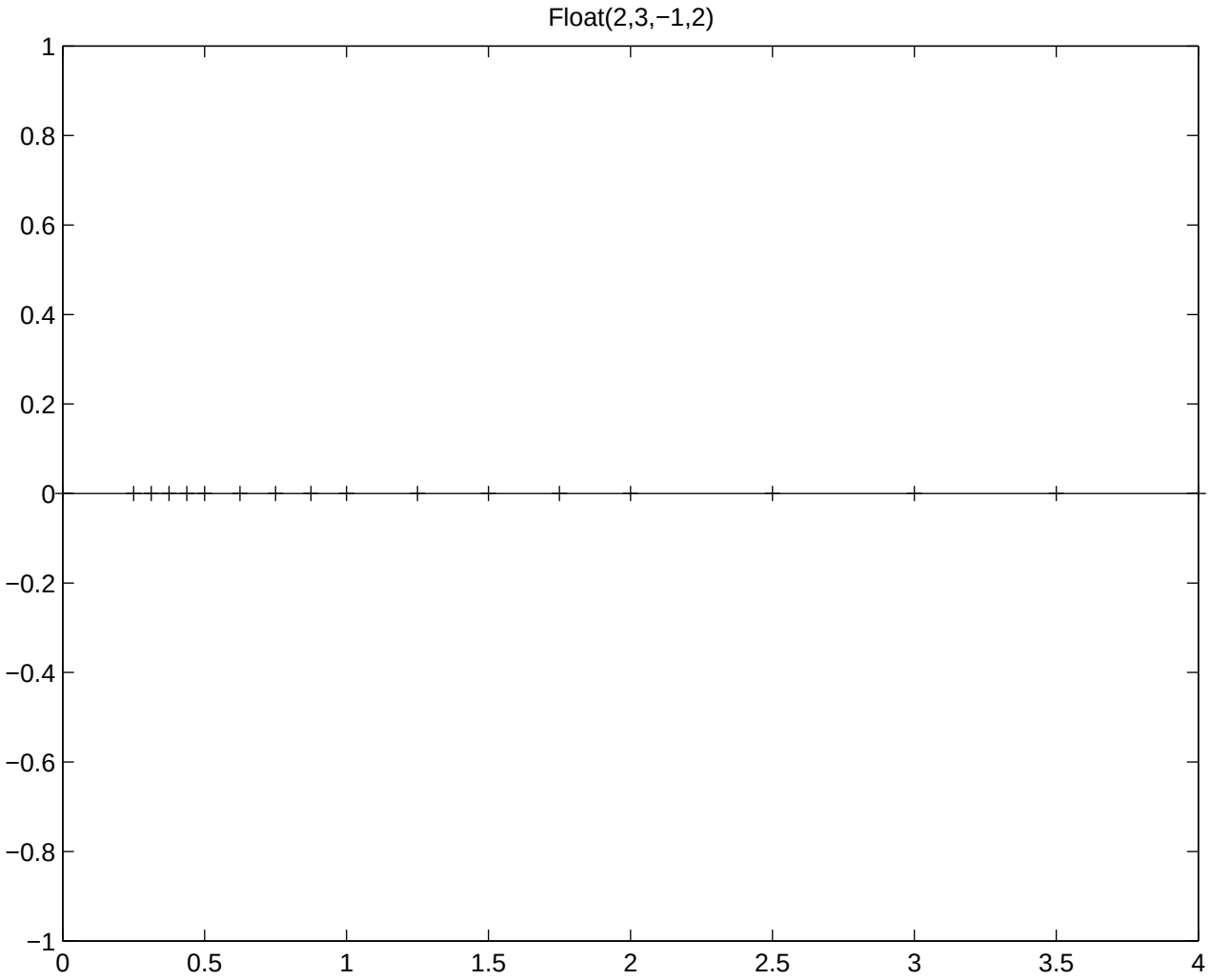
$\underline{e = 2}$

$(.100)_2 \cdot 2^2 = (\frac{1}{2} + \frac{0}{4} + \frac{0}{8}) \cdot 4 = 2 + \frac{0}{2} = \frac{4}{2}$

$(.101)_2 \cdot 2^2 = (\frac{1}{2} + \frac{0}{4} + \frac{1}{8}) \cdot 4 = 2 + \frac{1}{2} = \frac{5}{2}$

$(.110)_2 \cdot 2^2 = (\frac{1}{2} + \frac{1}{4} + \frac{0}{8}) \cdot 4 = 2 + \frac{2}{2} = \frac{6}{2}$

$(.111)_2 \cdot 2^2 = (\frac{1}{2} + \frac{1}{4} + \frac{1}{8}) \cdot 4 = 2 + \frac{3}{2} = \frac{7}{2}$


Float(2,3,−1,2)
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
0
0.5
1
1.5
2
2.5
3
3.5
4

Approssimazione di dati e funzioni

Richiamiamo i principali metodi di approssimazione polinomiale di un insieme di dati (x_i, y_i), $i = 0, ..., n$. Le ordinate y_i possono essere i valori assunti nei nodi x_i da una funzione f nota analiticamente, cioè $y_i = f(x_i)$, oppure possono rappresentare valori sperimentali. Nel primo caso, l'approssimazione mira a sostituire f con una funzione più semplice da trattare, per eseguire ad esempio poi operazioni di derivazione o integrazione. Nel secondo caso, lo scopo primario dell'approssimazione è di fornire una sintesi significativa dei dati disponibili che potrebbero essere numerosi, in modo da avere a disposizione una funzione approssimante, nel caso più semplice sarà appunto un polinomio algebrico, che rappresenti il fenomeno oggetto dell'analisi sperimentale.

Interpolazione polinomiale

Teorema.
Dati $n + 1$ punti distinti

$$x_0, x_1, ..., x_n,$$

e $n + 1$ corrispondenti valori

$$y_0, y_1, ..., y_n,$$

(eventualmente $y_i = f(x_i)$), allora $\exists!\ \pi_n \in \mathbb{P}_n$ tale che

$$\pi_n(x_i) = y_i, \quad \forall i = 0, ..., n.$$

Si dice che π_n è il polinomio interpolatore (o di interpolazione) di f rispetto ai dati (x_i, y_i).

Unicità del polinomio di interpolazione.
Per assurdo: $\exists p, q \in \mathbb{P}_n$, $p \neq q$, tali che $p(x_i) = q(x_i) = y_i$, $\forall i = 0, ..., n$. Allora $p-q \in \mathbb{P}_n$ e $p(x_i)-q(x_i) = 0$, $\forall i = 0, ..., n$, cioè in $n+1$ punti distinti $\Rightarrow p-q = 0$ ovunque, perchè, per il teorema fondamentale dell'algebra, l'unico polinomio di grado n che si annulla in $n + 1$ punti è il polinomio banale identicamente nullo. Dunque $p = q$.

Esistenza del polinomio di interpolazione.
Per dimostrare l'esistenza si procede in maniera costruttiva, proponendo tre metodi per ottenere l'espressione di π_n. Avendo a priori dimostrato l'unicità del polinomio interpolatore, si può concludere che le espressioni fornite da metodi di costruzione diversi differiranno solo per errori di arrotondamento dovuti alla rappresentazione dei numeri sul calcolatore.

1) Costruzione del polinomio di interpolazione mediante la matrice di Vandermonde.
Si impongono direttamente le $n+1$ condizioni di interpolazione per un generico polinomio π_n espresso mediante i coefficienti c_j, $j = 0, ..., n$, nella consueta base

di monomi $1, x, x^2, ..., x^n$:

$$\pi_n(x) = \sum_{j=0}^{n} c_j x^j;$$

$$\boxed{\pi_n(x_i)} = \sum_{j=0}^{n} c_j x_i^j = \boxed{y_i}\,, \qquad i = 0, ..., n;$$

$$\Downarrow$$

$$V\mathbf{c} = \mathbf{y},$$

con

$$V_{ij} = x_i^j, \quad \text{e} \quad \det(V) = \prod_{0 \le i < j \le n} (x_j - x_i).$$

Si osservi che se gli x_i sono tutti distinti, allora $\det(V) \neq 0$, cioè la matrice V è non singolare.

2) Costruzione del polinomio di interpolazione mediante il metodo di interpolazione di Lagrange.
Si costruisce una nuova base per $\mathbb{P}_n$, costituita dai polinomi di base di Lagrange $L_i(x),\ i = 0, ..., n$, che soddisfano le seguenti tre proprietà:

($\bullet$) $L_i(x) \in \mathbb{P}_n$;
($\bullet\bullet$) $L_i(x_j) = 0,\ i = 0, ..., n,\ i \neq j$;
($\bullet\bullet\bullet$) $L_i(x_i) = 1,\ i = 0, ..., n$;
[($\bullet\bullet$)+($\bullet\bullet\bullet$)$\Rightarrow L_i(x_j) = \delta_{ij}$, dove δ_{ij} è il simbolo di Kronecker]

$$\Downarrow$$

($\bullet$)+($\bullet\bullet$) $\Rightarrow L_i(x) = C(x - x_0)(x - x_1)...(x - x_{i-1})(x - x_{i+1})...(x - x_n),\ C \in \mathbb{R}$
($\bullet\bullet\bullet$) $\Rightarrow C = 1/[(x_i - x_0)(x_i - x_1)...(x_i - x_{i-1})(x_i - x_{i+1})...(x_i - x_n)]$

Si ottiene:

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

Di conseguenza il polinomio interpolatore esiste ed ha la seguente forma:

$$\boxed{\pi_n(x) = \sum_{i=0}^{n} y_i L_i(x)}$$

Infatti, $\forall k = 0, ..., n$:

$$\boxed{\pi_n(x_k)} = \sum_{i=0}^{n} y_i L_i(x_k) = y_0 L_0(x_k) + y_1 L_1(x_k) + ... + \boxed{y_k L_k(x_k)} + ... + y_n L_n(x_k) =$$

$$y_0\delta_{0k} + y_1\delta_{1k} + ... + \boxed{y_k\delta_{kk}} + ... + y_n\delta_{nk} = 0 + 0 + ... + \boxed{y_k} + ... + 0 = \boxed{y_k}$$

che è la condizione di interpolazione nel nodo $k-$esimo.

[Numero di operazioni $\approx 2n(n+1)$]

3) Costruzione del polinomio di interpolazione mediante il metodo di Newton o delle differenze divise.

Differenze divise.

$$f[x_i] := y_i, \qquad i = 0, ..., n.$$

$$f[x_i, ..., x_{i+k}] := \frac{f[x_{i+1}, ..., x_{i+k}] - f[x_i, ..., x_{i+k-1}]}{x_{i+k} - x_i}, \quad \forall k \geq 0 \text{ t.c. } i + k \leq n.$$

Teorema.
Posto

$$a_k := f[x_0, ..., x_k], \quad k = 0, ..., n,$$

si ha

$$\boxed{\pi_n(x) = a_0 + a_1(x - x_0) + ... + a_n(x - x_0)...(x - x_{n-1})}$$

[Numero operazioni $\approx \frac{n(n+1)}{2}$]

Proprietà.
Data una nuova coppia di valori (x_{n+1}, y_{n+1}), allora vale la relazione *ricorsiva*:

$$\pi_{n+1}(x) = \pi_n(x) + a_{n+1}(x - x_0)...(x - x_n),$$

dove $\pi_{n+1}(x)$ è il polinomio di grado $n+1$ che interpola i dati (x_i, y_i), $i = 0, ..., n+1$.

Osservazione.
Per passare da π_n a π_{n+1}, si richiede dunque solo il calcolo degli $n+1$ nuovi valori delle differenze divise

$$f[\underbrace{x_n, x_{n+1}}_{2 \text{ nodi}}], \ f[\underbrace{x_{n-1}, x_n, x_{n+1}}_{3 \text{ nodi}}], \ ..., f[\underbrace{x_0, x_1, ..., x_{n+1}}_{(n+1) \text{ nodi}}],$$

e infine si pone $a_{n+1} = f[x_0, x_1, ..., x_{n+1}]$.

Proprietà.
Il valore assunto dalla differenza divisa è invariante rispetto ad una permutazione degli indici dei nodi, cioè si ha:

$$f[x_0, x_1, ..., x_k] = f[x_{i_0}, x_{i_1}, ..., x_{i_k}]$$

per ogni permutazione $(i_0, i_1, .., i_k)$ dei $(k+1)$ indici $(0, 1, ..., k)$.

Algoritmo di Horner.
Valutazione efficiente del polinomio $\pi_n(x)$ nel punto x (es.: $n = 4$).

$\pi_4(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + a_3(x - x_0)(x - x_1)(x - x_2) + a_4(x - x_0)(x - x_1)(x - x_2)(x - x_3)$

$$\Downarrow$$

$\pi_4(x) = a_0 + (x - x_0)\{a_1 + (x - x_1)[a_2 + (x - x_2)\{a_3 + (x - x_3)a_4\}]\}$.

[Numero operazioni $\approx n$]

Stima dell'errore di interpolazione. Teorema.
Siano $x_0, x_1, ..., x_n$, $n+1$ nodi distinti, $x \neq x_i, i = 0, ..., n$, $f \in C^{n+1}(I_x)$, dove I_x è il più piccolo intervallo chiuso e limitato contenente i nodi $x_0, x_1, ..., x_n, x$. Allora l'errore di interpolazione nel punto x è dato da:

$$R_n(x) = f(x) - \pi_n(x) = \frac{\omega(x)}{(n+1)!} f^{(n+1)}(\xi),$$

con $\xi \in I_x$ e

$$\omega(x) = (x - x_0)(x - x_1)...(x - x_n) = \prod_{i=0}^{n} (x - x_i).$$

Dimostrazione.
Sia x un generico punto distinto dai nodi di interpolazione. Definiamo

$$R_n(x) = f(x) - \pi_n(x),$$

da cui

$$R_n(x_i) = f(x_i) - \pi_n(x_i) = 0, \quad 0 \leq i \leq n,$$

grazie alla condizione di interpolazione.
Introduciamo la funzione

$$G(z) = f(z) - \pi_n(z) - \frac{R_n(x)}{\omega(x)}\omega(z) = R_n(z) - \frac{R_n(x)}{\omega(x)}\omega(z),$$

con $G \in C^{n+1}(I_x)$, essendo infatti G somma di $f \in C^{n+1}(I_x)$ e di un polinomio. Si osservi che in questo contesto x è fissata e z è la variabile dipendente. Si ha:

$G(\ \boxed{z = x}\ ) = R_n(x) - \frac{R_n(x)}{\omega(x)}\omega(x) = 0.$
$G(\ \boxed{z = x_i}\ ) = 0,\ i = 0, ..., n$, essendo $R_n(x_i) = 0$ e $\omega(x_i) = 0$, $\forall i = 0, ..., n$.

Per il teorema di Rolle, poichè la funzione G si annulla in $n+2$ punti, la funzione G' si annulla in $n+1$ punti, la funzione G'' si annulla in n punti, e, per ricorsione, la funzione $G^{(n+1)}$ si annulla in un punto. Sia esso ξ. Calcolando la derivata $(n+1)-$esima di G rispetto alla variabile z si ottiene

$$G^{(n+1)}(z) = f^{(n+1)}(z) - \pi_n^{(n+1)}(z) - \frac{R_n(x)}{\omega(x)}\omega^{(n+1)}(z).$$

In particolare si ha:

$\pi_n^{(n+1)}(z) = 0$, essendo π_n un polinomio di grado n;
$\omega^{(n+1)}(z) = (n+1)!$

Dunque, essendo $G^{(n+1)}(\xi) = 0$ (per la scelta di ξ), si ha

$$f^{(n+1)}(\xi) - \frac{R_n(x)}{\omega(x)}(n+1)! = 0,$$

da cui la tesi

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega(x).$$

Corollario.
Nelle ipotesi del teorema precedente si ha

$$|R_n(x)| \leq \frac{\max_{\xi \in I_x} |f^{(n+1)}(\xi)|}{(n+1)!} \max_{t \in I_x} |\omega(t)|$$

Purtroppo in generale non si può dedurre dal teorema e dal corollario che l'errore tende a 0 per $n \to \infty$. Infatti esistono funzioni f per le quali l'errore può essere infinito, ossia

$$\lim_{n\to\infty} \max_{x \in I_x} |R_n(x)| = \infty.$$

A titolo di esempio è consuetudine ricordare il fenomeno di Runge, per cui se si interpola la funzione

$$f(x) = \frac{1}{x^2+1}$$

su un insieme di nodi equispaziati nell'intervallo [-5,5], si può dimostrare che

$$\lim_{n\to\infty} |R_n(x)| = \infty, \quad \forall x \text{ tale che } 3.64 < |x| < 5.$$

(si veda [A. Quarteroni, F. Saleri, *Introduzione al Calcolo Scientifico*, 2ª edizione, Springer-Verlag Italia, Milano, 2004 - Cap. 3, Esempio 3.2]).

Polinomi di Chebyshev

Il fenomeno di Runge richiamato nelle considerazioni conclusive della sezione riguardante l'interpolazione può essere evitato utilizzando opportune distribuzioni di nodi. In particolare in questa sezione introduciamo i nodi di Chebyshev che sono gli zeri dei polinomi di Chebyshev.

Definizione.
Sia $\omega(x)$ una funzione peso sull'intervallo (a, b) tale che:

$$\omega(x) \geq 0, \quad \int_a^b \omega(x)\mathrm{d}x < +\infty.$$

Indichiamo con $\{\varphi_k,\ k = 0, 1, ...\}$ una famiglia di polinomi algebrici, con

$$\text{grado}(\varphi_k) = k.$$

Diciamo che $\{\varphi_k,\ k = 0, 1, ...\}$ è una famiglia di polinomi ortogonali su (a, b) rispetto al peso $\omega(x)$ se:

$$\int_a^b \varphi_m(x)\varphi_n(x)\omega(x)\mathrm{d}x = 0, \quad m \neq n.$$

Teorema.
Se $\{\varphi_k,\ k = 0, 1, ...\}$ è una famiglia di polinomi ortogonali su (a, b) rispetto al peso $\omega(x)$, allora il polinomio φ_k ha k radici reali e distinte appartenenti all'intervallo aperto (a, b).

Polinomi di Legendre.

$\omega(x) = 1, \quad (a, b) = (-1, 1).$

Formula ricorsiva:
$P_0 = 1,\ P_1(x) = x,\ (n + 1)P_{n+1}(x) = (2n + 1)xP_n(x) - nP_{n-1}(x),\ n \geq 1.$

Polinomi di Chebyshev.

$\omega(x) = 1/\sqrt{1 - x^2}, \quad (a, b) = (-1, 1).$
$T_k(x) = \cos k\theta, \quad \theta = \arccos x, \quad k \geq 0$

Formula ricorsiva:
$T_0 = 1,\ T_1(x) = x,\ T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)\ ,\ n \geq 1.$

La formula ricorsiva si ottiene osservando che:

$T_{n+1}(x) = \cos[(n+1)\theta]$
$T_{n-1}(x) = \cos[(n-1)\theta]$

Sommando membro a membro e applicando la formula di addizione

$$\cos(\alpha \pm \beta) = \cos\alpha\cos\beta \mp \sin\alpha\sin\beta,$$

si ha:

$$T_{n+1}(x) + T_{n-1}(x) = \cos n\theta\cos\theta - \sin n\theta\sin\theta + \cos n\theta\cos\theta + \sin n\theta\sin\theta \Rightarrow$$
$$T_{n+1}(x) = 2\cos n\theta\cos\theta - T_{n-1}(x)$$

da cui si ottiene la formula ricorsiva sfruttando la definizione implicita di $T_n(x)$ e osservando banalmente che $\cos\theta = x$.

Zeri del polinomio $T_n(x)$ sull'intervallo $(-1, 1)$
Cerchiamo n ascisse $x_j,\ j = 1, ..., n$ tali che $T_n(x_j) = 0$, cioè $\cos n\theta_j = 0$, dove $\theta_j = \arccos x_j$. Dunque

$$n\theta_j = (2j-1)\frac{\pi}{2},\ j = 1, ..., n$$

($n\theta_j,\ j = 1, ..., n$, sono i primi n archi che sono multipli dispari di $\pi/2$):

$$x_j = \cos\theta_j = \cos\left(\frac{2j-1}{2n}\pi\right), \quad j = 1, ..., n.$$

Zeri del polinomio $T_n(x)$ sull'intervallo (a, b):

$$t_j = \frac{b-a}{2}x_j + \frac{b+a}{2},\ j = 1, ..., n$$

La formula si ottiene utilizzando l'equazione della retta che passa per i punti di coordinate $(-1, a),\ (1, b)$, e che trasforma quindi l'intervallo $[-1, 1]$ nell'intervallo $[a, b]$.

Teorema di Bernstein.
Se $f \in C^1[a, b]$, il polinomio π_n di grado n interpolante f negli $n+1$ zeri del polinomio di Chebyshev di grado $n+1$ converge uniformemente a f su $[a, b]$, per $n \to \infty$, cioè:

$$\lim_{n\to\infty} \max_{a\le x\le b} |f(x) - \pi_n(x)| = 0.$$

Altre proprietà dei polinomi di Chebyshev

1) Proprietà di ortogonalità.

$$\int_{-1}^{1} T_k(x)T_n(x)\ \frac{1}{\sqrt{1-x^2}}\ \mathrm{d}x = 0, \quad \text{se } k \neq n$$

$$\int_{-1}^{1} T_n(x)T_n(x)\ \frac{1}{\sqrt{1-x^2}}\ \mathrm{d}x = \pi, \quad \text{se } n = 0$$

$$\int_{-1}^{1} T_n(x)T_n(x)\ \frac{1}{\sqrt{1-x^2}}\ \mathrm{d}x = \frac{\pi}{2}, \quad \text{se } n > 0$$

2) $T_n(x)$ ha coefficiente direttivo 2^{n-1}, $n \geq 1$.

3) $T_n(-x) = (-1)^n T_n(x)$.

4) $T_n(1) = 1,\ \forall n \geq 0$.

5) Massimi e minimi interni: $\max |T_n(x)| = 1$ se $|\cos n\theta| = 1$.

$$n\theta_k = k\pi,\ \theta_k = \frac{k\pi}{n},\ x_k = \cos\frac{k\pi}{n},\ k = 1, ..., n-1.$$

Splines interpolanti

Si è osservato che per distribuzioni equispaziate di nodi di interpolazione, non è garantita la convergenza uniforme di π_n, polinomio di interpolazione, alla funzione f. Tuttavia l'interpolazione di Lagrange risulta ragionevolmente accurata per gradi bassi, a patto di interpolare su intervalli sufficientemente piccoli. È pertanto naturale suddividere l'intervallo di interpolazione in un certo numero di sottointervalli e applicare su ogni sottointervallo l'interpolazione semplice di Lagrange con grado basso, ottenendo una funzione interpolatrice che è globalmente continua e localmente un polinomio algebrico. All'interno di ciascun sottointervallo si potrà utilizzare la stima dell'errore dimostrata per l'interpolazione semplice.

Si possono definire poi interpolazioni composite che globalmente siano più che continue. Infatti in molte applicazioni, come ad esempio in *computer graphics*, è necessario utilizzare funzioni approssimanti che siano almeno derivabili con continuità. A questo scopo si introducono le funzioni splines.

Definizione.
Dato un insieme di punti (x_i, y_i), $i = 0, ..., n$, con $a = x_0 < x_1, ... < x_n = b$, una spline di grado p interpolante è una funzione: $x \to S^p(x)$ tale che:

1) su ogni $[x_{i-1}, x_i]$ S^p è un polinomio di grado p, $i = 1, ..., n$;

2) S^p e le sue prime $(p-1)$ derivate sono continue, cioè $S^p \in C^{p-1}[a, b]$;

3) $S^p(x_i) = y_i,\ i = 0, ..., n$.

Dalla definizione discende che un qualunque polinomio di grado p su $[a, b]$ è una spline. Tuttavia nella pratica una spline sarà rappresentata da un polinomio diverso su ciascun intervallo e, per questo motivo, potrebbe presentare una discontinuità della derivata $p-$esima nei nodi interni x_1,...,x_{n-1}. I nodi per i quali ciò avviene effettivamente sono detti nodi attivi.

Gradi di libertà (d.o.f.: degrees of freedom)
Numero di incognite: $N_x = n * (p + 1)$, essendo n il numero di sottointervalli e $p+1$ il numero di coefficienti necessari per rappresentare un polinomio di grado p su ciascun intervallo.

Numero di condizioni: $N_c = p*(n-1)+(n+1)$, essendo p il numero di condizioni di continuità che si devono imporre in ciascuno dei nodi interni x_i, $i = 1, ..., n-1$ e $n + 1$ il numero di condizioni di interpolazione.

d.o.f.: $N_x - N_c = n * (p + 1) - p * (n - 1) - (n + 1) = p - 1$.

Nel caso di $p > 1$ si dovranno dunque introdurre condizioni aggiuntive per eliminare i gradi di libertà.

• $p = 1$ (d.o.f.=0).

Stima dell'errore. Teorema.
$f \in C^2([a,b]) \Rightarrow$

$$|f(x) - S^1(x)| \leq \frac{1}{8}h^2 M,$$

dove:

$$h = \max_{0\leq i\leq n-1} h_i = \max_{0\leq i\leq n-1}(x_{i+1} - x_i), \quad M = \max_{a\leq t\leq b} |f''(t)|.$$

Dimostrazione.
Sia $x \in [a,b]$, con $x \neq x_i$, $\forall i = 0, ..., n$. Sia $[x_{i-1}, x_i]$ l'intervallo a cui appartiene x. Poichè su $[x_{i-1}, x_i]$ la funzione spline lineare $S^1(x)$ è un polinomio di grado 1 si può sfruttare la stima dell'errore di interpolazione:

$$f(x) - S^1(x) = \frac{(x - x_{i-1})(x - x_i)}{2!} f^{(2)}(t), \quad t \in [x_{i-1}, x_i].$$

Si ottiene pertanto la maggiorazione:

$$|f(x) - S^1(x)| \leq \frac{1}{2} \max_{x_{i-1}\leq \overline{x}\leq x_i} |(\overline{x} - x_{i-1})(\overline{x} - x_i)| \max_{x_{i-1}\leq t\leq x_i} |f^{(2)}(t)|$$

Ponendo

$$h_i := \quad x_i - x_{i-1}, \qquad \forall i = 1, ..., n,$$

$$h := \quad \max_{1\leq i\leq n} h_i,$$

e osservando che

$$\max_{x_{i-1}\leq \overline{x}\leq x_i} |(\overline{x} - x_{i-1})(\overline{x} - x_i)| = \left| \left(\frac{x_i + x_{i-1}}{2} - x_{i-1} \right) \left(\frac{x_i + x_{i-1}}{2} - x_i \right) \right| = \frac{h_i}{2} \cdot \frac{h_i}{2}$$

si ha

$$|f(x) - S^1(x)| \leq \frac{1}{8}h^2 \max_{x_0\leq t\leq x_n} |f^{(2)}(t)|.$$

Osservazione.
Le splines lineari sono ampiamente utilizzate nell'approssimazione della soluzione di problemi ai limiti. In tale contesto le splines S^1 sono più comunemente indicate come elementi finiti del primo ordine.

Splines cubiche

• $p = 3 :\ (d.o.f. :\ p - 1 = 2)$.

Condizioni aggiuntive:

• naturale: $(S^3)''(a) = (S^3)''(b) = 0$.
• vincolata: $(S^3)'(a) = f'(a),\ (S^3)'(b) = f'(b)$.
• periodica: $(S^3)'(a) = (S^3)'(b),\ (S^3)''(a) = (S^3)''(b)$.
• not-a-knot: $(S^3)'''$ continua in x_1 e x_{n-1} (MATLAB).

Nell'ultimo caso significa che x_1 e x_{n-1} non sono nodi attivi (*not-a-knot*), poichè l'espressione della spline nel primo intervallo coincide con l'espressione nel secondo intervallo, e analogamente per il penultimo e ultimo intervallo.

Stima dell'errore.
Sia $f \in C^4([a, b])$ e, per semplicità, sia $h = h_i,\ i = 1, ..., n$. Allora esistono delle costanti C_k indipendenti da h tali che:

$$\max_{a \le x \le b} |f^{(k)}(x) - (S^3)^{(k)}(x)| \le C_k h^{(4-k)} \max_{a \le t \le b} |f^{(4)}(t)|, \quad k = 0, 1, 2,$$

dove S^3 è la spline cubica vincolata interpolante f nei nodi x_i, $i = 0, ..., n$.

Si osserva quindi che, quando la funzione è sufficientemente regolare, si ha convergenza non solo della spline, ma anche delle sue derivate fino all'ordine 2.

Osservazione.
Le splines cubiche di grado 3 di tipo interpolatorio hanno un rilievo particolare, in quanto sono le splines di grado minimo che consentono di ottenere approssimazioni almeno di classe C^2 e sono sufficientemente regolari in presenza di piccole curvature.

Minimi quadrati discreti

Sia dato un insieme di punti $(x_i, y_i),\ i = 0, ..., n$, dove eventualmente $y_i = f(x_i)$. Si vuole determinare un polinomio

$$\hat{p}_m(x) = \sum_{j=0}^{m} \hat{a}_j x^j,$$

di grado m (in genere $m << n$) che renda minima la funzione

$$E(a_0, a_1, ..., a_m) = \boxed{\sum_{i=0}^{n} [y_i - p_m(x_i)]^2} = \sum_{i=0}^{n} [y_i - \sum_{j=0}^{m} a_j x_i^j]^2,$$

cioè tale per cui:

$$E(\hat{a}_0, \hat{a}_1, ..., \hat{a}_m) \leq E(a_0, a_1, ..., a_m),$$

al variare dei coefficienti a_0, a_1,...a_m dei polinomi

$$p_m(x) = \sum_{j=0}^{m} a_j x^j,$$

di grado m.
Diremo che $\hat{p}_m$ approssima l'insieme di dati nel senso dei minimi quadrati.
Si può dimostrare che il punto di minimo per la funzione $E(a_0, a_1, ..., a_m)$ si ottiene imponendo le seguenti condizioni:

$$\frac{\partial E(a_0, a_1, ..., a_m)}{\partial a_0} = 0, \quad a_j = \text{costante } \forall j \neq 0$$
$$\frac{\partial E(a_0, a_1, ..., a_m)}{\partial a_1} = 0, \quad a_j = \text{costante } \forall j \neq 1$$
$$...$$
$$\frac{\partial E(a_0, a_1, ..., a_m)}{\partial a_m} = 0, \quad a_j = \text{costante } \forall j \neq m$$

(∂= simbolo di derivata parziale).

Si osservi che se i nodi x_i sono tutti distinti, per $m = n$ si ha un classico problema di interpolazione, con $E(\hat{a}_0, \hat{a}_1, ..., \hat{a}_m) = 0$.

Casi particolari

$\boxed{m = 0}$ $p_0(x) = a_0$,

$$E(a_0) = \sum_{i=0}^{n} (y_i - a_0)^2.$$

$$\frac{dE(a_0)}{da_0} = \frac{\partial E(a_0)}{\partial a_0} = \sum_{i=0}^{n} 2(y_i - a_0)(-1) = 0 \Rightarrow \boxed{\hat{a}_0 = \frac{1}{n+1}\sum_{i=0}^{n} y_i}$$

dunque la soluzione $\hat{p}_0(x) = \hat{a}_0$ è data dalla media dei valori y_i.

$\boxed{m = 1}$ $p_1(x) = a_0 + a_1 x$,

$$E(a_0, a_1) = \sum_{i=0}^{n} [(y_i - a_0 - a_1 x_i)]^2.$$

$$\frac{\partial E(a_0, a_1)}{\partial a_0} = \sum_{i=0}^{n} 2[(y_i - a_0 - a_1 x_i)](-1) = 0 \Rightarrow \boxed{a_0(n+1) + a_1 \sum_{i=0}^{n} x_i = \sum_{i=0}^{n} y_i}$$

$$\frac{\partial E(a_0, a_1)}{\partial a_1} = \sum_{i=0}^{n} 2[(y_i - a_0 - a_1 x_i)](-x_i) = 0 \Rightarrow \boxed{a_0 \sum_{i=0}^{n} x_i + a_1 \sum_{i=0}^{n} x_i^2 = \sum_{i=0}^{n} x_i y_i}$$

La soluzione del sistema 2×2, noto come sistema delle equazioni normali:

$$\begin{bmatrix} (n+1) & \sum_{i=0}^{n} x_i \\ \sum_{i=0}^{n} x_i & \sum_{i=0}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} \hat{a}_0 \\ \hat{a}_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{n} y_i \\ \sum_{i=0}^{n} x_i y_i \end{bmatrix}$$

fornisce i coefficienti del polinomio $\hat{p}_1(x) = \hat{a}_0 + \hat{a}_1 x$, noto come retta dei minimi quadrati o retta di regressione.
Si può verificare che il punto di coordinate (M_x, M_y) appartiene alla retta di equazione $y = \hat{a}_0 + \hat{a}_1 x$, dove

$$M_x := \frac{1}{n+1}\sum_{i=0}^{n} x_i, \qquad M_y := \frac{1}{n+1}\sum_{i=0}^{n} y_i.$$

$\boxed{m > 1}$

$\forall k = 0, ..., m$:

$$\frac{\partial E(a_0, a_1, ..., a_m)}{\partial a_k} = 0, \ a_j = \text{costante}, \ j \neq k.$$

Si ottiene il sistema delle equazioni normali $A\hat{\mathbf{a}} = \mathbf{z}$, con

$$A_{k,j} = \sum_{i=0}^{n} x_i^{k+j}, \quad k, j = 0, ..., m,$$

$$z_k = \sum_{i=0}^{n} x_i^k y_i, \quad k = 0, ..., m.$$

Per esteso:

$$\begin{bmatrix} (n+1) & \sum_{i=0}^{n} x_i & \dots & \sum_{i=0}^{n} x_i^m \\ \sum_{i=0}^{n} x_i & \sum_{i=0}^{n} x_i^2 & \dots & \sum_{i=0}^{n} x_i^{m+1} \\ \dots & \dots & \dots & \dots \\ \sum_{i=0}^{n} x_i^m & \sum_{i=0}^{n} x_i^{m+1} & \dots & \sum_{i=0}^{n} x_i^{2m} \end{bmatrix} \begin{bmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \dots \\ \hat{a}_m \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{n} y_i \\ \sum_{i=0}^{n} x_i y_i \\ \dots \\ \sum_{i=0}^{n} x_i^m y_i \end{bmatrix}$$

(Si osservi la numerazione inusuale di righe e colonne a partire dall'indice 0).

Integrazione numerica

Approssimazione numerica di:

$$I(f) = \int_a^b f(x)\mathrm{d}x.$$

Motivazioni.

- Non sempre si riesce a trovare la forma esplicita della primitiva.
- Valutazione 'costosa' della primitiva.
- La funzione da integrare può essere data non in forma analitica, ma solo per punti.

Formule di quadratura.

In generale: se $\tilde{f}$ è un'approssimazione di $f \Rightarrow I(f) \approx I(\tilde{f})$.

$$I(f) := \int_a^b f(x)\mathrm{d}x \approx \boxed{\tilde{I}(f) := \sum_{i=1}^{n} f(x_i)\alpha_i}$$

x_i:nodi, α_i:pesi, $i = 1, ..., n$.

Dunque l'integrale definito si approssima mediante una formula di quadratura che è data dalla combinazione lineare di valori della funzione (ed eventualmente in formule più complesse anche delle derivate) in punti opportuni (nodi), moltiplicati per dei coefficienti opportuni (pesi).

Definizione. Si definisce grado di precisione di una formula di quadratura il massimo intero $r \geq 0$ tale che

$$\tilde{I}(f) = I(f), \quad \forall f \in \mathbb{P}_r.$$

Si può dimostrare che una formula di quadratura ha grado di precisione $r \Leftrightarrow$ $\boxed{I(x^k) = \tilde{I}(x^k),\ \forall k = 0, 1, ..., r}$.
Si osservi che

$$r = 0 \Leftrightarrow \sum_{i=1}^{n} \alpha_i = (b - a).$$

Formule di quadratura di tipo interpolatorio.

$$\tilde{I}(f) = \int_a^b \Pi_{n-1}(x)\mathrm{d}x = \int_a^b \left(\sum_{i=1}^{n} L_i(x) f(x_i) \right) \mathrm{d}x = \sum_{i=1}^{n} f(x_i) \int_a^b L_i(x)\mathrm{d}x$$

$$\Rightarrow \boxed{\alpha_i = \int_a^b L_i(x)\mathrm{d}x} \quad (L_i = \prod_{\substack{j=1 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} : \text{polinomi di Lagrange}).$$

Formule aperte: $a < x_1 < x_2 < ... < x_{n-1} < x_n < b$.
Formule chiuse: $a = x_1 < x_2 < ... < x_{n-1} < x_n = b$.

Formule di quadratura di Newton-Cotes (nodi equidistanti)

Formule aperte: $h = \frac{b-a}{n+1}$, $x_i = a + ih,\ i = 1, ..., n$.

Formule chiuse: $h = \frac{b-a}{n-1}$, $x_i = a + (i-1)h,\ i = 1, ..., n$, con $a = x_1$ e $b = x_n$.

Formula del punto medio (aperta, $n = 1$): Area di un rettangolo.

$$\tilde{f} = \pi_0,\ x_1 = \frac{a+b}{2},\ \alpha_1 = (b-a)$$

$$I(f) \approx \boxed{\tilde{I}_{PM}(f) := (b-a) f\left(\frac{b+a}{2}\right)}$$

$$I(f) - \tilde{I}_{PM}(f) = \frac{(b-a)^3}{24} f^{(2)}(t),\ t \in (a,b),\ \text{ se } f \in C^2[a,b].$$

Formula dei trapezi (chiusa, $n = 2$): Area di un trapezio.

$$\tilde{f} = \pi_1,\ x_1 = a,\ x_2 = b,\ \alpha_1 = \alpha_2 = \frac{b-a}{2}$$

$$I(f) \approx \boxed{\tilde{I}_T(f) := \frac{b-a}{2}\left[f(a) + f(b)\right]}$$

$$I(f) - \tilde{I}_T(f) = -\frac{(b-a)^3}{12} f^{(2)}(t),\ t \in (a,b),\ \text{ se } f \in C^2[a,b].$$

Formula di Cavalieri-Simpson (chiusa, $n = 3$): Area di un segmento parabolico.

$$\tilde{f} = \pi_2,\ x_1 = a,\ x_2 = \frac{a+b}{2},\ x_3 = b,\ \alpha_1 = \alpha_3 = \frac{b-a}{6},\ \alpha_2 = \frac{4}{6}(b-a)$$

$$I(f) \approx \boxed{\tilde{I}_{CS}(f) := \frac{b-a}{6}\left[f(a) + 4f\left(\frac{b+a}{2}\right) + f(b)\right]}$$

$$I(f) - \tilde{I}_{CS}(f) = -\frac{(b-a)^5}{2880} f^{(4)}(t),\ t \in (a,b),\ \text{ se } f \in C^4[a,b].$$

Formule di quadratura di Newton-Cotes composite.
Le formule di quadratura composite, che sono le formule più comunemente usate, si definiscono operando una preliminare suddivisione dell'intervallo di integrazione $[a,b]$ in sottointervalli e, utilizzando la proprietà additiva dell'integrale, si scrive l'integrale assegnato come una somma di integrali definiti su ciascun intervallo della suddivisione e si approssimano tali integrali definiti mediante

formule di quadratura semplici. I motivi che suggeriscono l'introduzione delle formule composite sono sostanzialmente gli stessi che suggeriscono l'introduzione dell'interpolazione composita.

Posto $M \geq 1$:

$$H = \frac{b-a}{M}, \quad a_i = a + iH,\ i = 0, ..., M,\ a = a_0,\ b = a_M.$$

($M = 1 \Rightarrow$ Formule di quadratura semplici).

• Formula del punto medio composita

$$\boxed{\tilde{I}_{PM}^{(c)}} = \boxed{H \sum_{i=1}^{M} f\left(a_i - \frac{H}{2}\right)}$$

Errore:

$$I(f) - \tilde{I}_{PM}^{(c)} = \quad \tfrac{b-a}{24} H^2 f^{(2)}(\eta),\ \eta \in (a,b) \quad \text{(formula classica)}$$

$$I(f) - \tilde{I}_{PM}^{(c)} = \quad \tfrac{H^2}{24}[f^{(1)}(b) - f^{(1)}(a)] \qquad \text{(formula asintotica).}$$

• Formula dei trapezi composita

$$\boxed{\tilde{I}_{T}^{(c)}} = \frac{H}{2} \sum_{i=1}^{M} [f(a_{i-1}) + f(a_i)] = \boxed{\frac{H}{2}\left[f(a) + f(b) + 2\sum_{i=1}^{M-1} f(a_i)\right]}$$

Errore:

$$I(f) - \tilde{I}_{T}^{(c)} = \quad -\tfrac{b-a}{12} H^2 f^{(2)}(\eta),\ \eta \in (a,b) \quad \text{(formula classica)}$$

$$I(f) - \tilde{I}_{T}^{(c)} = \quad \tfrac{H^2}{12}[f^{(1)}(a) - f^{(1)}(b)] \qquad \text{(formula asintotica).}$$

• Formula di C. Simpson composita

$$\boxed{\tilde{I}_{CS}^{(c)}} = \frac{H}{6} \sum_{i=1}^{M} \left[f(a_{i-1}) + 4f\left(a_i - \frac{H}{2}\right) + f(a_i)\right] =$$

$$\boxed{\frac{H}{6}\left[f(a) + f(b) + 2\sum_{i=1}^{M-1} f(a_i) + 4\sum_{i=1}^{M} f\left(a_i - \frac{H}{2}\right)\right]}$$

Errore:

$$I(f) - \tilde{I}_{CS}^{(c)} = \quad -\frac{b-a}{2880} H^4 f^{(4)}(\eta),\ \eta \in (a,b) \quad \text{(formula classica)}$$

$$I(f) - \tilde{I}_{CS}^{(c)} = \quad \frac{H^4}{2880}[f^{(3)}(a) - f^{(3)}(b)] \qquad \text{(formula asintotica).}$$

Traccia della dimostrazione della stima dell'errore per la formula dei trapezi compositi.

Partendo dalla stima dell'errore per la formula dei trapezi (semplice):

$$\int_a^b f(x)\mathrm{d}x - \tilde{I}_T(f) = -\frac{(b-a)^3}{12} f^{(2)}(\xi),\ \xi \in (a,b), \quad f \in C^2[a,b],$$

e applicandola a ciascun sottointervallo $[a_{j-1}, a_j]$, con $H = a_j - a_{j-1}$, si ha

$$\int_a^b f(x)\mathrm{d}x - \sum_{j=1}^{M} \frac{H}{2}[f(a_{j-1} + f(a_j)] = \boxed{\sum_{j=1}^{M} \left(-\frac{H^3}{12}\right) f''(\xi_j)}, \quad \xi_j \in [a_{j-1}, a_j].$$

Da qui in poi si può procedere in due diversi modi che portano alle due diverse stime:

1. Stima 'classica'.

$$= \boxed{-\frac{H^3}{12} \sum_{j=1}^{M} f''(\xi_j)} = -\frac{1}{12}\frac{(b-a)^3}{M^3} \sum_{j=1}^{M} f''(\xi_j) = -\frac{1}{12}\frac{(b-a)^2}{M^2}(b-a) \left[\frac{1}{M}\sum_{j=1}^{M} f''(\xi_j)\right]$$

Si osserva:

$$\min_{x\in[a,b]} f''(x) \le f''(\xi_j) \le \max_{x\in[a,b]} f''(x), \quad \forall j$$

Sommando rispetto a $j = 1, ..., M$:

$$M \min_{x\in[a,b]} f''(x) \le \sum_{j=1}^{M} f''(\xi_j) \le M \max_{x\in[a,b]} f''(x)$$

$$\min_{x\in[a,b]} f''(x) \le \underbrace{\frac{1}{M}\sum_{j=1}^{M} f''(\xi_j)}_{F} \le \max_{x\in[a,b]} f''(x)$$

Nell'ipotesi che $f \in C^2[a,b]$, allora $f'' \in C^0[a,b]$, dunque $\exists \xi \in [a,b]$ tale che $f''(\xi) = F$. Infatti grazie all'ipotesi di continuità di f'', si ha che f'' assume in almeno un punto dell'intervallo (qui abbiamo definito il punto con la variabile ξ) un qualsiasi valore compreso fra il minimo e il massimo valore assunto da f'' nell'intervallo (che qui abbiamo definito con la variabile F).
Concludendo si ha:

$$-\frac{1}{12}\frac{(b-a)^2}{M^2}(b-a) \underbrace{\left[\frac{1}{M}\sum_{j=1}^{M} f''(\xi_j)\right]}_{F} = \boxed{-\frac{1}{12}\frac{(b-a)^3}{M^2} f''(\xi)}, \quad \xi \in [a,b]$$

2. Stima 'asintotica'

$$\boxed{\sum_{j=1}^{M}\left(-\frac{H^3}{12}\right)f''(\xi_j)} = -\frac{H^2}{12}\underbrace{\sum_{j=1}^{M} Hf''(\xi_j)}_{\text{somma di Riemann}}$$

Si ha:

$$\sum_{j=1}^{M} Hf''(\xi_j) \longrightarrow \int_a^b f''(x)\mathrm{d}x, \quad \text{per } H \to 0$$

Ma

$$\int_a^b f''(x)\mathrm{d}x = f'(b) - f'(a).$$

Dunque, con buona approssimazione, si pone:

$$-\frac{H^2}{12}\sum_{j=1}^{M} Hf''(\xi_j) = \boxed{\frac{H^2}{12}[f'(a) - f'(b)]}\,.$$

Formule di quadratura di Gauss

Consideriamo il problema della costruzione di formule di quadratura in tre casi diversi.

1. Una formula di quadratura interpolatoria

$$I(f) = \int_a^b f(x)\mathrm{d}x \approx \tilde{I}(f) = \sum_{j=1}^{n} \alpha_j f(x_j)$$

che utilizza n punti distinti ha grado di precisione $\geq n-1$.

Vale anche il viceversa.

Una formula di quadratura che ha grado di precisione $\geq n-1$ è necessariamente interpolatoria, cioè dati i nodi x_1, x_2,...,x_n deve essere

$$\alpha_j = \int_a^b L_j(x)\mathrm{d}x$$

dove $L_j(x)$ è il $j-$esimo polinomio di Lagrange di grado $n-1$.

Infatti: $I(x^k) = \tilde{I}(x^k),\ 0 \leq k \leq n-1 \Rightarrow$

$$\sum_{j=1}^{n} \alpha_j x_j^k = \left(\int_a^b x^k \mathrm{d}x \right) = \frac{b^{k+1} - a^{k+1}}{k+1}, 0 \leq k \leq n-1$$

Il sistema $A\alpha = \mathbf{f}$, con

$$A_{kj} = x_j^k,\ 1 \leq j \leq n,\ 0 \leq k \leq n-1, \quad f_k = \frac{b^{k+1} - a^{k+1}}{k+1},\ 0 \leq k \leq n-1$$

ha per matrice A la matrice trasposta della matrice di Vandermonde ed ha pertanto determinante diverso da zero, essendo i nodi tutti x_j distinti e quindi il sistema ammette una ed una sola soluzione α. Dunque

$$\alpha_j = \int_a^b L_j(x)\mathrm{d}x,$$

essendo una soluzione che verifica il sistema, cioè tale per cui la formula di quadratura ha grado di precisione $\geq n-1$, è l'unica soluzione del sistema $A\alpha = \mathbf{f}$. Infatti, se per assurdo la soluzione fosse

$$\alpha_j \neq \int_a^b L_j(x)\mathrm{d}x$$

esisterebbero due soluzioni distinte in corrispondenza delle quali si verificherebbe $I(x^k) = \tilde{I}(x^k),\ 0 \leq k \leq n-1$, e dunque la soluzione del sistema con matrice A non singolare non sarebbe unica.

2. Nel caso 1. abbiamo considerato il problema di come determinare i pesi $\alpha_j,\ j = 1, ..., n$ di una formula di quadratura, una volta fissati i nodi $x_j,\ j = 1, ..., n$, in modo che la formula di quadratura abbia grado di precisione massimo.

   Ci si può porre il problema inverso: dati i pesi (per semplificare si considerino i pesi tutti uguali: $\alpha = \alpha_j, j = 1, ..., n$), determinare i nodi $x_j,\ j = 1, ..., n$, in modo che la formula di quadratura abbia grado di precisione massimo.

   Fissato $\alpha = \alpha_j = (b-a)/n$ in modo che $\sum_{j=1}^{n} \alpha_j = b - a$ e che quindi la formula di quadratura abbia grado di precisione almeno zero, si impongono le relazioni

$$I(x^k) = \tilde{I}(x^k),\ 1 \le k \le n$$

$$\alpha \sum_{j=1}^{n} x_j^k = \frac{b^{k+1} - a^{k+1}}{k+1},\ 1 \le k \le n$$

   che è un SISTEMA NON LINEARE di n equazioni in n incognite.

3. Nel caso più generale ci si pone il problema di come determinare i pesi $\alpha_j,\ j = 1, ..., n$ e i nodi $x_j,\ j = 1, ..., n$ di una formula di quadratura in modo che la formula di quadratura abbia grado di precisione massimo. Poichè si pretende che il grado di precisione sia il massimo possibile, dovrà essere almeno $\ge n-1$, dunque la formula dovrà esere di tipo interpolatorio, cioè, dati i nodi x_j i pesi si dovranno calcolare mediante le formule

$$\alpha_j = \int_a^b L_j(x)\mathrm{d}x.$$

   Consideriamo per semplicità il caso $(a, b) = (-1, 1)$.

   Ci si chiede quale è il massimo intero m per cui si ha

$$I(p_{n+m}) = \tilde{I}(p_{n+m}), \quad \forall p_{n+m} \in \mathbb{P}_{n+m}.$$

   Dividendo il *generico* polinomio $p_{n+m}(x)$ per $w(x) = \prod_{j=1}^{n}(x - x_j)$ si ha

$$p_{n+m}(x) = w(x)p_m(x) + q_{n-1}(x),$$

   dove $p_m(x)$ è il *generico* quoziente di grado m e $q_{n-1}(x)$ è il *generico* resto di grado $n - 1$.

   Si osservi in particolare che $p_{n+m}(x_j) = q_{n-1}(x_j),\ j = 1, ..., n$, essendo $w(x_j) = 0$.

   E' noto inoltre che, considerando una formula di quadratura di tipo interpolatorio, si ha

$$I(g) = \tilde{I}(g), \quad \forall g \in \mathbb{P}_{n-1}$$

   Dunque si ottiene:

$$\underbrace{I(p_{n+m})} = \int_{-1}^{1} p_{n+m}(x)\mathrm{d}x = \int_{-1}^{1} w(x)p_m(x)\mathrm{d}x + \int_{-1}^{1} q_{n-1}(x)\mathrm{d}x =$$

$$\int_{-1}^{1} w(x)p_m(x)\mathrm{d}x + \sum_{j=1}^{n} \alpha_j q_{n-1}(x_j) =$$

$$\int_{-1}^{1} w(x)p_m(x)\mathrm{d}x + \sum_{j=1}^{n} \alpha_j [w(x_j)p_m(x_j) + q_{n-1}(x_j)] =$$

$$\int_{-1}^{1} w(x)p_m(x)\mathrm{d}x + \underbrace{\tilde{I}(p_{n+m})}$$

Avremmo quindi

$$I(p_{n+m}) = \tilde{I}(p_{n+m}), \quad \forall p_{n+m} \in \mathbb{P}_{n+m}$$

se $w(x)$ fosse scelto in modo tale per cui

$$\int_{-1}^{1} w(x)p_m(x)\mathrm{d}x = 0, \quad \forall p_m \in \mathbb{P}_m.$$

Si osservi che il massimo valore di m per cui l'integrale definito può annullarsi è $n-1$. Infatti, se per assurdo fosse $m = n$, potremmo scegliere in particolare $p_m = w$ ed avere

$$\int_{-1}^{1} w(x)w(x)\mathrm{d}x = 0,$$

ma essendo gli x_j tutti distinti questa conclusione sarebbe assurda.

Resta quindi da stabilire come scegliere i nodi x_j in modo che

$$\int_{-1}^{1} w(x)p_m(x)\mathrm{d}x = 0, \ \forall p_m \in \mathbb{P}_m, \ m = 0, 1, ..., n-1.$$

Se i nodi $x_j, \ j = 1, ..., n$ sono gli zeri del polinomio di Legendre di grado n, e dunque $w(x)$ è il polinomio di Legendre P_n di grado n (definito a meno di una costante moltiplicativa), e se si considera lo sviluppo del generico polinomio p_m nella base dei polinomi di Legendre

$$p_m(x) = \sum_{k=0}^{m} c_k P_k(x), \ c_k \in \mathbb{R}, \ m \leq n-1$$

si ha:

$$\int_{-1}^{1} w(x)p_m(x)\mathrm{d}x = \int_{-1}^{1} P_n(x) \sum_{k=0}^{m} c_k P_k(x)\mathrm{d}x =$$

$$\sum_{k=0}^{m} c_k \int_{-1}^{1} P_n(x)P_k(x)\mathrm{d}x = 0$$

per l'ortogonalità dei polinomi di Legendre P_n e P_k, $\forall k < n$. Poichè quanto detto è vero per ogni $m < n$, si ha infine che il grado di precisione della formula di quadratura così ottenuta, detta di Gauss-Legendre, è $n + m = n + n - 1 = 2n - 1$.

• CASI PARTICOLARI.

$n = 1$: $x_1 = 0$, $\alpha_1 = 2$.

$n = 2$: $x_{1,2} = \pm\sqrt{1/3}$, $\alpha_{1,2} = 1$.

$n = 3$: $x_{1,3} = \pm\sqrt{3/5}$, $x_2 = 0$, $\alpha_{1,3} = 8/9$, $\alpha_2 = 5/9$.

• OSSERVAZIONE. La formula di quadratura di Gauss-Legendre può essere generalizzata per approssimare un integrale definito su un intervallo (a, b) qualsiasi. A tale scopo basta considerare la sostituzione

$$t = \frac{b-a}{2}x + \frac{b+a}{2}.$$

• **Integrazione di Gauss-Chebyshev**.

$$\int_{-1}^{1} f(x)\frac{1}{\sqrt{1-x^2}}\mathrm{d}x = \sum_{j=1}^{n} \alpha_j f(x_j)$$

$$x_j = \cos\left(\frac{2j-1}{2n}\pi\right), \quad \alpha_j = \frac{\pi}{n}, \quad j = 1, ..., n$$

Equazioni non lineari

Il calcolo degli zeri di una funzione f, o delle radici dell'equazione $f(x) = 0$, è un problema assai ricorrente nel calcolo scientifico (si pensi come semplice esempio al problema del calcolo delle radici di un polinomio). Solo in alcuni casi particolari esistono formule che permettono di calcolare gli zeri di una funzione in un numero finito di passi. I metodi numerici per la risoluzione di questo problema sono pertanto necessariamente metodi iterativi. A partire da uno o più dati iniziali, scelti convenientemente, essi generano una successione di valori x_k che, sotto opportune ipotesi, convergerà ad uno zero α della funzione f oggetto di studio.

Teorema.
Sia $f : \mathbb{R} \to \mathbb{R}$, $f \in C^0[a,b]$, $f(a)f(b) < 0 \Rightarrow \exists \alpha \in (a,b)$ tale che $f(\alpha) = 0$.

Metodo di bisezione.
• Passo 1: $a_0 = a,\ b_0 = b$.
$f(a_0)f(b_0) < 0 \Rightarrow \exists \alpha \in [a_0, b_0]$ tale che $f(\alpha) = 0$;
$\boxed{x_1 = \dfrac{a_0 + b_0}{2}}$, $|x_1 - a_0| = |x_1 - b_0| = \frac{b_0 - a_0}{2}$, $|x_1 - \alpha| < \frac{b-a}{2}$;
• Passo 2
$(*)$ $f(a_0)f(x_1) < 0$ oppure $(**)$ $f(x_1)f(b_0) < 0$;
Es: vale $(**) \Rightarrow \exists \alpha \in [x_1, b_0]$ tale che $f(\alpha) = 0$;
$a_1 = x_1,\ b_1 = b_0$, $\boxed{x_2 = \dfrac{a_1 + b_1}{2}}$, $|x_2 - a_1| = |x_2 - b_1| = \frac{b_1 - a_1}{2}$,
$|x_2 - \alpha| < \frac{b-a}{2^2}$ (analogamente se vale $(*)$).
$\vdots$
• Passo n
$\boxed{x_n = \dfrac{a_{n-1} + b_{n-1}}{2}}$, $|x_n - a_{n-1}| = |x_n - b_{n-1}| = \frac{b_{n-1} - a_{n-1}}{2}$, $|x_n - \alpha| < \frac{b-a}{2^n}$
Ci aspettiamo: $|x_n - \alpha| \approx \frac{1}{2}|x_{n-1} - \alpha|$.
Si osservi che l'errore non è monotono decrescente, ma lo è la maggiorazione $\frac{b-a}{2^n}$ disponibile per l'errore $|x_n - \alpha|$.

Definizione.
Una sequenza $\{x_n | n \geq 0\}$ si dice che converge a α con ordine p se, posto $e_n = x_n - \alpha$, si ha

$$\lim_{n\to\infty} e_n = 0$$

e

$$|e_{n+1}| \leq C|e_n|^p, \quad C > 0,\ n \geq 0.$$

Condizione sufficiente perchè il metodo sia di ordine p è per esempio il fatto che per una costante C_1 opportuna si verifichi:

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^p} = C_1, \quad C_1 > 0.$$

• $p = 1$ (convergenza lineare): $|e_n| \leq C|e_{n-1}|$.
Si ottiene: $|e_n| \leq C|e_{n-1}| \leq C^2|e_{n-2}| \leq ... \leq C^n|e_0|$.
La condizione $C < 1$ è sufficiente per garantire la convergenza.

• $p = 2$ (convergenza quadratica): $|e_n| \leq C|e_{n-1}|^2$.
Si ottiene: $|Ce_n| \leq |Ce_{n-1}|^2 \leq |Ce_{n-2}|^{2^2} \leq ... \leq |Ce_0|^{2^n}$.
La condizione $|Ce_0| < 1$, cioè $|e_0| < \frac{1}{C}$ è sufficiente per garantire la convergenza.
Dunque l'errore iniziale deve essere sufficientemente piccolo (per esempio si può controllare con un metodo del primo ordine).

Metodo di Newton-Raphson (o metodo delle tangenti).
Dato $x_n \in \mathbb{R}$, consideriamo la retta tangente a f in x_k:

$$y = f(x_k) + f'(x_k)(x - x_k).$$

Chiamiamo x_{k+1} l'ascissa del punto di intersezione della retta tangente con l'asse delle ascisse, cioè poniamo $y = 0$ e $x = x_{k+1}$ nell'equazione della retta.
Si ha:

$$\boxed{x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}}$$

Modifiche del metodo di Newton.
• *Metodo delle corde* (ordine 1).

$$\boxed{x_{k+1} = x_k - \frac{f(x_k)}{m}}, \ m \text{ costante.}$$

• *Metodo delle secanti* (ordine $p = (1 + \sqrt{5})/2$).

$$\boxed{x_{k+1} = x_k - \frac{f(x_k)}{f(x_k) - f(x_{k-1})}(x_k - x_{k-1})}$$

(Si approssima $f'(x_k)$ con un rapporto incrementale all'indietro).

Analisi di convergenza del metodo di Newton.
Sia $f \in C^2$ in un intervallo I che contiene una radice semplice α ($f(\alpha) = 0$, $f'(\alpha) \neq 0$), sia I t.c. $f'(x) \neq 0$, $\forall x \in I$, $x_k \in I$,$\forall k \geq \overline{k}$. Dallo sviluppo di Taylor si ha:

$$0 = f(\alpha) = f(x_k) + (\alpha - x_k)f'(x_k) + \frac{1}{2}(\alpha - x_k)^2 f''(z_k), \quad z_k \in [x_k, \alpha],$$

da cui, dividendo per $f'(x_k)$, si ottiene:

$$\frac{f(x_k)}{f'(x_k)} + \alpha - x_k = -\frac{1}{2}(\alpha - x_k)^2 \frac{f''(z_k)}{f'(x_k)} \Rightarrow \alpha - x_{k+1} = -\frac{1}{2}(\alpha - x_k)^2 \frac{f''(z_k)}{f'(x_k)}.$$

Passando al limite si ottiene:

$$\lim_{k\to\infty}\frac{\alpha-x_{k+1}}{(\alpha-x_k)^2}=-\frac{1}{2}\lim_{k\to\infty}\frac{f''(z_k)}{f'(x_k)}=-\frac{1}{2}\frac{f''(\alpha)}{f'(\alpha)},$$

oppure, passando ai valori assoluti si ottiene la maggiorazione:

$$|\alpha-x_{k+1}|=\frac{1}{2}(\alpha-x_k)^2\Big|\frac{f''(z_k)}{f'(x_k)}\Big|\le M(\alpha-x_k)^2,$$

con

$$M=\frac{1}{2}\,\frac{\max_{\{t\in I\}}|f''(t)|}{\min_{\{z\in I\}}|f'(z)|},$$

Dunque il metodo di Newton è del secondo ordine (per l'approssimazione di radici semplici). Si ha $|e_{k+1}|\le M|e_k|^2$, dunque il metodo converge se $|e_0|<\frac{1}{M}$.

Teorema.
Supponiamo che valgano le seguenti ipotesi:
1) $f\in C^2[a,b]$, $f(a)f(b)<0$;
2) $f'(x)\neq 0, \forall x\in[a,b]$;
3) $f''(x)\ge 0$ oppure $f''(x)\le 0$, $\forall x\in[a,b]$;
4) $\big|\frac{f(a)}{f'(a)}\big|<b-a \quad \big|\frac{f(b)}{f'(b)}\big|<b-a$.
Allora il metodo di Newton converge all'unica soluzione $\alpha\in[a,b]$ per ogni scelta di $x_0\in[a,b]$.

Dimostrazione.
Osserviamo che le ipotesi 1) e 2) garantiscono che esiste una ed un'unica soluzione $\alpha\in[a,b]$ tale che $f(\alpha)=0$.
Dimostriamo il teorema in una delle situazioni possibili, al variare del segno di $f(a)$, $f(b)$, f' e f'', scegliendo per esempio

$$f(a)<0,\ f(b)>0,\ f'>0,\ f''\le 0.$$

Sia $a\le x_0<\alpha$ e dunque $f(x_0)<0$.

[**Osservazione.**
L'ipotesi 4) garantisce che la tangente al grafico di f nei punti estremi dell'intervallo aventi ascissa a o b interseca l'asse x all'interno dell'intervallo $[a,b]$. In particolare nel caso preso in esame garantisce che $\forall x_0\in(\alpha,b]$ $x_1\in[a,\alpha)$ e dunque si ricade nel caso considerato. Infatti, si consideri la retta tangente al grafico di f nel punto di ascissa b:

$$y=f(b)+f'(b)(x-b)$$

Sia $\overline{\alpha}$ il punto in cui la retta interseca l'asse delle ascisse:

$$0=f(b)+f'(b)(\overline{\alpha}-b)\Rightarrow\overline{\alpha}-b=-\frac{f(b)}{f'(b).}$$

Passando ai moduli si ha:

$$b - \overline{\alpha} = |\overline{\alpha} - b| = \left| \frac{f(b)}{f'(b)} \right| < b - a.$$

Da $b - \overline{\alpha} < b - a$ si ottiene $a < \overline{\alpha}$.
Con procedimento analogo si dimostra che la retta tangente al grafico di f in un punto di ascissa x_0, con $\alpha < x_0 < b$, interseca l'asse x in un punto $\overline{\alpha}$, con $a < \overline{\alpha} < x_0$.]

Per la formula di Newton si ha:

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} > x_0$$

essendo $f(x_0) < 0$ e $f'(x_0) > 0$.
Dimostriamo per induzione che $x_k \leq \alpha$ e $x_{k+1} \geq x_k$. Poichè la proprietà è vera per $k = 0$ (infatti $x_0 < \alpha$ e $x_1 > x_0$) dimostriamo che, supposta vera per k, essa è verificata per $k + 1$. Si deve dunque verificare che $x_{k+1} \leq \alpha$ e $x_{k+2} \geq x_{k+1}$.
Per il teorema del valor medio di Lagrange:

$$-f(x_k) = f(\alpha) - f(x_k) = (\alpha - x_k) f'(x_k^*), \quad \text{con } x_k \leq x_k^* \leq \alpha.$$

Dal momento che $f''(x) \leq 0$, si ha che f' è non crescente e quindi $f'(x_k^*) \leq f'(x_k)$ per $x_k^* \geq x_k$, da cui

$$-f(x_k) = (\alpha - x_k) f'(x_k^*) \leq (\alpha - x_k) f'(x_k)$$

Essendo $f'(x_k) > 0$ si ha

$$-\frac{f(x_k)}{f'(x_k)} \leq \alpha - x_k$$

Dalla formula di Newton:

$$\boxed{x_{k+1}} = x_k - \frac{f(x_k)}{f'(x_k)} \boxed{\leq} x_k + (\alpha - x_k) = \boxed{\alpha}$$

Sfruttando ora $x_{k+1} < \alpha$, e le ipotesi iniziali per cui f è crescente e $f(\alpha) = 0$, si ha $f(x_{k+1}) < 0$, e dunque:

$$\boxed{x_{k+2}} = x_{k+1} - \frac{f(x_{k+1})}{f'(x_{k+1})} \boxed{\geq x_{k+1}}$$

essendo $f(x_{k+1}) < 0$ e $f'(x_{k+1}) > 0$.
Osservando infine che la successione $\{x_k\}$ è monotona crescente e limitata superiormente si deduce che è convergente. Passando al limite

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

$$\Downarrow$$

$$\alpha = \alpha - \frac{f(\alpha)}{f'(\alpha)}$$

Si ottiene quindi $f(\alpha) = 0$, cioè il metodo iterativo converge alla soluzione dell'equazione non lineare $f(x) = 0$.
Una possibile scelta per x_0 può essere l' estremo di Fourier (a oppure b), cioè l'estremo ove si verifica $f(a)f''(a) > 0$, oppure $f(b)f''(b) > 0$.

Test d'arresto.
- $\boxed{|f(x_k)| < \varepsilon_1}$ (residuo)
- $\boxed{|x_{k+1} - x_k| < \varepsilon_2}$ (errore)

Sfruttando il teorema del valor medio si ha:

$$f(x_k) = f(x_k) - f(\alpha) = f'(\xi_k)(x_k - \alpha) \Rightarrow \frac{f(x_k)}{f'(\xi_k)} = x_k - \alpha, \quad \xi_k \in [\alpha, x_k].$$

Dalla formula di Newton si ottiene quindi:

$$\boxed{x_{k+1} - x_k} = -\frac{f(x_k)}{f'(x_k)} \approx -\frac{f(x_k)}{f'(\xi_k)} = \boxed{\alpha - x_k}$$

$$|x_{k+1} - x_k| \approx |\alpha - x_k| = |e_k|$$

Esercizio.
Data $f(x) = a - \frac{1}{x}$, $a > 0$, si scriva il metodo di Newton per approssimare la radice $x = \frac{1}{a}$.
Posto

$$e_n = \frac{\left| x_n - \frac{1}{a} \right|}{\frac{1}{a}},$$

si verifichi che $e_n = e_{n-1}^2$. Determinare x_0 affinchè il metodo di Newton converga.

Equazioni non lineari: metodi a un passo

Si vuole risolvere l'equazione non lineare $f(x) = 0$ riconducendosi al problema della ricerca del punto fisso di una funzione $y = g(x)$, cioè si vuole trovare α tale che:

$$\boxed{f(\alpha) = 0 \iff \alpha = g(\alpha)}$$

Graficamente: si cercano le intersezioni di $y = g(x)$ con la bisettrice del primo e terzo quadrante

$$\alpha \text{ soluzione di } \begin{cases} y = & x \\ \\ y = & g(x) \end{cases}$$

Per esempio:

$\bullet f(x) = 0 \Leftrightarrow \underbrace{x + f(x)}_{g(x)} = x$; oppure, $f(x) = 0 \Leftrightarrow \underbrace{x + kf(x)}_{g(x)} = x, \quad k \neq 0$

$\bullet x - e^{x-2} = 0 \Leftrightarrow \boxed{x = \underbrace{e^{x-2}}_{g_1(x)}} \Leftrightarrow \ln x = \ln e^{x-2},\ x > 0 \Leftrightarrow \ln x = x - 2 \Leftrightarrow \boxed{x = \underbrace{2 + \ln x}_{g_2(x)}}$

Metodo iterativo:

$$\boxed{x_{n+1} = g(x_n)}, \quad \forall n \geq 0, \quad x_0 \text{ assegnato.}$$

Esempio.

Data $f(x) = x^2 - 4$ ($\alpha = \pm 2$); $x_0 = 3,\ x_{n+1} = g_i(x_n),\ n \geq 0,\ i = 1, 2, 3$:

$g_1(x) = x^2 + x - 4;\ x_0 = 3,\ x_1 = 8,\ x_2 = 68,\ x_3 = 4688,\ \ldots\ x_n \to +\infty$

$g_2(x) = \frac{4}{x};\ x_0 = 3,\ x_1 = \frac{4}{3},\ x_2 = 3,\ x_3 = \frac{4}{3},\ x_4 = 3, \ldots$

$g_3(x) = \frac{x}{2} + \frac{2}{x};\ x_0 = 3,\ x_1 = \underline{2}.1666\ldots,\ x_2 = \underline{2.00}6410257,\ x_3 = \underline{2.0000}1024\ldots$

Dunque si osserva che non tutte le funzioni di iterazione garantiscono la convergenza al punto fisso.

Proprietà.
Assegnato un procedimento iterativo ad un passo:

$$x_{n+1} = g(x_n),\ n \geq 0, \quad x_0 \text{ assegnato,}$$

g continua, i punti fissi della funzione di iterazione g sono *tutti e soli* i punti limite delle successioni $\{x_n\}$.

Teorema.
Si consideri la successione $x_{k+1} = g(x_k)$, per $k \geq 0,\ x_0$ assegnato e si supponga che valgano le seguenti ipotesi:

1) $\exists I = [a, b]$ tale che $\forall x \in I,\ g(x) \in I$;

2) $g \in C^1(I)$;

3) $\exists K < 1 :\ \forall x \in I, |g'(x)| \leq K$.

Allora:

4) g ha un unico punto fisso $\alpha \in I$;

5) la successione $\{x_k,\ k \geq 0\}$ converge ad α, $\forall x_0 \in I$;

6) $\lim_{k\to\infty} \frac{x_{k+1}-\alpha}{x_k-\alpha} = g'(\alpha)$.

Dimostrazione.
Si consideri la funzione $G(x) = x - g(x)$ e, sfruttando le ipotesi 1) e 2) si osservi che $G(a) = a - g(a) \leq 0$, $G(b) = b - g(b) \geq 0$ e dunque la funzione G ammette almeno una radice $\alpha \in I$, ovvero la funzione g ammette almeno un punto fisso $\alpha \in I$. Per quanto riguarda l'unicità del punto fisso α, si supponga per assurdo che esistano due punti fissi α_1 e $\alpha_2 \in I$, cioè $\alpha_1 = g(\alpha_1)$ e $\alpha_2 = g(\alpha_2)$. Per il teorema del valor medio di Lagrange si ha:

$$\alpha_2 - \alpha_1 = g(\alpha_2) - g(\alpha_1) = g'(t)(\alpha_2 - \alpha_1), \quad t \in I$$

da cui passando ai valori assoluti e sfruttando l'ipotesi 3) si ottiene:

$$\boxed{|\alpha_2 - \alpha_1|} = |g(\alpha_2) - g(\alpha_1)| = |g'(t)|\ |\alpha_2 - \alpha_1| \leq K|\alpha_2 - \alpha_1| \boxed{< |\alpha_2 - \alpha_1|}\,.$$

e dunque necessariamente $\alpha_1 = \alpha_2$.
Per quanto riguarda la convergenza, se si considera l'errore al passo $k + 1$ si osserva:

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

con t_k appartenente all'intervallo di estremi α e x_k. Passando ai valori assoluti e sfruttando ancora l'ipotesi 3) si ottiene:

$$|x_{k+1} - \alpha| = |g(x_k) - g(\alpha)| = |g'(t_k)|\ |x_k - \alpha| \leq K|x_k - \alpha| \leq K^2|x_{k-1} - \alpha| \leq ...$$

$$\leq K^{k+1}|x_0 - \alpha| \longrightarrow 0 \ \text{ per } k \to \infty.$$

Quindi $x_k \longrightarrow \alpha$, per $k \to \infty$.
Inoltre:

$$\lim_{k\to\infty} \frac{x_{k+1} - \alpha}{x_k - \alpha} = \lim_{k\to\infty} g'(t_k) = g'(\alpha).$$

Definizione. Si definisce *fattore asintotico di convergenza* la quantità $|g'(\alpha)|$ e *velocità asintotica di convergenza* la quantità

$$R = \log \frac{1}{|g'(\alpha)|} = -\log |g'(\alpha)|.$$

Corollario.
Se $|g'(x)| > 1$, $\forall x \in [a, b]$ il procedimento iterativo ad un passo:

$$x_{n+1} = g(x_n),\ n \geq 0,\ x_0 \neq \alpha$$

sarà localmente divergente, cioè divergerà a $\pm\infty$, oppure convergerà a un punto fisso $\notin [a, b]$.

Ordine di convergenza.
Sia $g \in C[a, b]$, $g' \in C[a, b]$, $g'' \in C[a, b]$,..., $g^{(p)} \in C[a, b]$, con

$$g(\alpha) = \alpha,\quad g'(\alpha) = ... = g^{(p-1)}(\alpha) = 0,\ g^{(p)}(\alpha) \neq 0$$

Allora il metodo $x_{n+1} = g(x_n),\ n \geq 0$, ha ordine p e

$$\lim_{n\to\infty} \frac{x_{n+1} - \alpha}{(x_n - \alpha)^p} = \frac{g^{(p)}(\alpha)}{p!}.$$

Si considera infatti lo sviluppo di Taylor:

$$g(x_n) = g(\alpha) + (x_n - \alpha)g'(\alpha) + \frac{1}{2!}(x_n - \alpha)^2 g''(\alpha) + ...$$

e, tenendo conto che $g(x_n) = x_{n+1}$ e $g(\alpha) = \alpha$, si ha:

$$x_{n+1} - \alpha = (x_n - \alpha)g'(\alpha) + \frac{1}{2!}(x_n - \alpha)^2 g''(\alpha) + ...$$

• Nell'esempio $f(x) = x^2 - 4$ si ha:

$g_1'(x) = 2x + 1,\ g_1'(2) = 5 > 1$

$g_2'(x) = -\frac{4}{x^2},\ g_2'(2) = -1$

$g_3'(x) = \frac{1}{2} - \frac{2}{x^2},\ g_3'(2) = 0$
$g_3''(x) = \frac{4}{x^3},\ g_3''(2) = \frac{1}{2} \neq 0$ (metodo del secondo ordine)

Il metodo di Newton come metodo di punto fisso.

$$g(x) = x - \frac{f(x)}{f'(x)} \Rightarrow g'(\alpha) = 0.$$

Infatti:

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

e, in particolare,

$$g'(\alpha) = 1 - \frac{[f'(\alpha)]^2 - \overbrace{f(\alpha)}^{=0} f''(\alpha)}{[f'(\alpha)]^2} = 1 - \frac{[f'(\alpha)]^2}{[f'(\alpha)]^2} = 0.$$

Si può verificare che $g''(\alpha) \neq 0$.
Se la radice α ha molteplicità $p > 1$, il metodo di Newton è del primo ordine con:

$$g'(\alpha) = 1 - \frac{1}{p}.$$

Infatti, se si considera $f(x) = (x-\alpha)^p h(x)$, con $h(\alpha) \neq 0$, si ha:

$$f'(x) = p(x-\alpha)^{p-1}h(x) + (x-\alpha)^p h'(x) = (x-\alpha)^{p-1}[ph(x) + (x-\alpha)h'(x)]$$

$$g(x) = x - \frac{(x-\alpha)^p h(x)}{(x-\alpha)^{p-1}[ph(x) + (x-\alpha)h'(x)]} = x - (x-\alpha)\frac{h(x)}{ph(x) + (x-\alpha)h'(x)}$$

Derivando:

$$g'(x) = 1 - \frac{h(x)}{ph(x) + (x-\alpha)h'(x)} - (x-a)\frac{d}{dx}\left[\frac{h(x)}{ph(x) + (x-\alpha)h'(x)}\right].$$

In particolare:

$$g'(\alpha) = 1 - \frac{\boxed{h(\alpha)}}{p\boxed{h(\alpha)} + \underbrace{(\alpha-\alpha)}_{=0}h'(\alpha)} - \underbrace{(\alpha-\alpha)}_{=0}\frac{d}{dx}\left[\frac{h(x)}{ph(x) + (x-\alpha)h'(x)}\right]_{x=\alpha} = 1 - \frac{1}{p}$$

Dunque

$$g'(\alpha) = 1 - \frac{1}{\boxed{p}} = 0 \iff p = 1$$

Il metodo di Newton modificato:

$$g(x) = x - \boxed{p}\frac{f(x)}{f'(x)},$$

ha ordine 2.

Osservazione.
Se α ha molteplicità p per f, allora α ha molteplicità $p-1$ per f' e α ha molteplicità 1 per $\Phi(x) := \frac{f(x)}{f'(x)}$. $\Rightarrow$ Si può applicare il metodo di Newton 'classico' alla funzione Φ, ottenendo nuovamente un metodo di ordine 2.

Test d'arresto.

$$\boxed{\alpha - x_n \approx \frac{1}{1 - g'(x_n)}(x_{n+1} - x_n)}$$

Infatti:

$$\boxed{x_{n+1} - x_n} = x_{n+1} - \alpha + \alpha - x_n = \underbrace{g(x_n) - g(\alpha)}_{\text{Teor. valor medio Lagrange}} + \underbrace{\alpha - x_n}_{} =$$

$$g'(t_n)(x_n - \alpha) + (\alpha - x_n) = -g'(t_n)(\alpha - x_n) + (\alpha - x_n) =$$

$$(1 - g'(t_n))(\alpha - x_n) \approx \boxed{(1 - g'(x_n))(\alpha - x_n)}$$

.

Richiami di algebra lineare

In questa sezione richiamiamo alcune notazioni, definizioni e risultati di algebra lineare essenziali nello sviluppo del corso ed in particolare nello studio di metodi numerici per la risoluzione di sistemi lineari e per l'approssimazione di autovalori e autovettori.

Norme vettoriali

Diciamo che l'applicazione $\|\cdot\|$ da $\boxed{V = \mathbb{R}^n}$ in $\mathbb{R}^+ \cup \{0\}$ è una norma vettoriale se sono soddisfatte le seguenti condizioni:

1.) $\|\mathbf{x}\| \geq 0,\ \forall \mathbf{x} \in V$ e $\|\mathbf{x}\| = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$.

2.) $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|,\ \forall \alpha \in \mathbb{R},\ \forall \mathbf{x} \in V$.

3.) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|,\ \forall \mathbf{x}, \mathbf{y} \in V$.

Definizione di $\|\cdot\|_p$ e $\|\cdot\|_\infty$ in V:

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}},\ 1 \leq p < \infty; \quad \|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|.$$

Proprietà. $\forall \mathbf{x} \in V$, $\lim_{p\to\infty} \|\mathbf{x}\|_p = \|\mathbf{x}\|_\infty$.

Disuguaglianza di Cauchy-Schwarz: $|(\mathbf{x}, \mathbf{y})| \leq \|\mathbf{x}\|_2\ \|\mathbf{y}\|_2,\ \forall \mathbf{x}, \mathbf{y} \in V$.

Disuguaglianza di Minkowsky: $\|\mathbf{x} + \mathbf{y}\|_p \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p,\ \forall \mathbf{x}, \mathbf{y} \in V$.

Sfera unitaria: $S_p = \{\mathbf{x} \in V : \|\mathbf{x}\|_p = 1\}$.

Casi particolari: Studio di $S_1,\ S_2,\ S_\infty$ (in $\mathbb{R}^2$).

Proprietà di equivalenza fra norme.
Tutte le norme definibili su V sono equivalenti, cioè, date due norme $\|\cdot\|$ e $|||\cdot|||$ su V, $\exists\ c_1,\ c_2 > 0$ tali che $\forall \mathbf{x} \in V$, si ha $c_1|||\mathbf{x}||| \leq \|\mathbf{x}\| \leq c_2|||\mathbf{x}|||$.

Proprietà di continuità.
Ogni norma vettoriale è una funzione continua delle componenti del vettore.

Autovalori e autovettori

Sia A una matrice quadrata di ordine n. Il numero $\lambda \in \mathbb{C}$ è detto autovalore di A se esiste un vettore $\mathbf{x} \neq \mathbf{0}$, tale che $A\mathbf{x} = \lambda\mathbf{x}$. Il vettore $\mathbf{x}$ è detto autovettore associato all'autovalore λ. L'insieme $\sigma(A)$ degli autovalori di A è detto spettro di A. L'autovalore λ è soluzione dell' equazione caratteristica

$$p_A(\lambda) := \det(A - \lambda I) = 0,$$

dove $p_A(\lambda)$ è il polinomio caratteristico.

- Una matrice è singolare $\Leftrightarrow$ ha almeno un autovalore λ nullo.
- $\det(A) = \prod_{i=1}^{n} \lambda_i$.
- $\mathrm{tr}(A) := \sum_{i=1}^{n} a_{ii} = \sum_{i=1}^{n} \lambda_i$ (traccia).

Norme di matrici

Diciamo che l'applicazione $\|\cdot\|$ da $\mathbb{R}^{n\times n}$ in $\mathbb{R}^+ \cup \{0\}$ è una norma di matrici se sono soddisfatte le seguenti condizioni:

1.) $\|A\| \geq 0,\ \forall A \in \mathbb{R}^{n\times n}$ e $\|A\| = 0 \Leftrightarrow A = 0$.

2.) $\|\alpha A\| = |\alpha|\|A\|,\ \forall \alpha \in \mathbb{R},\ \forall A \in \mathbb{R}^{n\times n}$.

3.) $\|A + B\| \leq \|A\| + \|B\|,\ \forall A, B \in \mathbb{R}^{n\times n}$.

4.) $\|AB\| \leq \|A\|\ \|B\|,\ \forall A, B \in \mathbb{R}^{n\times n}$ (condizione aggiuntiva).

Definizione di norma compatibile.
Una norma di matrice $\|\cdot\|_*$ è compatibile con una norma di vettore $\|\cdot\|$ se:

$$\|A\mathbf{x}\| \leq \|A\|_*\|\mathbf{x}\|, \quad \forall \mathbf{x} \in V,\ \forall A \in \mathbb{R}^{n\times n}.$$

Definizione di norma naturale.
Data una norma di vettore $\|\cdot\|$ definiamo

$$\|A\| = \sup\left\{\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}, \mathbf{x} \in \mathbb{R}^n \backslash \{\mathbf{0}\}\right\}$$

norma naturale o indotta dalla norma di vettore.
Si può dimostrare che

$$\|A\| = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|.$$

Casi particolari.

$$\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{n} |a_{ij}|; \quad \|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|; \quad \|A\|_2 = \sqrt{\rho(A^T A)}$$

dove

$$\rho(B) = \max_{1\le i\le n} |\lambda_i(B)|, \ \forall B \in \mathbb{R}^{n\times n}$$

è detto raggio spettrale della matrice B.

Proprietà.
Per ogni norma naturale $\|\cdot\|$ e per ogni matrice quadrata A si ha

$$\rho(A) \le \|A\|.$$

Matrici simmetriche e definite positive (d.p.)

Una matrice simmetrica $A \in \mathbb{R}^{n\times n}$ si dice definita positiva se:

$$\boxed{(A\mathbf{x}, \mathbf{x})} = \mathbf{x}^T A\mathbf{x} = \sum_{i=1}^{n} x_i \sum_{j=1}^{n} a_{ij}x_j = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j \boxed{\ge 0}, \ \forall \mathbf{x} \in V$$

$$\boxed{(A\mathbf{x}, \mathbf{x}) = 0 \iff \mathbf{x} = 0}$$

Criterio di Sylvester.
Una matrice A simmetrica di ordine n è d.p. $\Leftrightarrow \det(A_k) > 0$, $k = 1, 2, ...n$, dove A_k è la sottomatrice principale di ordine k, cioè formata dalle prime k righe e k colonne.

- A d.p. $\Rightarrow A$ non singolare.
- A d.p. $\Longleftrightarrow A$ simmetrica e $\lambda_i(A) > 0, \ i = 1, ..., n.$
- A d.p. $\Rightarrow a_{ii} > 0, \ i = 1, ..., n.$
- A d.p. $\Rightarrow |a_{ik}|^2 < a_{ii}a_{kk}, \ i \ne k.$
- A d.p. $\Rightarrow$ Il massimo elemento della matrice A giace sulla diagonale principale di A.
- Ogni sottomatrice principale di una matrice d.p. è d.p., quindi non singolare.

Matrici a dominanza diagonale (d.d.)

Una matrice A si dice a dominanza diagonale stretta o strettamente diagonalmente dominante per righe se

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\ne i}}^{n} |a_{ij}|, \ i = 1, ..., n.$$

Se la disuguaglianza precedente vale con il simbolo $\geq$, A si dice a dominanza diagonale debole (debolmente diagonalmente dominante) per righe.
Una matrice A si dice a dominanza diagonale stretta o strettamente diagonalmente dominante per colonne se

$$|a_{jj}| > \sum_{\substack{i=1 \\ i \neq j}}^{n} |a_{ij}|, \ j = 1, ..., n.$$

Se la disuguaglianza precedente vale con il simbolo $\geq$, A si dice a dominanza diagonale debole (debolmente diagonalmente dominante) per colonne.
Qualora non sia specificato, si intende dominanza diagonale per righe.
• A d.d. stretta $\Rightarrow$ A non singolare.

Matrici a banda

Si dice che una matrice A ha una banda superiore p se $a_{ij} = 0$ per $j > i + p$ e una banda inferiore q se $a_{ij} = 0$ per $i > j + q$. Se $p = q = 1$ la matrice si dice tridiagonale.

Matrici sparse

Una matrice A si dice sparsa se ha un numero elevato di elementi $a_{ij} = 0$. Quale sia la percentuale di elementi necessaria per far ritenere una matrice sparsa dipende ovviamente dal contesto. Comunemente una matrice $\in \mathbb{R}^{n \times n}$ è ritenuta sparsa quando il numero di elementi diversi da zero è di ordine $O(n)$. Le matrici a banda sono ritenute sparse se $p, q \ll n$.

Analisi degli errori: Condizionamento di un sistema lineare

Sia $A \in \mathbb{R}^{n\times n}$ una matrice quadrata non singolare e $\mathbf{b} \in \mathbb{R}^n$ un vettore , allora esiste ed è unico il vettore $\mathbf{x} \in \mathbb{R}^n$ tale che $A\mathbf{x} = \mathbf{b}$.

(•) **Analisi a priori in avanti**.
Sensibilità della soluzione di $A\mathbf{x} = \mathbf{b}$ a cambiamenti nei dati A e/o $\mathbf{b}$ (condizionamento di un sistema lineare).

(••) **Analisi a priori all'indietro**.
Supposta nota una soluzione approssimata $\widehat{\mathbf{x}}$ del sistema $A\mathbf{x} = \mathbf{b}$, si vuole determinare di quanto si dovrebbero perturbare A e $\mathbf{b}$ affinchè $\widehat{\mathbf{x}}$ sia la soluzione esatta di un sistema perturbato.

(• • •) **Analisi a posteriori**.
Supposta nota una soluzione approssimata $\widehat{\mathbf{x}}$ del sistema $A\mathbf{x} = \mathbf{b}$, si vuole determinare una stima dell'errore $\mathbf{x} - \widehat{\mathbf{x}}$ in funzione del residuo $\mathbf{r} := \mathbf{b} - A\widehat{\mathbf{x}}$ e di quantità note.

(•) **Analisi a priori in avanti**: perturbazione del dato $\mathbf{b}$.

(•1) $\mathbf{b} \Rightarrow \boxed{\boxed{\widehat{\mathbf{b}} := \mathbf{b} + \Delta\mathbf{b}}}$, allora $\mathbf{x} \Rightarrow \boxed{\boxed{\widehat{\mathbf{x}} := \mathbf{x} + \Delta\mathbf{x}}}$:

$A(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b} + \Delta\mathbf{b}$

$\boxed{A\mathbf{x}} + A\Delta\mathbf{x} = \boxed{\mathbf{b}} + \Delta\mathbf{b}$

$\Delta\mathbf{x} = A^{-1}\Delta\mathbf{b} \Rightarrow \underbrace{\|\Delta\mathbf{x}\| \leq \|A^{-1}\|\ \|\Delta\mathbf{b}\|}$

(•2) $\mathbf{b} = A\mathbf{x} \Rightarrow \|\mathbf{b}\| \leq \|A\|\ \|\mathbf{x}\| \Rightarrow \underbrace{\dfrac{1}{\|\mathbf{x}\|} \leq \dfrac{\|A\|}{\|\mathbf{b}\|}}$

$$(\bullet 1) \times (\bullet 2) \qquad \boxed{\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \|A\|\ \|A^{-1}\| \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}}$$

(•) **Analisi a priori in avanti**: perturbazione del dato A.

$A \Rightarrow \boxed{\boxed{\widehat{A} := A + \Delta A}}$, allora $\mathbf{x} \Rightarrow \boxed{\boxed{\widehat{\mathbf{x}} := \mathbf{x} + \Delta\mathbf{x}}}$:

$(A + \Delta A)(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b}$

$\boxed{A\mathbf{x}} + \Delta A\ \mathbf{x} + A\ \Delta\mathbf{x} + \Delta A\ \Delta\mathbf{x} = \boxed{\mathbf{b}}$

$\Delta\mathbf{x} = -A^{-1}\ \Delta A\ (\mathbf{x} + \Delta\mathbf{x}) \Rightarrow \|\Delta\mathbf{x}\| \leq \|A^{-1}\|\ \|\Delta A\|\ \|\mathbf{x} + \Delta\mathbf{x}\|$

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x} + \Delta\mathbf{x}\|} \leq \|A^{-1}\|\ \|\Delta A\| \Rightarrow \boxed{\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x} + \Delta\mathbf{x}\|} \leq \|A\|\ \|A^{-1}\| \frac{\|\Delta A\|}{\|A\|}}$$

Definizione: $\boxed{K(A) := \|A\| \, \|A^{-1}\|}$
(numero di condizionamento della matrice A).

Teorema di Wilkinson.
Sia A non singolare, ΔA e $\Delta \mathbf{b}$ perturbazioni di A e $\mathbf{b}$, con $\|\Delta A\| < 1/\|A^{-1}\|$. Allora:

$$\frac{\|\Delta \mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{K(A)}{1 - K(A)\frac{\|\Delta A\|}{\|A\|}} \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta \mathbf{b}\|}{\|\mathbf{b}\|} \right).$$

(• • •) **Analisi a posteriori**.

Sia $\widehat{\mathbf{x}}$ una soluzione fornita da un metodo numerico e $\mathbf{r} = \mathbf{b} - A\widehat{\mathbf{x}}$ il residuo associato.
Si dimostra, con procedimento analogo a quello del caso (•), la seguente disuguaglianza:

$$\boxed{\frac{\|\Delta \mathbf{x}\|}{\|\mathbf{x}\|} \leq K(A) \frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}}$$

Osservazioni su $K(A)$.

- Per ogni A non singolare e per ogni norma naturale di matrice si ha:

  $1 = \|I\| = \|AA^{-1}\| \leq \|A\| \, \|A^{-1}\| = K(A).$

- $\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} := \frac{\max_{i=1,\ldots,n}(|\lambda_i(A)|)}{\min_{i=1,\ldots,n}(|\lambda_i(A)|)} = \rho(A)\rho(A^{-1}) \leq \|A\| \, \|A^{-1}\| = K(A).$

- Se A è simmetrica:
  - $\rho(A) = \|A\|_2$;
  - $K_2(A) = \frac{\max_{i=1,\ldots,n} |\lambda_i(A)|}{\min_{i=1,\ldots,n} |\lambda_i(A)|}$.

- Se A è simmetrica e definita positiva:
  - $\rho(A) = \|A\|_2$;
  - $K_2(A) = \frac{\max_{i=1,\ldots,n} \lambda_i(A)}{\min_{i=1,\ldots,n} \lambda_i(A)}$.

Risoluzione numerica di sistemi lineari

Data una matrice quadrata $A \in \mathbb{R}^{n\times n}$ non singolare e un vettore $\mathbf{b} \in \mathbb{R}^n$, esiste ed è unico il vettore $\mathbf{x} \in \mathbb{R}^n$ tale che $A\mathbf{x} = \mathbf{b}$.

• Costo computazionale della regola di Cramer: $(n+1)!$ *flops*. $\Rightarrow$
Un calcolatore in grado di effettuare 10^9 *flops* al secondo impiegherebbe circa 10^{48} anni per risolvere un sistema lineare di sole 50 equazioni.

• **Metodi diretti**: in assenza di errori di arrotondamento forniscono la soluzione in un numero finito di operazioni.
• **Metodi iterativi**: la soluzione è ottenuta come limite di una successione di vettori soluzione di sistemi lineari più semplici.

Metodi diretti

Risoluzione di sistemi triangolari.
L: matrice triangolare inferiore ($l_{ij} = 0$ se $i < j$).

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

(sostituzione in avanti o forward substitution).

U: matrice triangolare superiore ($u_{ij} = 0$ se $i > j$).

$$U\mathbf{x} = \mathbf{b} \Rightarrow x_i = \left(b_i - \sum_{j=i+1}^{n} u_{ij}x_j \right) \frac{1}{u_{ii}}, \quad i = n, ..., 1$$

(sostituzione all'indietro o backward substitution).

[Numero operazioni $\approx \frac{n(n+1)}{2}$]

Il metodo di eliminazione di Gauss (E.G.).
Idea: Si trasforma il sistema $A\mathbf{x} = \mathbf{b}$ in un sistema equivalente triangolare superiore $U\mathbf{x} = \hat{\mathbf{b}}$ mediante combinazioni lineari di righe.

Passo 0: $A^{(1)} = A$

Passo 1: $A^{(1)} \to A^{(2)}$, $a_{i1}^{(2)} = 0,\ i = 2, ..., n$

Passo 2: $A^{(2)} \to A^{(3)}$, $a_{i2}^{(3)} = 0,\ i = 3, ..., n$

$\vdots$

Passo k: $A^{(k)} \to A^{(k+1)}$, $a_{ik}^{(k+1)} = 0,\ i = k+1, ..., n$

Passo 1: $i = 2, ..., n$

$$R_i^{(2)} = R_i^{(1)} - m_{i1} R_1^{(1)}$$

$$a_{ij}^{(2)} = a_{ij}^{(1)} - m_{i1} a_{1j}^{(1)},\ m_{i1} \text{ t.c. } 0 = a_{i1}^{(2)} = a_{i1}^{(1)} - m_{i1} a_{11}^{(1)} \ \Rightarrow\ m_{i1} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}}$$

Passo 2: $i = 3, ..., n$

$$R_i^{(3)} = R_i^{(2)} - m_{i2} R_2^{(2)}$$

$$a_{ij}^{(3)} = a_{ij}^{(2)} - m_{i2} a_{2j}^{(2)},\ m_{i2} \text{ t.c. } 0 = a_{i2}^{(3)} = a_{i2}^{(2)} - m_{i2} a_{22}^{(2)} \ \Rightarrow\ m_{i2} = \frac{a_{i2}^{(2)}}{a_{22}^{(2)}}$$

$\vdots$

Passo k: $i = k+1, ..., n$

$$R_i^{(k+1)} = R_i^{(k)} - m_{ik} R_k^{(k)}$$

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik} a_{kj}^{(k)},\ m_{ik} \text{ t.c. } 0 = a_{ik}^{(k+1)} = a_{ik}^{(k)} - m_{ik} a_{kk}^{(k)} \ \Rightarrow\ m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$$

• **Algoritmo E.G.**

$k = 1, ..., n-1$
  $i = k+1, ..., n$
    $m_{ik} = a_{ik}^{(k)} / a_{kk}^{(k)}$
    $j = k+1, ..., n$
      $a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik} a_{kj}^{(k)}$
      [si osservi che $a_{ik}^{(k+1)} = 0$ automaticamente, grazie alla scelta di m_{ik}]
    end j
    $b_i^{(k+1)} = b_i^{(k)} - m_{ik} b_k^{(k)}$
  end i
end k

[Numero operazioni $\approx$: $\frac{n^3}{3}$]

• **Fattorizzazione LU**.

Se $L = \{l_{ii} = 1,\ l_{ij} = m_{ij}, i = 1, ..., n,\ j < i\}$ e $U = A^{(n)}$, allora $\boxed{A = LU}$.

• **Applicazioni della fattorizzazione LU**.

◦ det(A)=det(L)det(U)=det(U)=$\prod_{i=1}^{n} u_{ii}$

◦ $A\mathbf{x} = \mathbf{b}$, $A = LU \Rightarrow$ Soluzione di due sistemi triangolari:

1)$L\mathbf{y} = \mathbf{b}$ (triangolare inferiore)
2)$U\mathbf{x} = \mathbf{y}$ (triangolare superiore).

◦ Se L_1U_1 e L_2U_2 sono due fattorizzazioni della matrice A, con L_i (risp. U_i) matrici triangolari inferiori (risp. superiori), allora le matrici L_i (risp. U_i) differiscono per una matrice diagonale.
Infatti, supponendo $L_1U_1 = L_2U_2$, moltiplicando a sinistra per L_2^{-1} e a destra per U_1^{-1}, si ha

$$\underbrace{L_2^{-1}L_1}_{\text{tr. inf.}} = \underbrace{U_2U_1^{-1}}_{\text{tr. sup.}},$$

da cui si deduce che $L_2^{-1}L_1 = U_2U_1^{-1} = D$, con D matrice diagonale. Risulta pertanto $L_1 = L_2D$ e $U_2 = DU_1$.

• **Tecnica del pivoting per righe.**
Se a un certo punto del procedimento di EG si ha $a_{kk}^{(k)} = 0$, non si può più procedere con il calcolo del moltiplicatore $m_{ik} = a_{ik}^{(k)}/a_{kk}^{(k)}$.
Dunque ad ogni passo k si può per esempio cercare la riga $\overline{r}$ per cui si verifica

$$|a_{\overline{r}\ k}^{(k)}| = \max_{k\le i\le n} |a_{ik}^{(k)}|$$

e scambiare la riga k con la riga $\overline{r}$.
Si osservi che la tecnica del pivoting permette di generare moltiplicatori m_{ik} per cui si verifica $|m_{ik}| \le 1$. È pertanto utilizzata, anche quando non si incontrano elementi pivotali $a_{kk}^{(k)}$ nulli, allo scopo di non amplificare gli errori di arrotondamento nel calcolo degli elementi $a_{ij}^{(k+1)}$.
In modo analogo si può definire la tecnica del pivoting per colonne

$$|a_{k\ \overline{c}}^{(k)}| = \max_{k\le j\le n} |a_{kj}^{(k)}|$$

(si scambia la colonna k con la colonna $\overline{c}$) o pivoting totale

$$|a_{\overline{r}\ \overline{c}}^{(k)}| = \max_{k\le i,j\le n} |a_{ij}^{(k)}|$$

(si scambia la riga k con la riga $\overline{r}$ e la colonna k con la colonna $\overline{c}$).

• **Risoluzione di sistemi tridiagonali**.
Sia data la matrice

$$A = \begin{bmatrix} a_1 & c_1 & 0 & ... & 0 \\ b_2 & a_2 & c_2 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & ... & b_{n-1} & a_{n-1} & c_{n-1} \\ 0 & ... & ... & b_n & a_n \end{bmatrix}$$

(matrice a banda con $p = q = 1$). Allora se la fattorizzazione di Gauss esiste, i fattori L e U sono due matrici bidiagonali (inferiore e superiore, rispettivamente) della forma

$$L = \begin{bmatrix} 1 & 0 & 0 & ... & 0 \\ \beta_2 & 1 & 0 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & ... & \beta_{n-1} & 1 & 0 \\ 0 & ... & ... & \beta_n & 1 \end{bmatrix}, \quad U = \begin{bmatrix} \alpha_1 & \gamma_1 & 0 & ... & 0 \\ 0 & \alpha_2 & \gamma_2 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & ... & 0 & \alpha_{n-1} & \gamma_{n-1} \\ 0 & ... & ... & 0 & \alpha_n \end{bmatrix}$$

I coefficienti incogniti $\alpha_i,\ \beta_i,\ \gamma_i$ delle matrici L e U possono essere determinati imponendo l'uguaglianza (elemento per elemento) $LU = A$. In tal modo si trova $\alpha_1 = a_1,\ \gamma_i = c_i,\ \forall i = 1, ..., n-1$, e:

$$\beta_i = \frac{b_i}{\alpha_{i-1}},\ \alpha_i = a_i - \beta_i \gamma_{i-1},\ i = 2, ..., n$$

Osservazioni
• Metodo di Cholesky per matrici simmetriche definite positive $A = LL^T$.
• Fenomeno del fill−in.

Metodi iterativi

Assegnato un vettore arbitrario $\mathbf{x}^{(0)} = \{x_1^{(0)}, x_2^{(0)}, ..., x_n^{(0)}\}$, si costruisce una successione

$$\mathbf{x}^{(k)} = \{x_1^{(k)}, x_2^{(k)}, ..., x_n^{(k)}\},$$

tale che

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

Si consideri la i-esima equazione del sistema lineare $A\mathbf{x} = \mathbf{b}$:

$$x_i = \frac{1}{a_{ii}}\Big\{b_i - \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}x_j\Big\}, \quad a_{ii} \neq 0, \quad i = 1, ..., n$$

Metodo di JACOBI (J).

$$\forall k \geq 0, \quad x_i^{(k+1)} = \frac{1}{a_{ii}}\Big\{b_i - \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}x_j^{(k)}\Big\}, \quad i = 1, ..., n$$

Metodo di GAUSS-SEIDEL (GS).

$$\forall k \geq 0, \quad x_i^{(k+1)} = \frac{1}{a_{ii}}\Big\{b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\Big\}, \quad i = 1, ..., n.$$

Osservazioni.

◦ $\forall$ passo k di (J) e (GS): n^2 *flops*.

◦ (J): aggiornamenti simultanei (parallelo).

◦ (GS): aggiornamenti successivi (sequenziale).

Alcuni risultati sullo studio della convergenza.

Teorema.
A diagonalmente dominante in senso forte $\Longrightarrow$ i metodi (J) e (GS) convergono.
(È una condizione SOLO sufficiente).

Dimostrazione [nel caso del metodo (J)].
Calcolando la differenza fra x_i e $x_i^{(k+1)}$ si ottiene:

$$e_i^{(k+1)} = x_i - x_i^{(k+1)} = -\frac{1}{a_{ii}} \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}(x_j - x_j^{(k)}) = -\sum_{\substack{j=1\\ j\neq i}}^{n} \frac{a_{ij}}{a_{ii}}\ e_j^{(k)}$$

$$|e_i^{(k+1)}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| |e_j^{(k)}| \leq \Big\{ \sum_{\substack{j=1 \\ j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| \Big\} \|\mathbf{e}^{(k)}\|_\infty$$

Posto

$$\mu = \max_{1\leq i\leq n} \mu_i, \quad \text{con} \quad \mu_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right|, \quad i = 1, ..., n$$

si ha $\mu_i < 1$, essendo A matrice diagonalmente dominante in senso forte, e dunque anche $\mu < 1$.
Segue che:

$$|e_i^{(k+1)}| \leq \mu \|\mathbf{e}^{(k)}\|_\infty, \quad i = 1, ..., n,$$

e dunque:

$$\|\mathbf{e}^{(k+1)}\|_\infty \leq \mu \|\mathbf{e}^{(k)}\|_\infty \leq \mu^2 \|\mathbf{e}^{(k-1)}\|_\infty \leq ... \leq \mu^{k+1} \|\mathbf{e}^{(0)}\|_\infty$$

Si conclude che

$$\lim_{k\to\infty} \|\mathbf{e}^{(k+1)}\|_\infty = 0$$

Proprietà.
Sia A simmetrica con $a_{ii} > 0, \forall i$. Allora (GS) converge $\iff$ A è definita positiva.

Definizione.
Una matrice quadrata A si dice convergente se

$$\lim_{k\to\infty} [A^k]_{ij} = 0, \quad \forall i, j = 1, ..., n,$$

dove $[A^k]_{ij}$ è il generico elemento della matrice $A^k := \underbrace{A \cdot A \cdot ... \cdot A}_{k}$.

Proprietà.
Sono condizioni equivalenti:

1) A è una matrice convergente.
2) $\lim_{k\to\infty} \|A^k\| = 0$ per una norma di matrice.
3) $\rho(A) < 1$.

Corollario.
A è convergente se $\exists$ una norma di matrice tale che $\|A\| < 1$.

Contesto generale dei metodi iterativi.
Si cerca una decomposizione (splitting) di A nella forma: $A = N - P$, con $\det(N) \neq 0$, $Ny = f$ sistema facilmente risolubile e $K(N) << K(A)$. Allora:

$$A\mathbf{x} = \mathbf{b} \iff N\mathbf{x} - P\mathbf{x} = \mathbf{b} \iff N\mathbf{x} = P\mathbf{x} + \mathbf{b}.$$

Metodo iterativo: $N\mathbf{x}^{(k+1)} = P\mathbf{x}^{(k)} + \mathbf{b} \iff \mathbf{x}^{(k+1)} = N^{-1}P\mathbf{x}^{(k)} + N^{-1}\mathbf{b}$.

Matrice di iterazione associata: $B = N^{-1}P$.

Relazione sull'errore. Posto $\mathbf{e}^{(k)} = \mathbf{x} - \mathbf{x}^{(k)}$, si ha:

$$N\mathbf{x} = P\mathbf{x} + \mathbf{b}$$

$$N\mathbf{x}^{(k+1)} = P\mathbf{x}^{(k)} + \mathbf{b}$$

$$\Downarrow$$

$$N\mathbf{x} - N\mathbf{x}^{(k+1)} = P\mathbf{x} + \cancel{\mathbf{b}} - P\mathbf{x}^{(k)} - \cancel{\mathbf{b}}$$

$$\Downarrow$$

$$N(\mathbf{x} - \mathbf{x}^{(k+1)}) = P(\mathbf{x} - \mathbf{x}^{(k)})$$

$$\Downarrow$$

$$\mathbf{e}^{(k+1)} = B\mathbf{e}^{(k)}.$$

Teorema.
Condizione necessaria e sufficiente per la convergenza è che B sia una matrice convergente.

Teorema.
Condizione necessaria e sufficiente per la convergenza è che $\rho(B) < 1$.

Corollario.
Condizione sufficiente per la convergenza è che esista una norma naturale di matrice $\|\cdot\|$ tale che $\|B\| < 1$.

Dimostrazione.
Dalla relazione ricorsiva sull'errore si ha:

$$\|\mathbf{e}^{(k+1)}\| = \|B\mathbf{e}^{(k)}\| \leq \|B\| \, \|\mathbf{e}^{(k)}\| \leq \|B^2\| \, \|\mathbf{e}^{(k-1)}\| \leq ...$$

$$\leq \|B^{(k+1)}\| \, \|\mathbf{e}^{(0)}\| \leq \|B\|^{(k+1)} \|\mathbf{e}^{(0)}\|,$$

da cui si ottiene che, se $\|B\| < 1$, si ha

$$\lim_{k\to\infty} \|\mathbf{e}^{(k)}\| = 0,$$

ovvero:

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

Si osservi che la convergenza è garantita per ogni $\mathbf{e}^{(0)}$, cioè per ogni scelta del vettore iniziale $\mathbf{x}^{(0)}$.

Applicazione ai metodi (J) e (GS).

• $\boxed{A = L + D + U}$, con:

$L_{ij} = A_{ij},\ i > j$ (triangolo inferiore)
$U_{ij} = A_{ij},\ i < j$ (triangolo superiore)
$D_{ii} = A_{ii},\ \forall i$ (diagonale).

• (J) $\Rightarrow$ matrice di iterazione $\boxed{B_J = -D^{-1}(L+U)}$

• (GS) $\Rightarrow$ matrice di iterazione $\boxed{B_{GS} = -(D+L)^{-1}U}$

• Calcolo degli autovalori della matrice di iterazione del metodo (J):

$$\boxed{\det(B_J - \lambda I) = 0} \iff \det[-D^{-1}(L+U) - \lambda I] = 0 \iff$$

$$\det[D^{-1}(L+U) + \lambda \underbrace{D^{-1}D}_{I}] = 0 \iff \det[D^{-1}(L+U+\lambda D)] = 0$$

$$\iff \underbrace{\det D^{-1}}_{\neq 0} \det(L+U+\lambda D) = 0 \iff \boxed{\det(L+U+\lambda D) = 0}$$

• Calcolo degli autovalori della matrice di iterazione del metodo (GS):

$$\boxed{\det(B_{GS} - \lambda I) = 0} \iff \det[-(D+L)^{-1}U - \lambda I] = 0 \iff$$

$$\det[(D+L)^{-1}U + \lambda \underbrace{(D+L)^{-1}(D+L)}_{I}] = 0 \iff$$

$$\det\{(D+L)^{-1}[U + \lambda(D+L)]\} = 0 \iff \underbrace{\det(D+L)^{-1}}_{\neq 0} \det[U + \lambda(D+L)] = 0$$

$$\iff \boxed{\det[U + \lambda(D+L)] = 0}$$

Definizione.
Si chiama velocità asintotica del metodo iterativo relativo ad una matrice d'iterazione B il numero:

$$R(B) = -\ln \rho(B).$$

Proposizione.
Il numero di iterazioni m necessarie per ridurre l'errore di un fattore ε verifica la disuguaglianza

$$m \geq \frac{-\ln \varepsilon}{R(B)}.$$

Osservazione. La proposizione afferma che se

$$m \geq \frac{-\ln \varepsilon}{R(B)},$$

allora

$$\frac{\|\mathbf{x} - \mathbf{x}^{(m)}\|}{\|\mathbf{x} - \mathbf{x}^{(0)}\|} \leq \varepsilon.$$

In particolare, scegliendo $\mathbf{x}^{(0)} = \mathbf{0}$ si ottiene che m è il numero di iterazioni necessarie affinchè l'errore relativo sia minore di ε, cioè:

$$\frac{\|\mathbf{x} - \mathbf{x}^{(m)}\|}{\|\mathbf{x}\|} \leq \varepsilon.$$

Si ottiene pertando una stima a priori dell'errore relativo al passo m.

Proprietà.
Se A è una matrice tridiagonale di dimensione n non singolare con $a_{ii} \neq 0$, $i = 1, ..., n$, allora

$$\rho(B_{GS}) = [\rho(B_J)]^2,$$

cioè i metodi di Jacobi e di Gauss-Seidel sono entrambi convergenti o entrambi divergenti. Nel caso di convergenza, il metodo di Gauss-Seidel converge asintoticamente due volte più velocemente di quello di Jacobi, cioè vale la relazione

$$R(B_{GS}) = 2R(B_J).$$

Infatti:

$$R(B_{GS}) = -\ln \rho(B_{GS}) = -\ln[\rho(B_J)]^2 = -2\ln \rho(B_J) = 2[-\ln \rho(B_J)] = 2R(B_J).$$

METODI ITERATIVI PER SISTEMI LINEARI: COMPLEMENTI

Il metodo del rilassamento successivo (o metodo SOR)

A partire dal metodo di Gauss-Seidel si introduce il metodo SOR (*Successive Over-Relaxation*). Assegnato un vettore arbitrario $\mathbf{x}^{(0)} = \{x_1^{(0)}, x_2^{(0)}, ..., x_n^{(0)}\}$, e un parametro di rilassamento ω, si definisce, componente per componente, il metodo SOR come segue:

$$x_i^{(k+1)} = \frac{\omega}{a_{ii}} \underbrace{\left\{ b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right\}}_{\text{Gauss}-\text{Seidel}} + (1-\omega) x_i^{(k)}, \quad i = 1, ..., n.$$

Casi particolari:

• $\omega = 0$: $x_i^{(k+1)} = x_i^{(k)},\ \forall i = 1, ..., n,\ \forall k \geq 0$, dunque dovrà essere $\omega > 0$

• $\omega = 1$: Gauss-Seidel (Esiste una versione del metodo SOR anche a partire dal metodo di Jacobi).

• Forma matriciale del metodo SOR:

$$\mathbf{x}^{(k+1)} = \omega D^{-1}(\mathbf{b} - L\mathbf{x}^{(k+1)} - U\mathbf{x}^{(k)}) + (1-\omega) I \mathbf{x}^{(k)} =$$

$$-\omega D^{-1} L \mathbf{x}^{(k+1)} + [(1-\omega) I - \omega D^{-1} U] \mathbf{x}^{(k)} + \omega D^{-1} \mathbf{b}$$

• Matrice d'iterazione:

$$B_\omega = (I + \omega D^{-1} L)^{-1} [(1-\omega) I - \omega D^{-1} U]$$

$$\det(B_\omega) = \det[(I + \omega D^{-1} L)^{-1}] \det[(1-\omega) I - \omega D^{-1} U]$$

Si osservi che $I + \omega D^{-1} L$ è una matrice triangolare inferiore con elementi uguali a 1 sulla diagonale, mentre $(1-\omega) I - \omega D^{-1} U$ è una matrice triangolare superiore con elementi uguali a $1 - \omega$ sulla diagonale. Dunque si ha:

$$\det(B_\omega) = (1-\omega)^n$$

$$|(1-\omega)^n| = |\det B_\omega| = \left| \prod_{i=1}^{n} \lambda_i(B_\omega) \right| \leq \left[\max_{i=1,...n} |\lambda_i(B_\omega)| \right]^n = [\rho(B_\omega)]^n$$

Pertanto, condizione necessaria per la convergenza del metodo SOR:

$$|1-\omega| < 1 \Rightarrow 0 < \omega < 2$$

PROPRIETÀ. Se A è simmetrica definita positiva, il metodo SOR converge se e solo se $0 < \omega < 2$.

PROPRIETÀ. Sia A una matrice tridiagonale, con $a_{ii} \neq 0$. Se tutti gli autovalori della matrice di iterazione B_J del metodo di Jacobi sono reali, allora il metodo di Jacobi e il metodo SOR, per $0 < \omega < 2$, convergono o divergono simultaneamente. Nel caso di convergenza esiste un valore ω_{ott}:

$$\omega_{\text{ott}} = \frac{2}{1 + \sqrt{1 - \rho^2(B_J)}}$$

tale per cui $\rho(B_\omega)$ raggiunge il minimo valore uguale a $(\omega_{\text{ott}} - 1)$,cioè:

$$\omega_{\text{ott}} - 1 = \rho(B_{\omega_{\text{ott}}}) = \min_{0<\omega<2} \rho(B_\omega)$$

Metodi di Richardson (Metodi del gradiente)

Per il sistema $A\mathbf{x} = \mathbf{b}$ si considera lo splitting della matrice A dato da

$$A = I - (I - A)$$

da cui si ottiene il metodo iterativo:

$$\underline{\mathbf{x}^{(k+1)}} = (I-A)\mathbf{x}^{(k)} + \mathbf{b} = \mathbf{x}^{(k)} + \mathbf{b} - A\mathbf{x}^{(k)} = \underline{\mathbf{x}^{(k)} + \mathbf{r}^{(k)}}, \text{ dove } \mathbf{r}^{(k)} = \mathbf{b} - A\mathbf{x}^{(k)}.$$

La matrice di iterazione è dunque $B = I - A$.

Se si introduce un opportuno parametro di rilassamento (o di accelerazione) $\alpha \neq 0$ si può generalizzare tale metodo e definire il metodo di Richardson:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha \mathbf{r}^{(k)}$$

con matrice di iterazione $B_\alpha = I - \alpha A$. (Questo corrisponde a considerare il metodo precedente applicato al sistema modificato $\alpha A\mathbf{x} = \alpha\mathbf{b}$, che è equivalente a quello assegnato $\forall \alpha \neq 0$).

$\underline{\alpha = \text{costante}}$ (indipendente da k)
metodo di Richardson stazionario o metodo del gradiente a parametro costante

$\underline{\alpha = \alpha_k}$ (variabile in funzione del passo k)
metodo di Richardson dinamico o metodo del gradiente a parametro dinamico

PROPRIETÀ. Nel caso del metodo di Richardson stazionario o metodo del gradiente a parametro costante, se la matrice A ha autovalori reali positivi, ordinati in modo che $0 < \lambda_1 < \lambda_2 < ... < \lambda_n$, allora il metodo converge se e solo se $0 < \alpha < 2/\lambda_n$. Inoltre esiste un valore α_{ott}:

$$\alpha_{\text{ott}} = \frac{2}{\lambda_1 + \lambda_n}$$

tale per cui $\rho(B_\alpha)$ raggiunge il minimo valore uguale a

$$\min_\alpha[\rho(B_\alpha)] = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}.$$

Se inoltre la matrice A è simmetrica e definita positiva è noto che

$$K_2(A) = \|A\|_2\|A^{-1}\|_2 = \frac{\lambda_n}{\lambda_1}$$

e dunque

$$\min_\alpha[\rho(B_\alpha)] = \frac{K_2(A) - 1}{K_2(A) + 1}.$$

Nel caso di metodo di Richardson dinamico o metodo del gradiente a parametro dinamico, una possibile strategia è quella di determinare α_k in modo da rendere minima la quantità $\|\mathbf{r}^{(k+1)}\|_2^2$, al variare del parametro reale α. Si ha:

$$\mathbf{r}^{(k+1)} = \mathbf{b} - A\mathbf{x}^{(k+1)} = \mathbf{b} - A\mathbf{x}^{(k)} - \alpha A\mathbf{r}^{(k)} = \mathbf{r}^{(k)} - \alpha A\mathbf{r}^{(k)}$$

e dunque:

$$\|\mathbf{r}^{(k+1)}\|_2^2 = (\mathbf{r}^{(k+1)}, \mathbf{r}^{(k+1)}) = (\mathbf{r}^{(k)} - \alpha A\mathbf{r}^{(k)}, \mathbf{r}^{(k)} - \alpha A\mathbf{r}^{(k)}) =$$
$$\alpha^2(A\mathbf{r}^{(k)}, A\mathbf{r}^{(k)}) - 2\alpha(A\mathbf{r}^{(k)}, \mathbf{r}^{(k)}) + (\mathbf{r}^{(k)}, \mathbf{r}^{(k)}) \equiv [p_2(\alpha)]$$

Il valore di α che rende minima la quantità $\|\mathbf{r}^{(k+1)}\|_2^2 \equiv [p_2(\alpha)]$ è dunque dato dall'ascissa α_k del vertice della parabola $p_2(\alpha)$:

$$\alpha_k = \frac{(A\mathbf{r}^{(k)}, \mathbf{r}^{(k)})}{(A\mathbf{r}^{(k)}, A\mathbf{r}^{(k)})}$$

Riassumendo, il metodo di Richardson dinamico o metodo del gradiente a parametro dinamico, può essere così descritto:

$\mathbf{x}^{(0)}$ assegnato, $\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}$

$n \geq 0$:

$$\mathbf{z}^{(k)} = A\mathbf{r}^{(k)}$$

$$\alpha_k = \frac{(\mathbf{z}^{(k)}, \mathbf{r}^{(k)})}{(\mathbf{z}^{(k)}, \mathbf{z}^{(k)})}$$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k\mathbf{r}^{(k)}$$

$$\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k\mathbf{z}^{(k)}$$

$$\text{STOP se } \frac{\|\mathbf{r}^{(k+1)}\|_2}{\|\mathbf{b}\|_2} < \varepsilon$$

Esercizio

Sia dato il sistema lineare $A\mathbf{x} = \mathbf{b}$, con $A \in \mathbb{R}^{n\times n}$ matrice simmetrica definita positiva. Siano $\lambda_i,\ i = 1, ..., n$ gli autovalori di A, con

$$0 < m = \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n = M.$$

1) Fornire una condizione necessaria e sufficiente su m e M affinchè risulti convergente il metodo iterativo

$$(*) \qquad \mathbf{x}^{(k+1)} = -A\mathbf{x}^{(k)} + \mathbf{x}^{(k)} + \mathbf{b}.$$

2) Fornire una condizione necessaria e sufficiente su ω affinchè risulti convergente il metodo iterativo

$$(**) \qquad \mathbf{x}^{(k+1)} = \omega[(I - A)\mathbf{x}^{(k)} + \mathbf{b}] + (1 - \omega)\mathbf{x}^{(k)}.$$

3) Trovare in funzione di m e M il valore ottimale ω_{ott}, cioè il valore di ω per cui la il raggio spettrale della matrice di iterazione associata al metodo definito al punto 2) sia minimo. Calcolare la corrispondente velocità di convergenza ottimale in funzione del numero di condizionamento della matrice A nella norma $\|\cdot\|_2$.

Traccia dello svolgimento.

1) Matrice di iterazione associata al metodo $(*)$: $B = I - A$.
CNS per la convergenza: $\rho(B) = \rho(I - A) < 1$.

Problema di autovalori per $B = I - A$:
$(I - A)\mathbf{x} = \mu\mathbf{x}, \quad [\mu \in \sigma(B)]$
$I\mathbf{x} - A\mathbf{x} = \mu\mathbf{x}$
$A\mathbf{x} = (1 - \mu)\mathbf{x}$
$\lambda = 1 - \mu \Rightarrow \mu = 1 - \lambda, \quad [\lambda \in \sigma(A)]$.
CNS per la convergenza del metodo $(*)$:
$|1 - \lambda| < 1 \Leftrightarrow 0 < \lambda < 2, \ \ \forall\lambda \in \sigma(A) \Rightarrow 0 < m \leq M < 2$.

2) Il metodo $(**)$ si può riscrivere come:

$$\mathbf{x}^{(k+1)} = (\omega I - \omega A + I - \omega I)\mathbf{x}^{(k)} + \omega\mathbf{b} = (I - \omega A)\mathbf{x}^{(k)} + \omega\mathbf{b}.$$

Matrice di iterazione associata al metodo $(**)$: $B_\omega = I - \omega A$.
CNS per la convergenza: $\rho(B_\omega) = \rho(I - \omega A) < 1$.

Problema di autovalori per $B_\omega = I - \omega A$:

$(I - \omega A)\mathbf{x} = \eta\mathbf{x}, \quad [\eta \in \sigma(B_\omega)]$

$I\mathbf{x} - \omega A\mathbf{x} = \eta\mathbf{x}$

$\omega A\mathbf{x} = (I - \eta I)\mathbf{x}$

$A\mathbf{x} = \frac{1-\eta}{\omega} I\mathbf{x}$

$\lambda = \frac{1-\eta}{\omega} \Rightarrow \eta = 1 - \lambda\omega, \quad [\lambda \in \sigma(A)].$

CNS per la convergenza del metodo $(**)$:

$|1 - \lambda\omega| < 1 \Leftrightarrow 0 < \lambda\omega < 2, \ \ \forall\lambda \in \sigma(A) \Leftrightarrow 0 < \omega < \frac{2}{\lambda} \ \Rightarrow 0 < \omega < \frac{2}{M}.$

3) Il valore ottimale ω_{ott} è il valore di ω per cui $\rho(B_\omega)$ è minimo al variare di $\omega \in (0, 2/M)$:

$$\min_{\omega\in(0,2/M)} \rho(B_\omega) = \rho(B_{\omega_{\text{ott}}}).$$

Per definizione di raggio spettrale si ha:

$$\rho(B_\omega) = \rho(I - \omega A) = \max_{\eta\in\sigma(B_\omega)} |\eta| = \max_{i=1,\dots,n} |1 - \lambda_i\omega|.$$

Rappresentiamo in Figura 1 le curve $f_i(\omega) = |1 - \lambda_i\omega|$, $i = 1, ..., n$, con $0 < \omega < \frac{2}{M}$, tenendo conto che:

$$f_n\Big(\frac{2}{M}\Big) = \Big|1 - \lambda_n\frac{2}{M}\Big| = \Big|1 - M\frac{2}{M}\Big| = 1.$$

Consideriamo poi per esempio,

$$\frac{1}{\lambda_j} < \frac{1}{\lambda_i} < \frac{2}{M},$$

dove $\lambda_i < \lambda_j$ $(i < j)$, allora:

$$f_i\Big(\frac{2}{M}\Big) = \lambda_i\frac{2}{M} - 1 < \lambda_j\frac{2}{M} - 1 = f_j\Big(\frac{2}{M}\Big).$$

Nel caso invece in cui per qualche indice j si ha

$$\frac{2}{M} < \frac{1}{\lambda_j},$$

si ha $f_j(\omega) = |1 - \lambda_j\omega| = 1 - \lambda_j\omega$, $\forall\omega \in (0, 2/M)$.

In entrambi i casi si ha:

$\rho(B_\omega) = |1 - m\omega| = 1 - m\omega$, se $\omega \le \omega^*$,

$\rho(B_\omega) = |1 - M\omega| = M\omega - 1$, se $\omega \ge \omega^*$,

dove ω^* è il valore di ω per cui $f_1(\omega^*) = f_n(\omega^*)$, cioè tale per cui:

$$1 - m\omega = M\omega - 1 \Rightarrow \omega = \frac{2}{M+m},$$

che è anche il valore in corrispondenza del quale la quantità $\rho(B_\omega)$ assume il minimo valore, ossia:

$$\rho(B_{\omega^*}) = \rho(B_{\omega_{\text{ott}}}) = \min_{\omega \in (0,2/M)} \rho(B_\omega).$$

Si ha inoltre:

$$\rho(B_{\omega_{\text{ott}}}) = M\omega_{\text{ott}} - 1 = M\frac{2}{M+m} - 1 = \frac{M-m}{M+m} = \frac{\frac{M}{m} - 1}{\frac{M}{m} + 1},$$

oppure,

$$\rho(B_{\omega_{\text{ott}}}) = 1 - m\omega_{\text{ott}} = 1 - m\frac{2}{M+m} = \frac{M-m}{M+m} = ...$$

Essendo A matrice simmetrica e definita positiva si ha che

$$K_2(A) = \|A\|_2 \ \|A^{-1}\|_2 = \rho(A)\rho(A^{-1}) = \frac{M}{m}$$

e dunque

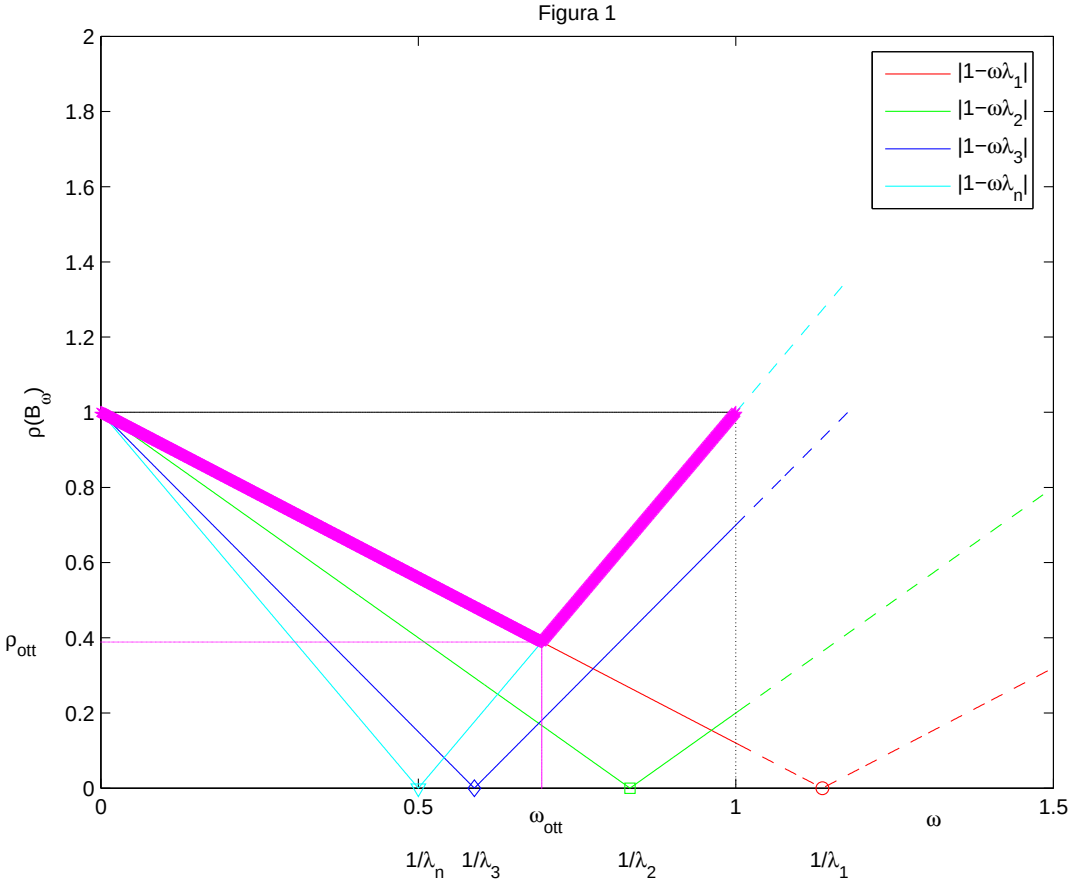
$$\rho(B_{\omega_{\text{ott}}}) = \frac{K_2(A) - 1}{K_2(A) + 1}.$$

Si osservi che per la velocità asintotica del metodo iterativo $(**)$, definita da $R(B_\omega) = -\ln \rho(B_\omega)$, si ha:

$$\lim_{\text{"}K_2(A) \to 1^+\text{"}} R(B_\omega) = \infty \quad \text{(velocità infinita)}$$

$$\lim_{\text{"}K_2(A) \to +\infty\text{"}} R(B_\omega) = 0 \quad \text{(velocità nulla).}$$

Figura 1


2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
ρ(B_ω)
ρ_ott
0
0.5
1
1.5
ω_ott
ω
1/λ_n
1/λ_3
1/λ_2
1/λ_1
|1−ωλ_1|
|1−ωλ_2|
|1−ωλ_3|
|1−ωλ_n|

Localizzazione di autovalori

Sia A una matrice quadrata di ordine n a coefficienti reali; il numero $\lambda \in \mathbb{C}$ è detto autovalore di A se esiste un vettore $\mathbf{x} \in \mathbb{C}$ non nullo tale che

$$A\mathbf{x} = \lambda\mathbf{x}.$$

Il vettore $\mathbf{x}$ è detto autovettore associato all'autovalore λ e l'insieme degli autovalori di A è detto spettro di A ed è denotato con il simbolo $\sigma(A)$. E' noto che $\sigma(A) = \sigma(A^T)$. Inoltre si definisce raggio spettrale della matrice A

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

Gli autovalori λ sono le soluzioni dell'equazione caratteristica

$$p_A(\lambda) \equiv \det(A - \lambda I) = 0,$$

dove $p_A(\lambda)$ è il polinomio caratteristico di grado n rispetto alla variabile λ. Dunque il problema della ricerca degli autovalori di una matrice è un PROBLEMA NON LINEARE, che viene approssimato con metodi numerici di tipo iterativo.
Una localizzazione preliminare della posizione degli autovalori della matrice A nel piano complesso può risultare utile per accelerare la convergenza dei metodi iterativi per l'approssimazione degli autovalori.
Una prima stima si ottiene ricordando che

$$\rho(A) \leq \|A\|,$$

per ogni norma naturale di matrice.

Localizzazioni più accurate sono fornite dai teoremi di Gershgorin.

• Primo teorema di Gershgorin

Sia $A \in \mathbb{R}^{n\times n}$ e λ un autovalore di A. Allora:

$$\lambda \in \bigcup_{i=1}^{n} Z_i, \quad Z_i = \{z \in \mathbb{C} : \ |z - a_{ii}| \leq \underbrace{\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|}_{r_i}\},$$

dove gli Z_i sono detti cerchi di Gershgorin.

Dimostrazione
Sia $\lambda \in \sigma(A)$ e sia $\mathbf{x}$ un autovettore corrispondente a λ. Sia k l'indice della componente del vettore $\mathbf{x}$ per cui si verifica

$$|x_k| = \max_{1\leq j\leq n} |x_j| = \|\mathbf{x}\|_\infty.$$

Allora dalla relazione $A\mathbf{x} = \lambda\mathbf{x}$, considerando l'equazione $k-$esima, si ha

$$\sum_{j=1}^{n} a_{kj}x_j = \lambda x_k,$$

$$\sum_{\substack{j=1\\ j\neq k}}^{n} a_{kj}x_j + a_{kk}x_k = \lambda x_k,$$

$$(\lambda - a_{kk})x_k = \sum_{\substack{j=1\\ j\neq k}}^{n} a_{kj}x_j,$$

$$|\lambda - a_{kk}|\ |x_k| \leq \sum_{\substack{j=1\\ j\neq k}}^{n} |a_{kj}|\ |x_j|,$$

$$|\lambda - a_{kk}| \leq \sum_{\substack{j=1\\ j\neq k}}^{n} |a_{kj}| \underbrace{\frac{|x_j|}{|x_k|}}_{\leq 1} \leq \sum_{\substack{j=1\\ j\neq k}}^{n} |a_{kj}| = r_k,$$

dunque $\lambda \in Z_k$.

• OSSERVAZIONE. Poichè $\sigma(A) = \sigma(A^T)$, applicando il teorema prima alla matrice A e successivamente alla matrice trasposta A^T, risulta che gli autovalori di A appartengono all'intersezione delle due regioni ottenute.

• Secondo teorema di Gershgorin
Se l'unione M_1 di k cerchi di Gershgorin è disgiunta dall'unione M_2 dei rimanenti $n-k$ cerchi di Gershgorin, allora k autovalori (ciascuno contato con la propria molteplicità) appartengono a M_1 e $n-k$ autovalori appartengono a M_2.

Approssimazione numerica di autovalori e autovettori

Il metodo delle potenze.
Non sempre è necessario conoscere lo spettro di A, cioè l'insieme di tutti i suoi autovalori; spesso, come nel caso dello studio della convergenza di metodi iterativi per la risoluzione numerica di sistemi lineari o nello studio del condizionamento di una matrice, ci si può limitare ad individuare quelli estremi, cioè quelli di modulo massimo e/o modulo minimo.
Sia $A \in \mathbb{R}^{n\times n}$ matrice con autovalori $\lambda_i,\ i = 1, ..., n$ tali che

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq ... \geq \lambda_n,$$

cioè si suppone che l'autovalore di modulo massimo λ_1 sia distinto dai restanti autovalori di A.
Supponiamo che esista per ipotesi una base di autovettori $\{\mathbf{x}_1,\ \mathbf{x}_2, ...,\ \mathbf{x}_n\}$, dove $\mathbf{x}_j$ è l'autovettore associato all'autovalore λ_j, cioè $A\mathbf{x}_j = \lambda_j\mathbf{x}_j$, $j = 1, ..., n$.
Sia $\mathbf{z}_0 \in \mathbb{C}^n$ un vettore di componenti complesse assegnato. Dunque

$$\mathbf{z}_0 = \sum_{j=1}^{n} \alpha_j \mathbf{x}_j.$$

Il metodo delle potenze consiste nel porre:

$$\mathbf{z}_1 = A\mathbf{z}_0,\ \mathbf{z}_2 = A\mathbf{z}_1 = AA\mathbf{z}_0 = A^2\mathbf{z}_0, ...,\ \mathbf{z}_k = A^k\mathbf{z}_0.$$

Utilizzando l'espressione dello sviluppo del vettore $\mathbf{z}_0$ nella base di autovettori e considerando la relazione $A\mathbf{x}_j = \lambda_j\mathbf{x}_j$ che lega l'autovalore j-esimo al corrispondente autovettore, si ha:

$$\mathbf{z}_k = A^k\mathbf{z}_0 = A^k \sum_{j=1}^{n} \alpha_j\mathbf{x}_j = \sum_{j=1}^{n} \alpha_j A^k\mathbf{x}_j = \sum_{j=1}^{n} \alpha_j\lambda_j^k\mathbf{x}_j =$$

$$\lambda_1^k\alpha_1\mathbf{x}_1 + \lambda_2^k\alpha_2\mathbf{x}_2 + ... + \lambda_n^k\alpha_n\mathbf{x}_n =$$

$$\lambda_1^k \left[\alpha_1\mathbf{x}_1 + \underbrace{\left(\frac{\lambda_2}{\lambda_1}\right)^k \alpha_2\mathbf{x}_2 + \left(\frac{\lambda_3}{\lambda_1}\right)^k \alpha_3\mathbf{x}_3 + ... + \left(\frac{\lambda_n}{\lambda_1}\right)^k \alpha_n\mathbf{x}_n}_{\to 0 \text{ se } k\to\infty} \right]$$

Dunque al crescere di k, $\mathbf{z}_k \approx \lambda_1^k\alpha_1\mathbf{x}_1$, cioè $\mathbf{z}_k$ tende a disporsi nella direzione dell'autovettore $\mathbf{x}_1$ associato all'autovalore λ_1.
Consideriamo ora il quoziente di Rayleigh che fornisce in generale l'autovalore λ a partire da un autovalore associato $\mathbf{x}$:

$$A\mathbf{x} = \lambda\mathbf{x} \implies (A\mathbf{x}, \mathbf{x}) = \lambda(\mathbf{x}, \mathbf{x}), \implies \lambda = \frac{(A\mathbf{x}, \mathbf{x})}{(\mathbf{x}, \mathbf{x})}$$

Dunque il quoziente di Rayleigh associato al vettore $\mathbf{z}_k \approx \lambda_1^k \alpha_1 \mathbf{x}_1$ tende all'autovalore di modulo massimo λ_1. Infatti:

$$\sigma_k = \frac{(A\mathbf{z}_k, \mathbf{z}_k)}{(\mathbf{z}_k, \mathbf{z}_k)} \approx \frac{(A\lambda_1^k \alpha_1 \mathbf{x}_1, \lambda_1^k \alpha_1 \mathbf{x}_1)}{(\lambda_1^k \alpha_1 \mathbf{x}_1, \lambda_1^k \alpha_1 \mathbf{x}_1)} = \frac{(\lambda_1^{k+1} \alpha_1 \mathbf{x}_1, \lambda_1^k \alpha_1 \mathbf{x}_1)}{(\lambda_1^k \alpha_1 \mathbf{x}_1, \lambda_1^k \alpha_1 \mathbf{x}_1)} = \frac{\lambda_1^{2k+1} \alpha_1^2 (\mathbf{x}_1, \mathbf{x}_1)}{\lambda_1^{2k} \alpha_1^2 (\mathbf{x}_1, \mathbf{x}_1)} = \lambda_1.$$

Si osservi tuttavia che:

- Se $|\lambda_1| < 1 \Rightarrow \ \|\mathbf{z}_k\|_2 \approx \|\lambda_1^k \alpha_1 \mathbf{x}_1\|_2 \to 0$;
- Se $|\lambda_1| > 1 \Rightarrow \|\mathbf{z}_k\|_2 \approx \|\lambda_1^k \alpha_1 \mathbf{x}_1\|_2 \to \infty$.

Si introduce quindi la seguente normalizzazione dei vettori:

$$\mathbf{y}_k = \frac{\mathbf{z}_k}{\|\mathbf{z}_k\|_2},$$

e si pone

$$\mathbf{z}_{k+1} = A\mathbf{y}_k.$$

Di conseguenza si ha:

$$\sigma_k = \frac{(A\mathbf{y}_k, \mathbf{y}_k)}{(\mathbf{y}_k, \mathbf{y}_k)} = (\mathbf{z}_{k+1}, \mathbf{y}_k),$$

essendo $(\mathbf{y}_k, \mathbf{y}_k) = \|\mathbf{y}_k\|_2 = 1$.

L'algoritmo è allora il seguente:

Inizializzazione: $\mathbf{z}_0 \neq \mathbf{0}$

$$\forall k \geq 0$$
$$\mathbf{y}_k = \frac{\mathbf{z}_k}{\|\mathbf{z}_k\|_2}$$
$$\mathbf{z}_{k+1} = A\mathbf{y}_k$$
$$\sigma_k = (\mathbf{z}_{k+1}, \mathbf{y}_k)$$

Test d'arresto, per esempio :$|\sigma_k - \sigma_{k-1}| < \varepsilon$, $k \geq 1$, oppure $\|\mathbf{z}_{k+1} - \sigma_k \mathbf{y}_k)\|_2 < \varepsilon$.

Osservazione 1. Il metodo delle potenze è convergente anche nel caso in cui l'autovalore di modulo massimo abbia molteplicità algebrica > 1.

Osservazione 2. Quando invece l'autovalore di modulo massimo non è unico, i risultati di convergenza non sono più applicabili. Si veda come esempio l'applicazione del metodo delle potenze alla matrice $A = [0\ 1;\ 1\ 0]$, che ha autovalori $+1$ e -1.

Osservazione 3. L'errore $\|\mathbf{y}_k - \mathbf{x}_1\|_2$ è proporzionale al rapporto $(\lambda_2/\lambda_1)^k$. Nel caso in cui la matrice A sia reale e simmetrica si può dimostrare che l'errore è proporzionale al rapporto $(\lambda_2/\lambda_1)^{2k}$.

Il metodo delle potenze inverse.
Una prima generalizzazione del metodo delle potenze consiste nell'applicare il metodo – nel caso in cui $\det(A) \neq 0$, cioè $\exists A^{-1}$ –, alla matrice inversa A^{-1} per approssimare l'autovalore di modulo minimo di A, essendo gli autovalori di A^{-1} i reciproci degli autovalori di A.

L'algoritmo è il seguente:

Inizializzazione: $\mathbf{z}_0 \neq \mathbf{0}$

$$\forall k \geq 0$$

$$\mathbf{y}_k = \frac{\mathbf{z}_k}{\|\mathbf{z}_k\|_2}$$

$$\mathbf{z}_{k+1} = A^{-1}\mathbf{y}_k \Longrightarrow A\mathbf{z}_{k+1} = \mathbf{y}_k$$

$$\mu_k = (\mathbf{z}_{k+1}, \mathbf{y}_k), \qquad \left(\lambda_n \approx \frac{1}{\mu_k}\right).$$

Cenni sulla risoluzione numerica di equazioni differenziali ordinarie (ODE)

Problema di Cauchy.

$$\begin{cases} y'(x) = f(x, y(x)) & x \in [x_0, T] \\ y(x_0) = y_0 & \text{(condizione iniziale)} \end{cases}$$

Formulazione integrale.

$$\int_{x_0}^{x} y'(t)\mathrm{d}t = \int_{x_0}^{x} f(t, y(t))\mathrm{d}t$$
$$y(x) - y(x_0) = \int_{x_0}^{x} f(t, y(t))\mathrm{d}t$$

Approssimazione numerica.

Nodi di discretizzazione in $[x_0, T]$:

$h > 0$, $x_j = x_0 + jh$, $j = 0, 1, ..., N$, $x_N \leq T$; [es.: $h = (T - x_0)/N$]

Soluzione esatta nei nodi: $y(x_j)$.

Soluzione approssimata/numerica: $u_j \approx y(x_j)$, $u_0 = y(x_0) = y_0$.

Definizione.
Un metodo numerico per l'approssimazione del problema di Cauchy si dice a un passo se $\forall n \geq 0$, u_{n+1} dipende solo da u_n e non da u_j, $j < n$. In caso contrario si dirà a più passi o multistep.

Definizione.
Un metodo numerico per l'approssimazione del problema di Cauchy si dice esplicito se u_{n+1} ri ricava direttamente in funzione dei valori nei punti precedenti x_j, $j \leq n$. Un metodo è implicito se u_{n+1} dipende implicitamente da se stesso attraverso la funzione f.

Il metodo di Eulero esplicito.

1) Costruzione geometrica.

   Si considera la retta r tangente al grafico della soluzione $y(x)$ nel punto di coordinate $(x_0, y(x_0))$ e si approssima il valore di $y(x_1)$ con il valore assunto dalla retta r nel punto di ascissa x_1.

   Equazione della retta tangente: $r(x) = y(x_0) + y'(x_0)(x - x_0)$

$$r(\ \boxed{x = x_1}\ ) = y(x_0) + y'(x_0)(x_1 - x_0) = y(x_0) + hf(x_0, y(x_0)),$$

da cui si ricava l'approssimazione

$$y(x_1) \approx u_1 = u_0 + hf(x_0, u_0)$$

Ripetendo lo stesso procedimento sugli intervalli $[x_1, x_2]$, $[x_2, x_3]$... si ottiene la formula del metodo di Eulero esplicito per i punti successivi:

$$u_0 = y_0, \quad \boxed{u_{n+1} = u_n + hf(x_n, u_n)} \; n \geq 0$$

2) Costruzione basata sullo sviluppo di Taylor (in avanti):

$$y(x) = y(x_0) + (x - x_0)y'(x_0) + \frac{1}{2!}(x - x_0)^2 y''(t_0), \quad x_0 < t_0 < x.$$

Calcolo dello sviluppo in $x = x_1$, dunque $(x_1 - x_0) = h$:

$$y(x_1) = y(x_0) + hy'(x_0) + \frac{1}{2}h^2 y''(t_0), \quad x_0 < t_0 < x_1$$

$$y(x_1) \approx u_1 = y(x_0) + hy'(x_0) = y(x_0) + hf(x_0, y(x_0)) = u_0 + hf(x_0, u_0)$$

da cui, $\forall n$, la formula del metodo di Eulero esplicito per i punti successivi.

3) Costruzione basata sulla differenziazione numerica (approssimazione delle derivate).

Per definizione:

$$y'(x_0) = \lim_{x \to x_0} \frac{y(x) - y(x_0)}{x - x_0},$$

da cui si può considerare

$$y'(x_0) = f(x_0, y(x_0)) \approx \frac{y(x_1) - y(x_0)}{x_1 - x_0} \Rightarrow y(x_1) \approx y(x_0) + hf(x_0, y(x_0))...$$

4) Costruzione basata sull'integrazione numerica.

Data la formulazione integrale del problema di Cauchy

$$\int_{x_0}^{x_1} y'(t)\mathrm{d}t = \int_{x_0}^{x_1} f(t, y(t))\mathrm{d}t$$

$$y(x_1) - y(x_0) = \int_{x_0}^{x_1} f(t, y(t))\mathrm{d}t$$

consideriamo un' approssimazione dell'integrale al secondo membro basata su un'interpolazione della funzione integranda con un polinomio di grado zero rispetto al nodo x_0:

$$\int_{x_0}^{x_1} g(t)\mathrm{d}t \approx hg(x_0), \quad \text{formula del rettangolo (sinistra)}$$

Si ottiene:

$$y(x_1) - y(x_0) \approx hf(x_0, y(x_0)) \;\Rightarrow\; y(x_1) = y(x_0) + hf(x_0, y(x_0))....$$

Il metodo di Eulero implicito.

$$u_0 = y_0, \quad \boxed{u_{n+1} = u_n + hf(x_{n+1}, u_{n+1})} \; n \geq 0$$

1) Costruzione geometrica.

   Si approssima il valore di $y(x_1)$ con il valore assunto dalla retta r passante per il punto di coordinate $(x_0, f(x_0))$ e avente come coefficiente angolare $y'(x_1) = f(x_1, y(x_1))$..

$$r(\ \boxed{x = x_1}\ ) = y(x_0) + hf(x_1, y(x_1)),$$

2) Costruzione basata sullo sviluppo di Taylor (all'indietro):

$$y(x) = y(x_1) + (x - x_1)y'(x_1) + \frac{1}{2!}(x - x_1)^2 y''(t_0), \quad x < t_0 < x_1.$$

   Calcolo dello sviluppo in $x = x_0$, dunque $(x_0 - x_1) = -h$:

$$y(x_0) = y(x_1) - hy'(x_1) + \frac{1}{2!}h^2 y''(t_0), \quad x_0 < t_0 < x_1$$
$$y(x_1) \approx y(x_0) + hy'(x_1) = y(x_0) + hf(x_1, y(x_1))...$$

3) Costruzione basata sulla differenziazione numerica (approssimazione delle derivate).

   Per definizione:

$$y'(x_1) = \lim_{x \to x_1} \frac{y(x) - y(x_1)}{x - x_1}$$

   da cui si può considerare

$$y'(x_1) = f(x_1, y(x_1)) \approx \frac{y(x_0) - y(x_1)}{x_0 - x_1} \Rightarrow y(x_1) \approx y(x_0) + hf(x_1, y(x_1))...$$

4) Costruzione basata sull'integrazione numerica.

   Consideriamo la seguente approssimazione dell'integrale basata su un'interpolazione della funzione integranda con un polinomio di grado zero rispetto al nodo x_1:

$$\int_{x_0}^{x_1} g(t)\mathrm{d}t \approx hg(x_1) \quad \text{formula del rettangolo (destra)}$$

   Si ottiene:

$$y(x_1) - y(x_0) \approx hf(x_1, y(x_1) \Rightarrow y(x_1) = y(x_0) + hf(x_1, y(x_1)).... $$

Il metodo dei trapezi o di Crank-Nicolson.

1) Sommando membro a membro le formule dei metodi di Eulero esplicito e implicito si ha:

$$2u_{n+1} = 2u_n + h[f(x_n, u_n) + f(x_{n+1}, u_{n+1})]$$

$$\boxed{u_{n+1} = u_n + \frac{h}{2}[f(x_n, u_n) + f(x_{n+1}, u_{n+1})]}$$

2) Data la formulazione integrale del problema di Cauchy, approssimando l'integrale con il metodo dei trapezi si ha:

$$y(x_1) - y(x_0) \approx \frac{h}{2}[f(x_0, y(x_0) + f(x_1, y(x_1)]....$$

Il metodo di Heun.

Si ottiene a partire dal metodo di Crank-Nicolson sostituendo nella $f(x_{n+1}, u_{n+1})$ al posto di u_{n+1} una sua approssimazione ottenuta con un passo del metodo di Eulero esplicito a partire da u_n:

$$\boxed{u_{n+1} = u_n + \frac{h}{2}[f(x_n, u_n) + f(x_{n+1}, \underbrace{u_n + hf(x_n, u_n)}_{\approx u_{n+1}})]}$$

Il metodo di Eulero modificato.

Data la formulazione integrale del problema di Cauchy, approssimando l'integrale con il metodo del punto medio si ha:

$$y(x_1) - y(x_0) \approx h\left[f\left(x_0 + \frac{h}{2}, y\left(x_0 + \frac{h}{2}\right)\right)\right]$$

da cui si ottiene la formula

$$u_1 - u_0 = hf\left(x_0 + \frac{h}{2}, u_{0+\frac{1}{2}}\right), \quad u_{0+\frac{1}{2}} \approx y\left(x_0 + \frac{h}{2}\right)$$

Al posto di $u_{0+\frac{1}{2}}$ si sostituisce una sua approssimazione ottenuta con mezzo passo del metodo di Eulero esplicito a partire da u_0:

$$u_1 - u_0 = hf\left(x_0 + \frac{h}{2}, \underbrace{u_0 + \frac{h}{2}f(x_0, u_0)}_{\approx u_{0+\frac{1}{2}}}\right)$$

In generale:

$$\boxed{u_{n+1} = u_n + hf\left(x_n + \frac{h}{2}, \underbrace{u_n + \frac{h}{2} f(x_n, u_n)}_{\approx u_{n+\frac{1}{2}}}\right)}$$

Il metodo del punto medio (metodo a due passi).

Data la formulazione integrale del problema di Cauchy

$$\int_{x_0}^{x_2} y'(t)\mathrm{d}t = \int_{x_0}^{x_2} f(t, y(t))\mathrm{d}t$$

$$y(x_2) - y(x_0) = \int_{x_0}^{x_2} f(t, y(t))\mathrm{d}t,$$

consideriamo l'approssimazione dell'integrale basata sul metodo del punto medio:

$$\int_{x_0}^{x_2} g(t)\mathrm{d}t \approx 2h[g(x_1)].$$

Si ottiene:

$$y(x_2) - y(x_0) \approx 2hf(x_1, y(x_1) \Rightarrow y(x_2) \approx y(x_0) + 2hf(x_1, y(x_1)....$$

e quindi, dato $u_0 = y_0$ e calcolato u_1 con un metodo a un passo, la formula del punto medio è data da:

$$\boxed{u_{n+1} = u_{n-1} + 2hf(x_n, u_n)} \quad n \geq 1$$

Il $\theta-$metodo.

Si ottiene mediante una combinazione lineare dei metodi di Eulero esplicito e implicito, con coefficienti $(1-\theta)$ e θ, con $\theta \in [0, 1]$:

$$\boxed{u_{n+1} = u_n + h[(1-\theta)f(x_n, u_n) + \theta f(x_{n+1}, u_{n+1})]}$$

$\theta = 0$: Eulero esplicito.
$\theta = 1$: Eulero implicito.

Metodi di Runge-Kutta di ordine 2.

Al variare del parametro $\alpha \in (0, 1]$, si definisce la famiglia di metodi:

$$\boxed{u_{n+1} = u_n + h(1-\alpha)f(x_n, u_n) + h\alpha f\left(x_n + \frac{h}{2\alpha}, u_n + \frac{h}{2\alpha}f(x_n, u_n)\right)}$$

$\alpha = \frac{1}{2}$: Heun.
$\alpha = 1$: Eulero modificato.

Cenni sull'analisi dei metodi numerici a un passo

Consistenza.

Un metodo esplicito a un passo si può scrivere in forma compatta come

$$u_0 = y_0, \quad u_{n+1} = u_n + h\Phi(x_n, u_n, f(x_n, u_n), h), \quad 0 \le n \le N-1.$$

Sostituendo le soluzioni esatte $y(x_n)$ e $y(x_{n+1})$ nello schema numerico, esso sarà soddisfatto a meno di un residuo ε_{n+1}, avendo preteso che la soluzione esatta verifichi lo schema numerico:

$$u_0 = y_0, \quad y(x_{n+1}) = y(x_n) + h\Phi(x_n, y(x_n), f(x_n, y(x_n)), h) + \underbrace{\varepsilon_{n+1}}_{\text{residuo}}, \quad 0 \le n \le N-1$$

Scrivendo il residuo nella forma

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

la quantità $\tau_{n+1}(h)$ è detta errore locale di troncamento associato al nodo x_{n+1}.
Si definisce poi errore globale di troncamento la quantità:

$$\tau(h) = \max_{0 \le n \le N-1} |\tau_{n+1}(h)|.$$

Definizione.

Un metodo numerico si dice consistente se

$$\lim_{h \to 0} \tau(h) = 0,$$

cioè l'errore globale di troncamento è infinitesimo rispetto a h. In particolare si dirà che uno schema numerico ha ordine di consistenza p se

$$\tau(h) = O(h^p).$$

Consistenza del metodo di Eulero esplicito.

$$u_{n+1} = u_n + hf(x_n, u_n)$$
$$y(x_{n+1}) - y(x_n) - hf(x_n, y(x_n)) = \varepsilon_{n+1}$$

Sfruttando lo sviluppo di Taylor di y con centro in x_n e incremento h si ha:

$$\underbrace{y(x_n) + hy'(x_n) + \frac{1}{2}h^2y''(t_n)}_{y(x_{n+1})} - y(x_n) - hy'(x_n) =$$

$$\frac{1}{2}h^2y''(t_n) = \varepsilon_{n+1} = h\tau_{n+1}(h), \quad x_n < t_n < x_{n+1}$$

Dunque:

$$\tau_{n+1}(h) = \frac{h}{2}y''(t_n)$$
$$\tau(h) = \frac{h}{2}\max_{0\le n\le N-1}|y''(x_n)| \Rightarrow \ \tau(h) = O(h).$$

Definizione di convergenza.

Un metodo numerico si dice convergente con ordine di convergenza p se

$$|u_n - y(x_n)| \le Ch^p, \quad \forall n \text{ tale che } 0 \le n \le N, \ C > 0$$

Concetto di *zero*-stabilità (stabilità su intervalli limitati).

Si dice che un metodo numerico è *zero*-stabile se, in un intervallo limitato e fissato (x_0, T), a piccole perturbazioni sui dati corrispondono piccole perturbazioni sulla soluzione, quando $h \to 0$.
Per una definizione rigorosa si veda: [A. Quarteroni, R. Sacco e F. Saleri, *Matematica Numerica*, Springer-Verlag Italia, Milano 2000].

Assoluta stabilità (stabilità su intervalli illimitati).

Il concetto di assoluta stabilità ha a che fare con il comportamento asintotico della soluzione numerica u_n, per $x_n \to \infty$.
Consideriamo il problema modello

$$(Pm) \qquad \begin{cases} y'(x) = -\lambda y(x) & x \in [0, \infty), \quad \lambda > 0 \\ y(0) = 1 & \text{(condizione iniziale)} \end{cases}$$

la cui soluzione esatta $y(x) = e^{-\lambda x}$ tende a 0 per $x \to \infty$.

Definizione.

Un metodo numerico per l'approssimazione di (Pm) si dice assolutamente stabile se

$$|u_n| \to 0 \quad \text{per} \quad x_n \to +\infty.$$

1) **Assoluta stabilità del metodo di Eulero esplicito.**

$$u_{n+1} = u_n - h\lambda u_n = (1-h\lambda)u_n = (1-h\lambda)^2 u_{n-1} = ... = (1-h\lambda)^{n+1} u_0$$

$$\Rightarrow u_{n+1} = (1-h\lambda)^{n+1}$$

$$|u_n| \to 0 \text{ per } x_n \to +\infty \Leftrightarrow |1-h\lambda| < 1 \Leftrightarrow -1 < 1-h\lambda < 1 \Leftrightarrow$$

$$0 < h\lambda < 2 \Leftrightarrow h < \frac{2}{\lambda}$$

2) **Assoluta stabilità del metodo di Eulero implicito.**

$$u_{n+1} = u_n - h\lambda u_{n+1} \;\Rightarrow\; (1+h\lambda)u_{n+1} = u_n \;\Rightarrow\; u_{n+1} = \frac{1}{1+h\lambda} u_n =$$

$$\frac{1}{(1+h\lambda)^2} u_{n-1} = ... = \frac{1}{(1+h\lambda)^{n+1}} u_0 = \frac{1}{(1+h\lambda)^{n+1}}$$

$$|u_n| \to 0 \text{ per } x_n \to +\infty \Leftrightarrow \frac{1}{|1+h\lambda|} < 1, \;\Leftrightarrow\; \forall\, h\lambda \;(h, \lambda > 0)$$

3) **Assoluta stabilità del metodo di Heun.**

$$u_{n+1} = u_n + \frac{h}{2}\left[-\lambda u_n - \lambda(u_n + hf(x_n, u_n))\right] = u_n + \frac{h}{2}\left[-\lambda u_n - \lambda(u_n - h\lambda u_n)\right] =$$

$$u_n + \frac{h}{2}[-\lambda u_n - \lambda u_n + h\lambda^2 u_n] = \left(1 - h\lambda + \frac{h^2\lambda^2}{2}\right) u_n = \left(1 - h\lambda + \frac{h^2\lambda^2}{2}\right)^2 u_{n-1} =$$

$$... = \left(1 - h\lambda + \frac{h^2\lambda^2}{2}\right)^{n+1} u_0 = \left(1 - h\lambda + \frac{h^2\lambda^2}{2}\right)^{n+1}$$

$$|u_n| \to 0 \text{ per } x_n \to +\infty \Leftrightarrow \left|1 - h\lambda + \frac{h^2\lambda^2}{2}\right| < 1 \Leftrightarrow -1 < 1 - h\lambda + \frac{h^2\lambda^2}{2} < 1$$

$$\begin{cases} 1 - h\lambda + \frac{h^2\lambda^2}{2} > -1 \\ 1 + h\lambda + \frac{h^2\lambda^2}{2} < 1 \end{cases} \Rightarrow \begin{cases} h^2\lambda^2 - 2h\lambda + 4 > 0 \\ h^2\lambda^2 - 2h\lambda < 0 \end{cases} \Rightarrow \begin{cases} \forall\, h\lambda \\ 0 < h\lambda < 2 \end{cases}$$