Calcolo Numerico - Appunti

260236

agosto 2024

Prefazione

Ogni sezione di teoria presente in questi appunti, è stata ricavata dalle seguenti risorse:

- Calcolo Scientifico: Esercizi e problemi risolti con MATLAB e Octave. [2]
- Slide del corso. [1]

Altro:

GitHub repository

Indice

1 Equazioni non lineari

1.1 Introduzione

Il **calcolo degli zeri di una funzione** f reale di variabile reale o delle **radici dell'equazione** $f(x) = 0$, è un problema assai ricorrente nel Calcolo Scientifico.

In generale, *non è possibile* approntare metodi numerici che calcolino gli zeri di una generica funzione in un numero finito di passi. I metodi numerici per la risoluzione di questo problema sono pertanto necessariamente *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 studiata.

1.2 Il metodo di bisezione (o iterativo)

Sia f una funzione continua in $[a, b]$ tale che $f(a) f(b) < 0$. Per cui, vale il teorema degli zeri di una funzione continua, ossia f ammette almeno uno zero in (a, b).

Si supponga che ci sia un solo zero, indicato con α e nel caso in cui ce ne sia più di uno, individuare un intervallo tale che ne contenga solo uno.

Il **metodo di bisezione** (o **iterativo**) è una strategia che si suddivide nei seguenti passaggi:

1. **Dimezzare l'intervallo di partenza**;
2. **Selezionare tra i due sotto-intervalli ottenuti quello nel quale f cambia di segno agli estremi**;
3. **Applicare ricorsivamente questa procedura all'ultimo intervallo selezionato**.

Matematicamente parlando, dato $I^{(0)} = (a, b)$, e più in generale, $I^{(k)}$ il sotto-intervallo selezionato al passo k-esimo, si sceglie come $I^{(k+1)}$ il semi-intervallo di $I^{(k)}$ ai cui estremi f cambia di segno.

Questa procedura garantisce che ogni sotto-intervallo selezionato $I^{(k)}$ conterrà α. Questo poiché la successione $\left\{x^{(k)}\right\}$ dei punti medi dei sotto-intervalli $I^{(k)}$ dovrà ineluttabilmente convergere a α, in quanto la **lunghezza dei sotto-intervalli tende a** 0 per k che **tende all'infinito**.

Formalizziamo questa idea con un piccolo algoritmo. Ponendo:

$$a^{(0)} = a, \quad b^{(0)} = b, \quad I^{(0)} = \left(a^{(0)}, b^{(0)}\right), \quad x^{(0)} = \frac{a^{(0)} + b^{(0)}}{2}$$

Al passo $k \geq 1$ il metodo di bisezione calcolerà il semi-intervallo $I^{(k)} = \left(a^{(k)}, b^{(k)}\right)$ dell'intervallo $I^{(k-1)} = \left(a^{(k-1)}, b^{(k-1)}\right)$, nel seguente modo (si ricorda che α è lo zero che si sta cercando):

1. Calcolo $x^{(k-1)} = \dfrac{a^{(k-1)} + b^{(k-1)}}{2}$
2. Se $f\left(x^{(k-1)}\right) = 0$:
   (a) Allora $\alpha = x^{(k-1)}$ e l'algoritmo termina.
3. Altrimenti, se $f\left(a^{(k-1)}\right) \cdot f\left(x^{(k-1)}\right) < 0$:
   (a) Si pone $a^{(k)} = a^{(k-1)}$
   (b) Si pone $b^{(k)} = x^{(k-1)}$
   (c) Si incrementa $k + 1$ e si ripete ricorsivamente.
4. Altrimenti, se $f\left(x^{(k-1)}\right) \cdot f\left(b^{(k-1)}\right) < 0$:
   (a) Si pone $a^{(k)} = x^{(k-1)}$
   (b) Si pone $b^{(k)} = b^{(k-1)}$
   (c) Si incrementa $k + 1$ e si ripete ricorsivamente.

Esempio 1

Data la funzione $f(x) = x^2 - 1$, si parta da $a^{(0)} = -0.25$ e $b^{(0)} = 1.25$, e si applichi il metodo di bisezione:

1. Con $a^{(0)} = -0.25$ e $b^{(0)} = 1.25$:
   (a) Si calcola il punto medio:
   $$x^{(0)} = \frac{a^{(0)} + b^{(0)}}{2} = \frac{-0.25 + 1.25}{2} = 0.5$$
   (b) Si calcola la funzione con il punto medio come parametro:
   $$f(0.5) = 0.5^2 - 1 = -0.75$$
   (c) Dato che la funzione nel punto medio non è uguale a zero, l'algoritmo deve continuare. Per farlo, bisogna sostituire il punto medio con uno dei due estremi. Per decidere quale dei due sostituire, è necessario capire in quale cambia valore la funzione. Si verifica inizialmente con $a^{(0)}$:
   $$\begin{aligned} f\left(a^{(0)}\right) f\left(x^{(0)}\right) < 0 &= f(-0.25)\, f(0.5) < 0 \\ &= (-0.9375 \cdot -0.75) < 0 \\ &= 0.703125 \text{ ✗} \end{aligned}$$

(d) Si procede con l'algoritmo, provando adesso la $b^{(0)}$:

$$\begin{aligned} f\left(x^{(0)}\right) f\left(b^{(0)}\right) < 0 &= f(0.5)\, f(1.25) < 0 \\ &= (-0.75 \cdot 0.5625) < 0 \\ &= -0.421875 \text{ ✓} \end{aligned}$$

(e) Si pone $a^{(1)} = x^{(0)} = 0.5$

(f) Si pone $b^{(1)} = b^{(0)} = 1.25$

(g) Si incrementa k, $k = k + 1 = 0 + 1 = 1$

2. Con $a^{(1)} = 0.5$ e $b^{(1)} = 1.25$:

(a) Si calcola il punto medio:

$$x^{(1)} = \frac{a^{(1)} + b^{(1)}}{2} = \frac{0.5 + 1.25}{2} = 0.875$$

(b) Si calcola la funzione con il punto medio come parametro:

$$f(0.875) = 0.875^2 - 1 = -0.234375$$

(c) Dato che la funzione nel punto medio non è uguale a zero, l'algoritmo deve continuare:

$$\begin{aligned} f\left(a^{(1)}\right) f\left(x^{(1)}\right) < 0 &= f(0.5)\, f(-0.234375) < 0 \\ &= (-0.75 \cdot -0.945068359375) < 0 \\ &= 0.70880126953125 \text{ ✗} \end{aligned}$$

(d) Si procede con l'algoritmo:

$$\begin{aligned} f\left(x^{(1)}\right) f\left(b^{(1)}\right) < 0 &= f(-0.234375)\, f(1.25) < 0 \\ &= (-0.945068359375 \cdot 0.5625) < 0 \\ &= -0.5316009521484375 \text{ ✓} \end{aligned}$$

(e) Si pone $a^{(2)} = x^{(1)} = 0.875$

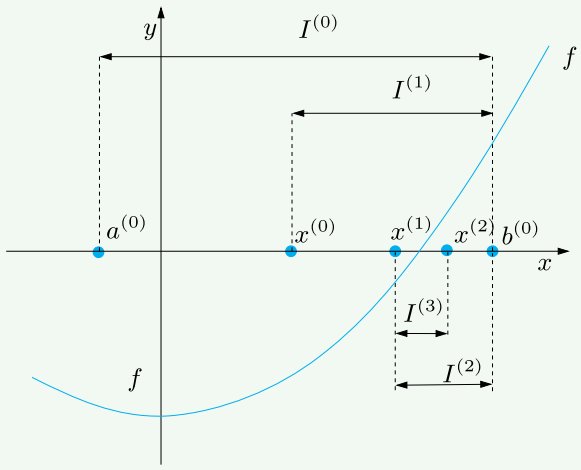
(f) Si pone $b^{(2)} = b^{(1)} = 1.25$

(g) Si incrementa k, $k = k + 1 = 1 + 1 = 2$

Si omettono i restanti calcoli per $k = 2, k = 3$, ma si lasciano qua di seguito i risultati:

- $I^{(2)} = (0.875, 1.25)$ e $x^{(2)} = 1.0625$
- $I^{(3)} = (0.875, 1.0625)$ e $x^{(2)} = 0.96875$

Nella seguente figura si possono vedere le iterazioni effettuate:



Iterazioni effettuate. [2]

Si noti che ogni intervallo $I^{(k)}$ contiene lo zero α. Inoltre, la successione $\{x^{(k)}\}$ converge necessariamente allo zero α in quanto ad ogni passo l'ampiezza $\left|I^{(k)}\right| = b^{(k)} - a^{(k)}$ dell'intervallo $I^{(k)}$ si dimezza.

Il valore $I^{(k)}$ può essere riassunto come:

$$\left|I^{(k)}\right| = \left(\frac{1}{2}\right)^{k} \cdot \left|I^{(0)}\right|$$

E di conseguenza l'**errore al passo *k*** può essere calcolato come:

$$\left|e^{(k)}\right| = \left|x^{(k)} - \alpha\right| < \frac{1}{2} \cdot \left|I^{(k)}\right| = \left(\frac{1}{2}\right)^{k+1} \cdot (b-a)$$

Inoltre, data una certa **tolleranza** ε, per **garantire che l'errore al passo** k **sia minore della tolleranza data** (ovvero, $\left|e^{(k)}\right| < \varepsilon$), basta applicare la seguente formula:

$$k_{\min} > \log_2\left(\frac{b-a}{\varepsilon}\right) - 1 \tag{1}$$

Dove $k_{\min}$ rappresenta il **numero <u>minimo</u> di iterazioni prima di trovare un intero che soddisfi la disuguaglianza**.

Possibile svantaggio

Il metodo di bisezione **non garantisce una riduzione monotona dell'errore**, ma solo il dimezzamento dell'ampiezza dell'intervallo all'interno del quale si cerca lo zero. Infatti, **non viene tenuto conto del reale andamento di** f e questo può provocare il mancato coinvolgimento di approssimazioni di α accurate.

1.3 Il metodo di Newton

Il **metodo di Newton** sfrutta la funzione f maggiormente rispetto al metodo di bisezione, usando i suoi valori e la sua derivata.

Si ricorda che la retta tangente alla curva $(x, f(x))$ nel punto $x^{(k)}$ è:

$$y(x) = f\left(x^{(k)}\right) + f'\left(x^{(k)}\right)\left(x - x^{(k)}\right)$$

Cercando un $x^{(k+1)}$ tale che la **retta tangente in quel punto sia uguale a zero** $y\left(x^{(k+1)}\right) = 0$, allora si trova:

$$x^{(k+1)} = x^{(k)} - \frac{f\left(x^{(k)}\right)}{f'\left(x^{(k)}\right)} \qquad k \geq 0 \tag{2}$$

Purché la derivata prima nel punto $x^{(k)}$ sia diversa da zero, cioè $f'\left(x^{k}\right) \neq 0$.

Questa equazione consente di calcolare una successione di valori $x^{(k)}$ a partire da un dato iniziale $x^{(0)}$. In altre parole, il **metodo di Newton calcola lo zero di f sostituendo localmente a f la sua retta tangente**.

A differenza del metodo di bisezione, tale **metodo converge allo zero in un solo passo quando la funzione f è lineare**, ovvero nella forma $f(x) = a_1 x + a_0$.

Limitazione

La **convergenza** del metodo di Newton non è garantita **per ogni scelta** di $x^{(0)}$, ma **soltanto** per valori di $x^{(0)}$ **sufficientemente vicini** ad α, ovvero **appartenenti ad un intorno** $I(\alpha)$ sufficientemente piccolo di α.

Alcune osservazioni a seguito anche di questa limitazione:

- A seguito di questa limitazione, risulta evidente che se $x^{(0)}$ è stato scelto opportunamente e se lo zero α è semplice ($f'(\alpha) \neq 0$), allora il metodo converge.
- Nel caso in cui f è derivabile con continuità pari a due, allora si ottiene la seguente convergenza:

  $$\lim_{k\to\infty} \frac{x^{(k+1)-\alpha}}{\left(x^{(k)} - \alpha\right)^2} = \frac{f''(\alpha)}{2f'(\alpha)} \tag{3}$$

  Il significato è: se $f'(\alpha) \neq 0$ il metodo di Newton converge almeno quadraticamente o con **ordine 2**.

  In parole povere, **per k sufficientemente grande, l'errore al passo $(k+1)$-esimo si comporta come il quadrato dell'errore al passo k-esimo, moltiplicato per una costante indipendente da k.**

- Se lo zero α ha molteplicità m maggiore di 1, ovverosia:

  $$f'(\alpha) = 0, \ldots, f^{(m-1)}(\alpha) = 0$$

Allora il metodo di Newton è ancora convergente, purché $x^{(0)}$ sia scelto opportunamente e $f'(x) \neq 0 \ \forall x \in I(\alpha) \setminus \{\alpha\}$. Tuttavia in questo caso l'ordine di convergenza è pari a 1. In tal caso, l'ordine 2 può essere ancora recuperato usando la seguente relazione al posto dell'equazione 2 ufficiale:

$$x^{(k+1)} = x^{(k)} - m \cdot \frac{f\left(x^{(k)}\right)}{f'\left(x^{(k)}\right)} \qquad k \geq 0 \tag{4}$$

Purché $f'\left(x^{(k)}\right) \neq 0$. Naturalmente, questo **metodo di Newton modificato** richiede una conoscenza a priori di m.

1.3.1 Come arrestare il metodo di Newton

Data una tolleranza fissa ε, esistono due tecniche applicabili per capire quando è necessario fermarsi ed evitare di continuare ad iterare:

- La **differenza fra due iterate consecutive**, il quale si arresta in corrispondenza del più piccolo intero $k_{\min}$ per il quale:

$$\left|x^{(k_{\min})} - x^{(k_{\min}-1)}\right| < \varepsilon \tag{5}$$

  (test sull'incremento).

- Un'altra tecnica applicata anche per altri metodi iterativi è il **residuo** al passo k, il quale è definito come:

$$r^{(k)} = f\left(x^{(k)}\right)$$

  Che è nullo quando $x^{(k)}$ è uno zero di f. In questo modo, il metodo viene arrestato alla prima iterata $k_{\min}$:

$$\left|r^{(k_{\min})}\right| = \left|f\left(x^{(k_{\min})}\right)\right| < \varepsilon \tag{6}$$

  Da notare che tale tecnica fornisce una **stima accurata dell'errore** solo quando $|f'(x)|$ è circa pari a 1 in un intorno di I_α dello zero α cercato.

  Attenzione! Se la derivata non è circa pari a 1 in un intorno dello zero cercato, la tecnica porterà:

  - Ad una **sovrastima** dell'errore se $|f'(x)| \gg 1$ per $x \in I_\alpha$
  - Ad una **sottostima** dell'errore se $|f'(x)| \ll 1$ per $x \in I_\alpha$

1.4 Il metodo delle secanti

Nel caso in cui la funzione f non sia nota, il metodo di Newton non può essere applicato. Per fortuna, arriva in soccorso il **metodo delle secanti**, il quale esegue una valutazione di $f'\left(x^{(k)}\right)$ andando a sostituire quest'ultima con un **rapporto incrementale calcolato su valori di f già noti**.

Più formalmente, assegnati due punti $x^{(0)}$ e $x^{(1)}$, per $k \geq 1$ si calcola:

$$x^{(k+1)} = x^{(k)} - \left(\frac{f\left(x^{(k)}\right) - f\left(x^{(k-1)}\right)}{x^{(k)} - x^{(k-1)}} \right)^{-1} \cdot f\left(x^{(k)}\right) \tag{7}$$

Quando converge?

Il metodo delle secanti converge a seguito di certe condizioni:

- **Converge ad α**, se:
  - α radice semplice[1];
  - $I(\alpha)$ è un opportuno intorno di α;
  - $x^{(0)}$ e $x^{(1)}$ sono sufficientemente vicini ad α
  - $f'(x) \neq 0 \ \forall x \in I(\alpha) \setminus \{\alpha\}$
- **Converge con ordine p super-lineare**, se:
  - $f \in \mathcal{C}^2(I(\alpha))$
  - $f'(\alpha) \neq 0$

  Ovvero, esiste una costante $c > 0$ tale che:

$$\left| x^{(k+1)} - \alpha \right| \leq c \left| x^{(k)} - \alpha \right|^p \qquad p = \frac{1+\sqrt{5}}{2} \approx 1.618\ldots \tag{8}$$

- **Convergenza lineare**, se:
  - Radice α è multipla.

  Come succederebbe usando il metodo di Newton.

[1] $f'(\alpha) \neq 0$

1.5 I sistemi di equazioni non lineari

Di solito i metodi presentati nelle pagine precedenti vengono inseriti in dei sistemi. Nella realtà ci sono varie condizioni che influiscono sul sistema in analisi. È per questo motivo che si introducono i sistemi.

Si consideri un generale sistema di equazioni non lineari:

$$\begin{cases} f_1(x_1, x_2, \dots, x_n) = 0 \\ f_2(x_1, x_2, \dots, x_n) = 0 \\ \vdots \\ f_n(x_1, x_2, \dots, x_n) = 0 \end{cases}$$

Dove $f_1, \dots, f_n$ sono funzioni non lineari. Si pongono i seguenti vettori:

- $\mathbf{f} \equiv (f_1, \dots, f_n)^T$
- $\mathbf{x} \equiv (x_1, \dots, x_n)^T$

Con l'obbiettivo di riscrivere il sistema in maniera più agevole:

$$\mathbf{f}(\mathbf{x}) = \mathbf{0}$$

Esempio 2: esempio di sistema non lineare

Un esempio banale di sistema non lineare:

$$\begin{cases} f_1(x_1, x_2) = x_1^2 + x_2^2 - 1 = 0 \\ f_2(x_1, x_2) = \sin\left(\pi \dfrac{x_1}{2}\right) + x_2^3 = 0 \end{cases}$$

Prima di estendere i metodi di Newton e delle secanti si introduce la matrice Jacobiana.

Definizione 1: matrice Jacobiana

Senza entrare troppo nel gergo matematico (non è l'obbiettivo del corso), la **matrice Jacobiana** di una funzione è quella **matrice i cui elementi sono le derivate parziali prime della funzione**.

$$\mathbf{J_f} = \begin{pmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_m}{\partial x_1} & \cdots & \dfrac{\partial f_m}{\partial x_n} \end{pmatrix} \tag{9}$$

Che può essere riscritto in modo più leggibile come:

$$(\mathbf{J_f})_{ij} \equiv \frac{\partial f_i}{\partial x_j} \qquad i, j = 1, \dots, n \tag{10}$$

Dove rappresenta la derivata parziale della funzione f_i rispetto a x_j.

Il metodo di Newton e delle secanti può essere esteso sfruttando la matrice Jacobiana:

- Il **metodo di Newton** usando un sistema di equazioni non lineari diventa: dato $\mathbf{x}^{(0)} \in \mathbb{R}^n$, per $k = 0, 1, \ldots$, fino a convergenza

$$\begin{aligned} &\text{risolvere} && \mathbf{J_f}\left(\mathbf{x}^{(k)}\right) \boldsymbol{\delta}\mathbf{x}^{(k)} = -\mathbf{f}\left(\mathbf{x}^{(k)}\right) \\ &\text{porre} && \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \boldsymbol{\delta}\mathbf{x}^{(k)} \end{aligned} \tag{11}$$

  Se ne deduce che venga richiesto ad ogni passo la soluzione di un sistema lineare di matrice $\mathbf{J_f}\left(\mathbf{x}^{(k)}\right)$.

- Il **metodo delle secanti** usando un sistema di equazioni non lineari si basa sulla matrice Jacobiana e sul metodo di Broyden.

  L'**idea di base** è sostituire le matrici Jacobiane $\mathbf{J_f}\left(\mathbf{x}^{(k)}\right)$ (per $k \geq 0$) con delle matrici chiamate B_k, definite ricorsivamente a partire da una matrice B_0 che sia una approssimazione di $\mathbf{J_f}\left(\mathbf{x}^{(0)}\right)$.

  Dato $\mathbf{x}^{(0)} \in \mathbb{R}^n$, data $\mathrm{B}_0 \in \mathbb{R}^{n \times n}$ per $k = 0, 1, \ldots$, fino a convergenza

$$\begin{aligned} &\text{risolvere} && \mathrm{B}_k \boldsymbol{\delta}\mathbf{x}^{(k)} = -\mathbf{f}\left(\mathbf{x}^{(k)}\right) \\ &\text{porre} && \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \boldsymbol{\delta}\mathbf{x}^{(k)} \\ &\text{porre} && \boldsymbol{\delta}\mathbf{f}^{(k)} = \mathbf{f}\left(\mathbf{x}^{(k+1)}\right) - \mathbf{f}\left(\mathbf{x}^{(k)}\right) \\ &\text{calcolare} && \mathrm{B}_{k+1} = \mathrm{B}_k + \frac{\left(\boldsymbol{\delta}\mathbf{f}^{(k)} - \mathrm{B}_k \boldsymbol{\delta}\mathbf{x}^{(k)}\right) \boldsymbol{\delta}\mathbf{x}^{(k)^T}}{\boldsymbol{\delta}\mathbf{x}^{(k)^T} \boldsymbol{\delta}\mathbf{x}^{(k)}} \end{aligned} \tag{12}$$

  Da notare che non si chiede alla successione $\{\mathrm{B}_k\}$ così costruita di convergere alla vera matrice Jacobiana $\mathbf{J_f}(\boldsymbol{\alpha})$ ($\boldsymbol{\alpha}$ è la radice del sistema); questo risultato non è garantito tuttavia.

1.6 Iterazioni di punto fisso

Esempio 3: esempio di introduzione

Con una calcolatrice si può facilmente verificare che applicando ripetutamente la funzione cos partendo dal numero 1 si genera la seguente successione di numeri reali:

$$\begin{aligned} x^{(1)} &= \cos(1) &&= 0.54030230586814 \\ x^{(2)} &= \cos\left(x^{(1)}\right) &&= 0.85755321584639 \\ &\vdots \\ x^{(10)} &= \cos\left(x^{(9)}\right) &&= 0.74423735490056 \\ &\vdots \\ x^{(20)} &= \cos\left(x^{(19)}\right) &&= 0.73918439977149 \end{aligned}$$

Che tende al valore $\alpha = 0.73908513$.

Con l'esempio di introduzione è possibile capire il punto fisso. Essendo per costruzione $x^{(k+1)} = \cos\left(x^{(k)}\right)$ per $k = 0, 1, \ldots$ (con $x^{(0)} = 1$), α è tale che $\cos(\alpha) = \alpha$. Quindi, α viene detto punto fisso della funzione coseno.

Perché è interessante?

Se α è un punto fisso per il coseno, allora esso è uno zero della funzione $f(x) = x - \cos(x)$ ed il metodo appena proposto potrebbe essere usato per il calcolo degli zeri di f.

Non tutte le funzioni hanno un punto fisso

Non tutte le funzioni ammettono punti fissi. Ad esempio, ripetendo l'esperimento dell'esempio con una funzione esponenziale a partire da $x^{(0)} = 1$, dopo soli 4 passi si giunge ad una situazione di *overflow* (figura 1, pagina 15).

Definizione 2

Data una funzione $\phi : [a, b] \to \mathbb{R}$, trovare $\alpha \in [a, b]$ tale che:

$$\alpha = \phi(\alpha)$$

Se tale α esiste, viene detto un **punto fisso** di ϕ e lo si può determinare come limite della seguente successione:

$$x^{(k+1)} = \phi\left(x^{(k)}\right) \qquad k \geq 0 \tag{13}$$

Dove $x^{(0)}$ è un dato iniziale. L'algoritmo viene chiamato **iterazioni di punto fisso** e la funzione ϕ è detta **funzione di iterazione**.

Dalla definizione, si deduce che l'esempio introduttivo è un algoritmo di iterazioni di punto fisso per la funzione $\phi(x) = \cos(x)$.

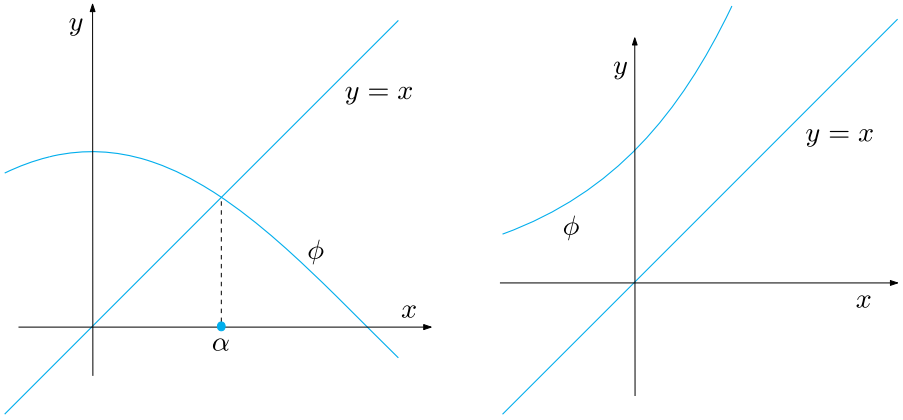


Figura 1: La funzione $\phi(x) = \cos(x)$ (sx) ammette un solo punto fisso, mentre la funzione $\phi(x) = e^x$ (dx) non ne ammette alcuno.

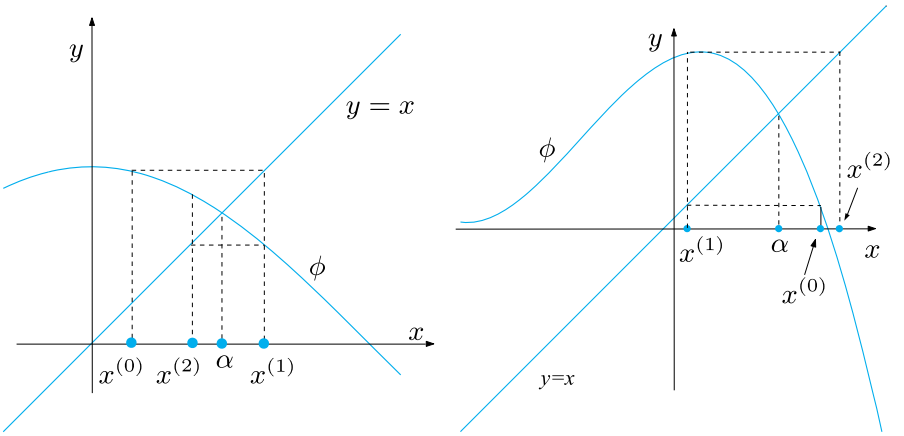


Figura 2: Rappresentazione delle prime iterazioni di punto fisso per due funzioni di iterazione. Le iterazioni convergono verso il punto fisso α (sx), mentre si allontanano da α (dx).

Definizione 3: quando una funzione ha un punto fisso?

Si consideri la successione (formula) 13 a pagina 14.

1. Si supponga che $\phi(x)$ sia continua nell'intervallo $[a,b]$ e che $\phi(x) \in [a,b]$ per ogni $x \in [a,b]$; allora **esiste almeno un punto fisso** $\alpha \in [a,b]$.

2. Si supponga inoltre che esista un valore L minore di 1 tale per cui:

$$|\phi(x_1) - \phi(x_2)| \leq L\,|x_1 - x_2|$$

Per ogni x_1, x_2 appartenente all'insieme $[a,b]$. Con tale supposizione, allora ϕ ha un **unico punto fisso** $\phi \in [a,b]$ e la successione definita nell'equazione 13 a pagina 14 converge a α, qualunque sia il dato iniziale $x^{(0)}$ in $[a,b]$.

La supposizione scritta in precedenza può essere riassunta in un'equazione:

$$\exists L < 1 \;\; t.c. \;\; |\phi(x_1) - \phi(x_2)| \leq L\,|x_1 - x_2| \quad \forall x_1, x_2 \in [a,b] \quad (14)$$

Nella pratica è però spesso **difficile delimitare a priori l'ampiezza dell'intervallo** $[a,b]$; in tal caso è utile il seguente risultato di convergenza locale:

Teorema 1 (di Ostrowski)**.** *Sia α un punto fisso di una funzione ϕ continua e derivabile con continuità in un opportuno intorno $\mathcal{J}$ di α. Se risulta $|\phi'(\alpha)| < 1$, allora esiste un $\delta > 0$ in corrispondenza del quale la successione $\left\{x^{(k)}\right\}$ converge ad α, per ogni $x^{(0)}$ tale che $\left|x^{(0)} - \alpha\right| < \delta$. Inoltre si ha:*

$$\lim_{k\to\infty} \frac{x^{(k+1)} - \alpha}{x^{(k)} - \alpha} = \phi'(\alpha) \quad (15)$$

Dal teorema si deduce che le iterazioni di punto fisso convergono almeno linearmente cioè che, per k sufficientemente grande, l'errore del passo $k+1$ si comporta come l'errore al passo k moltiplicato per una costante, $\phi'(\alpha)$ nel teorema, indipendente da k ed il cui valore assoluto è minore di 1. Per questo motivo la costante viene chiamata **fattore di convergenza** e la convergenza sarà tanto più rapida quanto più piccola è tale costante.

Definizione 4: quando il metodo di punto fisso è convergente

Si suppongano valide le ipotesi del teorema di Ostrowski 1. Se, inoltre, ϕ è derivabile con continuità due volte e se:

$$\phi'(\alpha) = 0 \qquad \phi''(\alpha) \neq 0$$

Allora il metodo di punto fisso (eq. 13) è convergente di ordine 2 e si ha:

$$\lim_{k\to\infty} \frac{x^{(k+1)} - \alpha}{\left(x^{(k)} - \alpha\right)^2} = \frac{1}{2}\phi''(\alpha) \quad (16)$$

Un'ultima osservazione interessante:

- Nel caso in cui $|\phi'(\alpha)| > 1$, se $x^{(k)}$ è sufficientemente vicino ad α, in modo tale che $\left|\phi'\left(x^{(k)}\right)\right| > 1$, allora $\left|\alpha - x^{(k+1)}\right| > \left|\alpha - x^{(k)}\right|$, e **non è possibile che la successione converga al punto fisso**.
- Nel caso in cui $|\phi'(\alpha)| = 1$ **non si può trarre alcuna conclusione** perché potrebbero verificarsi sia la convergenza sia la divergenza, a seconda delle caratteristiche della funzione di punto fisso.

2 Metodi risolutivi per sistemi lineari e non lineari

2.1 Metodi diretti per sistemi lineari

2.1.1 Metodo delle sostituzioni in avanti e all'indietro

Perché sono importanti i metodi numerici?

Si consideri il seguente **sistema lineare**:

$$Ax = b$$

Dove:

- $A \in \mathbb{R}^{n \times n}$ di componenti a_{ij} e $b \in \mathbb{R}^n$ sono valori noti.
- $x \in \mathbb{R}^n$ è il vettore delle incognite.
- La costante n rappresenta il numero di equazioni lineari delle incognite x_j.

Con queste caratteristiche, è possibile rappresentare la i-esima equazione nel seguente modo:

$$\sum_{j=1}^{n} a_{ij}x_j = b_i \quad \rightarrow \quad a_{1i}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n = b_i \qquad \forall i = 1, \ldots, n$$

La **soluzione esatta** del sistema, chiamata **formula di Cramer**, è:

$$x_j = \frac{\det(A_j)}{\det(A)} \tag{17}$$

Con $A_j = |a_1 \ \ldots \ a_{j-1} \ b \ a_{j+1} \ \ldots \ a_n|$ e a_i le colonne di A. Ovviamente la soluzione **esiste ed è unica se il determinante** della matrice A è **diverso da zero**:

$$\det(A) \neq 0$$

Purtroppo questo metodo è **inutilizzabile** poiché il calcolo di un determinante richiede all'incirca $n!$ (fattoriale di n) operazioni.

Risulta evidente che sia necessario uno studio approfondito di **metodi numerici che si traducano in algoritmi efficienti** da farli eseguire su calcolatori. Nelle seguenti pagine si introducono i primi due algoritmi "efficienti".

Definizione 1

Il seguente algoritmo rappresenta il **metodo delle sostituzioni in avanti**. Dati:

- $L \in \mathbb{R}^{n\times n}$ matrice triangolare inferiore non singolare (cioè con determinante diverso da zero $\det(L) \neq 0$)
- $\mathbf{b} \in \mathbb{R}^n$ vettore termine noto

La soluzione è data da $Lx = \mathbf{b}$ con $x \in \mathbb{R}^n$. Più in generale si ha:

$$x_i = \frac{b_i - \sum_{j=1}^{i=1} L_{ij}\, x_j}{L_{ii}} \tag{18}$$

Per comprendere meglio la definizione del metodo di sostituzioni in avanti, è possibile visualizzare in modo generale la matrice triangolare inferiore L (non singolare):

$$L_{n\times n} = \begin{bmatrix} L_{11} & 0 & 0 & 0 & 0 \\ & \ddots & 0 & 0 & 0 \\ & & \ddots & 0 & 0 \\ & L_{ij} & & \ddots & 0 \\ & & & & L_{nn} \end{bmatrix}$$

Dove ovviamente n è la grandezza della matrice quadrata. Dalla matrice, è possibile rappresentare le prime tre iterazioni, ovvero x_1, x_2 e x_3:

- La prima riga è:
$$x_1 = \frac{b_1}{L_{11}}$$
Si può scrivere anche in linea nel seguente modo: $L_{11}x_1 = b_1$.
- La seconda riga:
$$x_2 = \frac{b_2 - (L_{21}\cdot x_1 + L_{22}\cdot \cancelto{0}{x_2})}{L_{22}}$$
In cui x_1 è il risultato del punto precedente e x_2 è il risultato che attualmente si sta calcolando, quindi uguale a zero.
- La terza riga:
$$x_3 = \frac{b_3 - (L_{31}\cdot x_1 + L_{32}\cdot x_2 + L_{33}\cdot \cancelto{0}{x_3})}{L_{33}}$$

Il **numero di operazioni** richieste dal metodo delle sostituzioni in avanti è dato da 1 sottrazione, $i-1$ moltiplicazioni, $i-2$ addizioni e 1 divisione:

$$\#op. = \sum_{i=1}^{n} (i-1) + (i-2) + 1 + 1 = \sum_{i=1}^{n} (2i-1) = n^2 \tag{19}$$

Per completezza si presenta anche il metodo delle sostituzioni all'indietro.

Definizione 2

Il seguente algoritmo rappresenta il **metodo delle sostituzioni all'indietro**. Dati:

- $U \in \mathbb{R}^{n\times n}$ matrice triangolare superiore non singolare (cioè con determinante diverso da zero $\det(U) \neq 0$)
- $\mathbf{b} \in \mathbb{R}^n$ vettore termine noto

La soluzione è data da $Ux = \mathbf{b}$ con $x \in \mathbb{R}^n$. Più in generale si ha:

$$x_i = \frac{b_i - \sum_{j=i+1}^{n} U_{ij}\, x_j}{U_{ii}} \tag{20}$$

Come per il metodo precedente, anche in questo caso è utile visualizzare la matrice generale:

$$U_{n\times n} = \begin{bmatrix} U_{11} & & & & \\ 0 & \ddots & & U_{ij} & \\ 0 & 0 & \ddots & & \\ 0 & 0 & 0 & \ddots & \\ 0 & 0 & 0 & 0 & U_{nn} \end{bmatrix}$$

Anche da questa matrice è possibile rappresentare le iterazioni:

- La prima riga è:
$$x_1 = \frac{b_1 - (U_{12}\cdot x_2 + U_{13}\cdot x_3 + U_{14}\cdot x_4)}{U_{11}}$$
- La seconda riga è:
$$x_2 = \frac{b_2 - (U_{23}\cdot x_3 + U_{24}\cdot x_4)}{U_{22}}$$
- La terza riga è:
$$x_3 = \frac{b_3 - (U_{34}\cdot x_4)}{U_{33}}$$
- L'ultima riga è:
$$x_4 = \frac{b_4}{U_{44}}$$

Si deduce ovviamente che l'ultima riga può essere generalizzata nel seguente modo:

$$x_n = \frac{b_n}{U_{nn}}$$

Infine, il **numero di operazioni** è il medesimo del metodo delle sostituzioni in avanti, quindi n^2.

2.1.2 Fattorizzazione LRU: MEG e Cholesky

Sia $A \in \mathbb{R}^{n\times n}$. Si supponga che esistano due opportune matrici L ed U, triangolare inferiore e superiore, rispettivamente, tali che:

$$A = LU \tag{21}$$

L'equazione viene chiamata **fattorizzazione LU** (o **decomposizione LU**) di A.

La fattorizzazione LU è stata introdotta poiché se A **non è singolare** (quindi il determinante è diverso da zero) tali matrici devono essere **anch'esse non singolari**; questo assicura che i loro **elementi diagonali siano non nulli**. Da questa osservazione, si ottiene un risultato interessante perché la risoluzione di $Ax = b$ è *equivalente* alla risoluzione dei due seguenti sistemi triangolari:

$$Ly = b \qquad Ux = y \tag{22}$$

Dove y rappresenta la soluzione dell'equazione 18 a pagina 19, ovvero la risoluzione del metodo delle sostituzioni in avanti. Analogamente, la x rappresenta la soluzione dell'equazione 20 a pagina 20, ovvero la risoluzione del metodo delle sostituzioni all'indietro.

Chiaro, ma che algoritmi esistono per calcolare la fattorizzazione LU?

Esistono principalmente due algoritmi: il **Metodo di Eliminazione di Gauss (MEG)** e la **Fattorizzazione di Cholesky**.

- Senza entrare troppo nel dettaglio, la fattorizzazione LU viene chiamata anche fattorizzazione di Gauss poiché è dimostrato che è possibile applicare l'algoritmo di Gauss, ovvero il **Metodo di Eliminazione di Gauss (MEG)**.

  Il **MEG** è possibile **applicarlo** per alcuni tipi di matrici:

  1. Le **matrici a dominanza diagonale stretta**. Una matrice è detta **matrice a dominanza diagonale per righe** se:

  $$|a_{ii}| \geq \sum_{j=1 \wedge j\neq i}^{n} |a_{ij}|\,, \qquad i = 1, \ldots, n \tag{23}$$

  Una matrice è detta **matrice a dominanza diagonale per colonne** se:

  $$|a_{ii}| \geq \sum_{j=1 \wedge j\neq i}^{n} |a_{ji}|\,, \qquad i = 1, \ldots, n \tag{24}$$

  Quando nelle precedenti disuguaglianze è possibile sostituire il segno $\geq$ con quello $>$ si può dire che la matrice A è a dominanza diagonale **stretta**.

  2. Le **matrici reali simmetriche**[2]**e definite positive**[3].

[2] Una matrice è simmetrica se coincide con la sua matrice trasposta.

[3] Una matrice viene **definita positiva** se:

$$\forall \mathbf{x} \in \mathbb{R}^n \qquad \text{con } \mathbf{x} \neq \mathbf{0}, \qquad \mathbf{x}^T \mathrm{A} \mathbf{x} > 0$$

Il **calcolo dei coefficienti** dei fattori L ed U **richiede** circa $\frac{2n^3}{3}$ **operazioni**.

- Se la matrice A, cioè la matrice usata nella definizione, è **simmetrica** e **definita positiva**, è possibile trovare la **fattorizzazione di Cholesky**:

$$A = R^T R \tag{25}$$

Dove R è una matrice triangolare superiore con elementi positivi sulla diagonale. Inoltre, tale fattorizzazione è **unica**.

Il **calcolo della matrice** R **richiede** circa $\frac{n^3}{3}$ **operazioni** (cioè la **metà di quelle richieste per calcolare le due matrici della fattorizzazione LU**).

2.1.3 La tecnica del pivoting

Si introduce un metodo che consenta di portare a compimento il processo di fattorizzazione LU per una qualunque matrice A non simmetrica ($\det(A) \neq 0$).

La tecnica si basa sulla **permutazione** (cioè sullo scambio) opportuno di righe della matrice di partenza A. Purtroppo non è noto a priori quali siano le righe che dovranno essere tra loro scambiate; tuttavia questa decisione può essere presa ad ogni passo durante il quale si generino elementi nulli.

Dato che lo scambio tra righe comporta un cambiamento del pivot, questa tecnica viene chiamata **pivoting per righe**. La fattorizzazione che si trova restituisce la matrice A di partenza a meno di una permutazione fra le righe. Precisamente:

$$PA = LU \tag{26}$$

Dove P è un'opportuna **matrice di permutazione**. Ovvero, è una matrice uguale all'identità all'inizio del processo di fattorizzazione e se durante l'applicazione le righe di A vengono scambiate, allora deve essere fatto uno scambio analogo sulle righe di P. Per cui, alla fine sarà necessario risolvere i seguenti sistemi triangolari:

$$L\mathbf{y} = P\mathbf{b} \qquad U\mathbf{x} = \mathbf{y} \tag{27}$$

Dove y rappresenta la soluzione dell'equazione 18 a pagina 19, ovvero la risoluzione del metodo delle sostituzioni in avanti. Analogamente, la x rappresenta la soluzione dell'equazione 20 a pagina 20, ovvero la risoluzione del metodo delle sostituzioni all'indietro.

2.1.4 Errori di arrotondamento nel MEG

Prima di introdurre l'errore di arrotondamento nel Metodo di Eliminazione di Gauss, è necessario capire il problema alla fonte.

In generale, un elaboratore memorizza i numeri nel seguente modo:

$$x = (-1)^s \cdot (0.a_1 a_2 \dots a_t) \cdot \beta^e = (-1)^s \cdot m \cdot \beta^{e-t} \qquad \text{con } a_1 \neq 0 \tag{28}$$

Dove:

- s è il **segno** e può essere uguale a 0 o 1.
- β è la **base** e può essere un numero intero positivo maggiore od uguale a 2.
- m è la **mantissa**, un intero la cui **lunghezza** t è il **numero massimo di cifre** a_i (con $0 \leq a_i \leq \beta - 1$) **memorizzabili**.
- e è l'**esponente** ed è un intero.

I numeri con un formato identico all'equazione 28 sono detti numeri **floating-point normalizzati** essendo variabile la posizione del punto decimale. Inoltre, le cifre $a_1 a_2 \dots a_p$ (con $p \leq t$) vengono generalmente chiamate le **prime p cifre significative di x**.

Si noti che la condizione $a_1 \neq 0$ nell'equazione 28 impedisce che lo stesso numero possa avere più rappresentazioni (per esempio $0.1 \cdot 10^0$ uguale a $0.01 \cdot 10^1$).

L'insieme $\mathbb{F}$ è dunque l'**insieme dei numeri *floating point*** ed è completamente caratterizzato dalla base β, dal numero di cifre significative t e dall'intervallo (L, U) (con $L < 0$ ed $U > 0$) di variabilità dell'esponente e.

Sostituendo ad un numero reale $x \neq 0$ il suo rappresentante *floating point* $fl(x)$ in $\mathbb{F}$, è inevitabile un **errore di arrotondamento** uguale a:

$$\frac{|x - fl(x)|}{|x|} \leq \frac{1}{2}\epsilon_M \tag{29}$$

Dove:

- $\epsilon_M = \beta^{1-t}$ è la **epsilon macchina**, ovvero la distanza fra 1 ed il più piccolo numero *floating-point* maggiore di 1.
- $|x|$ è l'**errore relativo**.
- $|x - fl(x)|$ è l'**errore assoluto**.
- Il numero $u = \frac{1}{2}\epsilon_M$ è l'**unità di arrotondamento** poiché rappresenta il **massimo errore relativo che la macchina può commettere nella rappresentazione di un numero reale**.

Date queste osservazioni, si possono ricavare anche il **più grande** ed il **più piccolo numero positivo** di $\mathbb{F}$:

$$x_{\min} = \beta^{L-1} \qquad x_{\max} = \beta^U \cdot \left(1 - \beta^{-t}\right)$$

- Se un numero è minore del numero più piccolo positivo, allora si ha una situazione di **underflow**.
- Se un numero è maggiore del numero più grande positivo, allora si ha una situazione di **overflow**.

Quindi che cosa accade in una somma tra valori molto grandi?

Ottima domanda. Quando si sommano tra loro numeri che hanno all'incirca lo stesso module, ma segno opposto, il risultato della somma può essere assai impreciso e ci si riferisce a questa situazione con l'espressione **cancellazione di cifre significative**.

Risulta quindi necessario fare una distinzione. L'aritmetica (o la logica) utilizzata dal calcolatore viene chiamata **aritmetica floating-point** (quella spiegata fin ora); al contrario, l'**aritmetica esatta** si basa sulla effettuazione esatta delle operazioni elementari (quindi senza tener conto degli errori di arrotondamento) su operandi noti esattamente (e non attraverso la loro rappresentazione *floating-point*).

Va bene, ma perché è importante considerare questo aspetto?

Nonostante gli errori di arrotondamento sono generalmente piccoli, se ripetuti all'interno di algoritmi lunghi e complessi, possono portare ad effetti catastrofici (vedi per esempio l'incidente del razzo Ariane 5).

Adesso si prova a definire in modo formale questo comportamento.

- Con e_a si identificano i **tipi di errore che si manifestano a seguito di una serie di errori di arrotondamento**.
- Con e_t si identifica l'**errore di troncamento**. Tali errori sono assenti soltanto in quei modelli matematici che sono già di dimensione finita (per esempio nella risoluzione di un sistema lineare).
- Con e_c si identifica l'**errore computazionale**, ovvero l'insieme degli errori e_a e e_t.

Indicando con x la soluzione esatta del modello matematico e con $\widehat{x}$ la soluzione ottenuta al termine del processo numerico, allora l'**errore computazionale assoluto** sarà dunque:

$$e_c^{ass} = |x - \widehat{x}| \tag{30}$$

Nel caso in cui l'errore relativo fosse diverso da zero ($x \neq 0$):

$$e_c^{rel} = \frac{|x - \widehat{x}|}{|x|} \tag{31}$$

Relazione tra gli errori di arrotondamento e MEG

L'introduzione fatta nelle pagine precedenti è servita perché è possibile trovare errori di arrotondamento con il prodotto LU, la quale non ritorna esattamente la matrice A.

Come appena accennato, il **prodotto LU produce un errore di arrotondamento**. Esso può essere **ridotto usando la tecnica del pivoting** che consente di contenerlo.

Inoltre, quando si risolve numericamente il sistema lineare $\mathrm{A}\mathbf{x} = \mathbf{b}$ si può trovare la **soluzione esatta** $\widehat{\mathbf{x}}$ di un **sistema perturbato** della forma:

$$(\mathrm{A} + \delta\mathrm{A})\,\widehat{\mathbf{x}} = \mathbf{b} + \delta\mathbf{b} \tag{32}$$

Dove $\delta\mathrm{A}$ e $\delta\mathbf{b}$ sono rispettivamente una matrice ed un vettore di perturbazione che dipendono dallo specifico metodo numerico impiegato nella risoluzione del sistema.

Usando le norme si ha:

$$\frac{||\mathbf{x} - \widehat{\mathbf{x}}||}{||\mathbf{x}||} \leq \frac{\lambda_{\max}}{\lambda_{\min}} \cdot \frac{||\delta\mathbf{b}||}{||\mathbf{b}||} \tag{33}$$

Dove l'errore relativo sulla soluzione dipende dall'errore relativo sui dati attraverso la seguente constante (≥ 1):

$$K\,(A) = \frac{\lambda_{\max}}{\lambda_{\min}} \tag{34}$$

Essa viene chiamata **numero di condizionamento (spettrale) della matrice** A. Ovviamente, si ricorda che tale matrice A deve essere simmetrica e definita positiva.
Se la matrice A è una matrice simmetrica e definita positiva e δA una matrice non nulla simmetrica e definita positiva tale che $\lambda_{\max}\,(\delta A) < \delta_{\min}\,(A)$, allora vale:

$$\frac{||\mathbf{x} - \widehat{\mathbf{x}}||}{||\mathbf{x}||} \leq \frac{K\,(A)}{1 - \dfrac{\lambda_{\max}\,(\delta A)}{\lambda_{\min}\,(A)}} \cdot \left(\frac{\lambda_{\max}\,(\delta A)}{\lambda_{\max}\,(A)} + \frac{||\delta\mathbf{b}||}{||\mathbf{b}||} \right) \tag{35}$$

Infine, se le matrici A e δA non sono simmetriche e definite positive e δA è tale che $||\delta\mathrm{A}||_2 \cdot \left|\left|\mathrm{A}^{-1}\right|\right|_2 < 1$, vale la seguente stima:

$$\frac{||\mathbf{x} - \widehat{\mathbf{x}}||}{||\mathbf{x}||} \leq \frac{K_2\,(A)}{1 - \dfrac{K_2\,(A) \cdot ||\delta\mathrm{A}||_2}{||\mathrm{A}||_2}} \cdot \left(\frac{||\delta\mathrm{A}||_2}{||\mathrm{A}||_2} + \frac{||\delta\mathbf{b}||}{||\mathbf{b}||} \right) \tag{36}$$

Essendo $||\delta\mathrm{A}||_2 = \sqrt{\lambda_{\max}\,(A^T A)}$ e:

$$K_2\,(A) = ||\delta\mathrm{A}||_2 \cdot \left|\left|\delta\mathrm{A}^{-1}\right|\right|_2 \tag{37}$$

Il **numero di condizionamento in norma 2**.

- Se $K_2(\mathrm{A})$ (o $K(\mathrm{A})$) è "piccolo" la matrice A viene detta **ben condizionata** ed a **piccoli errori sui dati corrisponderanno errori dello stesso ordine di grandezza sulla soluzione**.
- Se $K_2(\mathrm{A})$ (o $K(\mathrm{A})$) è "grande" la matrice A viene detta **mal condizionata** ed *potrebbe accadere* che a **piccole perturbazioni sui dati corrispondano grandi errori sulla soluzione**.

Infine, volendo l'equazione 33 può essere riscritta introducendo il **residuo r**:

$$\mathbf{r} = \mathbf{b} - \mathrm{A}\widehat{\mathbf{x}} \tag{38}$$

Chiaramente se $\widehat{\mathbf{x}}$ fosse la **soluzione esatta**, il **residuo sarebbe il vettore nullo**.

L'**efficacia** del residuo dipende dalla grande del numero di condizionamento di A. Infatti, sempre dall'equazione 33 si ricava:

$$\frac{||\mathbf{x} - \widehat{\mathbf{x}}||}{||\mathbf{x}||} \leq K_2(\mathrm{A}) \cdot \frac{||\mathbf{r}||}{||\mathbf{b}||} \tag{39}$$

- Se $K_2(\mathrm{A})$ è "piccolo" si avrebbe la **certezza che l'errore sarà piccolo quando lo è il residuo**.
- Se $K_2(\mathrm{A})$ è "grande" non si può avere la certezza che l'errore sarà piccolo quando lo è il residuo.

2.1.5 Il pivoting totale

Si opera un **pivoting totale** quando la ricerca del pivot è estesa alla sottomatrice $A^{(k)}$ costituita dagli elementi $a_{ij}^{(k)}$ con $i, j = k, \ldots, n$. A differenza della tecnica del pivoting introdotta nel capitolo 2.1.3 a pagina 22, il parziale prevede il **coinvolgimento delle righe e delle colonne**, e conduce alla costruzione di due matrici di permutazione, chiamate P e Q, una sulle righe, l'altra sulle colonne:

$$\mathrm{PAQ} = \mathrm{LU} \tag{40}$$

Inoltre, la **soluzione** del sistema $A\mathbf{x} = \mathbf{b}$ è ottenuta attraverso la **risoluzione di due sistemi triangolari e di una permutazione**:

$$\mathrm{L}\mathbf{y} = \mathrm{P}\mathbf{b} \qquad \mathrm{U}\mathbf{x}^* = \mathbf{y} \qquad \mathbf{x} = \mathrm{Q}\mathbf{x}^* \tag{41}$$

Dal punto di vista **computazionale**, la tecnica del pivoting totale ha un **costo superiore rispetto a quello parziale** (capitolo 2.1.3 a pagina 22) in quanto ad ogni passo della fattorizzazione devono essere svolti molti più confronti. Tuttavia, può **apportare dei vantaggi in termini di risparmio di memoria e di stabilità**.

Quindi riassumendo:

Come funziona?

Può essere visto come un pivoting parziale, ma in cui c'è un coinvolgimento anche delle colonne e non solo delle righe.

Svantaggi

Più costoso rispetto al pivoting parziale poiché opera su più fronti e quindi devono essere eseguiti più confronti.

Vantaggi

Non sempre, ma spesso si possono ottenere risparmi di memoria e un'alta stabilità.

2.1.6 Il *fill-in* di una matrice

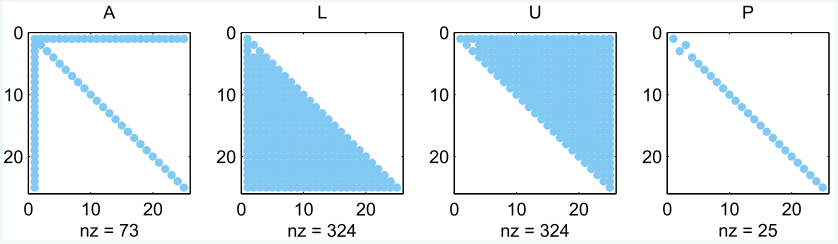
Un altro problema che è possibile riscontrare durante la fattorizzazione LU è il **fill-in**.

Come si manifesta il fenomeno del *fill-in*?

Durante la fattorizzazione LU, non è certo che le matrici L ed U ottenute mantengano la struttura del corrispondente triangolo della matrice A iniziale. Al contrario, il **processo di fattorizzazione tende** generalmente a **riempire le matrici** L ed U. Tale fenomeno **dipende fortemente dalla struttura e dal valore dei singoli elementi non nulli** della matrice A.

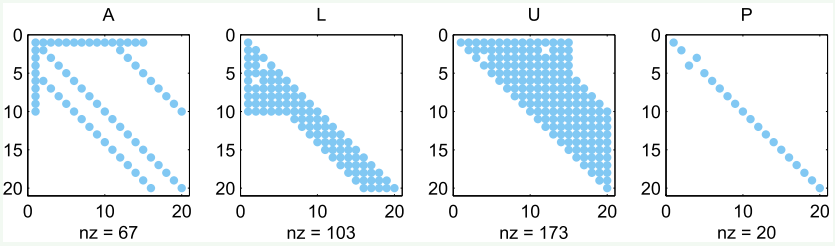
Esempio 1

Qua di seguito è possibile osservare un fenomeno di *fill-in* su una generica matrice.



Esempio 2

Un altro esempio di *fill-in* su una generica matrice. In questo caso, gli elementi non nulli della prima riga e della prima colonna di A inducono un riempimento totale delle corrispondenti colonne in U e righe in L, rispettivamente, mentre gli elementi non nulli sopra e sotto le diagonali di A comportano un riempimento delle diagonali superiori di U ed inferiori di L comprese tra quella principale e quelle non nulle di A.



Come risolvere il *fill-in*?

Per ovviare al problema del *fill-in*, si possono adottare **tecniche di riordinamento che permutano righe e colonne della matrice prima di realizzare la fattorizzazione**. Tuttavia, in alcuni casi la **sola applicazione della tecnica di pivoting totale** (paragrafo 2.1.5 pagina 27) consente di raggiungere lo stesso obiettivo.

Prima di introdurre un nuovo esempio, si forniscono due definizioni.

Definizione 3

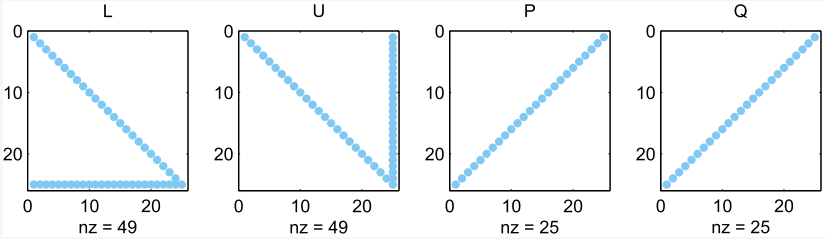
Una matrice quadrata di dimensione n è detta **matrice sparsa** se ha un numero di elementi non nulli dell'ordine di n (e non di n^2).

Definizione 4

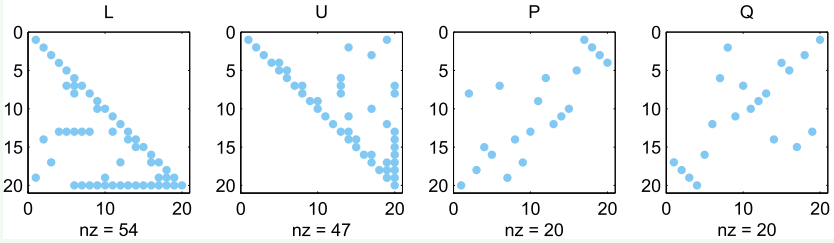
Si chiama **pattern** di una matrice sparsa l'insieme dei suoi elementi non nulli.

Esempio 3

Nella seguente figura sono mostrati i pattern delle matrici ottenute dalla fattorizzazione con pivotazione totale. Come si può vedere, il fenomeno di *fill-in* è contenuto.



Un altro esempio.



2.2 Metodi iterativi per sistemi lineari

Un **metodo iterativo** per la risoluzione del sistema lineare $Ax = b$ con:

- $A \in \mathbb{R}^{n \times n}$
- $b \in \mathbb{R}^n$
- $x \in \mathbb{R}^n$
- $\det(A) \neq 0$

Consiste nel costruire una successione di vettori del tipo:

$$\left\{\mathbf{x}^{(k)},\ k \geq 0\right\}$$

Di $\mathbb{R}^n$ che **converge** alla soluzione esatta $\mathbf{x}$, ossia tale che:

$$\lim_{k \to \infty} \mathbf{x}^{(k)} = \mathbf{x} \tag{42}$$

Per un qualunque vettore iniziale $\mathbf{x}^{(0)} \in \mathbb{R}$, ossia la **convergenza non deve dipendere dalla scelta di** $\mathbf{x}^{(0)}$.

Inoltre, si definisce ricorsivamente:

$$\mathbf{x}^{(k+1)} = \mathrm{B}\mathbf{x}^{(k)} + \mathbf{g} \qquad k \geq 0 \tag{43}$$

Essendo:

- B una matrice opportuna (dipendente da A)
- $\mathbf{g}$ un vettore opportuno (dipendente da A e da $\mathbf{b}$)

E come si possono scegliere questi valori?

La scelta della matrice B e del vettore $\mathbf{g}$ è piuttosto semplice. Devono essere **rispettate due proprietà**:

- **Condizione di consistenza**. I valori devono garantire che:

  $$\mathbf{x} = \mathrm{B}\mathbf{x} + \mathbf{g} \tag{44}$$

  Essendo $\mathbf{x} = \mathrm{A}^{-1}\mathbf{b}$, necessariamente dovrà aversi $\mathbf{g} = (\mathrm{I} - \mathrm{B})\,\mathrm{A}^{-1}\mathbf{b}$.

- **Condizione di convergenza**. Prima di dare la condizione, è necessario comprendere alcune cose. Prima di tutto, la seguente espressione:

  $$\mathbf{e}^{(k)} = \mathbf{x} - \mathbf{x}^{(k)}$$

  Viene identificata come l'**errore al passo** k. Adesso sottraendo l'equazione 43 dalla condizione di consistenza 44 si ottiene:

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

  Per tale ragione B viene detta **matrice di iterazione** del metodo 43.

  E adesso, si presenta la condizione di convergenza:

– La matrice B è simmetrica[4] e definita positiva[5], allora:

$$\left|\left|\mathbf{e}^{(k+1)}\right|\right| = \left|\left|\mathrm{B}\mathbf{e}^{(k)}\right|\right| \leq \rho\left(\mathrm{B}\right)\left|\left|\mathbf{e}^{(k)}\right|\right| \qquad \forall k \geq 0 \tag{45}$$

Dove $\rho\left(\mathrm{B}\right)$ è il **raggio spettrale** di B ed è il massimo degli autovalori di B. Si tenga conto che nel caso di matrici simmetriche e definite positive esso coincide con il massimo autovalore.

– Iterando a ritroso il punto precedente, si ottiene:

$$\left|\left|\mathbf{e}^{(k)}\right|\right| \leq \left[\rho\left(\mathrm{B}\right)\right]^k \left|\left|\mathbf{e}^{(0)}\right|\right| \qquad k \geq 0 \tag{46}$$

Si noti che se $\rho\left(\mathrm{B}\right) < 1$, allora $\mathbf{e}^{(k)} \to \mathbf{0}$ per $k \to \infty$ per ogni possibile $\mathbf{e}^0$ (e, conseguentemente, per ogni $\mathbf{x}^{(0)}$).

In generale, la condizione sufficiente e necessaria di convergenza è la seguente.

Definizione 5: condizione sufficiente e necessaria di convergenza

Un metodo iterativo della forma dell'equazione 43, la cui matrice di iterazione B soddisfi la condizione di consistenza (eq. 44), è **convergente** per ogni $\mathbf{x}^{(0)}$ *se e soltanto se* $\rho\left(\mathrm{B}\right) < 1$ (raggio spettrale minore di 1).

Inoltre, minore è $\rho\left(\mathrm{B}\right)$, minore è il numero di iterazioni necessarie per ridurre l'errore iniziale di un dato fattore.

Perché introdurre i metodi iterativi?

L'esigenza di introdurre i metodi iterativi sorge nel momento in cui si ragiona sulla quantità di tempo spesa da un calcolatore per eseguire la fattorizzazione LU su matrici di grandi dimensioni. Difatti, con matrici con ordini di 10^7, sono necessari circa 11 giorni.

[4] Quindi corrisponde con la sua trasposta
[5] Definizione a pagina: 21

2.2.1 Il metodo di Jacobi

Se i coefficienti diagonali della matrice A non sono nulli, allora:

$$D = \text{diag}\,(a_{11}, a_{22}, \ldots, a_{nn})$$

Ovvero D è la matrice diagonale costruita a partire dagli elementi diagonali di A.

Il **metodo di Jacobi** corrisponde alla forma:

$$\mathrm{D}\mathbf{x}^{(k+1)} = \mathbf{b} - (\mathrm{A} - \mathrm{D})\,\mathbf{x}^{(k)} \qquad k \geq 0$$

Che per componenti assume la forma:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1, j\neq i}^{n} a_{ij} x_j^{(k)} \right) \qquad i = 1, \ldots, n \tag{47}$$

Dove $k \geq 0$ e $\mathbf{x}^{(0)} = \left(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}\right)^T$ è il vettore iniziale.

Quando converge?

In due casi:

1. Se la matrice $\mathrm{A} \in \mathbb{R}^{n\times n}$ è a dominanza diagonale stressa per righe[6], allora il metodo di Jacobi converge.
2. Con la seguente definizione (si veda il paragrafo 2.2.2 per comprendere meglio le definizioni):

Definizione 6: convergenza di Jacobi e Gauss-Seidel

Se $A \in \mathbb{R}^{n\times n}$ è una **matrice tridiagonale**, quindi una matrice quadrata che al di fuori della diagonale principale e delle linee immediatamente al di sopra e al di sotto di essa (la prima sovradiagonale e la prima sottodiagonale), ha solo valori nulli. Un esempio

[6] In algebra lineare una **matrice a dominanza diagonale stretta per righe** o **matrice diagonale dominante stretta per righe** è una matrice quadrata $A \in \mathbb{C}^{n\times n}$ di ordine n i cui elementi diagonali sono maggiori (non stretta sarebbe maggiori o uguali) in valore assoluto della somma di tutti i restanti elementi della stessa riga in valore assoluto:

$$|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}| \tag{48}$$

In senso non stretto:

$$|a_{ii}| \geq \sum_{j=1, j\neq i}^{n} |a_{ij}| \tag{49}$$

di amtrice tridiagonale:

$$\begin{bmatrix} 1 & 4 & 0 & 0 \\ 3 & 4 & 1 & 0 \\ 0 & 2 & 3 & 4 \\ 0 & 0 & 1 & 3 \end{bmatrix}$$

Se tale matrice è non singolare, quindi $\det(A) \neq 0$, con valori della diagonale diversi da zero, allora i metodi di Jacobi e di Gauss-Seidel sono **entrambi convergenti o entrambi divergenti**.

- Se convergono, la velocità dei metodi:

  $$\text{Gauss-Seidel} \gg \text{Jacobi}$$

  Precisamente il raggio spettrale della matrice di iterazione del metodo di Gauss-Seidel è il quadrato del raggio spettrale di quella del metodo di Jacobi.

HPC curiosity: è interessante notare che il metodo di Jacobi viene facilmente parallelizzato per aumentare le prestazioni di calcolo.

2.2.2 Il metodo di Gauss-Seidel

Con il metodo di Jacobi, ogni componente $x_i^{(k+1)}$ del nuovo vettore $\mathbf{x}^{(k+1)}$ viene calcolata indipendentemente dalle altre. Questa può essere la base di partenza per costruire un nuovo metodo più ottimizzato, poiché se per il calcolo di $x_i^{(k+1)}$ venissero usate le nuove componenti già disponibili $x_j^{(k+1)}$, $j = 1, \ldots, i-1$, assieme con le vecchie $x_j^{(k)}$, $j \geq i$.

Supponendo che gli elementi della diagonale non siano nulli, per $k \geq 0$, il **metodo di Gauss-Seidel**:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right) \qquad i = 1, \ldots, n \tag{50}$$

HPC curiosity: a differenza di Jacobi, questo metodo non può essere parallelizzato, ma è necessario operare in modo sequenziale.

Quando converge?

In tre casi:

1. Se la matrice $A \in \mathbb{R}^{n \times n}$ è a dominanza diagonale stressa per righe (vedi pag. 32), allora il metodo di Gauss-Seidel converge.
2. Se la matrice A è simmetrica (uguale alla sua trasposta) e definita positiva (vedi pag. 21), allora Gauss-Seidel converge.
3. Con la definizione di convergenza data per il metodo di Jacobi a pagina 32

Da notare: **non esistono risultati generali che consentono di affermare che il metodo di Gauss-Seidel converga sempre più rapidamente di quello di Jacobi**, a parte il caso della definizione a pagina 21.

2.2.3 Il metodo di Richardson

Una tecnica generale per costruire un metodo iterativo è basata sulla seguente **decomposizione additiva** (o **splitting**) della matrice A:

$$A = P - (P - A)$$

In cui P è un'opportuna matrice non singolare ($\det(P) \neq 0$) chiamata **precondizionatore** di A. Di conseguenza:

$$P\mathbf{x} = (P - A)\,\mathbf{x} + \mathbf{b}$$

È un sistema purché si ponga:

$$\begin{aligned} B &= P^{-1}(P - A) = I - P^{-1}A \\ \mathbf{g} &= P^{-1}\mathbf{b} \end{aligned}$$

Questa identità suggerisce la definizione del seguente metodo iterativo:

$$P\left(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right) = \mathbf{r}^{(k)} \qquad k \geq 0$$

In cui:

$$\mathbf{r}^{(k)} = \mathbf{b} - A\mathbf{x}^{(k)} \tag{51}$$

Denota il vettore residuo alla k-esima iterazione. Una generalizzazione di questo metodo iterativo è

$$P\left(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right) = \alpha_k \mathbf{r}^{(k)} \qquad k \geq 0 \tag{52}$$

Dove $\alpha_k \neq 0$ è un parametro che può cambiare ad ogni iterazione e che, a priori, servirà a migliorare le proprietà di convergenza della successione $\left\{\mathbf{x}^{(k)}\right\}$.

L'equazione 52 è nota come **metodo di Richardson** o **metodo di Richardson dinamico**; se $\alpha_k = \alpha$ per ogni $k \geq 0$ esso si dice **metodo di Richardson stazionario**.

Questo metodo richiede ad ogni passo di trovare il cosiddetto **residuo precondizionato $\mathbf{z}^{(k)}$** dato dalla soluzione del sistema lineare:

$$P\mathbf{z}^{(k)} = \mathbf{r}^{(k)} \tag{53}$$

Quindi, la nuova iterata è definita da $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{z}^{(k)}$.

Per questa ragione la matrice P deve essere scelta in modo tale che il costo computazionale richiesto dalla risoluzione di 53 sia modesto (ogni matrice P diagonale, tridiagonale o triangolare andrebbe bene a questo scopo).

Quando converge?

Per studiare la convergenza, si definisce la **norma dell'energia** associata alla matrice A:

$$||\mathbf{v}||_A = \sqrt{\mathbf{v}^T A \mathbf{v}} \qquad \forall \mathbf{v} \in \mathbb{R}^n \tag{54}$$

E si definisce la seguente definizione.

Definizione 7

Sia $A \in \mathbb{R}^{n \times n}$.
Per ogni matrice non singolare ($\det \neq 0$) $P \in \mathbb{R}^{n \times n}$ il **metodo di Richardson stazionario converge** se e solo se:

$$|\lambda_i|^2 < \frac{2}{\alpha} \mathrm{Re}\lambda_i \qquad \forall i = 1, \dots, n$$

In cui λ_i sono gli autovalori della matrice risultato di $P^{-1}A$.

In particolare, se gli autovalori della matrice risultato di $P^{-1}A$ sono reali, allora esso converge se e solo se:

$$0 < \alpha\lambda_i < 2 \qquad \forall i = 1, \dots, n$$

Se A e P sono entrambe simmetriche (quindi uguali alle loro rispettive trasposte) e definite positive (definizione pagina 21) il metodo di Richardson stazionario **converge per ogni possibile scelta di $\mathbf{x}^{(0)}$** se e solo se:

$$0 < \alpha < \frac{2}{\lambda_{\max}}$$

Dove $\lambda_{\max}$ (che è maggiore di zero) è l'autovalore massimo della matrice risultato di $P^{-1}A$.

E ancora, il raggio spettrale $\rho(B_\alpha)$ della matrice di iterazione $B_\alpha = I - \alpha P^{-1}A$ è il minimo quando $\alpha = \alpha_{\text{opt}}$, dove:

$$\alpha_{\text{opt}} = \frac{2}{\lambda_{\min} + \lambda_{\max}} \tag{55}$$

Dove ovviamente $\lambda_{\min}$ è l'autovalore minimo della matrice risultato di $P^{-1}A$.

Infine, se $\alpha = \alpha_{\text{opt}}$, vale la seguente **stima di convergenza**:

$$\left|\left|\mathbf{e}^{(k)}\right|\right|_A \leq \left(\frac{K\left(P^{-1}A\right) - 1}{K\left(P^{-1}A\right) + 1}\right)^k \left|\left|\mathbf{e}^{(0)}\right|\right|_A \qquad k \geq 0 \tag{56}$$

Si noti che la massima velocità di convergenza è data anche dal raggio spettrale:

$$\rho_{\text{opt}} = \frac{K\left(P^{-1}A\right) - 1}{K\left(P^{-1}A\right) + 1} \tag{57}$$

2.2.4 Il metodo del Gradiente e del Gradiente Coniugato

Si definisce la funzione $\Phi : \mathbb{R}^n \to \mathbb{R}$:

$$\Phi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{x}^T\mathbf{b} \tag{58}$$

Nel caso in cui la matrice A è simmetrica[7] e definita positiva[8], Φ è una funzione convessa, cioè tale che ogni per ogni $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ e per ogni $\alpha \in [0, 1]$ vale:

$$\Phi(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha\Phi(\mathbf{x}) + (1-\alpha)\Phi(\mathbf{y}) \tag{59}$$

Ed ammette un unico punto stazionario $\mathbf{x}^*$ che è anche un **punto di minimo locale ed assoluto**. Quindi:

$$\mathbf{x}^* = \underset{\mathbf{x}\in\mathbb{R}^n}{\operatorname{argmin}}\, \Phi(\mathbf{x}) \tag{60}$$

È l'unica soluzione dell'equazione:

$$\nabla\Phi(\mathbf{x}) = A\mathbf{x} - \mathbf{b} = \mathbf{0} \tag{61}$$

Risolvere il **problema di minimo** dell'equazione 60 **equivale** pertanto a **risolvere il sistema lineare**[9].

Per un generico $\overline{\mathbf{x}} \in \mathbb{R}^n$ diverso da $\mathbf{x}^*$, $\nabla\Phi(\mathbf{x})$ è un vettore non nullo di $\mathbb{R}^n$ che individua la direzione lungo cui avviene:

- La **massima crescita** di Φ;
- Di conseguenza $-\nabla\Phi(\overline{\mathbf{x}})$ individua la direzione di **massima decrescita** di Φ a partire da $\overline{\mathbf{x}}$.

Inoltre, ricordando che il vettore residuo nel punto $\overline{\mathbf{x}}$ è $\overline{\mathbf{r}} = \mathbf{b} - A\overline{\mathbf{x}}$, grazie alla direzione lungo cui avviene la massima crescita 61, si ha che il **vettore residuo** è uguale a:

$$\overline{\mathbf{r}} = -\nabla\Phi(\overline{\mathbf{x}}) \tag{62}$$

Cioè il residuo **individua una possibile direzione lungo cui muoversi per avvicinarsi al punto di minimo $\mathbf{x}^*$**.

In generale, un vettore $\mathbf{d}$ rappresenta una **direzione di discesa per Φ** nel punto $\overline{\mathbf{x}}$ se si verificano le seguenti condizioni:

$$\begin{aligned} \nabla\Phi(\overline{\mathbf{x}}) \neq \mathbf{0} \quad &\Rightarrow \quad \mathbf{d}^T\nabla\Phi(\overline{\mathbf{x}}) < 0 \\ \nabla\Phi(\overline{\mathbf{x}}) = \mathbf{0} \quad &\Rightarrow \quad \mathbf{d} = \mathbf{0} \end{aligned} \tag{63}$$

Il **residuo** è pertanto una **direzione di discesa**. È banale dire che quindi per **avvicinarsi al punto di minimo** ci si **muoverà lungo opportune direzioni di discesa**. I **metodi di discesa** sono definiti nel seguente modo.

[7] Quindi corrisponde con la sua trasposta

[8] Definizione a pagina: 21

[9] Il gradiente è un argomento del calcolo differenziale vettoriale e il vettore gradiente di una funzione scalare punta secondo la direzione di massima crescita della funzione stessa. Se si è un po' arrugginiti a riguardo, si può dare un'occhiata su Wikipedia o su YouMath

Definizione 8: metodi di discesa

Assegnato un vettore $\mathbf{x}^{(0)} \in \mathbb{R}^n$, per $k = 0, 1, \dots$ fino a convergenza:

1. Si determina una direzione di discesa $\mathbf{d}^{(k)} \in \mathbb{R}^n$
2. Si determina un passo $\alpha_k \in \mathbb{R}$
3. Si pone $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)}$

Ma quindi quale è la differenza tra Gradiente e Gradiente Coniugato?

I metodi del gradiente e del gradiente coniugato sono metodi di discesa che si **differenziano nella scelta delle direzioni di discesa**, mentre la scelta dei passi α_k è *comune* ad entrambi. Per cui, qua di seguito si esami la scelta dei passi, supponendo di aver già calcolato la nuova direzione $\mathbf{d}^{(k)}$.

La **scelta del passo** α_k è comune ad entrambi, quindi:

$$\alpha_k = \underset{\alpha \in \mathbb{R}^n}{\operatorname{argmin}} \, \Phi\left(\mathbf{x}^{(k)} + \alpha \mathbf{d}^{(k)}\right) \tag{64}$$

Ovvero un passo α_k tale che $\Phi\left(\mathbf{x}^{(k+1)}\right)$ sia il **minimo** di Φ lungo la direzione $\mathbf{d}^{(k)}$ passante per $\mathbf{x}^{(k)}$.

Definizione metodo del gradiente

Adesso si può dare la definizione del Gradiente. Il **metodo del gradiente** è caratterizzato dalla scelta:

$$\mathbf{d}^{(k)} = \mathbf{r}^{(k)} = -\nabla\Phi\left(\mathbf{x}^{(k)}\right) \qquad k = 0, 1, \dots \tag{65}$$

Quindi la **direzione di discesa ad ogni passo k è la direzione opposta a quella del gradiente** (come si vede dal meno davanti a ∇) della funzione Φ.

L'**algoritmo del metodo del gradiente** è: dato $\mathbf{x}^{(0)} \in \mathbb{R}^n$, si definisce $\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}$ e si calcola:

$$\begin{aligned} &\text{per } k = 0, 1, \dots \\ &\qquad \alpha_k = \frac{\left(\mathbf{r}^{(k)}\right)^T \mathbf{r}^{(k)}}{\left(\mathbf{r}^{(k)}\right)^T A\mathbf{r}^{(k)}} \\ &\qquad \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{r}^{(k)} \\ &\qquad \mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k A\mathbf{r}^{(k)} \end{aligned} \tag{66}$$

Quando converge il metodo del gradiente?

Se $A \in \mathbb{R}^{n\times n}$ è simmetrica[10] e definita positiva[11] il metodo del gradiente converge alla soluzione del sistema lineare $A\mathbf{x} = \mathbf{b}$ per ogni dato iniziale $\mathbf{x}^{(0)} \in \mathbb{R}^n$ e inoltre vale la seguente **stima dell'errore**:

$$\left\|\mathbf{e}^{(k)}\right\|_A \leq \left(\frac{K(A)-1}{K(A)+1}\right)^k \left\|\mathbf{e}^{(0)}\right\|_A \qquad k \geq 0 \tag{67}$$

Dove $||\cdot||$ è la **norma dell'energia** definita a pagina 36 e $K(A)$ è il numero di condizionamento (spettrale) della matrice A definito a pagina 25.

Ma si può fare di meglio! Ed ecco che arriva il metodo del gradiente coniugato...

È possibile costruire delle direzioni di discesa "migliori" al fine di **giungere a convergenza in un numero di iterazioni inferiore rispetto a quelle del metodo del gradiente**. Partendo dunque dal metodo del gradiente e sfruttando la definizione ricorsiva dei vettori $\mathbf{x}^{(k)}$, si può ottenere:

$$\begin{aligned}\mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)} \\ &= \mathbf{x}^{(k-1)} + \alpha_{k-1}\mathbf{d}^{(k-1)} + \alpha_k \mathbf{d}^{(k)} \\ &= \dots \\ &= \mathbf{x}^{(0)} + \sum_{j=0}^{k} \alpha_j \mathbf{d}^{(j)}\end{aligned} \tag{68}$$

Si deduce che il vettore $\left(\mathbf{x}^{(k+1)} - \mathbf{x}^{(0)}\right) \in \mathbb{R}^n$ è una combinazione lineare delle direzioni di discesa $\left\{\mathbf{d}^{(j)}\right\}_{j=0}^{k}$.

Il **metodo del gradiente coniugato** costruisce un sistema di direzioni di discesa in $\mathbb{R}^n$ che siano tutte linearmente indipendenti fra di loro (quindi $\left\{\mathbf{d}^{(k)}\right\}_{k=0}^{n-1}$ sarà una base di $\mathbb{R}^n$) e tali che i valori $\{\alpha_k\}_{k=0}^{n-1}$ siano proprio i coefficienti dello sviluppo di $\left(\mathbf{x}^* - \mathbf{x}^{(0)}\right)$ rispetto alla base $\left\{\mathbf{d}^{(k)}\right\}_{k=0}^{n-1}$. Questo comporta che l'$n$-simo termine $\mathbf{x}^{(n)}$ della successione (eq. 68) coincide con la soluzione esatta $\mathbf{x}^*$.

Quando converge il metodo del gradiente coniugato?

Sia $A \in \mathbb{R}^{n\times n}$ una matrice simmetrica e definita positiva. Il metodo del gradiente coniugato per risolvere un sistema lineare **converge al più in n iterazioni** (in aritmetica esatta). Inoltre, il residuo $\mathbf{r}^{(k)}$ alla k-esima iterazione (con $k < n$) è ortogonale a $\mathbf{d}^{(j)}$ per $j = 0, \dots, k-1$ e:

$$\left\|\mathbf{e}^{(k)}\right\|_A \leq \frac{2c^k}{1+c^{2k}} \left\|\mathbf{e}^{(0)}\right\|_A \tag{69}$$

[10] Quindi corrisponde con la sua trasposta
[11] Definizione a pagina: 21

Con:

$$c = \frac{\sqrt{K(A)} - 1}{\sqrt{K(A)} + 1} \tag{70}$$

In aritmetica esatta il metodo del gradiente coniugato è pertanto un **metodo diretto in quanto termina dopo un numero finito di iterazioni**!

Il precondizionamento migliora il metodo del gradiente

Le iterazioni del metodo del gradiente, quando convergono alla soluzione esatta $\mathbf{x}^*$, sono caratterizzate da una **velocità di convergenza dipendente dal numero di condizionamento** della matrice A: **più grande è il numero di condizionamento di A e tanto più lenta sarà la convergenza**. Quindi, invece di risolvere il sistema lineare, si può risolvere il cosiddetto **sistema precondizionato** (equivalente):

$$P_L^{-1} A P_R^{-1} \widehat{\mathbf{x}} = P_L^{-1} \mathbf{b} \qquad \text{con } \widehat{\mathbf{x}} = P_R \mathbf{x} \tag{71}$$

Dove P_L e P_R sono opportune matrici non singolari tali che:

1. La matrice:
$$P = P_L P_R \tag{72}$$
è detta **precondizionatore**, è simmetrica e definita positiva;
2. $K\left(P^{-1}A\right) \ll K(A)$
3. La risoluzione di sistemi lineari della forma:
$$P_L \mathbf{s} = \mathbf{t} \qquad P_R \mathbf{v} = \mathbf{w} \tag{73}$$
è computazionalmente molto meno costosa della risoluzione del sistema originario $A\mathbf{x} = \mathbf{b}$.

Se si sceglie P_R uguale alla matrice identità, la matrice $P = P_L$ è detta **precondizionatore sinistro**, ed il sistema precondizionato assume la forma:

$$P^{-1} A \mathbf{x} = P^{-1} \mathbf{b} \tag{74}$$

Analogamente, se P_L è la matrice identità, la matrice $P = P_R$ è detta **precondizionatore destro**, ed il sistema precondizionato assume la forma:

$$A P^{-1} \widehat{\mathbf{x}} = \mathbf{b} \qquad \text{con } \widehat{\mathbf{x}} = P\mathbf{x} \tag{75}$$

Il **metodo del gradiente precondizionato** si ottiene applicando il metodo del gradiente al sistema precondizionato in cui si prenda $P_R = P_L^T$ e quindi $P_R^{-1} = P_L^{-T} \equiv \left(P_L^{-1}\right)^T$. Più precisamente, riscrivendo:

$$\underbrace{P_L^{-1} A P_L^{-T}}_{\widehat{A}} \underbrace{P_L^T \mathbf{x}}_{\widehat{\mathbf{x}}} = \underbrace{P_L^{-1} \mathbf{b}}_{\widehat{\mathbf{b}}} \tag{76}$$

2.3 Metodi numerici per sistemi non lineari

2.3.1 Introduzione

Molto spesso i problemi ingegneristici sono *non lineari*. Per risolvere tali problemi in modo ottimale, si adotta l'approssimazione numerica, la quale porta a dover risolvere un **sistema di equazioni non lineari**.

In generale, si ha il seguente problema:

$$\mathbf{f}(\mathbf{x}) = \mathbf{0} \longleftrightarrow \begin{cases} f_1(x_1, \ldots, x_j, \ldots, x_n) = 0 \\ \ldots \\ f_i(x_1, \ldots, x_j, \ldots, x_n) = 0 \\ \ldots f_n(x_1, \ldots, x_j, \ldots, x_n) = 0 \end{cases}$$

Dove:

- $\mathbf{x} \in \mathbb{R}^n$ è il vettore delle incognite x_j
- $\mathbf{f}$ è una funzione del tipo $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$
- f_i sono funzioni che esprimono relazioni non lineari fra le incognite

Questo caso è un **problema poiché presenta due principali difficoltà** dal punto di vista numerico:

1. Si è di fronte ad un **sistema**, per cui non si hanno equazioni singole.
2. Inoltre, il sistema è **non lineare**, e questo significa che deve essere linearizzato per poterlo risolvere.

2.3.2 Metodo di Newton

Il metodo di newton per sistemi lineari era stato introdotto nel paragrafo 1.3 (pagina 9). Può essere riscritto in modo equivalente per sistemi non lineari nel seguente modo.

Se la funzione $f : \mathbb{R}^n \to \mathbb{R}$ con $n \geq 1$, è di classe $C^2(\mathbb{R}^n)$, e se è possibile calcolare le sue derivate prime e seconde, allora è possibile applicare il metodo di Newton all'equazione vettoriale $\mathbf{f}(\mathbf{x}) = \nabla f(\mathbf{x}) = \mathbf{0}$ per calcolare il punto di minimo $\mathbf{x}^*$.

Dunque, il **metodo di Newton per sistemi non lineari** si formula nel seguente modo. Dato $\mathbf{x}^{(0)} \in \mathbb{R}^n$, per $k = 0, 1, \ldots$ e fino a convergenza, si deve:

$$\begin{aligned} &\text{risolvere} \quad \mathrm{H}\left(\mathbf{x}^{(k)}\right) \delta\mathbf{x}^{(k)} = -\nabla\mathbf{f}\left(\mathbf{x}^{(k)}\right) \\ &\text{ponendo} \quad \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \delta\mathbf{x}^{(k)} \end{aligned} \tag{77}$$

È interessante notare che la matrice Hessiana utilizzata ne metodo di Newton è uguale alla matrice Jacobiana $J_{\mathbf{f}}\left(\mathbf{x}^{(k)}\right)$ valutata nel punto $\mathbf{x}^{(k)}$.

Dato un punto $z \in \mathbb{R}^n$, si introduce la matrice Jacobiana relativa a $\mathbf{f}$:

$$J(\mathbf{z}) = \nabla\mathbf{f}(\mathbf{z}) \tag{78}$$

Di componenti:

$$j_{il} = \frac{\partial f_i(\mathbf{z})}{\partial x_l} \qquad i, l = 1, \ldots, n \tag{79}$$

Per ogni $\mathbf{x} \in \mathbb{R}^n$ si ha quindi $J(\mathbf{z}) \in \mathbb{R}^{n \times n}$. Avendo definito la matrice Jacobiana non singolare (determinante diverso da zero), si può riscrivere il metodo di Newton in modo alternativo nel seguente modo:

$$\begin{aligned} &\text{risolvere} \quad \mathrm{J}\left(\mathbf{x}^{(k)}\right) \delta\mathbf{x}^{(k)} = -\mathbf{f}\left(\mathbf{x}^{(k)}\right) \\ &\text{ponendo} \quad \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \delta\mathbf{x}^{(k)} \end{aligned} \tag{80}$$

Algoritmo

Ad ogni iterazione k, il primo passo del metodo di Newton consiste nella risoluzione di un sistema lineare di dimensione n:

1. Risolvere il sistema lineare $n \times n$:

$$\mathrm{J}\left(\mathbf{x}^{(k)}\right) \delta\mathbf{x}^{(k)} = -\mathbf{f}\left(\mathbf{x}^{(k)}\right)$$

   Infatti $\mathrm{J}\left(\mathbf{x}^{(k)}\right)$ è una matrice non singolare di coefficienti noti:

$$\frac{\partial f_i\left(\mathbf{x}^{(k)}\right)}{\partial x_l}$$

   E $-\mathbf{f}\left(\mathbf{x}^{(k)}\right)$ è il termine noto.

2. Una volta risolto il sistema lineare e ottenuto $\delta\mathbf{x}^{(k)}$, si aggiorna la variabile $\mathbf{x}^{(k+1)}$ mediante il secondo passo del metodo:

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

Come detto, questo procedimento avviene ad ogni iterazione k!

Il metodo di Newton, può essere considerato un:

- **Metodo locale** se:
  - Esiste un $\delta > 0$
  - È vera la condizione:
  $$\left|\left|\alpha - \mathbf{x}^{(0)}\right|\right| < \delta$$
  Per cui:
  $$\lim_{k\to\infty} \left|\left|\alpha - \mathbf{x}^{(k)}\right|\right| = 0$$
- **Metodo del secondo ordine** se:
  - Newton converge
  - La matrice Jacobiana J è derivabile

  Per cui:
  $$\frac{\left|\left|\alpha - \mathbf{x}^{(k+1)}\right|\right|}{\left|\left|\alpha - \mathbf{x}^{(k)}\right|\right|^2} \leq C$$

Si ricorda che α sono le radici di $\mathbf{f}$.

Criteri d'arresto

Dato che il metodo di Newton è un metodo iterativo, è necessario introdurre opportuni criteri di arresto. Quindi:

- Criterio sul **residuo**:
$$\left|\left|\mathbf{f}\left(\mathbf{x}^{(k)}\right)\right|\right| < \varepsilon \tag{81}$$
- Criterio sull'**incremento**:
$$\left|\left|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right|\right| < \varepsilon \tag{82}$$

Scegliendo un'opportuna tolleranza ε.

Costi computazionali e varianti ottimizzate

Il metodo di Newton per sistemi non lineari, richiede come **costo computazionale**:

$$\texttt{CPU TIME} = \#\text{ iterazioni} \times (C_{cos} + C_{sl}) \tag{83}$$

Dove:

- C_{cos} è il **tempo** necessario per **costruire la matrice Jacobiana** $J\left(\mathbf{x}^{(k)}\right)$.
- C_{sl} è il **tempo** di **risoluzione** del corrispondente **sistema lineare**.

Si possono fare alcune osservazioni riguardo questo calcolo:

- La matrice Jacobiana $J\left(\mathbf{x}^{(k)}\right)$ deve essere ricostruita ad ogni iterazione.
- Generalmente, il numero di iterazioni (# iterazioni) è molto basso poiché il metodo di Newton è di ordine 2.

In ogni caso, a causa di un tempo di processo alto nei casi di sistemi lineari con ordini maggiori, esistono due principali alternative:

1. **Aggiornamento Jacobiana ogni p iterazioni**. Dato un numero intero $p \geq 2$, l'idea è quella di aggiornare la matrice Jacobiana $J\left(\mathbf{x}^{(k)}\right)$ solo ogni p volte.

   Così facendo, il tempo di costruzione della matrice Jacobiana C_{cos} viene ridotto, dato che si riduce a p iterazioni.

   In questo metodo, si definisce $C_{\text{cos}-\text{Newton}}$ il tempo di costruzione della matrice Jacobiana che si avrebbe applicando il semplice metodo di Newton (e non tale variante!). E la variabile C_{cos} nella formula 83 diventa:

   $$C_{\text{cos}} = \frac{C_{\text{cos}-\text{Newton}}}{p} \tag{84}$$

   Ovviamente p sono il numero di iterazioni.

   Infine, utilizzando la fattorizzazione LU per la risoluzione del sistema lineare, il costo della fattorizzazione della matrice Jacobiana $J\left(\mathbf{x}^{(k)}\right)$ è suddiviso su p iterazioni. Di fatto, ad ogni iterazione mediamente si ha un costo di:

   $$\frac{2n^3}{3p}$$

   Tuttavia, così facendo, il numero di iterazioni è nettamente maggiore a quello del secondo ordine. Un numero di iterazioni troppo elevato rischia di perdere la convergenza. Per questo motivo, si dovrebbe rispettare la seguente relazione:

   $$\#\text{iterazioni} \times \frac{(C_{\text{cos}} + C_{\text{sl}})}{p} < \#\text{iterazioni} \times (C_{\text{cos}} + C_{\text{sl}}) \tag{85}$$

2. **Inexact Newton**. L'idea è quella di sostituire ad ogni iterazione k la matrice Jacobiana $J\left(\mathbf{x}^{k}\right)$ con una sua approssimazione $\tilde{J}^k$ che sia in generale facilmente costruibile e tale che il sistema lineare associato sia facilmente risolvibile.

   Tale tecnica è un **metodo esatto**, quindi raggiunta la convergenza la soluzione è α (la parola *inexact* si riferisce alla matrice Jacobiana!).

3 Approssimazione di funzioni e di dati

3.1 Interpolazione

In molte applicazioni concrete si conosce una funzione solo attraverso i suoi valori in determinati punti. Si supponga di conoscere $n+1$ coppie di valori $\{x_i, y_i\}$ con $i = 0, \ldots, n$, dove i punti x_i, tutti distinti, vengono chiamati **nodi**.

In tal caso, può apparire naturale richiedere che la funzione approssimante $\tilde{f}$ soddisfi le seguenti uguaglianze:

$$\tilde{f}(x_i) = y_i \qquad i = 0, 1, \ldots, n \tag{86}$$

Una tale funzione $\tilde{f}$ viene chiamata **interpolatore** dell'insieme di dati $\{y_i\}$ e le equazioni del tipo 86 sono le **condizioni di interpolazione**.

Esistono vari tipi di interpolatori:

- L'**interpolatore polinomiale**:

$$\tilde{f}(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$

- L'**interpolatore trigonometrico**:

$$\tilde{f}(x) = a_{-M} e^{-iMx} + \cdots + a_0 + \cdots + a_M e^{iMx}$$

  Dove M è un intero pari a $\frac{n}{2}$ se n è pari, $\frac{(n+1)}{2}$ se n è dispari, e i è l'unità immaginaria.

- L'**interpolatore razionale**:

$$\tilde{f}(x) = \frac{a_0 + a_1 x + \cdots + a_k x^k}{a_{k+1} + a_{k+2} x + \cdots + a_{k+n+1} x^n}$$

3.2 Interpolazione Lagrangiana

L'interpolazione Lagrangiana è un'interpolazione di tipo polinomiale.

Definizione 1: interpolazione Lagrangiana

Per ogni insieme di coppie $\{x_i, y_i\}$, $i = 0, \dots, n$, con i nodi x_i distinti fra loro, esiste un unico polinomio di grado minore od uguale a n, che viene indicato con $\prod_n$ e viene chiamato **polinomio interpolatore** dei valori y_i nei nodi x_i, tale che:

$$\prod_n (x_i) = y_i \qquad i = 0, \dots, n \tag{87}$$

Quando i valori $\{y_i, i = 0, \dots, n\}$, rappresentano i valori assunti da una funzione continua f (ovvero $y_i = f(x_i)$), $\prod_n$ è detto **polinomio interpolatore** di f (in breve, interpolatore di f) e viene indicato con $\prod_n f$.

Quindi un *interpolatore* è una **funzione che assume il valore dei dati in corrispondenza dei nodi** x_i.

Come ottenere il polinomio interpolatore?

Per ogni k compreso tra 0 e n si costruisce un polinomio di grado n, denotato $\varphi_k(x)$, il quale interpola i valori y_i tutti nulli fuorché quello per $i = k$ per il quale $y_k = 1$, ovvero:

$$\varphi_k \in \mathbb{P}_n \qquad \varphi_k(x_j) = \delta_{jk} = \begin{cases} 1 & \text{se } j = k \\ 0 & \text{altrimenti} \end{cases}$$

Dove δ_{jk} è il simbolo di Kronecker. Si può dunque definire la formula dei **polinomi caratteristici di Lagrange** $\varphi_k(x)$:

$$\varphi_k(x) = \prod_{j=0, j\neq k}^{n} \frac{x - x_j}{x_k - x_j} \qquad k = 0, \dots, n \tag{88}$$

Che non è altro che:

$$\varphi_k(x) = \frac{(x - x_0)(x - x_1) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n)}{(x_k - x_0)(x_k - x_1) \cdots (x_k - x_{k-1})(x_k - x_{k+1}) \cdots (x - x_n)}$$

Essi sono dati dal prodotto di n termini di primo grado, perciò sono dei **polinomi di grado** n.

Esempio 1: polinomi caratteristici di Lagrange

Si prenda:

- $n = 2$
- $x_0 = -1$
- $x_1 = 0$
- $x_2 = 1$

I 3 polinomi caratteristici di Lagrange sono dati da:

$$\begin{aligned}\varphi_0(x) &= \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{1}{2}x(x-1)\\ \varphi_1(x) &= \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = -(x+1)(x-1)\\ \varphi_2(x) &= \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{1}{2}x(x+1)\end{aligned}$$

Come atteso, si nota che i 3 polinomi caratteristici di Lagrange sono dei polinomi di grado 2 e soddisfano la proprietà $\varphi_k(x_i) = \delta_{ik}$

Quanto detto finora, può essere generalizzato nel seguente modo:

$$\prod_n(x) = \sum_{j=0}^{n} y_j \varphi_j(x) \tag{89}$$

Proprietà interpolatore Lagrangiano

Le proprietà sono:

1. $\prod_n(x)$ è un **interpolatore**. Difatti, valutandolo per un generico nodo x_i, si ottiene:
$$\prod_n(x_i) = \sum_{j=0}^{n} y_j \varphi_j(x_i) = \sum_{j=0}^{n} y_j \delta_{ij} = y_i$$

2. $\prod_n(x)$ è un **polinomio di grado** n. Infatti, esso è dato dalla somma dei polinomi di grado n $y_j\varphi_j(x)$.

3. $\prod_n(x)$ è l'**unico polinomio di grado** m **interpolante gli** $n+1$ **dati** (x_i, y_i), $i = 0, \dots, n$.

   Si supponga esista un altro interpolatore polinomiale di grado n $\psi_n(x)$ che interpola i dati (x_i, y_i), $i = 0, \dots, n$. Si introduce il seguente polinomio di grado n:
   $$D(x) = \prod_n(x) - \psi_n(x)$$

   Allora vale il seguente risultato:
   $$D(x_i) = \prod_n(x_i) - \psi_n(x_i) = 0 \qquad \forall i = 0, \dots, n$$

   Di conseguenza, $D(x)$ ha $n+1$ zeri. Essendo un polinomio di grado n, si deve avere $D(x) \equiv 0$, da cui segue l'unicità.

3.2.1 Accuratezza (errore) nel caso di approssimazione di funzioni

Si consideri il caso di approssimazione dei valori di una funzione $f(x)$. Si ha dunque il seguente risultato.

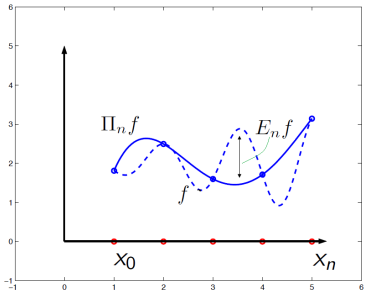
Definizione 2: quantificazione dell'errore

Sia I un intervallo limitato, e si considerino $n+1$ nodi di interpolazione distinti $\{x_i, i = 0, \ldots, n\}$ in I. Sia f derivabile con continuità fino all'ordine $n+1$ in I. Allora $\forall x \in I$, $\exists \xi_x \in I$ tale che:

$$E_n f(x) = f(x) - \prod_n f(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \prod_{i=0}^{n} (x - x_i) \tag{90}$$

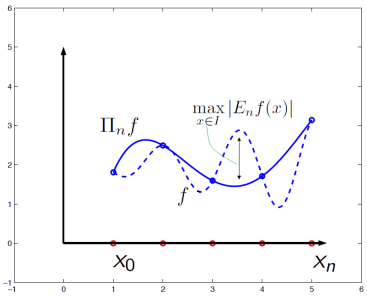
In altre parole, la definizione rappresenta come si può **quantificare l'errore** che si commette sostituendo ad una funzione f il suo polinomio interpolatore $\prod_n f$ (in questo caso Lagrangiano).

Nel seguente grafico si può notare che la funzione $E_n f(x)$ si annulla in corrispondenza dei nodi x_i. Inoltre, è in generale diverso da zero lontano dai nodi, ovvero con $x \neq x_i$.



Nel caso di nodi equispaziati, vale la seguente **stima dell'errore massimo dell'interpolatore Lagrangiano**:

$$\max_{x \in I} |E_n f(x)| \leq \frac{\left| \max_{x \in I} f^{(n+1)(x)} \right|}{4(n+1)} \cdot h^{n+1} \tag{91}$$



3.2.2 Convergenza dell'interpolatore Lagrangiano

All'aumento delle informazioni a disposizione $(n+1)$, idealmente si vorrebbe un **miglioramento dell'accuratezza di una funzione approssimante**. In particolare, si vorrebbe il seguente risultato:

$$\lim_{n\to\infty} \max_{x\in I} |E_n f(x)| = 0$$

Dall'equazione 91, a pagina 48, si può notare che il denominatore della frazione e il termine h^{n+1} tendono a zero quando n tende a infinito. Al contrario, il numeratore non è certo se sia limitato per n che tende a infinito.

Difatti, possono esistere casi in cui si ha:

$$\lim_{n\to\infty} \left| \max_{x\in I} f^{(n+1)}(x) \right| = \infty$$

Che portano a:

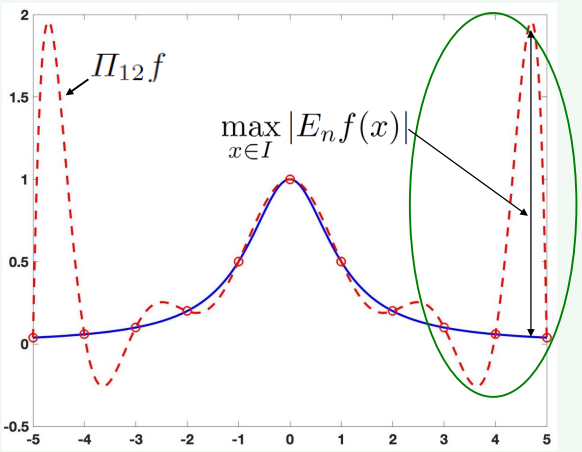
$$\lim_{n\to\infty} \max_{x\in I} |E_n f(x)| = \infty$$

Ovvero alla **non convergenza dell'interpolatore Lagrangiano**.

Esempio 2: fenomeno di Runge

Il **fenomeno di Runge** è un evento che si scatena quando **l'interpolatore Lagrangiano non raggiunge la convergenza**.
In particolare, in presenza di questo fenomeno, la funzione errore E_n presenta delle oscillazioni ai nodi estremi che crescono con il crescere di n.
Un esempio di interpolatore $\prod_{12} f$ (in linea continua) calcolato su 13 nodi equispaziati nel caso della funzione di Runge (in linea tratteggiata):

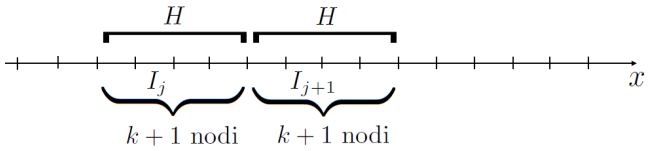
$$f(x) = \frac{1}{(1+x^2)}$$



3.3 Interpolazione Lagrangiana composita

Una miglioria che è possibile fare all'interpolazione Lagrangiana, è quella di introdurre un **interpolatore continuo dato dall'unione di tanti interpolatori Lagrangiani di basso ordine**, ovvero $k \ll n$.

Tali singoli interpolatori locali sono costruiti sugli intervalli disgiunti I_j ognuno composto da $k+1$ nodi e di lunghezza $H = kh$:



Tale interpolatore globale, chiamato **interpolazione Lagrangiana composita**, viene indicato con $\prod_k^H(x)$. Se si vuole enfatizzare che l'interpolatore è stato costruito per approssimare una certa funzione f, allora si utilizza la notazione $\prod_k^H f(x)$.

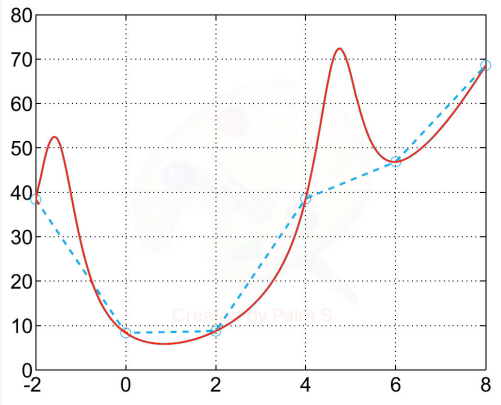
Esempio 3: interpolatore Lagrangiano composito lineare

Si consideri il caso $k = 1$. La funzione rappresentata in linea continua è:

$$f(x) = x^2 + \frac{10}{(\sin(x) + 1.2)}$$

Ed il suo interpolatore lineare composito rappresentato con la linea tratteggiata:

$$\prod_1^H f(x)$$



Vista la sua semplicità, questo interpolatore è molto utilizzato nelle applicazioni.

3.3.1 Accuratezza (errore) e convergenza dell'interpolatore Lagrangiano composito

A differenza dell'interpolazione Lagrangiana, all'aumentare del numero di informazioni $n+1$ che si hanno a disposizione, l'accuratezza dell'interpolatore Lagrangiano composito subisce un *miglioramento*.

All'aumentare del valore di n, verranno aumentati anche il numero di intervalli I_j su cui costruire gli interpolatori Lagrangiani locali, **senza variare il gradi locale k che rimarrà costante**. Questa tecnica riesce ad evitare il fenomeno di Runge, ovverosia che l'interpolatore Lagrangiano non raggiunge la convergenza.

Riguardo alla convergenza, si introduce l'**errore puntuale**:

$$E_k^H f(x) = f(x) - \prod_k^H f(x) \tag{92}$$

Il quale è dato dal massimo degli errori degli interpolatori Lagrangiani di grado k su ogni I_j. In caso di nodi equispaziati, si ottiene che la **stima dell'errore dell'interpolatore Lagrangiano composito** tende a zero nel caso in cui H tende a zero:

$$\begin{aligned}
\max_{x\in I}\left|E_k^H f(x)\right| &\leq \max_j \frac{\max_{x\in I_j}\left|f^{(k+1)}(x)\right|}{4(k+1)}\cdot h^{(k+1)} \\
&\leq \frac{\max_{x\in I}\left|f^{(k+1)}(x)\right|}{4(k+1)}\cdot h^{(k+1)} \\
&\quad \text{sostituendo } h \text{ con } \frac{H}{k} \\
&\leq \underbrace{\frac{\max_{x\in I}\left|f^{(k+1)}(x)\right|}{4(k+1)\,k^{(k+1)}}}_{\text{Indipendente da } H}\cdot H^{(k+1)}
\end{aligned} \tag{93}$$

Oltre ad un errore, questo dimostra la **convergenza dell'interpolatore Lagrangiano composito**.

Se H e k sono importanti, come devono essere scelti?

Dipende dall'origine dei dati:

- Se i dati provengono da una funzione $f(x)$ nota sull'intervallo $[a, b]$ che genera $y_i = f(x_i)$ con $i = 0, \ldots, n$, allora:
  - Si **decide il valore di** k da utilizzare cercando di non superare il valore 3 per non incorrere nel fenomeno di Runge (non convergenza!);
  - Si **sceglie il valore dell'ampiezza degli intervalli** H in modo da avere l'errore desiderato in base alla stima dell'errore riportato nell'equazione 93;
  - Si **partiziona l'intervallo** $[a, b]$ in intervallo di ampiezza H e su **ognuno di essi si considerano** $k + 1$ **nodi**.
- Se i dati provengono da misure, il numero di questi $n + 1$ è fissato. Per cui le operazioni da fare possono essere una delle seguenti:
  - Un'**interpolazione composita lineare**:

$$k = 1 \qquad H = \frac{(b - a)}{n}$$

  - Un'**interpolazione composita quadratica**:

$$k = 2 \qquad H = \frac{2(b - a)}{n}$$

3.4 Interpolazione sui nodi di Chebyshev

L'interpolazione Lagrangiana composita consente di evitare il fenomeno di Runge. Tuttavia, l'interpolazione Lagrangiana può essere modificata **posizionando i nodi in precise posizioni che garantiscono la stabilità al crescere di** n. Negli interpolatori presentati, tutti si basavano su nodi equispaziati. Nel seguente capitolo si mostrano i nodi di Chebyshev, ovvero un'ubicazione specifica dei nodi nello spazio.

Solo il fenomeno di Runge può essere evitato?

La risposta chiaramente è no. L'applicazione di questa tecnica consente:

- Di **utilizzare** *sempre* con successo il **polinomio interpolatore Lagrangiano** di grado n.
- Di evitare la non convergenza, ovvero il fenomeno di Runge.

Nodi di Chebyshev

Si consideri il caso in cui il dominio sia $[-1, 1]$ e $n+1$ dati ubicati in corrispondenza delle seguenti ascisse:

$$\widehat{x}_i = -\cos\left(\frac{\pi i}{n}\right) \qquad i = 0, \ldots, n \tag{94}$$

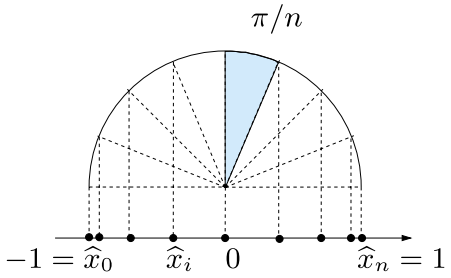


Figura 3: Distribuzione dei nodi di Chebyshev nell'intervallo $[-1, 1]$.

Questi **nodi di Chebyshev** sono **ottenuti come proiezioni sull'asse** x **di punti sulla circonferenza unitaria individuati da settori circolari ottenuti con lo stesso angolo** $\frac{\pi}{n}$.

E se viene applicato l'interpolatore Lagrangiano sui nodi di Chebyshev?

Come detto all'inizio, si può modificare direttamente l'interpolatore Lagrangiano. Quindi, si consideri l'interpolatore Lagrangiano di grado n costruito sui nodi di Chebyshev. Tale interpolatore si ottiene applicando le equazioni presentate

precedentemente, con l'unico vincolo di **prendere in considerazione $\widehat{x}_i$ come nodi su cui costruire gli n polinomi caratteristici di Lagrange**:

$$\begin{aligned} \widehat{\varphi}_k(x) &= \prod_{j=0, j\neq k}^{n} \frac{x - \widehat{x}_i}{\widehat{x}_k - \widehat{x}_i} \\ \prod_n^C(x) &= \sum_{j=0}^{n} y_j \widehat{\varphi}_j(x) \end{aligned} \tag{95}$$

La variabile $\prod_n^C(x)$ rappresenta l'**interpolatore Lagrangiano sui nodi di Chebyshev**. Si indica con la rappresentazione $\prod_n^C f(x)$ quando viene applicato a dati ottenuti dalla valutazione di una funzione f, ovvero $y_i = f(x_i)$.

3.4.1 Convergenza dell'interpolazione sui nodi di Chebyshev

Teorema 2. *Si supponga che la funzione $f(x)$ ammetta derivata continua fino all'ordine $s+1$ compreso, ovverosia $f \in C^{(s+1)}([-1,1])$.*

*Allora, si ha il seguente risultato di **convergenza per l'interpolazione sui nodi di Chebyshev**:*

$$\max_{x\in[-1,1]} \left| f(x) - \prod_n^C f(x) \right| \leq \tilde{C} \frac{1}{n^s} \tag{96}$$

Per un'opportuna costante C.

Dal precedente teorema, si possono fare 3 osservazioni:

1. Se $s \geq 1$ allora si è sicuri che ci sarà **sempre convergenza**:

$$\lim_{n\to\infty} \max_{x\in[-1,1]} \left| f(x) - \prod_n^C f(x) \right| = 0 \tag{97}$$

2. La **velocità di convergenza aumenta all'aumentare di** s.

3. Se il teorema è valido $\forall s > 0$, allora la **velocità di convergenza è *esponenziale***. Questa è la migliore stima che è possibile ottenere, dato che è più veloce di qualsiasi altro $\frac{1}{n^s}$:

$$\max_{x\in[-1,1]} \left| f - \prod_n^C f(x) \right| \leq \tilde{C} e^{-n} \tag{98}$$

3.4.2 Generalizzazione dell'intervallo

Nel caso in cui il dominio di interesse non sia più da $[-1, 1]$ (come viene utilizzato nelle precedenti equazioni), ma un intervallo generale $[a, b]$, la "conversione" può avvenire mappando i nodi $\widehat{x}_j$ da $[-1, 1]$ a $[a, b]$:

$$x = \psi(\widehat{x}) = \frac{a+b}{2} + \frac{b-a}{2}\,\widehat{x} \tag{99}$$

Applicando la **generalizzazione dell'intervallo ai nodi di Chebyshev**, si ottiene:

$$x_j^C = \psi(\widehat{x}_j) = \frac{a+b}{2} + \frac{b-a}{2}\,\widehat{x}_j \qquad j = 0, \ldots, n \tag{100}$$

Di conseguenza, il **polinomio di Lagrange applicato sui nodi di Chebyshev in un intervallo generale** diventa:

$$\varphi_k^C(x) = \prod_{j=0, j\neq k}^{n} \frac{x - x_j^C}{x_k^C - x_j^C} \tag{101}$$

3.5 Interpolazione trigonometrica

Nel caso di funzioni periodiche, ad esempio con periodo 2π e con $f(0) = f(2\pi)$, si ha il valore delle ascisse pari a:

$$x_j = \frac{2\pi j}{(n+1)} \qquad j = 0, \ldots, n$$

In questo caso, l'interpolatore Lagrangiano non approssima in modo ottimale la funzione f lontana dai nodi, dato che è un polinomio di grado n (per cui in generale non periodico!).

Supponendo n pari e $M = \frac{n}{2}$, si può costruire l'**interpolatore trigonometrico** $\tilde{f}(x)$ come la **combinazione lineare di seni e coseni**:

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{M} \left[a_k \cos(kx) + b_k \sin(kx)\right] \tag{102}$$

Per opportuni coefficienti complessi incogniti a_k, per $k = 0, \ldots, M$ e b_k per $k = 1, \ldots, M$. Come si può vedere, la funzione periodica $\tilde{f}(x)$ è di periodo 2π caratterizzata da $M + 1$ **frequenze** e **coefficienti** a_k e b_k con $k = 0, \ldots, M$.

Sfruttando la nota formula di **Eulero**:

$$e^{ikx} = \cos(kx) + i \sin(kx)$$

L'**interpolatore trigonometrico** può essere riscritto in modo più compatto:

$$\tilde{f}(x) = \sum_{k=-M}^{M} c_k e^{ikx} \tag{103}$$

Inoltre, dall'equazione 102 i valori a_k, b_k con $k = 0, \ldots, M$ sono:

$$\begin{cases} a_k = c_k + c_{-k} \\ b_k = i(c_k - c_{-k}) \end{cases}$$

E nell'interpolazione trigonometrica con Eulero 103 si ha:

$$\begin{cases} c_k = \frac{1}{2}(a_k - ib_k) \\ c_{-k} = \frac{1}{2}(a_k + ib_k) \end{cases}$$

Quindi, le funzioni interpolatrici trigonometriche $\tilde{f}(x)$ sono $2M + 1 = n + 1$ equazioni nelle $n + 1$ incognite c_k.

La **distanza** (costante) fra due nodi $h = \frac{2\pi}{(n+1)}$ come:

$$x_j = jh \qquad j = 0, \ldots, n$$

Si impone inoltre il **vincolo di interpolazione**:

$$\tilde{f}(x_j) = \sum_{k=-M}^{M} c_k e^{ikjh} = f(x_j) \qquad j = 0, \ldots, n \tag{104}$$

A differenza dell'interpolazione Lagrangiana, i coefficienti della combinazione lineare c_k con $k = -M, \ldots, M$, sono incogniti. Si hanno di conseguenza $n + 1$ equazioni nelle $2M + 1 = n + 1$ incognite c_k. Tale equazione è una base di partenza per le formule nei paragrafi successivi.

3.5.1 Trasformata Discreta di Fourier

Dato un intero $m \in [-M, M]$, si moltiplica all'equazione di base 104, a pagina 56, a sinistra e a destra per $e^{-imx_j} \Rightarrow e^{imjh}$ e si sommano su j:

$$\sum_{j=0}^{n} \sum_{k=-M}^{M} c_k e^{ikjh} e^{-imjh} = \sum_{j=0}^{n} f(x_j) e^{-imjh} \tag{105}$$

In cui si può esporre una relazione esplicita fra i coefficienti e i valori noti della funzione $f(x_j)$:

$$c_m = \frac{1}{n+1} \sum_{j=0}^{n} f(x_j) e^{-imjh} \qquad m = -M, \ldots, M \tag{106}$$

La funzione $f(x_j)$ viene chiamata **Trasformata discreta di Fourier (DFT)**. **Queste sono $n+1$ equazioni nelle incognite** $f(x_j)$.

Parlando di DTF, è necessario introdurre anche lo **spazio fisico** e delle **frequenze**. Dati i vettori $\mathbf{c} \in \mathbb{R}^{n+1}$ di componenti c_k e $\mathbf{f} \in \mathbb{R}^{n+1}$ di componenti $f(x_j)$, è possibile riscrivere la trasformata di Fourier nello **spazio delle frequenze** k come:

$$c_k = \frac{1}{n+1} \sum_{j=0}^{n} f(x_j) e^{-ikjh} \qquad k = -M, \ldots, M$$

E in forma matriciale nel seguente modo:

$$\mathbf{c} = T\mathbf{f} \tag{107}$$

Con la matrice T composta da elementi:

$$T_{kj} = \frac{1}{n+1} e^{-ikjh}$$

Nel caso in cui si vuole passare dallo **spazio delle frequenze** $f(x_j)$ allo **spazio fisico** c_k, si può usare il vincolo di interpolazione (eq. 104, pag. 56):

$$\tilde{f}(x_j) = \sum_{k=-M}^{M} c_k e^{ikjh} \qquad j = 0, \ldots, n$$

E in forma matriciale nel seguente modo:

$$\mathbf{f} = T^{-1}\mathbf{c} \tag{108}$$

Con la matrice T^{-1} composta da elementi:

$$\left(T^{-1}\right)_{kj} = e^{ikjh}$$

La **trasformazione dallo spazio delle frequenze allo spazio fisico** è chiamata **Trasformata Discreta di Fourier inversa (Inverse Discrete Fourier Transform, IDFT)**.

3.5.2 Fast Fourier Transform (FFT)

Il calcolo dei coefficienti c_k può essere fatto ancora più rapidamente utilizzato la **Trasformata rapida di Fourier (Fast Fourier Transform, FFT)**.

Si consideri una generica matrice quadrata A di dimensione $(n+1) \times (n+1)$ e un vettore $\mathbf{x}$ di dimensione $n+1$. È possibile calcolare la componente w_j del vettore $\mathbf{w} = A\mathbf{x}$ come:

$$w_j = \sum_{k=1}^{n} A_{jk} x_k \tag{109}$$

Nella FFT, la matrice T ha una struttura particolare chiamata di Toeplitz. Questo comporta che l'elemento w_j del vettore $\mathbf{w} = T\mathbf{x}$ è determinato solo dagli elementi della matrice T, ovverosia T_{km} (con $j = km$). Una struttura tipica della matrice è:

$$T = \frac{1}{n+1} \begin{bmatrix} e^{-ih} & e^{-2ih} & e^{-3ih} & \dots & \dots & e^{-nih} \\ e^{-2ih} & e^{-4ih} & e^{-6ih} & \dots & \dots & e^{-2nih} \\ e^{-3ih} & e^{-6ih} & e^{-9ih} & \dots & \dots & e^{-3nih} \\ \vdots & & \ddots & & & \vdots \\ \vdots & & \ddots & & & \vdots \\ e^{-nih} & e^{-2nih} & e^{-3nih} & \dots & \dots & e^{-n^2ih} \end{bmatrix}$$

Il **costo computazione** del vettore $\mathbf{w}$, a causa della forma particolare della matrice di Toeplitz, è uguale a $n \log_2 n$ operazioni.

3.5.3 Espressione Lagrangiana dell'interpolatore trigonometrico

L'interpolatore trigonometrico può essere scritto in forma Lagrangiana, ovvero con coefficienti nella combinazione lineare dati dal risultato dei vari $f(x_j)$, nel seguente modo:

$$\tilde{f}(x) = \sum_{j=0}^{n} f(x_j) \varphi_j^T(x) \tag{110}$$

Dove il polinomio φ non è altro che:

$$\varphi_j^T(x_k) = \delta_{jk} \tag{111}$$

E in questo caso la convergenza?

La convergenza è possibile esprimerla tramite il seguente teorema.

Teorema 3. *Si supponga che la funzione $f(x)$ ammetta derivata continua e periodica di periodo 2π fino all'ordine $s+1$ compreso.*

*Allora, si ha il seguente risultato di **convergenza per l'interpolatore trigonometrico in forma Lagrangiana**:*

$$\max_{x \in [0,2\pi]} \left| f(x) - \tilde{f}(x) \right| \leq C \frac{1}{n^s} \tag{112}$$

Per un'opportuna costante C.

Dal precedente teorema, se ne deriva che la **convergenza è esponenziale qualora il precedente risultato valga per** $\forall s > 0$.

3.5.4 Fenomeno dell'aliasing e teorema di Shannon

Se il numero di dati $n+1$ non è sufficientemente elevato, l'**interpolatore trigonometrico non è in grado di descrivere le frequenze k più alte**. Questo problema viene chiamato **fenomeno di aliasing**.

Nel mondo reale, l'occhio umano campiona con una frequenza massima le informazioni luminose che provengono dal mondo esterno. Ma se tale campionamento non è sufficientemente frequente da catturare la frequenza massima del fenomeno esterno, allora l'immagine riprodotta dal cervello (che di fatto agisce in questo caso come un interpolatore trigonometrico) sarà caratterizzata da frequenze diverse (aliasing). E questo è anche il motivo per negli elicotteri si vedono le pale girare molto lentamente o addirittura in senso opposto.

In particolare, sia $k_{\max}$ la **frequenza massima** della funzione $f(x)$. Se $n \leq 2k_{\max}$, la **frequenza massima dell'interpolatore trigonometrico** $\tilde{f}(x)$ è minore di $k_{\max}$ (aliasing).

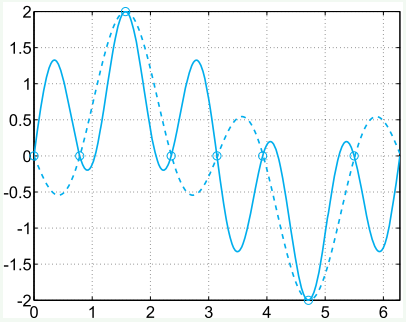
Esempio 4: aliasing

Gli effetti dell'aliasing si possono vedere confrontando, per esempio, le seguenti due funzioni:

- In linea continua:

$$f(x) = \sin(x) + \sin(5x)$$

- Linea tratteggiata, l'interpolatore trigonometrico $\tilde{f}(x)$ con $M = 3$



Quindi, il **teorema di Shannon** afferma che n deve essere:

$$n > 2k_{\max} \tag{113}$$

3.6 Il metodo dei minimi quadrati

Perché è necessario un altro metodo?

Al crescere del grado del polinomio dell'interpolazione di Lagrange, esso non garantisce una maggiore accuratezza nell'approssimazione di una funzione. Per aumentare l'accuratezza, è stato introdotto l'interpolazione Lagrangiana composita. Purtroppo, quest'ultima non è ottima per estrapolare informazioni da dati noti, ovvero **per generare nuove valutazioni in punti che giacciono al di fuori dell'intervallo di interpolazione**.

Quindi, i precedenti metodi esposti non sono ottimi:

- **Interpolazione Lagrangiana**: potrebbe soffrire del fenomeno di Runge, a causa del numero elevato di dati.
- **Interpolazione Lagrangiana composita**: terrebbe in conto solo gli ultimi $k+1$ dati, per cui non utilizzerebbe tutta la storia a disposizione e non darebbe un'espressione analiticamente semplice.
- **Interpolazione su nodi Chebyshev**: nella realtà è difficile da applicare poiché i dati a disposizione sono equispaziati.

Ragionamento per arrivare alla definizione del metodo

Si supponga di avere a disposizione un insieme di dati e ciascun elemento è formato da una coppia:

$$\{(x_i, y_i)\,,\ i = 0, \dots, n\}$$

In cui gli y_i potrebbero essere i valori che una funzione assume nei nodi x_i: $q(x_i) = y_i$.

L'obbiettivo è cercare un *polinomio globale* di grado m (con $m \geq 1$) molto più basso rispetto a n ($m \ll n$). Tale polinomio ha l'obbiettivo di **minimizzare lo *scarto quadratico medio***, ovverosia la somma dei quadrati degli errori nei nodi[12].

Dato lo **spazio dei polinomi di grado** m:

$$\mathbb{P}_m = \{p_m \ : \ \mathbb{R} \to \mathbb{R} \ : \ p_m(x) = b_0 + b_1 x + \cdots + b_m x^m\}$$

Allora, il (problema) **metodo dei minimi quadrati** consiste nel cercare il polinomio $q \in \mathbb{P}_m$ tale che:

$$\sum_{i=0}^{n} [y_i - q(x_i)]^2 \leq \sum_{i=0}^{n} [y_i - p_m(x_i)]^2 \qquad \forall p_m \in \mathbb{P}_m \tag{114}$$

Riscrivendo il polinomio $q(x)$ come:

$$q(x) = a_0 + a_1 x + \cdots + a_m x^m$$

[12] Si ricorda che l'errore di un nodo è la distanza tra la funzione lineare e il suo corrispettivo interpolatore. A pagina 48 è possibile vederlo visivamente.

Dove i coefficienti a sono incogniti, è possibile riformulare il metodo dei minimi quadrati (eq. 114). Quindi determinare $a_0, a_1, \ldots, a_m$ tali che:

$$\psi(a_0, a_1, \ldots, a_m) = \min_{\{b_i, i=0,\ldots,m\}} \psi(b_0, b_1, \ldots, b_m)$$

Dove:

$$\psi(b_0, b_1, \ldots, b_m) = \sum_{i=0}^{n} [y_i - (b_0 + b_1 x_i + \cdots + b_m x_i^m)]^2$$

Nel caso in cui si abbia una **retta di regressione** (cioè $m = 1$), il problema si riduce a:

$$\psi(b_0, b_1) = \sum_{i=0}^{n} \left[y_i^2 + b_0^2 + b_1^2 x_i^2 + 2b_0 b_1 x_i - 2b_0 y_i - 2b_1 x_i y_i\right]$$

Se ne ricava che il grafico della funzione ψ è un **paraboloide convesso** il cui punto di **minimo** (a_0, a_1) si trova imponendo le sue derivate parziali uguale a zero:

$$\frac{\partial \psi}{\partial b_0}(a_0, a_1) = 0 \qquad \frac{\partial \psi}{\partial b_1}(a_0, a_1) = 0$$

Che equivale a risolvere il seguente sistema di 2 equazioni e incognite:

$$\sum_{i=0}^{n} [a_0 + a_1 x_i - y_i] = 0 \qquad \sum_{i=0}^{n} \left[a_0 x_i + a_1 x_i^2 - x_i y_i\right] = 0$$

Ponendo $D = (n+1)\sum_{i=0}^{n} x_i^2 - \left(\sum_{i=0}^{n} x_i\right)^2$, la soluzione è:

$$\begin{aligned} a_0 &= \frac{1}{D}\left[\sum_{i=0}^{n} y_i \sum_{j=0}^{n} x_j^2 - \sum_{j=0}^{n} x_j \sum_{i=0}^{n} x_i y_i\right] \\ a_1 &= \frac{1}{D}\left[(n+1)\sum_{i=0}^{n} x_i y_i - \sum_{j=0}^{n} x_j \sum_{i=0}^{n} y_i\right] \end{aligned}$$

Il corrispondente polinomio:

$$q(x) = a_0 + a_1 x \tag{115}$$

È noto come **retta dei minimi quadrati**, o **retta di regressione**.

Attenzione

Da notare, anche se parzialmente scontato, che nel caso in cui m sia uguale a n ($m = n$), **il problema si riduce all'interpolatore Lagrangiano**.

4 Integrazione numerica

4.1 Introduzione

Alcuni problemi ingegneristici richiedono il calcolo di integrali, talvolta pure molto complessi. Non sempre si riesce a trovare in forma esplicita la primitiva di una funzione, anche nel caso in cui la si conosca, potrebbe essere difficile valutarla (come ad esempio $f(x) = \cos(4x)\cos(3\sin(x))$).

In tutti questi casi, è necessario utilizzare metodi numerici in gradi di restituire un valore approssimato della quantità di interesse, indipendentemente da quanto complessa sia la funzione da integrare o da differenziare. In questo capitolo, vengono presentati alcuni metodi per l'**approssimazione numerica di integrali di funzioni**, i quali vengono chiamate **formule di quadratura**.

Alcune notazioni e definizioni di introduzione

Si consideri l'integrale:

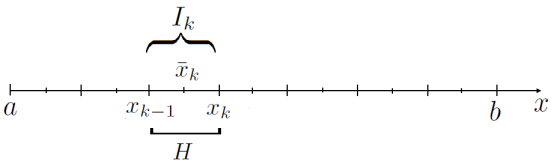
$$I(f) = \int_a^b f(x)\,\mathrm{d}x \tag{116}$$

Si suddivida l'intervallo $[a, b]$ in M intervalli I_k di ampiezza costante H:

$$[x_{k-1}, x_k] \quad k = 1, \ldots, M \qquad x_k = a + kH \quad k = 0, \ldots, M$$

Si introduce inoltre su ogni intervallo i punti medi:

$$\overline{x}_k = \frac{x_{k-1} + x_k}{2}$$



Si consideri una **formula di quadratura** per il calcolo approssimato dell'integrale nell'equazione 116 e sia $I_H(f)$ il valore approssimato ottenuto.

La **formula di quadratura** è di **ordine** p se l'*errore* soddisfa la seguente stima:

$$E_H = |I(f) - I_H(f)| \leq CH^p \tag{117}$$

Inoltre, la formula di quadratura ha **grado di esattezza pari ad** r se essa risulta esatta quando applicata **ai polinomi di grado minore o uguale ad** r, ovverosia se:

$$E_H = |I(f) - I_H(f)| = 0 \qquad \forall f \in \mathbb{P}^r(a, b) \tag{118}$$

4.2 Formula del punto medio composita

Data la proprietà di additività dell'integrale:

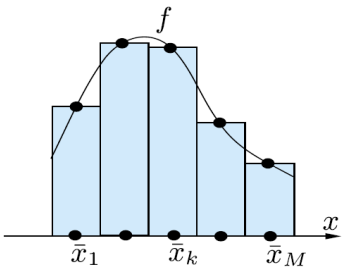
$$I(f) = \sum_{k=1}^{M} \int_{I_k} f(x)\ \mathrm{d}x$$

Si possono approssimare il valore dell'integrale su ogni intervallo e dopodiché sommare i contributi.

Per cui, l'idea del **punto medio composita** è quella di approssimare il valore dell'integrale della funzione:

$$\int_{I_k} f(x)\ \mathrm{d}x$$

Con l'area del rettangolo con altezza $f(\overline{x}_k)$:



Dato che l'altezza dei rettangoli è sempre pari ad H, si avrà la seguente approssimazione per $I(f)$:

$$I_{pm}(f) = H \sum_{k=1}^{M} f(\overline{x}_k) \tag{119}$$

Dove *pm* indica *punto medio*.

Inoltre, la **stima dell'errore del punto medio composita** è:

$$|I(f) - I_{pm}(f)| \leq \max_x |f''(x)|\ \frac{b-a}{24} H^2 \tag{120}$$

Si possono fare due **osservazioni** interessanti su questa stima dell'errore:

1. La formula del **punto medio composita** è di **ordine 2**.
2. La formula del **punto medio composita** ha **grado di esattezza pari a 1**. Dato che è stata assunta la continuità della derivata seconda di f, allora $f'' = 0$ in ogni x se f è una retta, ovverosia con $r = 1$.

Dall'ultima osservazione, risulta chiaro che la **formula del punto medio composita è esatta se applicata alle rette**.

4.3 Formula dei trapezi composita

Data la proprietà di additività dell'integrale:

$$I(f) = \sum_{k=1}^{M} \int_{I_k} f(x)\ \mathrm{d}x$$

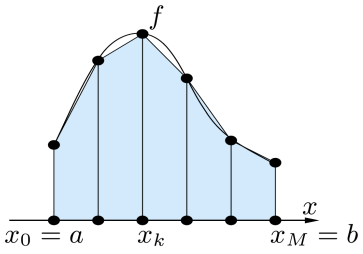
Si vuole approssimare l'area sottesa da f nell'intervallo I_k considerando il trapezio costruito sui punti x_{k-1} e x_k e sulle corrispondenti ordinate.

La formula dei **trapezi composita** è:

$$I_{tr}(f) = \frac{H}{2} \sum_{k=1}^{M} (f(x_{k-1}) + f(x_k)) \tag{121}$$

Per comodità di implementazione, è possibile riscriverla anche come:

$$I_{tr}(f) = \frac{H}{2} (f(a) + f(b)) + H \sum_{k=1}^{M-1} f(x_k)$$



Da notare che la formula dei trapezi composita equivale all'**integrale esatto dell'interpolatore Lagrangiano composito di grado 1**:

$$I_{tr}(f) = \int_a^b \left(\prod_1^H f(x) \right) \mathrm{d}x \tag{122}$$

Su ogni intervallo I_k si può considerare l'errore di interpolazione Lagrangiana.

La **stima dell'errore per la formula dei trapezi composita** è:

$$|I(f) - I_{tr}(f)| \leq \frac{1}{12} \max_x |f''(x)| (b-a) H^2 \tag{123}$$

Si possono fare due **osservazioni** riguardo alla stima dell'errore:

1. La **formula dei trapezi composita** è di **ordine 2**.
2. La **formula dei trapezi composita** ha **grado di esattezza pari a 1**.

Da notare che sia grado che ordine sono uguali al punto medio composita. Per cui, **è meglio scegliere questo metodo o il punto medio?** La scelta risiede principalmente sull'**avere a disposizione le coordinate dei punti medi o dei punti estremi degli intervalli**.

4.4 Formula di Simpson composita

Data la proprietà di additività dell'integrale:

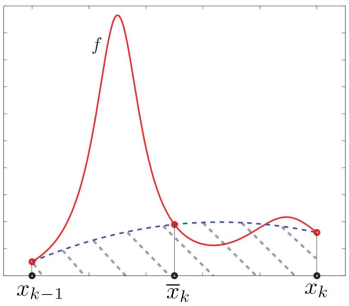
$$I(f) = \sum_{k=1}^{M} \int_{I_k} f(x)\ \mathrm{d}x$$

Si vuole approssimare l'area sottesa da f nell'intervallo I_k considerando l'area sottesa dalla parabola che interpola i punti x_{k-1}, $\overline{x}_k$ e x_k.

Integrando tali parabole su ogni intervallo I_k, si ottiene la **formula di Simpson composita**:

$$I_{sim}(f) = \frac{H}{6} \sum_{k=1}^{M} \left(f(x_{k-1}) + 4f(\overline{x}_k) + f(x_k) \right) \tag{124}$$

L'area sottesa alla parabole interpolante è quella tratteggiata in grigio:



Da notare che per la costruzione della forma di Simpson composita equivale all'**integrale esatto dell'interpolatore Lagrangiano composito di grado 2**:

$$I_{sim}(f) = \int_a^b \prod_{2}^{H} f(x)\ \mathrm{d}x \tag{125}$$

Assumendo che f abbia la derivata quarta continua su $[a, b]$, allora la **stima dell'errore per la formula dei trapezi composita** è:

$$|I(f) - I_{sim}(f)| \leq \frac{b-a}{2880} \max_x \left| f^{(iv)}(x) \right| H^4 \tag{126}$$

Si possono fare due **osservazioni** riguardo alla stima dell'errore:

1. La **formula di Simpson** è di **ordine 4**.
2. La **formula di Simpson** ha **grado di esattezza pari a 3**.

Per cui, tale metodo integra i polinomi che sono globalmente di:

- Grado $r = 1$ (rette)
- Grado $r = 2$ (parabole)
- Grado $r = 3$ (cubiche)

5 Approssimazione numerica di ODE

5.1 Problema di Cauchy

Un'equazione differenziale ordinaria ammette in generale infinite soluzioni. Per fissarne una è necessario imporre una condizione che prescriva il valore assunto dalla soluzione in un punto dell'intervallo di integrazione. In questa sezione ci si occuperà della risoluzione dei **problemi di Cauchy**, ossia di problemi nella seguente forma.

Problema di Cauchy

Sia $I = [t_0, T]$ l'intervallo temporale di interesse. Trovare la funzione vettoriale $\mathbf{y} : I \to \mathbb{R}^n$ che soddisfa:

$$\begin{cases} \mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)) \\ \mathbf{y}(t_0) = \mathbf{y}_0 \end{cases} \tag{127}$$

Con $\mathbf{f} : I \times \mathbb{R}^n \to \mathbb{R}^n$ funzione vettoriale assegnata.

Problema di Cauchy *scalare*

Per semplicità, verrà considerato soltanto il **caso scalare**, ovvero in cui $n = 1$, ma i metodi di approssimazione che verranno introdotti saranno facilmente estendibili e applicabili anche con $n > 1$. Sia $I = [t_0, T]$ l'intervallo temporale di interesse. Trovare $y : I \subset \mathbb{R} \to \mathbb{R}$ tale che:

$$\begin{cases} y'(t) = f(t, y(t)) & \forall t \in I \\ y(t_0) = y_0 \end{cases} \tag{128}$$

Con $f : I \times \mathbb{R} \to \mathbb{R}$ funzione assegnata.

Definizione 1: esistenza e unicità della soluzione continua

Si supponga che la funzione $f(t, y)$ sia:

1. Limitata e continua rispetto ad entrambi gli argomenti.
2. Lipschitziana rispetto al secondo argomento, ossia esista una costante L positiva (detta costante di Lipschitz) tale per cui:

$$|f(t, y_1) - f(t, y_2)| \leq L |y_1 - y_2| \qquad \forall t \in I, \ \forall y_1, y_2 \in \mathbb{R} \tag{129}$$

Allora la **soluzione** del problema di Cauchy **esiste**, è **unica** ed è di **classe** C^1 su I.

5.2 Approssimazione di derivate

Al fine di approssimare il problema di Cauchy, è necessario prima introdurre un altro problema: **siano noti i valori $v(t_n)$ di una funzione v in corrispondenza di noti t_n, $n = 0, \dots, N$, e si vuole approssimare i valori della sua derivata prima $v'(t)$ negli stessi nodi**.

Approssimazione in avanti della derivata prima

Partendo dallo sviluppo in serie di Taylor:

$$v(t_{n+1}) = v(t_n) + hv'(t_n) + \frac{h^2}{2}v''(t_n) + \frac{h^3}{6}v'''(t_n) + O(h^4)$$

E ricordando che una quantità è un $O(h^p)$ se vale:

$$\lim_{h\to 0} \frac{O(h^p)}{h^p} < +\infty$$

Ovverosia se va a 0 con la stessa velocità o più velocemente di h^p. Scrivendo la derivata $v'(t_n)$ come:

$$v'(t_n) = \frac{v(t_{n+1}) - v(t_n)}{h} - \underbrace{\frac{h}{2}v''(t_n) - \frac{h^2}{6}v'''(t_n) + O(h^3)}_{O(h)}$$

Questo consente di introdurre un'**approssimazione in avanti della derivata prima**:

$$D^+ v(t_n) = \frac{v(t_{n+1}) - v(t_n)}{h} \tag{130}$$

Il cui **errore** è dato da:

$$\left| v'(t_n) - D^+ v(t_n) \right| = O(h) \tag{131}$$

Approssimazione all'indietro della derivata prima

In modo analogo, considerando il seguente sviluppo di Taylor:

$$v(t_{n-1}) = v(t_n) + hv'(t_n) + \frac{h^2}{2}v''(t_n) + \frac{h^3}{6}v'''(t_n) + O(h^4)$$

Usando i passaggi analoghi a quelli di prima, si arriva all'**approssimazione all'indietro della derivata prima**

$$D^- v(t_n) = \frac{v(t_n) - v(t_{n-1})}{h} \tag{132}$$

Il cui **errore** è dato da:

$$\left| v'(t_n) - D^- v(t_n) \right| = O(h) \tag{133}$$

Approssimazione centrata della derivata prima

Per completezza, si riporta anche l'**approssimazione centrata della derivata prima**:

$$D^c v(t_n) = \frac{v(t_{n+1}) - v(t_{n-1})}{2h} \tag{134}$$

E in questo caso l'**errore** commesso è:

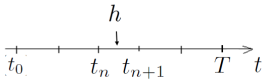
$$|v'(t_n) - D^c v(t_n)| \, 0O\left(h^2\right) \tag{135}$$

Infine, si evince che è un metodo di ordine 2.

E riguardo a Cauchy?

L'approssimazione numerica del problema di Cauchy si suddivide in 4 passi:

1. Si suddivide l'intervallo temporale in nodi:



2. Si scrive il problema di Cauchy per il generico nodo t_n valido per ogni $n = 1, \ldots, N = \frac{T}{h}$:
$$y'(t_n) = f(t_n, y(t_n))$$

3. Si sostituisce per ogni n a $y'(t_n)$ una delle approssimazioni presentate: in avanti (eq. 130), all'indietro (eq. 132) o centrata (eq. 134).

4. Si denota, per ogni n, con u_n la soluzione del nuovo problema approssimato ottenuto, la quale è una candidata per essere una buona approssimazione di $y(t_n)$.

La **soluzione approssimata** è composta in generale come:

$$\{u_0 = y_0, u_1, u_2, \ldots, u_n\}$$

5.3 I metodi di Eulero in avanti e all'indietro

Dato il problema di Cauchy, esso vale per ogni t nell'intervallo I, quindi in particolare anche in t_n:

$$y'(t_n) = f(t_n, y(t_n))$$

Approssimando la derivata sinistra con l'approssimazione in avanti introdotta a pagina 67, si ottiene:

$$D^+ y(t_n) = \frac{y(t_{n+1}) - y(t_n)}{h} \approxeq f(t_n, y(t_n))$$

Da notare che l'espressione assomiglia ad un'uguaglianza poiché l'espressione a sinistra è un'approssimazione della derivata e di conseguenza della funzione f.

L'idea è quella di introdurre come soluzione numerica u_n con $n = 1, \ldots, N$ la quale è una successione di valori che risolve in maniera esatta la precedente relazione.

$$\begin{aligned} \frac{y(t_{n+1}) - y(t_n)}{h} &\approxeq f(t_n, y(t_n)) \\ &\downarrow \\ \frac{u_{n+1} - u_n}{h} &= f(t_n, u_n) \end{aligned}$$

Da qui si ricava comodamente il **metodo di Eulero in avanti**:

$$u_{n+1} = u_n + hf(t_n, u_n) \qquad n = 0, \ldots, N-1 \tag{136}$$

In maniera analoga, partendo dal problema di Cauchy:

$$y'(t_{n+1}) = f(t_{n+1}, y(t_{n+1}))$$

Approssimando la derivata a sinistra con l'approssimazione all'indietro introdotto a pagina 67, si ottiene:

$$D^- y(t_{n+1}) = \frac{y(t_{n+1}) - y(t_n)}{h} \approxeq f(t_{n+1}, y(t_{n+1}))$$

Introducendo una soluzione numerica u_n con $n = 1, \ldots, N$ la quale è una successione di valori che risolve in maniera esatta la precedente relazione.

$$\begin{aligned} \frac{y(t_{n+1}) - y(t_n)}{h} &\approxeq f(t_{n+1}, y(t_{n+1})) \\ &\downarrow \\ \frac{u_{n+1} - u_n}{h} &= f(t_{n+1}, u_{n+1}) \end{aligned}$$

Da qui si ricava comodamente il **metodo di Eulero all'indietro**:

$$u_{n+1} = u_n + hf(t_{n+1}, u_{n+1}) \qquad n = 0, \ldots, N-1 \tag{137}$$

Perché il metodo di Eulero in avanti viene chiamato esplicito?

Per il metodo di Eulero in avanti, la condizione iniziale $u_0 = y_0$ è ben nota. Dopodiché, viene effettuato un **calcolo in sequenza**, ovverosia u_1, u_2, e così via. Tuttavia, non è un metodo iterativo poiché la soluzione è significativa per ogni n, essendo l'approssimazione della y a diversi istanti temporali.

Ovviamente i due metodi calcolano la nuova u_{n+1} in modo differente. Infatti, il metodo di Eulero in avanti calcola:

$$u_{n+1} = u_n + hf\left(t_n, u_n\right)$$

Il **nuovo valore è calcolato in maniera esplicita**. Difatti è sufficiente valutare la funzione f, anche se non lineare, per t_n, u_n ottenendo quindi un'espressione esplicita per il termine a destra. Questo è il motivo per cui il metodo di Eulero in avanti viene chiamato anche **metodo di Eulero esplicito**.

Perché il metodo di Eulero all'indietro viene chiamato implicito?

Per il metodo di Eulero all'indietro:

$$u_{n+1} = u_n + hf\left(t_{n+1}, u_{n+1}\right)$$

Non si ha un'espressione esplicita per il nuovo valore u_{n+1} **perché quest'ultimo compare anche a destra dell'uguale**!

Si pone come incognita $x = u_{n+1}$. Se la funzione f non è lineare rispetto a y, allora ad ogni passo temporale è necessario risolvere il **problema di ricerca degli zeri** della funzione:

$$g\left(x\right) = x - u_n - hf\left(t_{n+1}, x\right) \tag{138}$$

Se la funzione $g\left(x\right)$ è derivabile, allora è necessario introdurre ad ogni istante temporale un metodo iterativo, come quello di Newton per esempio. In ogni caso, a causa del fatto che u_{n+1} non può in generale essere determinato per via esplicita, il metodo di Eulero all'indietro viene chiamato anche **metodo di Eulero implicito**.

In generale, si possono **dividere i metodi numerici** per le equazioni ordinarie in **metodi espliciti**, la cui **soluzione al nuovo passo temporale è calcolata mediante un'espressione esplicita**, e **metodi impliciti** che invece **richiedono la soluzione di un'equazione non lineare**.

5.3.1 Assoluta stabilità

È stabile il metodo di Eulero?

Per discutere della stabilità, è necessario introdurre il concetto di **assoluta stabilità**. Un metodo numerico viene detto **assolutamente stabile** per un determinato valore di h maggiore di zero, se:

$$|u_{n+1}| \leq C_{AS} |u_n| \qquad C_{AS} < 1 \tag{139}$$

O, equivalentemente, se si è su un intervallo illimitato:

$$\lim_{n \to +\infty} |u_n| = 0$$

Adesso si analizzano i metodi singolarmente.

Assoluta stabilità: metodo di Eulero in avanti

Dato il **problema modello lineare**:

$$\begin{cases} y'(t) = \lambda y(t) & \lambda < 0 \\ y(0) = 1 \end{cases}$$

La cui soluzione esatta è: $y(t) = e^{\lambda t}$. Applicando il metodo di Eulero in avanti, si ottiene:

$$u_{n+1} = u_n + h\lambda u_n = (1 + h\lambda) u_n \longrightarrow C_{AS} = |1 + h\lambda|$$

Da cui l'assoluta stabilità è **garantita se**:

$$C_{AS} < 1 \longrightarrow -1 < 1 + h\lambda < 1$$

Ovvero h deve rispettare la seguente condizione:

$$h\lambda > -2 \Rightarrow h < -\frac{2}{\lambda} \tag{140}$$

Da notare che $h\lambda < 0$ è sempre vera.

Esempio 1: stabilità metodo di Eulero in avanti

Data la soluzione di un problema con $\lambda = -1$, ottenuto con il metodo di Eulero in avanti, nella seguente figura si può vedere come:

- Linea tratteggiata: rappresentata $h = \frac{30}{14}$. Quindi, essendo:

  $$h < -\frac{2}{\lambda} \Rightarrow \frac{30}{14} < -\frac{2}{-1} \Rightarrow \frac{30}{14} < 2 \text{ ✗}$$

  La soluzione numerica oscilla e infine esplode.

- Linea continua: rappresentata $h = \frac{30}{16}$. Quindi, essendo:

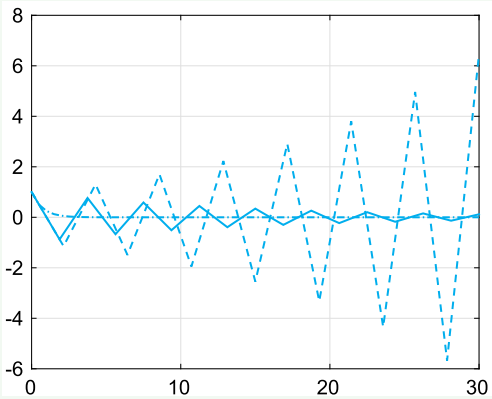
  $$\frac{30}{16} < 2 \text{ ✓}$$

La soluzione numerica oscilla ma non esplode. La sua oscillazione è dovuta al fatto che è molto vicina al limite 2.

- Linea tratto-punto: rappresentata $h = \frac{1}{2}$. Quindi, essendo:

$$\frac{1}{2} < 2 \checkmark$$

La soluzione numerica non oscilla.



Assoluta stabilità: metodo di Eulero all'indietro

Dato il solito problema modello lineare:

$$\begin{cases} y'(t) = \lambda y(t) & \lambda < 0 \\ y(0) = 1 \end{cases}$$

La cui soluzione esatta è: $y(t) = e^{\lambda t}$. Applicando il metodo di Eulero all'indietro, si ottiene:

$$u_{n+1} = u_n + h\lambda u_{n+1} \rightarrow u_{n+1} = \frac{1}{1-h\lambda} u_n \rightarrow C_{AS} = \left| \frac{1}{1-h\lambda} \right|$$

In questo caso, l'**assoluta stabilità è sempre garantita**, poiché:

$$C_{AS} < 1 \quad \forall h$$

Quindi, per il metodo di Eulero all'indietro si può essere certi che è **incondizionatamente assolutamente stabile**.

5.3.2 Problemi di Cauchy generali

Il concetto di stabilità introdotto nel paragrafo precedente riguarda il *problema modello lineare*. Per cui, per i problemi di Cauchy, in generale, che stabilità si ha?

Un metodo numerico produce una soluzione stabile se a *piccole* perturbazioni sul dato iniziale si producono perturbazioni *piccole* e decrescenti per n crescente.

Notando che per il *problema modello lineare* si ha:

$$f(y) = \lambda y \;\rightarrow\; f'(y) = \lambda$$

Ha senso introdurre la **stima dell'intervallo dei valori** di h che garantiscono una **soluzione stabile** per il ***metodo di Eulero in avanti***:

$$h < -\frac{2}{\overline{\lambda}} \qquad \overline{\lambda} = -\max_{t,y}\left|\frac{\partial f}{\partial y}\right| \qquad \frac{\partial f}{\partial y} < 0 \tag{141}$$

Essa rappresenta una situazione di cautela, la quale potrebbe essere molto restrittiva per il problema di Cauchy, ma sicuramente garantisce una soluzione numerica stabile.

In generale, vale la proprietà che un **metodo numerico assolutamente stabile per il *problema modello lineare*** sotto una condizione su h (piccolo a sufficienza) **produce una soluzione numerica stabile anche per il problema di Cauchy generico sotto la stessa condizione opportunatamente adattata**.

5.4 Il metodo di Crank-Nicolson

Dato il problema di Cauchy:

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

Dal teorema fondamentale del calcolo integrale limitato all'intervallo $[t_n, t_{n+1}]$ si ottiene:

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(t, y(t)) \, \mathrm{d}t$$

Adesso si applica una formula di quadratura a disposizione (punto medio composita, trapezi composita, Simpson composita) per approssimare l'integrale a destra dell'uguale e ottenere così un metodo di approssimazione per il problema di Cauchy.

Ed ecco che nasce il **metodo di Crank-Nicolson (CN)** applicando la formula dei trapezi composita (par. 4.3, pag. 64) considerando solo un intervallo $[t_n, t_{n+1}]$:

$$u_{n+1} = u_n + \frac{h}{2}\left(f(t_n, u_n) + f(t_{n+1}, u_{n+1})\right) \tag{142}$$

Esso è un **metodo implicito** e richiede ad ogni passo n di determinare la radice della funzione:

$$g(x) = x - \frac{h}{2} f(t_{n+1}, x) - u_n - \frac{h}{2} f(t_n, u_n)$$

Riguardo l'**assoluta stabilità**, si ha:

$$u_{n+1} = u_n + \frac{h\lambda}{2}(u_n + u_{n+1}) \;\rightarrow\; u_{n+1} = \frac{2 + h\lambda}{2 - h\lambda} u_n$$

Di conseguenza si ottiene:

$$C_{AS} = \left| \frac{2 + h\lambda}{2 - h\lambda} \right|$$

Che è sempre minore di 1. Per cui il metodo di Crank-Nicolson è **incondizionatamente assolutamente stabile**.

5.5 Il metodo di Heun

Partendo dal metodo di Crank-Nicolson introdotto nel paragrafo 5.4 a pagina 74:

$$u_{n+1} = u_n + \frac{h}{2}\left(f\left(t_n, u_n\right) + f\left(t_{n+1}, u_{n+1}\right)\right)$$

E al posto della variabile u_{n+1} a destra si introduce la sua approssimazione ottenuta con il metodo di Eulero in avanti, si ottiene il **metodo di Heun**:

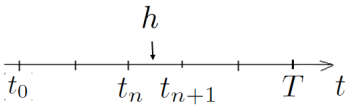
$$u_{n+1} = u_n + \frac{h}{2}\left(f\left(t_n, u_n\right) + f\left(t_{n+1}, u_n + hf\left(t_n, u_n\right)\right)\right) \tag{143}$$

Esso è un **metodo esplicito** con **condizione di assoluta stabilità uguale a quella del metodo di Eulero in avanti** (si veda a pagina 71 l'assoluta stabilità di tale metodo).

5.6 Convergenza

Per discutere l'**accuratezza della soluzione numerica** trovata da ciascun metodo introdotto per determinati valori di h, è necessario introdurre la definizione di **convergenza**:

$$|y(t_n) - u_n| = O(h) \qquad n = 0, \ldots, N \tag{144}$$



Modificando la convergenza appena introdotta (eq. 144), essa rappresenta la **stima dell'errore di convergenza**:

$$|y(t_n) - u_n| = O(h^p) \qquad n = 0, \ldots, N \tag{145}$$

Inoltre si dice che il **metodo numerico è di ordine** p. È interessante notare che all'**aumentare di** p, la **velocità di convergenza aumenta**.

Utilizzando λ del coefficiente del *problema modello lineare*, oppure:

$$\overline{\lambda} = -\max_{t,y} \left| \frac{\partial f}{\partial y} \right| \qquad \text{con } \frac{\partial f}{\partial y} < 0$$

Nella tabella 1 sono riassunti i metodi introdotti in questo capitolo.

Metodo	**Espl./Impl.**	**Convergenza**	**Assoluta stabilità**
Eulero in avanti	Esplicito	$O(h)$	$h < -\frac{2}{\lambda}$
Eulero all'indietro	Implicito	$O(h)$	Sempre garantita
Crank-Nicolson	Implicito	$O(h^2)$	Sempre garantita
Heun	Esplicito	$O(h^2)$	$h < -\frac{2}{\lambda}$

Tabella 1: Proprietà di convergenza e stabilità dei metodi per approssimare le ODE.

5.6.1 Consistenza dei metodi di Eulero

Innanzitutto si introduce l'**errore di troncamento locale** τ_n, il quale si **commette introducendo la soluzione esatta** y nello schema numerico. Per il **metodo di Eulero in avanti** si ha:

$$\tau_n = \left| f\left(t_n, y\left(t_n\right)\right) - \frac{y\left(t_{n+1}\right) - y\left(t_n\right)}{h} \right| \tag{146}$$

E per il **metodo di Eulero all'indietro** si ha:

$$\tau_n = \left| f\left(t_{n+1}, y\left(t_{n+1}\right)\right) - \frac{y\left(t_{n+1}\right) - y\left(t_n\right)}{h} \right| \tag{147}$$

Un metodo numerico è **consistente** se vale:

$$\lim_{h \to 0} \tau_n = 0 \qquad \forall n \tag{148}$$

Inoltre, se vale:

$$\tau_n = O\left(h^p\right) \qquad \forall n \tag{149}$$

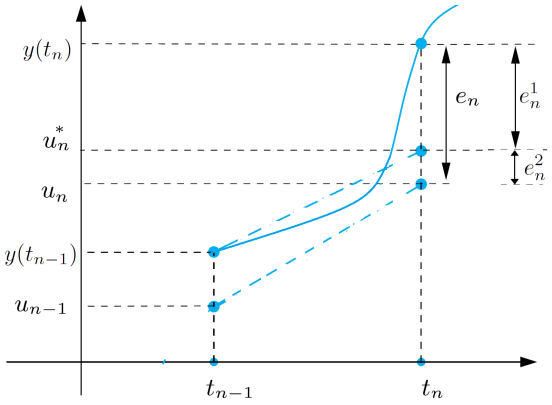
Si dice che l'**ordine di consistenza è** p. Riguardo i ***metodi di Eulero***, essi sono **consistenti del primo ordine**.

Si noti che l'**errore di troncamento locale** τ_n non è l'errore necessario per verificare la convergenza: $\left|y\left(t_n\right) - u_n\right|$. L'errore:

$$e_n = \left|y\left(t_n\right) - u_n\right|$$

È dato da due contributi:

- L'**errore** e_n^1 è **dovuto localmente al metodo numerico**, ottenuto partendo dalla vera soluzione $y\left(t_{n-1}\right)$.
- L'**errore** e_n^2 è **dovuto alla *propagazione degli errori*** commessi agli istanti precedenti



Riguardo al **metodo di Eulero in avanti** l'**errore** è dato da:

- Primo contributo dovuto localmente al metodo numerico:

$$e_n^1 = |y(t_n) - y(t_{n-1}) - hf(t_{n-1}, y(t_{n-1}))| = h\tau_{n-1} \tag{150}$$

  Da notare che il primo errore, a meno di un fattore h, è dato dall'errore di troncamento locale. Questo evidenzia come la **consistenza da sola non basti per avere la convergenza**.

- Secondo contributo dovuto alla propagazione degli errori:

$$e_n^* = |u_n^* - u_n| \tag{151}$$

Supponendo che valga la **condizione di assoluta stabilità** con $\overline{\lambda} = -\lambda_{\max}$:

$$h < \frac{2}{\lambda_{\max}}$$

Dunque si ha:

$$\begin{gathered} |e_n| \leq \left|e_n^1\right| + \left|e_n^2\right| \leq h\tau_{n-1} + |e_{n-1}| \\ \downarrow \\ |e_n| \leq nhO(h) = (t_n - t_0) O(h) \end{gathered} \tag{152}$$

Il **metodo di Eulero in avanti** è dunque **convergente** e del **primo ordine**. Lo stesso risultato vale anche per il **metodo di Eulero all'indietro**. In generale, è possibile affermare che, per un metodo convergente, l'**ordine di convergenza è uguale all'ordine di consistenza**.

La **convergenza** è una **proprietà della soluzione numerica** (o meglio all'errore), per n fissato, in ogni nodo e facendo tendere h a 0. Invece, l'assoluta stabilità analizza il comportamento della soluzione numerica per h fissato e al crescere di n.

In generale, i **metodi espliciti** (come Eulero in avanti e Heun) **garantiscono l'assoluta stabilità solo per un valore di h piccolo**. I **metodi impliciti** godono invece di ottime proprietà di **assoluta stabilità**; spesso questa è **incondizionata** (come per Eulero all'indietro e Crank-Nicolson).

6 Laboratorio

6.1 Introduzione al linguaggio MATLAB

L'introduzione al linguaggio di programmazione MATLAB sarà molto rapido. Si assume dunque che l'interfaccia grafica sia familiare e che concetti base di programmazione (per esempio "che cos'è una variabile?") siano ben noti.

In MATLAB, l'assegnazione di scalari a delle variabili è classica, quindi si utilizza il simbolo uguale: `a = 1` (assegnazione del valore `1` alla variabile `a`). Inoltre, il linguaggio è *case sensitive*, di conseguenza la variabile `a` è diversa dalla variabile `A`. Alcuni comandi utili e generali:

- `help` *`nome-comando`*, per avere informazioni in più riguardo al comando *`nome-comando`*;
- `clear` *`nome-variabile`*, per rimuovere la variabile *`nome-variabile`* dalla memoria. Se non viene inserito il *`nome-variabile`*, vengono rimosse tutte le variabili dalla memoria.
- `who`, per visualizzare le variabili attualmente in memoria.
- `clc`, per ripulire la *Command Window*.

Tabella 2: Argomenti trattati.

Well-known variables

Esistono alcune variabili che sono ben note e hanno valori prestabiliti. Tra le più importanti:

- `pi`, che rappresenta il π e MATLAB gli assegna il valore `3.1416`
- `i`, che rappresenta l'unità immaginaria e MATLAB gli assegna il valore `0.0000 + 1.0000i`
- `eps`, che rappresenta il più piccolo valore rappresentabile nel calcolatore (PC) attualmente in uso. Solitamente, `eps` ritorna il valore `2.2204e-16`.

Questo tipo di variabili possono essere ridefinite, ma non è una *good practice*.

Cambiare il formato delle variabili: `format`

Il comando `format` è utilizzato per cambiare il formato con cui sono rappresentate le variabili. MATLAB non cambia la precisione della variabile (quindi non si ottiene una precisione maggiore dopo la virgola), ma modifica soltanto la rappresentazione. Di default MATLAB utilizza una rappresentazione di tipo `short`. Tra i più utilizzati (di default `pi` è uguale a `3.1416`):

- `default` per reimpostare la rappresentazione di default.
- Decimale:
  - `short`, rappresentazione a 5 cifre:

```
>> format short
>> pi

ans =
    3.1416
```

  - `long`, rappresentazione a 15 cifre:

```
>> format long
>> pi

ans =
   3.141592653589793
```

- *Floating point*:
  - `short e`, rappresentazione a 5 cifre floating point:

```
>> format short e
>> pi

ans =
   3.1416e+00
```

  - `long e`, rappresentazione a 15 cifre floating point:

```
>> format long e
>> pi

ans =
     3.141592653589793e+00
```

Altri formati si possono trovare nella documentazione ufficiale.

Assegnamento di vettori e matrici

- **Vettore riga**, si può creare utilizzando uno spazio tra i valori o una virgola `,`:

```
>> a = [1 2 3 4]

a =
     1     2     3     4

>> b = [1, 2, 3, 4]

b =
     1     2     3     4
```

- **Vettore colonna**, si crea usando il punto e virgola `;`:

```
>> a = [1; 2; 3; 4]

a =
     1
     2
     3
     4
```

Talvolta può essere utile la generazione automatica di un vettore riga (sono ammessi anche i valori negativi e con la virgola ovviamente):

- **Vettore riga generato linearmente**, si crea usando i due punti e specificando il valore di inizio e il valore di fine:

```
>> a = [1 : 4]

a =
     1     2     3     4
```

- **Vettore riga generato usando un passo**, si crea usando i due punti e specificando (in ordine) il valore di inizio, il "salto", e il valore di fine. Nel caso in cui il salto sia troppo grande e si superi il valore di fine, MATLAB prenderà il primo valore ammissibile:

```
>> % Generazione con passo 1
>> a = [1 : 1 : 4]

a =
     1     2     3     4

>> % Generazione con passo 2
>> a = [1 : 2 : 5]

a =
     1     3     5

>> % Generazion con passo 2 (fine non raggiunta)
>> a = [1 : 2 : 6]

a =
     1     3     5

>> % ... ma cambiando l'upper bound
```

```
>> a = [1 : 2 : 7]

a =
     1     3     5     7
```

- **Vettore riga generato con valori uniformemente distanziati**, si crea usando la funzione `linspace`, la quale accetta tre parametri:
  - `x1`, valore di partenza.
  - `x2`, valore di fine.
  - `n`, numero di valori da generare; se non specificato, di default è `100`; se il valore inserito è zero o minore, viene creato un vettore vuoto.

```
>> % Vettore riga identico a: a = [1 : 4]
>> linspace(1, 4, 4)

ans =
     1     2     3     4

>> % Chiedendo piu' valori, MATLAB andra' ad utilizzare i
    decimali
>> linspace(1, 4, 6)
ans =

     1.000000000000000e+00     1.600000000000000e+00
     2.200000000000000e+00     2.800000000000000e+00
     3.400000000000000e+00     4.000000000000000e+00
```

Le matrici possono essere create a mano o usando la combinazione delle tecniche viste in precedenza:

- **Matrice**, le righe si creano usando gli spazi e le colonne si creano usando il punto e virgola:

```
>> a = [1 2 3 4; 5 6 7 8; 9 10 11 12]

a =
     1     2     3     4
     5     6     7     8
     9    10    11    12

>> a = [1, 2, 3, 4; 5, 6, 7, 8; 9, 10, 11, 12]

a =
     1     2     3     4
     5     6     7     8
     9    10    11    12
```

- **Matrice creata usando la generazione lineare dei vettori**, si possono utilizzare le tecniche precedenti e i punti e virgola:

```
>> % Usando a = [x : y]
>> a = [1 : 4; 5 : 8; 9 : 12]

a =
     1     2     3     4
     5     6     7     8
     9    10    11    12

```

```
>> % Usando a = [x : y : z]
>> a = [1 : 2 : 7; 9 : 2 : 15; 17 : 2 : 23]
a =

     1     3     5     7
     9    11    13    15
    17    19    21    23

>> % Usando linspace
>> a = [linspace(1, 4, 4); linspace(5, 8, 4); linspace(9, 12,
    4)]

a =
     1     2     3     4
     5     6     7     8
     9    10    11    12
```

Operazioni su vettori e matrici

- **Trasposizione**, la classica operazione eseguita con le matrici o vettori, si esegue con la keyword `'` oppure usando la funzione `transpose`:

```
>> a = [1 2 3 4]

a =
     1     2     3     4

>> a'

ans =
     1
     2
     3
     4

>> transpose(a)

ans =
     1
     2
     3
     4
```

- **Somma e sottrazione**
  - Tra **vettore e scalare**:

```
>> a = [1 2 3 4]

a =
     1     2     3     4

>> a + 1

ans =
     2     3     4     5

>> a - 1

ans =
     0     1     2     3
```

  - Tra **vettore e matrice**:

```
>> a = [1 2 3 4]

a =
     1     2     3     4

>> b = [1 2 3 4; 5 6 7 8; 9 10 11 12]

b =
     1     2     3     4
     5     6     7     8
     9    10    11    12

>> a + b

ans =
```

```
     2     4     6     8
     6     8    10    12
    10    12    14    16

>> a - b

ans =
     0     0     0     0
    -4    -4    -4    -4
    -8    -8    -8    -8
```

– Tra **matrice e scalare**:

```
>> b = [1 2 3 4; 5 6 7 8; 9 10 11 12]

b =
     1     2     3     4
     5     6     7     8
     9    10    11    12

>> b + 1

ans =
     2     3     4     5
     6     7     8     9
    10    11    12    13

>> b - 1

ans =
     0     1     2     3
     4     5     6     7
     8     9    10    11
```

– Tra **matrice e matrice**:

```
>> b = [1 2 3 4; 5 6 7 8; 9 10 11 12]

b =
     1     2     3     4
     5     6     7     8
     9    10    11    12

>> c = [13 14 15 16; 17 18 19 20; 21 22 23 24]

c =
    13    14    15    16
    17    18    19    20
    21    22    23    24

>> b + c

ans =
    14    16    18    20
    22    24    26    28
    30    32    34    36

>> b - c

ans =
   -12   -12   -12   -12
   -12   -12   -12   -12
   -12   -12   -12   -12
```

- **Prodotto**
  - **Prodotto matriciale**:

```
>> b = [1 2 3 4; 5 6 7 8; 9 10 11 12]

b =
     1     2     3     4
     5     6     7     8
     9    10    11    12

>> c = [13 14 15; 16 17 18; 19 20 21; 22 23 24]

c =
    13    14    15
    16    17    18
    19    20    21
    22    23    24

>> b * c

ans =
   190   200   210
   470   496   522
   750   792   834
```

  - **Prodotto punto per punto**, in MATLAB è possibile moltiplicare ogni cella di una matrice (o vettore) per la corrispettiva cella della matrice (o vettore) moltiplicata. La keyword utilizzata è `.*`:

```
>> b = [1 2 3 4; 5 6 7 8; 9 10 11 12]

b =
     1     2     3     4
     5     6     7     8
     9    10    11    12

>> c = [13 14 15 16; 17 18 19 20; 21 22 23 24]

c =
    13    14    15    16
    17    18    19    20
    21    22    23    24

>> b .* c

ans =
    13    28    45    64
    85   108   133   160
   189   220   253   288

>> d = [1 2 3 4]

d =
     1     2     3     4

>> b .* d

ans =
     1     4     9    16
     5    12    21    32
     9    20    33    48
```

- **Potenza**
  - **Potenza matriciale**:

```
>> b = [1 2 3; 4 5 6; 7 8 9]

b =
     1     2     3
     4     5     6
     7     8     9

>> b^2

ans =
    30    36    42
    66    81    96
   102   126   150
```

  - **Potenza punto per punto**, come per il prodotto, è possibile elevare al quadrato ogni valore della matrice (o vettore):

```
>> b = [1 2 3; 4 5 6; 7 8 9]

b =
     1     2     3
     4     5     6
     7     8     9

>> b.^2

ans =
     1     4     9
    16    25    36
    49    64    81
```

Funzioni intrinseche per vettori e matrici

Qua di seguito si elencano le funzioni più importanti da utilizzare per i vettori e le matrici.

- `size`, restituisce la dimensione del vettore o della matrice nel formato *`righe colonne`*. Specificando anche un valore (o vettore) come parametro, la funzione restituisce la dimensione (un vettore contenente le dimensioni richieste) nella "dimensione" richiesta:

```
>> a = [1 2 3 4]

a =
     1     2     3     4

>> size(a)

ans =
     1     4

>> size(a, 2)
ans =

     4

>> b = [1 2 3; 4 5 6; 7 8 9]

b =
     1     2     3
     4     5     6
     7     8     9

>> size(b)

ans =
     3     3

>> size(b, [2, 3])

ans =
     3     1
```

- `length`, restituisce la lunghezza del vettore e per le matrici restituisce il numero degli elementi per ogni riga:

```
>> a = [1 2 3 4]

a =
     1     2     3     4

>> length(a)

ans =
     4

>> b = [1 2 3 4 5 6; 7 8 9 10 11 12]

b =
     1     2     3     4     5     6
     7     8     9    10    11    12

```

```
>> length(b)

ans =
     6
```

- `max`, `min`, calcolano rispettivamente il massimo e il minimo valore delle componenti di un vettore; per le matrici viene presa in considerazione ogni colonna e calcolato il massimo o minimo:

```
>> a = [ 1 2 3 4]

a =
     1     2     3     4

>> max(a)

ans =
     4

>> min(a)

ans =
     1

>> b = [7 1 1 4; 2 3 9 10; 8 1 7 1]

ans =
     7     1     1     4
     2     3     9    10
     8     1     7     1

>> max(b)

ans =
     8     3     9    10

>> min(b)

ans =
     2     1     1     1
```

- `sum`, `prod`, calcola rispettivamente la somma e il prodotto degli elementi che compongono il vettore; nel caso di una matrice, viene presa in considerazione ogni colonna e calcolata la somma o il prodotto. Inoltre, i due comandi possono prendere un argomento in più per eseguire il calcolo in una dimensione specifica (cosa sensata con le matrici):

```
>> a = [ 1 2 3 4]

a =
     1     2     3     4

>> sum(a)

ans =
    10

>> prod(a)

ans =
    24
```

```
>> b = [7 1 1 4; 2 3 9 10; 8 1 7 1]

b =
     7     1     1     4
     2     3     9    10
     8     1     7     1

>> sum(b) % per colonne

ans =
    17     5    17    15

>> sum(b, 2) % per righe

ans =
    13
    24
    17

>> prod(b)

ans =
   112     3    63    40
```

- `norm`, la norma di un vettore o di una matrice. Passando un vettore o un matrice, viene calcolata di default la norma euclidea (norma 2):

$$||\mathrm{v}||_2 = \sqrt{\sum_{i=2}^{\texttt{length(v)}} \mathrm{v}_i^2}$$

  Passando un valore aggiuntivo, esso rappresenterà l'ordine della norma:

$$||\mathrm{v}||_n = \left(\sum_{i=2}^{\texttt{length(v)}} |\mathrm{v}_i|^n \right)^{\frac{1}{n}}$$

  Infine, con `inf` viene calcolata la norma infinito:

$$||\mathrm{v}||_\infty = \max_{1 \leq i \leq \texttt{length(v)}} |\mathrm{v}_i|$$

```
>> a = [1 2 3 4]

a =
     1     2     3     4

>> norm(a)

ans =
     5.477225575051661e+00

>> norm(a, 2)

ans =
     5.477225575051661e+00

>> norm(a, 3)
```

```
17
18 ans =
19      4.641588833612779e+00
20
21 >> norm(a, inf)
22
23 ans =
24      4
25
26 >> b = [7 1 1 4; 2 3 9 10; 8 1 7 1]
27
28 b =
29      7      1      1      4
30      2      3      9     10
31      8      1      7      1
32
33 >> norm(b, 2)
34
35 ans =
36      1.711222312384884e+01
37
38 >> norm(b, inf)
39
40 ans =
41     24
```

- `abs`, rappresenta il valore assoluto e restituisce il vettore o matrice dopo aver applicato il valore assoluto a ciascun elemento:

```
1  >> a = [-1 -2 -3 -4]
2
3  a =
4      -1     -2     -3     -4
5
6  >> abs(a)
7
8  ans =
9       1      2      3      4
10
11 >> b = [7 1 1 -4; 2 3 -9 10; 8 1 -7 1]
12 b =
13
14      7      1      1     -4
15      2      3     -9     10
16      8      1     -7      1
17
18 >> abs(b)
19
20 ans =
21      7      1      1      4
22      2      3      9     10
23      8      1      7      1
```

- `diag`, estrae la diagonale di una matrice esistente, oppure ne crea una con i valori dati come input. Inoltre, può creare una matrice con la diagonale spostata a seconda del valore dato (si veda l'esempio):

```
>> b = [7 1 1 4; 2 3 9 10; 8 1 7 1]

b =
     7     1     1     4
     2     3     9    10
     8     1     7     1

>> diag(b)

ans =
     7
     3
     7

>> diag(b, 1)

ans =
     1
     9
     1

>> diag(b, -1)

ans =

     2
     1

>> diag([1 2 3])

ans =
     1     0     0
     0     2     0
     0     0     3

>> diag([1 2 3], -1)

ans =
     0     0     0     0
     1     0     0     0
     0     2     0     0
     0     0     3     0
```

Funzioni matematiche elementari

Qua di seguito una lista di alcune funzioni matematiche elementari. Gli esempi e la sintassi non verranno mostrati poiché è sempre la medesima:

funzione(*parametro*)

Funzione	Comando
Radice quadrata	`sqrt`
Esponenziale	`exp`
Logaritmo Naturale	`log`
Logaritmo In Base 2	`log2`
Logaritmo In Base 10	`log10`
Seno	`sin`
Arcoseno	`asin`
Coseno	`cos`
Arcocoseno	`acos`
Tangente	`tan`

Tabella 3: Funzioni matematiche elementari.

Iterazione con il ciclo for

In MATLAB il ciclo for viene eseguito con la seguente sintassi.

```
for index = values
   statements
end
```

Di seguito si riporta un ciclo `for` che itera sulla diagonale secondaria di una matrice:

```
>> b = [7 1 1 4; 2 3 9 10; 8 1 7 1]

b =
     7     1     1     4
     2     3     9    10
     8     1     7     1

>> res = []

res =
     []

>> for i = 1 : size(b, 1)
    res(i) = b(i, size(b, 1) - i + 1);
end

>> res

res =
     1     3     8
```

Funzioni per definire vettori o matrici particolari

In queste pagine vengono presentate alcuni funzioni utili che consentono di creare matrici o vettori "particolari".

- **Vettore/Matrice nulla**, con la funzione `zeros` è possibile creare una matrice o un vettore di tutti zeri. I parametri ammessi corrispondono alla dimensione del vettore o matrice:

```
>> zeros(1, 4)

ans =
     0     0     0     0

>> zeros(4, 1)

ans =
     0
     0
     0
     0

>> zeros(4, 4)

ans =
     0     0     0     0
     0     0     0     0
     0     0     0     0
     0     0     0     0
```

- **Vettore unario/Matrice unaria**, con la funzione `ones` è possibile creare una matrice o un vettore di tutti uni. I parametri ammessi corrispondono alla dimensione del vettore o matrice:

```
>> ones(1, 4)

ans =
     1     1     1     1

>> ones(4, 1)

ans =
     1
     1
     1
     1

>> ones(4, 4)

ans =
     1     1     1     1
     1     1     1     1
     1     1     1     1
     1     1     1     1
```

- **Matrice identità**, con la funzione `eye` è possibile creare una matrice identità. I parametri ammessi corrispondono alla dimensione del vettore o matrice:

```
>> eye(3)

ans =
     1     0     0
     0     1     0
     0     0     1

>> eye(2, 3)

ans =
     1     0     0
     0     1     0

>> eye(1, 4)

ans =
     1     0     0     0
```

- **Matrice/Vettore riga di numeri casuali interi e non**, con il comando `rand` si genera una matrice di numeri casuali nell'intervallo $[0, 1]$ con la virgola, mentre con il comando `randi` si genera una matrice di numeri casuali interi (primo parametro deve essere specificato il range dei valori):

```
>> rand(3, 5)

ans =
    0.9157    0.6557    0.9340    0.7431    0.1712
    0.7922    0.0357    0.6787    0.3922    0.7060
    0.9595    0.8491    0.7577    0.6555    0.0318

>> % Matrice di valori interi random da 1 a 5
>> randi([1, 5], 3, 4)

ans =
     2     5     5     2
     1     4     1     4
     1     2     3     4

>> % Errore! L'intervallo e' sbagliato
>> randi([-1, -50], 3, 4)
Error using randi
First input must be a positive scalar integer value IMAX, or
    two integer values [IMIN IMAX] with IMIN less than or
    equal to IMAX.

>> randi([-50, -1], 3, 4)

ans =
   -41   -18   -37   -42
   -26   -15   -17   -45
   -28   -13   -18   -26
```

6.1.1 Esercizio

Creare una funzione (file) chiamato `mat_hilbert.m` che fornisca la matrice di Hilbert avente una generica dimensione `n`. Ogni cella della matrice di Hilbert deve rispettare la seguente condizione:

$$a_{ij} = \frac{1}{i + j - 1}$$

Dopo aver creato la funzione, utilizzare la funzione nativa di MATLAB `hilb`, per verificare il risultato ottenuto.

Soluzione

Il codice non ha bisogno di grandi spiegazioni. Vi è un controllo iniziale per verificare l'argomento inserito dall'utente e successivamente due cicli `for` per popolare la matrice:

```
function hilbert_matrix = mat_hilbert(n)

if n < 0
    error("n can't be 0 or less than zero")
end

hilbert_matrix = zeros(n);
for i = 1 : n
    for j = 1 : n
        hilbert_matrix(i, j) = 1 / (i + j - 1);
    end
end
```

Il risultato:

```
mat_hilbert(5)

ans =

    1.0000    0.5000    0.3333    0.2500    0.2000
    0.5000    0.3333    0.2500    0.2000    0.1667
    0.3333    0.2500    0.2000    0.1667    0.1429
    0.2500    0.2000    0.1667    0.1429    0.1250
    0.2000    0.1667    0.1429    0.1250    0.1111

hilb(5)

ans =

    1.0000    0.5000    0.3333    0.2500    0.2000
    0.5000    0.3333    0.2500    0.2000    0.1667
    0.3333    0.2500    0.2000    0.1667    0.1429
    0.2500    0.2000    0.1667    0.1429    0.1250
    0.2000    0.1667    0.1429    0.1250    0.1111
```

6.2 Zeri di funzione

6.2.1 Grafici di funzione

In MATLAB una funzione $f(x)$ viene memorizzata come un vettore. In particolare, il vettore y ottenuto valutando f nel vettore delle ascisse x. Per cui la rappresentazione della funzione $f(x)$ è di fatto la rappresentazione del vettore y contro il vettore x.

Per introdurre i concetti di funzione e grafici di funzione, si presentano qua di seguito alcuni esempi di caso d'uso.

Definire le seguenti variabili:

- x***: vettore di estremi*** 0 ***e*** 10 ***con passo*** 0.1
- $y = e^x + 1$

Il vettore delle ascisse x può essere costruito banalmente con il seguente costrutto:

```
x = [0 : 0.1 : 10];
```

Per quanto riguarda la **funzione**, si utilizza la keyword per indicare che f ha come input un valore (`x`) e rappresenta la funzione `exp(x)+1`. In questo caso, la funzione si dice anonima. Per dichiarare funzioni esplicite, si rimanda alla documentazione ufficiale.

```
f = @(x) exp(x) + 1
```

Una volta definita una **funzione**, per **valutarla in uno o più punti**, si utilizzerà banalmente la sintassi matematica:

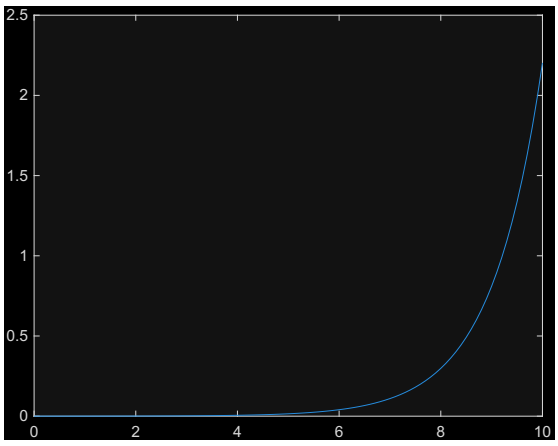
```
f(2)

ans =

    8.3891

f(0:3)

ans =

    2.0000    3.7183    8.3891   21.0855
```

Da notare che se l'argomento è un vettore, allora il risultato sarà un vettore della medesima lunghezza del vettore dato in input.

Utilizzando le variabili precedentemente definite, disegnare il grafico della funzione $y = e^x + 1$ nell'intervallo $[0, 10]$.

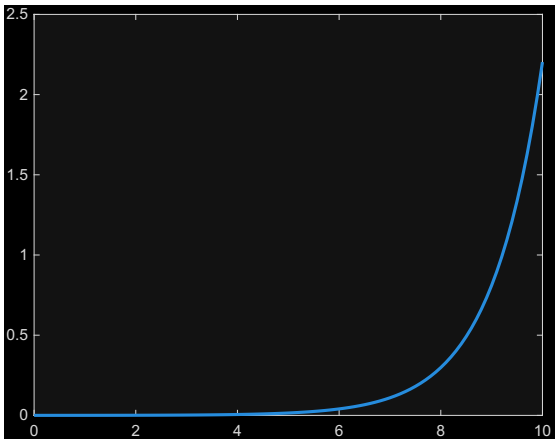
Per disegnare il grafico si utilizza il comando `plot`. Di default questa funzione disegna i valori in un piano cartesiano usando segmenti rettilinei (retta spezzata):

```
y = f(x);
plot(x, y)
```



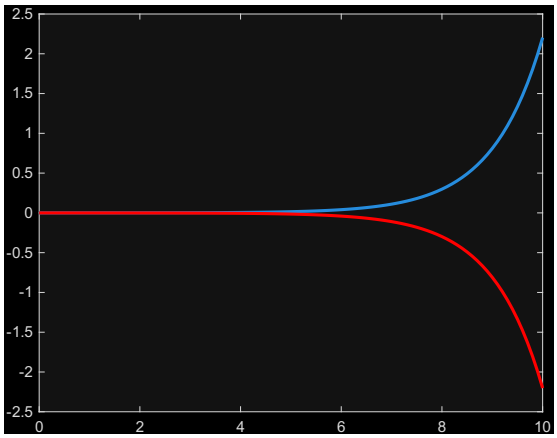
Per evitare che MATLAB sovrascriva la figura nella finestra aperta, è possibile numerarle usando la funzione `figure` (e.g. `figure(1); plot(x,y); figure(2); plot(0:3, 0:3)`).

La funzione `plot` accetta determinati valori per modificare il grafico finale. Nella documentazione ufficiale è possibile trovare l'intera lista e alcuni esempi. Scrivendo `plot(x, f(x), 'linewidth', 2)`, il parametro `'linewidth'` consente di definire lo spessore delle curve. Il valore che viene specificato in questo caso è 2 e il risultato:



Usando il comando `hold on` per fare un confronto tra i vari grafici e invocando di nuovo la funzione `plot` ma con parametri differenti, si ottiene:

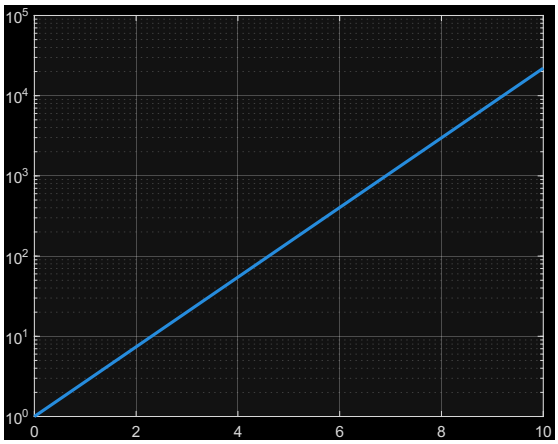
```
figure(1)
plot(x, f(x), 'linewidth', 2)
hold on
plot(x, -f(x), 'r', 'linewidth', 2)
```



Disegnare il grafico in scala semi-logaritmica (logaritmica solo per le ordinate) della funzione $y = e^x$ nell'intervallo $[0, 10]$. È possibile prevedere come sarà il grafico in scala semi-logaritmica della funzione $y = e^{2x}$? Verificare la risposta tracciando sulla medesima finestra le due funzioni utilizzando colori diversi per i due grafici.

Per disegnare il grafico in scala semi-logaritmica (logaritmica sulle ordinate) si utilizza il comando `semilogy` e si aggiunge anche la griglia:

```
semilogy(x, exp(x), 'linewidth', 2)
grid on
```

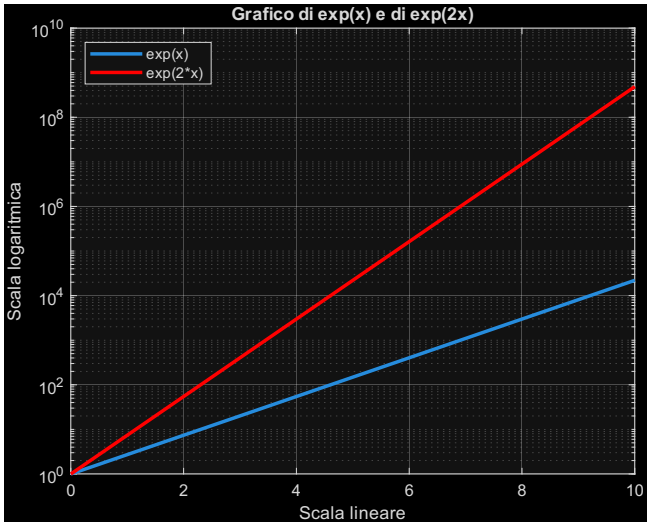


È una retta poiché $\log_{10}(y) = \log_{10}(e^x) = x \log_{10}(e)$.

- Il comando `semilogy` è l'equivalente di `plot` ma traccia un **grafico con l'asse delle ordinate in scala logaritmica**.
- Il comando `semilogx` traccia un **grafico con l'asse delle ascisse logaritmico**.
- Il comando `loglog` traccia un grafico in cui entrambi gli assi sono in scala logaritmica.

Passando alla risoluzione dell'esercizio, dato che $\log_{10}\left(e^{2x}\right) = 2x\log_{10}(e)$, disegnando in scala semi-logaritmica la funzione $y = e^{2x}$, si otterrà una retta con pendenza doppia rispetto alla retta precedentemente disegnata.

```
hold on
semilogy(x, exp(2 * x), 'r', 'linewidth', 2)
% oppure in un solo comando senza usare hold on
% semilogy(x, exp(x), 'b', x, exp(2*x), 'r', 'linewidth', 2)
title('Grafico di exp(x) e di exp(2x)')
xlabel('Scala lineare')
ylabel('Scala logaritmica')
grid on
legend('exp(x)', 'exp(2*x)', 'Location', 'NorthWest')
```



Il comando `legend` attribuisce alle curve disegnate da `plot` le stringhe di testo che gli vengono passate. Attenzione che alcune stringhe, come `'Location'` e `'NorthWest'`, vengono interpretate dalla funzione come comandi veri e propri. In questo caso si chiede di inserire una legenda in alto a sinistra.

6.3 Risoluzione di Sistemi di Equazioni Lineari

6.3.1 Metodi diretti

Si consideri la matrice di dimensione $n \times n$:

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & -1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix}$$

E $\mathbf{b}$ il vettore di dimensione n:

$$\mathbf{b} = [2, 0, 0, \ldots, 0]^T$$

1. Si ponga $n = 20$ e si assegnino in MATLAB la matrice A e il vettore dei termini noti $\mathbf{b}$.

```
n = 20;
% crea il vettore colonna composto da soli uno
R = ones(n, 1);
% crea una matrice di valori negativi (-1)
% sulla diagonale principale
% e sommala alla matrice creata come:
% - vettore R specificando il range per evitare
%   una matrice troppo grande;
% - -1 per indicare un livello sotto la diagonale principale:
A = -diag(R) + diag(R(1:n-1), -1);
% si riempi la prima riga della matrice A di uni
A(1, :) = 1;
% si crea un altro vettore di zeri
b = zeros(n, 1);
% e si sostituisce il primo valore con un 2
b(1) = 2;
```

   La funzione `diag` ha un parametro particolare, vedi la documentazione.

2. Si calcoli la fattorizzazione LU della matrice A, mediante la funzione MATLAB `lu`. Verificare che la tecnica del pivoting non è stata usata in questo caso.

```
% la funzione lu puo' essere utilizzata nel seguente modo
[L, U, P] = lu(A);
% in cui L ed U sono la fattorizzazione,
% mentre la matrice P indica la matrice permutazione.
% quest'ultima potrebbe avere delle permutazioni sulle righe
% dovute alla tecnica del pivoting.
% per controllare si puo' procedere controllando manualmente,
% quindi stampando la matrice P e controllare che sia
% una matrice identita',
% oppure invocare la funzione eye e verificare che siano
% uguali con un if statement
if P == eye(n)
    disp('pivoting effettuato!');
end
```

   Si veda a pagina 22 la spiegazione della matrice di permutazione.

3. Scrivere una funzione MATLAB `fwsub.m` che, dati in ingresso una matrice triangolare inferiore $L \in \mathbb{R}^{n \times n}$ e un vettore $\mathbf{f} \in \mathbb{R}^n$ restituisca in uscita il vettore $\mathbf{x}$, soluzione del sistema $L\mathbf{x} = \mathbf{f}$, calcolata mediante l'algoritmo della sostituzione in avanti (*forward substitution*). L'intestazione della funzione sarà ad esempio: `[x] = fwsub(L, f)`.

   Analogamente, scrivere la funzione `bksub.m` che implementi l'algoritmo della sostituzione all'indietro (*backward substitution*) per matrici triangolari superiori (U). Per controllare che le matrici L e U passate a `fwsub.m` e `bksub.m` siano effettivamente triangolari, è possibile utilizzare i comandi MATLAB `triu` e `tril` che, data una matrice, estraggono rispettivamente la matrice triangolare superiore e la matrice triangolare inferiore.

   Per creare una funzione, in MATLAB viene utilizzata la seguente sintassi:

```
function output_params = function_name(input_params)
    % Statements
end
```

   Introdotta la sintassi, si introduce il codice della funzione `fwsub.m`:

```
% si dichiara la funzione fwsub, che ha come input A e b
% e restituisce come output x
function x = fwsub(A,b)

    % ~ algoritmo di sostituzione in avanti ~
    % A: matrice quadrata triangolare inferiore
    % b: termine noto
    % x: soluzione del sistema Ax=b

    % si controlla che la matrice sia quadrata e
    % per farlo si calcola le dimensioni di b
    n = length(b);
    % se il numero di righe (size(A,1)) della matrice A
    % o se il numero di colonne (size(A,2)) della matrice A
    % sono diverse da n, allora le dimensioni non sono ammesse
    if (size(A, 1) ~= n || size(A, 2) ~= n)
        error("Dimensioni non ammesse");
    end

    % inoltre si controlla che la matrice sia una matrice
    % triangolare inferiore;
    % si utilizza la funzione tril per ottenere la
    % matrice triangolare inferiore
    if (A ~= tril(A))
        error("La matrice non e' triangolare inferiore");
    end

    % infine, si controllare che la matrice sia NON singolare,
    % ovvero che il determinante deve essere diverso da zero
    if (det(A) == 0)
        error("La matrice e' singolare");
    end

    % adesso l'algoritmo puo' iniziare;
    % si inizializza una matrice risultato, ovvero x,
    % nella quale verranno salvati i risultati
    x = zeros(n,1);
    % si applica la formula della sostituzione in avanti,
    % prima per la posizione (1,1)
    x(1) = b(1) / A(1,1);
    % e dopodiche' per tutte le posizioni della matrice
```

```
    for i = 2:n
        x(i) = (b(i) - A(i, 1:i-1) * x(1:i-1)) / A(i,i);
    end
```

Analogamente, si presenta il codice della funzione `bksub.m`:

```
% la funzione bksub avra' la stessa signature della funzione
% fwsub, ma la formula chiaramente sara' differente
function x = bksub(A, b)

    % ~ algoritmo di sostituzione all'indietro ~
    % A: matrice quadrata triangolare superiore
    % b: termine noto
    % x: soluzione del sistema Ax = b

    % 1. si esegue lo stesso controllo della funzione
    %    fwsub, si verifica che A sia quadrata
    n = length(b);
    if (size(A, 1) ~= n || size(A, 2) ~= n)
        error("Dimensioni non ammesse");
    end

    % 2. si controlla che sia effettivamente una
    %    matrice triangolare superiore,
    %    usando questa volta la funzione triu
    if (A ~= triu(A))
        error("La matrice non e' triangolare superiore");
    end

    % 3. l'ultimo controllo riguarda la "non singolarita'"
    %    ovvero, determinante diverso da zero
    if (det(A) == 0)
        error("La matrice e' singolare");
    end

    % 4. si parte con l'algoritmo e per farlo si inizia
    %    con l'ultima posizione;
    %    ovviamente si inizializza la matrice risultato
    x = zeros(n,1);
    x(n) = b(n) / A(n,n);
    % 5. si ricorda la sintassi del for statement
    %    initVal : step : endVal
    %    l'indice parte con un valore uguale a initVal,
    %    incrementa o decrementa a seconda dello step,
    %    termina quando raggiunge la condizione endVal
    for i = n-1 : -1 : 1
        x(i) = (b(i) - A(i, i+1:n) * x(i+1:n)) / A(i,i);
    end
```

4. Risolvere numericamente, utilizzando le funzioni `fwsub.m` e `bksub.m` implementate al punto precedente, i due sistemi triangolari necessari per ottenere la soluzione del sistema di partenza $A\mathbf{x} = \mathbf{b}$ mediante la fattorizzazione LU.

   Si utilizza la tecnica del pivoting e l'equazione 27 a pagina 22:

```
y = fwsub(L, P*b);
x = bksub(U, y);
```

5. Si calcoli la norma 2 dell'errore relativo

$$||\mathbf{err_{rel}}|| = \frac{||\mathbf{x} - \widehat{\mathbf{x}}||}{||\mathbf{x}||}$$

E la norma 2 del residuo normalizzata:

$$||\mathbf{r}|| = \frac{||\mathbf{b} - A\widehat{\mathbf{x}}||}{||\mathbf{b}||}$$

Sapendo che la soluzione esatta è il vettore di componenti:

$$\mathbf{x}(i) = \frac{2}{n} \qquad i = 1, \ldots, n$$

Si commenti il risultato ottenuto basandosi sul valore del numero di condizionamento della matrice A (si utilizzino i comandi `norm` e `cond`).

Il comando `norm` è stato spiegato a pagina 90.

```
x_ex = 2 / n * ones(n, 1);

err_rel = norm(x_ex - x) / norm(x_ex)
% il risultato: 5.1554e-16

r_nor = norm(b - A*x) / norm(b)
% il risultato: 2.7318e-16

K = cond(A)
% il risultato: 28.4998
```

6. Si ripeta il punto precedente per $n = 10, 20, 40, 80, 160$. Si rappresentino su un grafico in scala semi-logaritmica gli andamenti dell'errore relativo, del residuo normalizzato (si usa dire residuo normalizzato per la norma normalizzata del residuo) e del numero di condizionamento in funzione di n. Commentare il grafico ottenuto.

```
% si crea il vettore n
N = [10 20 40 80 160];
% e si inizializzano le variabili
K = [];
err_rel = [];
r_nor = [];

% per ogni n, si applicano i pezzi di codice precedenti
for n = N
    R = ones(n, 1);
    A = -diag(R) + diag(R(1:n-1), -1);
    A(1, :) = 1;

    b = zeros(n, 1);
    b(1) = 2;

    [L, U, P] = lu(A);

    y = fwsub(L, P*b);
    x_1 = bksub(U, y);

    x_ex = 2 / n * ones(n, 1);
    err_rel = [err_rel; norm(x_ex - x_1) / norm(x_ex)];
    r_nor = [r_nor; norm(b - A*x_1) / norm(b)];
```

```
25      K = [K; cond(A)];
26  end
27
28  % Semilog plot (y-axis has log scale)
29  semilogy(N, err_rel, '-s', N, r_nor, '-o', N, K, '-x')
30  legend('errore rel.', 'residuo norm.', 'n. di condizionamento'
        )
31  xlabel('dimensione n')
32  ylabel('err, r, K')
33  grid on
```

La seguente figura mostra l'andamento dell'errore relativo, del residuo normalizzato e del numero di condizionamento in funzione di n, in scala semi-logaritmica. Si noti che sia il residuo normalizzato sia l'errore relativo sono molto piccoli, dall'ordine di 10^{-16}, conseguenza del fatto che il numero di condizionamento $K(A)$ è in questo caso relativamente piccolo.

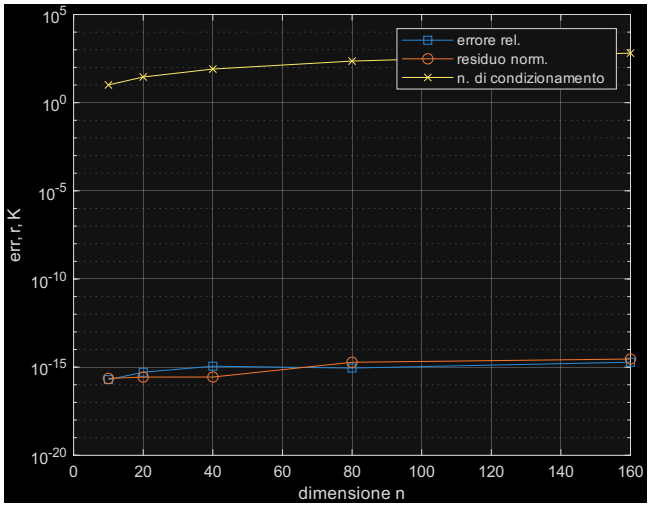


Figura 4: Andamento dell'errore relativo, del residuo normalizzato e del numero di condizionamento in funzione di n.

6.3.2 Metodi iterativi

I metodi iterativi stazionari sono considerati in genere nella seguente forma:

$$\mathbf{x}^{(k+1)} = B\mathbf{x}^{(k)} + \mathbf{f} \qquad k \geq 0$$

Dove B è detta matrice di iterazione. B e $\mathbf{f}$ identificano il metodo.

6.3.2.1 Metodo di Jacobi

Si consideri la matrice diagonale D degli elementi diagonali di A. Tale matrice è facilmente invertibile, se gli $a_{ii} \neq 0, i = 1, \ldots, n$, in quanto:

$$D = \begin{pmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ 0 & a_{22} & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & a_{nn} \end{pmatrix} \Longrightarrow D^{-1} = \begin{pmatrix} \frac{1}{a_{11}} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{a_{22}} & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & \frac{1}{a_{nn}} \end{pmatrix}$$

E il metodo può essere scritto direttamente in forma matriciale:

$$\mathbf{x}^{(0)} \text{ assegnato}$$
$$\mathbf{x}^{(k+1)} = B_J\mathbf{x}^{(k)} + \mathbf{f}_J$$

Dove $B_J = I - D^{-1}A = D^{-1}(D - A)$ è la matrice di iterazione di Jacobi e $\mathbf{f}_J = D^{-1}\mathbf{b}$

6.3.2.2 Metodo di Gauss-Seidel

Questo metodo si differenza dal metodo di Jacobi per il fatto che considera, oltre alla matrice D, anche le due matrici $-E$ e $-F$ triangolari superiore e inferiore della matrice A, ovvero:

$$-E = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ a_{21} & 0 & 0 & \cdots & 0 \\ \vdots & a_{32} & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ a_{n1} & a_{n2} & \cdots & a_{nn-1} & 0 \end{bmatrix} \qquad -F = \begin{bmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & 0 & a_{23} & \cdots & a_{2n} \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & a_{n-1n} \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

Dunque, il seguente algoritmo o le seguenti istruzioni:

$$\mathbf{x}^{(0)} \text{ assegnato}$$
$$\mathbf{x}^{(k+1)} = B_{GS}\mathbf{x}^{(k)} + \mathbf{f}_{GS}$$

Dove $B_{GS} = (D - E)^{-1} F$ è la matrice d'iterazione di Gauss-Seidel e $\mathbf{f}_{GS} = (D - E)^{-1}\mathbf{b}$.

6.3.2.3 Esercizio

Si considerino la matrice:

$$A = \begin{bmatrix} 9 & -3 & 1 & & & & \\ -3 & 9 & -3 & 1 & & & \\ 1 & -3 & 9 & -3 & 1 & & \\ & 1 & -3 & 9 & -3 & 1 & \\ & & 1 & -3 & 9 & -3 & 1 \\ & & & 1 & -3 & 9 & -3 \\ & & & & 1 & -3 & 9 \end{bmatrix}$$

E il termine noto:

$$\mathbf{b} = \begin{bmatrix} 7 & 4 & 5 & 5 & 5 & 4 & 7 \end{bmatrix}^T$$

1. Costruire la matrice A (utilizzando i comandi Matlab `diag` e `ones`) e determinare il numero di elementi non nulli tramite il comando `nnz`. La matrice A è a dominanza diagonale per righe? È simmetrica e definita positiva?

   È stato richiesto di utilizzare i comandi `diag` e `ones` per costruire la matrice A. Quindi per farlo, si controlla rapidamente la documentazione dei due comandi:

```
>> help ones
 ones - Create array of all ones
    This MATLAB function returns the scalar 1.

    Syntax
      X = ones
      X = ones(n)
      X = ones(sz1,...,szN)
      X = ones(sz)

      X = ones(___,typename)
      X = ones(___,'like',p)

    Input Arguments
      n - Size of square matrix
        integer value
      sz1,...,szN - Size of each dimension
        two or more integer values
      sz - Output size
        row vector of integer values
      typename - Output class
        'double' (default) | 'single' | 'logical' | 'int8' | '
    uint8' | ...
      p - Prototype
        variable

>> help diag
 diag - Create diagonal matrix or get diagonal elements of
    matrix
    This MATLAB function returns a square diagonal matrix with
     the elements
    of vector v on the main diagonal.

    Syntax
      D = diag(v)
      D = diag(v,k)
```

```
34
35      x = diag(A)
36      x = diag(A,k)
37
38    Input Arguments
39      v - Diagonal elements
40        vector
41      A - Input matrix
42        matrix
43      k - Diagonal number
44        integer
```

Adesso si osservi la matrice A. È possibile notare che la diagonale principale ha tutti i valori uguale a 9, mentre sopra e sotto la diagonale principale, altre due diagonali con valore pari a -3 e allo stesso modo due diagonali con valore uguale a 1.

Usando entrambi i comandi, si può giungere al seguente risultato parziale:

```
>> diag(9*ones(1, n)) + diag(-3*ones(1,n-1), 1) + diag(1*ones
    (1,n-2), 2)

ans =

         9    -3     1     0     0     0     0
         0     9    -3     1     0     0     0
         0     0     9    -3     1     0     0
         0     0     0     9    -3     1     0
         0     0     0     0     9    -3     1
         0     0     0     0     0     9    -3
         0     0     0     0     0     0     9
```

Ed eseguendo con la stessa logica anche sotto la diagonale principale, si ottiene la matrice A richiesta:

```
n = 7;
A = diag(9*ones(1, n)) + ... % diagonale principale
    diag(-3*ones(1,n-1), 1) + diag(-3*ones(1,n-1), -1) + ...
    diag(1*ones(1,n-2), 2) + diag(1*ones(1,n-2), -2)

% ans =
%
%      9    -3     1     0     0     0     0
%     -3     9    -3     1     0     0     0
%      1    -3     9    -3     1     0     0
%      0     1    -3     9    -3     1     0
%      0     0     1    -3     9    -3     1
%      0     0     0     1    -3     9    -3
%      0     0     0     0     1    -3     9
```

Il numero di elementi non nulli si calcola con il comando `nnz`:

```
>> help nnz
 nnz - Number of nonzero matrix elements
    This MATLAB function returns the number of nonzero
    elements in matrix X.

    Syntax
      N = nnz(X)

    Input Arguments
      X - Input matrix
```

```
        matrix

>> nnz(A)

ans =

    29
```

E infine, per confermare che la matrice sia a dominanza diagonale per righe, simmetrica e definita positiva:

```
% applicando la definizione
n = 7;
A = diag(9*ones(1, n)) + ... % diagonale principale
    diag(-3*ones(1,n-1), 1) + diag(-3*ones(1,n-1), -1) + ...
    diag(1*ones(1,n-2), 2) + diag(1*ones(1,n-2), -2);
A_diag = diag(abs(A));
A_no_diag = diag(-3*ones(1,n-1), 1) + ...
            diag(-3*ones(1,n-1), -1) + ...
            diag(1*ones(1,n-2), 2) + diag(1*ones(1,n-2), -2);
% definizione:
% https://it.wikipedia.org/wiki/Matrice_a_diagonale_dominante
for index = n
    if not(A_diag(index) >= sum(abs(A_no_diag(index:index, 1:n
    ))))
        error("La matrice non e' a diagonale dominante")
    end
end

% oppure in modo piu' rapido:
if (2 * abs(diag(A)) - sum(abs(A), 2) > 0)
    disp("La matrice e' a diagonale dominante")
end

% una matrice e' simmetrica se e' uguale alla sua transposta
if not(A == transpose(A))
    disp("La matrice non e' simmetrica")
end

% il metodo piu' efficiente per controllare se una matrice
% e' simmetrica e definita positiva, e' con l'utilizzo
% della fattorizzazione di Cholesky
% (studiata nella parte di teoria)
% source: https://rb.gy/uko7gs
try chol(A);
    disp("La matrice e' simmetrica e definita positiva")
catch ME
    error("La matrice non e' simmetrica definita positiva")
end
```

2. Si calcolino le matrici di iterazione:

$$B_J = D^{-1}(D - A)$$
$$B_{GS} = (D - E)^{-1} F$$

Associate rispettivamente ai metodi di Jacobi e Gauss-Seidel e i relativi raggi spettrali. La condizione necessaria e sufficiente per la convergenza del metodo iterativo è soddisfatta in entrambi i casi?

Le matrici di iterazione dei due metodi si calcolano a partire dalla definizione:

```
D = diag(diag(A));
Bj = D\(D-A);   % matrice di iterazione di Jacobi

E=-tril(A,-1);
F=-triu(A,1);
Bgs=(D-E)\F; % matrice di iterazione di Gauss-Seidel

rho_j = max(abs(eig(Bj)))
rho_gs = max(abs(eig(Bgs)))
```

Si noti l'istruzione `D = diag(diag(A))`; il comando interno estrae la diagonale principale di A, restituendo un vettore, il quale viene elaborato dal comando più esterno che crea una seconda matrice quadrata identica alla dimensione di A ma con solo la diagonale principale.

Dal calcolo del raggio spettrale delle matrici si può concludere che in questo caso entrambi i metodi convergono, in quanto l'autovalore massimo risulta in modulo strettamente minore di 1. Si osservi che il raggio spettrale della matrice di iterazione del metodo di Gauss-Seidel è più basso di quello della matrice del metodo di Jacobi.

3. Scrivere la funzione Matlab che implementi il metodo di Jacobi inversione *matriciale* per il sistema lineare $A\mathbf{x} = \mathbf{b}$. L'intestazione della funzione sarà la seguente:

$$\texttt{[x,k] = jacobi(A,b,x0,toll,nmax).}$$

Il processo iterativo si arresta quando:

$$\frac{||\mathbf{r}^{(k)}||}{||\mathbf{b}||} \leq \texttt{toll}$$

(criterio d'arresto del residuo normalizzato).

```
function [x,k]=jacobi(A,b,x0,toll,nmax)

% Metodo di Jacobi
%
% A: matrice del sistema
% b: termine noto
% x0: vettore iniziale
% toll: tolleranza sul residuo normalizzato
% nmax: massimo numero di iterazioni
%
% x: soluzione ottenuta
% k: numero di iterazioni effettuate

n = size(b,1);

% Controlliamo che la matrice A sia quadrata e che,
% insieme al guess iniziale x0,
% abbia dimensioni compatibili con b.
if ((size(A,1) ~= n) || (size(A,2) ~= n) || (size(x0,1) ~= n))
  error('Dimensioni incompatibili')
end

% Controlliamo che la matrice A non abbia elementi
% diagonali nulli.
if (prod(diag(A)) == 0)
```

```
  error('res_normore: elementi diagonali nulli')
end

% Estraiamo la matrice D da A e calcoliamo la matrice
% d'iterazione e il termine noto g
D = diag(diag(A));
Bj = eye(n) - D\A;
g = D\b;

%Inizializziamo x come x0, calcoliamo il residuo
% e l'res_normore normalizzato
x = x0;
r = b - A*x;
res_norm = norm(r) / norm(b);

% Inizializziamo l'indice d'iterazione
k = 0;

while (res_norm > toll && k < nmax)
    k = k + 1;

    % Calcoliamo il nuovo x
    x=Bj*x+g;

    % Calcoliamo residuo e res_normore
    r = b - A*x;
    res_norm = norm(r)/norm(b);
end
```

4. Scrivere una funzione Matlab che implementi il metodo di Gauss-Seidel inversione *matriciale* per il sistema lineare $A\mathbf{x} = \mathbf{b}$. L'intestazione della funzione sarà la seguente:

`[x,k] = gs(A,b,x0,toll,nmax).`

```
function [x,k]=gs(A,b,x0,toll,nmax)

% Metodo di Gauss-Seidel
%
% A: matrice del sistema
% b: termine noto
% x0: vettore iniziale
% toll: tolleranza sul residuo normalizzato
% nmax: massimo numero di iterazioni
%
% x: soluzione ottenuta
% k: numero di iterazioni effettuate

n = size(b,1);

% Controlliamo che la matrice A sia quadrata e che,
% insieme al guess iniziale x0,
% abbia dimensioni compatibili con b.
if (( size(A,1)~=n) || (size(A,2)~=n) || (size(x0,1) ~= n) )
  error('dimensioni incompatibili')
end

% Controlliamo che la matrice A non abbia
% elementi diagonali nulli.
if (prod(diag(A)) == 0)
    error('errore: elementi diagonali nulli')
```

```
end

% Decomponiamo la matrice in D,E e F,
% e calcoliamo la matrice d'iterazione e il termine g
D=diag(diag(A));
E=-tril(A,-1);
F=-triu(A,1);
Bgs=(D-E)\F;
g=(D-E)\b;

% Inizializziamo x come x0, calcoliamo il residuo
% e l'errore normalizzato
x = x0;
r = b - A * x;
err = norm(r) / norm(b);

% Inizializziamo l'indice d'iterazione
k = 0;

while ( err > toll && k < nmax )
  k = k + 1;

  % Calcoliamo il nuovo x
  x=Bgs*x+g;

  % Calcoliamo residuo e errore
  r = b - A*x;
  err = norm(r)/norm(b);
end
```

5. Costruire il termine noto $\mathbf{b}$. Utilizzando le funzioni costruite nei punti 3 e 4, risolvere il sistema $A\mathbf{x} = \mathbf{b}$ ponendo $x^{(0)} = [0, 0, \ldots, 0]^T$, `toll` $= 10^{-6}$ e `nmax` $= 1000$. Confrontare il numero di iterazioni necessarie per arrivare a convergenza per i due metodi e commentare i risultati ottenuti.

   Il metodo di Gauss-Seidel converge più velocemente alla soluzione esatta in accordo con il corrispondente raggio spettrale che è più basso di quello della matrice del metodo di Jacobi.

```
b = transpose([7 4 5 5 5 4 7]);
toll = 1e-6;
x0 = zeros(n, 1);
nmax = 1000;

[xJ, kJ] = jacobi(A, b, x0, toll, nmax);
[xGS, kGS] = gs(A, b, x0, toll, nmax);


xJ
kJ
xGS
kGS

% xJ =
%
%      1.0000
%      1.0000
%      1.0000
%      1.0000
%      1.0000
%      1.0000
```

```
%      1.0000
%
%
% kJ =
%
%      49
%
%
% xGS =
%
%      1.0000
%      1.0000
%      1.0000
%      1.0000
%      1.0000
%      1.0000
%      1.0000
%
%
% kGS =
%
%      12
```

6.3.2.4 Metodo di Richardson

Il metodo di Richardson stazionario è basato sulla seguente legge. Dati $\mathbf{x}^{(0)}$ e $\alpha \in \mathbb{R}$ si calcola:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha \mathbf{r}^{(k)} \qquad k \geq 0 \tag{153}$$

Dove $\alpha \neq 0$ è un parametro costante per ogni iterazione. Questo metodo richiede, ad ogni passo k, di calcolare il *residuo* $\mathbf{r}^{(k)}$ definito come:

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

Il metodo di Richardson *stazionario* converge solo per $0 < \alpha < \frac{2}{\lambda_{\max}}$, in cui $\lambda_{\max}$ è il massimo degli autovalori della matrice A. Inoltre, è possibile calcolare un valore di α ottimale che massimizza la velocità di convergenza. Questo valore è dato da:

$$\alpha_{\text{opt}} = \frac{2}{\lambda_{\min} + \lambda_{\max}}$$

In cui $\lambda_{\min}$ è il minimo degli autovalori della matrice A. Per questo valore, la velocità di convergenza è data da:

$$\rho_{\text{opt}} = \frac{K(A) - 1}{K(A) + 1}$$

Dove $K(A)$ è il numero di condizionamento, definito anche come:

$$K(A) = \left|\left|A^{-1}\right|\right| \cdot ||A||$$

Dove $||\cdot||$ è una opportuna norma introdotta per la matrice. Vale sempre $K(A) \geq 1$. Se la matrice A è simmetrica e definita positiva, utilizzando la sua norma 2, vale:

$$K(A) = \left|\left|A^{-1}\right|\right|_2 \cdot ||A||_2 = \frac{\lambda_{\max}}{\lambda_{\min}}$$

Si noti che, dalla sua definizione, il metodo di Richardson (eq. 153) può essere riscritto nella seguente forma (utilizzando la definizione di $\mathbf{r}$):

$$\mathbf{x}^{(k+1)} = (I - \alpha A)\mathbf{x}^{(k)} + \alpha \mathbf{b} \qquad k \geq 0 \tag{154}$$

Segue che il metodo di Richardson è un metodo iterativo caratterizzato dalla matrice di iterazione $B_\alpha = I - \alpha A$ e da $\mathbf{f} = \alpha \mathbf{b}$.

6.3.2.5 Precondizionamento

Il numero di condizionamento di una matrice governa il rapporto tra l'errore relativo commesso dalla soluzione numerica nella risoluzione di un sistema lineare e il corrispondente residuo normalizzata, alla iterata k:

$$\frac{\left|\left|\mathbf{x}^{(k)} - \mathbf{x}\right|\right|}{\left|\left|\mathbf{x}^{(k)}\right|\right|} \leq K(A) \frac{\left|\left|\mathbf{r}^{(k)}\right|\right|}{||\mathbf{b}||}$$

Dove:

- $\mathbf{x}^{(k)}$ è la soluzione numerica
- $\mathbf{r}^{(k)}$ è il residuo, ovvero $\mathbf{b} - A\mathbf{x}^{(k)}$
- $\mathbf{x}$ è la soluzione esatta del sistema lineare

Inoltre, nel caso del metodo di Richardson, si ottiene:

$$\left|\left|\mathbf{e}^{(k+1)}\right|\right| \leq \frac{K(A) - 1}{K(A) + 1} \cdot \left|\left|\mathbf{e}^{(k)}\right|\right|$$

Come migliore stima ottenibile usando α_{opt}.

Per problemi ben condizionati ($K(A)$ non molto più grande di 1), la soluzione del problema con piccoli residui non differisce molto dalla soluzione del problema originale; al contrario, in problemi con la matrice mal condizionata ($K(A) \gg 1$) a piccoli residui possono corrispondere grandi errori e la convergenza è molto lenta.

L'idea del precondizionamento consiste nel cercare di ridurre il numero di condizionamento della matrice del sistema, pre-moltiplicandola per una matrice P^{-1} (P è chiamata *precondizionatore*). Si ottiene così il sistema equivalente:

$$P^{-1}A\mathbf{x}^{(k)} = P^{-1}\mathbf{b}$$

Ovviamente il precondizionatore è efficace se $K\left(P^{-1}A\right) \ll K(A)$ e se la soluzione del sistema lineare in P che sorge ad ogni iterazione non è troppo onerosa. La prima proprietà è solitamente verificata quando $P^{-1} \approx A^{-1}$, cioè quando P e A hanno uno spettro simile, ma dovendo tener conto della seconda condizione è opportuno scegliere P con una struttura speciale che mantenga basso il costo computazionale (ad esempio diagonale o triangolare).

Ad esempio, nel caso di Richardson, per alleviare la dipendenza della convergenza ottimale dal numero di condizionamento, si introduce la tecnica di precondizionamento, che consiste nel sostituire A con $P^{-1}A$. Questo metodo richiede, ad ogni iterata, di trovare il cosiddetto *residuo precondizionato* $\mathbf{z}^{(k)}$ dato dalla soluzione del sistema lineare:

$$P\mathbf{z}^{(k)} = \mathbf{r}^{(k)} \tag{155}$$

Di conseguenza, la nuova iterata è definita da $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha\mathbf{z}^{(k)}$. Si noti che la matrice P, oltre ad essere non singolare (determinante diverso da zero),

simmetrica (coincide con la sua trasposta) e definita positiva, deve essere scelta in modo tale che il costo computazionale richiesto dalla risoluzione del sistema (eq. 155) sia basso. Per il caso precondizionato, valgono i precedenti risultati su α_{opt} e ρ_{opt} a patto che si considerino gli autovalori di $P^{-1}A$ invece di quelli di A.

6.3.2.6 Metodo del gradiente

È possibile generalizzare il metodo di Richardson precondizionato tramite l'introduzione di un parametro di accelerazione dinamico α_k:

$$P\left(\mathbf{x}^{(k+1)-\mathbf{x}^{(k)}}\right) = \alpha_k \mathbf{r}^{(k)} \qquad k \geq 0,\ \alpha_k \neq 0$$

Con P il precondizionatore. Lo scopo dell'utilizzo di α_k è quello di poter calcolare facilmente il parametro di accelerazione evitando il calcolo (spesso oneroso) degli autovalori di A come per α_{opt}. Questo metodo è detto *metodo di Richardson dinamico*. Se α_k è scelto in modo ottimale, allora il metodo è detto del *gradiente* se $P = I$, oppure del *gradiente precondizionato* se $P \neq I$.

A partire dal metodo di Richardson, è possibile ottenere i metodi del *gradiente* ($P = I$) e del *gradiente precondizionato* ($P \neq I$) tramite l'aggiunta del parametro α_k: dato $\mathbf{x}^{(0)}$ assegnato, si ponga $\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}$ e per $k = 0, 1, \ldots$

$$\begin{aligned} P\mathbf{z}^{(k)} &= \mathbf{r}^{(k)} \\ \alpha_k &= \frac{\left(\mathbf{z}^{(k)}\right)^T \mathbf{r}^{(k)}}{\left(\mathbf{z}^{(k)}\right)^T A\mathbf{z}^{(k)}} \\ \mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} + \alpha_k \mathbf{z}^{(k)} \\ \mathbf{r}^{(k+1)} &= \mathbf{r}^{(k)} + \alpha_k A\mathbf{z}^{(k)} \end{aligned} \tag{156}$$

6.3.2.7 Esercizi su Richardson e gradiente

Si considerino la matrice:

$$A = \begin{bmatrix} 9 & -3 & 1 & & & & \\ -3 & 9 & -3 & 1 & & & \\ 1 & -3 & 9 & -3 & 1 & & \\ & 1 & -3 & 9 & -3 & 1 & \\ & & 1 & -3 & 9 & -3 & 1 \\ & & & 1 & -3 & 9 & -3 \\ & & & & 1 & -3 & 9 \end{bmatrix}$$

E il termine noto:

$$\mathbf{b} = \begin{bmatrix} 7 & 4 & 5 & 5 & 5 & 4 & 7 \end{bmatrix}^T$$

1. Costruire la matrice A (utilizzando i comandi Matlab `diag` e `ones`).

```
n = 7;
A = diag(9*ones(1, n)) + ... % diagonale principale
    diag(-3*ones(1,n-1), 1) + diag(-3*ones(1,n-1), -1) + ...
    diag(1*ones(1,n-2), 2) + diag(1*ones(1,n-2), -2)

% ans =
%
%      9    -3     1     0     0     0     0
%     -3     9    -3     1     0     0     0
%      1    -3     9    -3     1     0     0
%      0     1    -3     9    -3     1     0
%      0     0     1    -3     9    -3     1
%      0     0     0     1    -3     9    -3
%      0     0     0     0     1    -3     9
```

2. Calcolare l'intervallo di valori di α per cui il metodo di Richardson stazionario non precondizionato converge. Determinare il valore ottimale di α per avere massima velocità di convergenza.

```
% calcolo di eigenvalues e eigenvectors
% V: right eigenvectors (matrice quadrata)
% D: eigenvalues (matrice diagonale)
[V,D] = eig(A);

% si ottiene l'autovalore massimo
lambda_max = max(diag(D))
% lambda_max =
%
%    16.0403

% il limite (suppenendo lambda_min = 0)
Lim = 2/lambda_max
% Lim =
%
%     0.1247

% si calcola infine alpha opt
lambda_min = min(diag(D))
alpha_opt = 2/(lambda_min+lambda_max)
% lambda_min =
%
%     4.9042
%
```

```
%
% alpha_opt =
%
%     0.0955
```

3. Scrivere una funzione Matlab che implementi il metodo di Richardson stazionario non precondizionato per il sistema lineare $A\mathbf{x} = \mathbf{b}$. I parametri in ingresso richiesti dalla funzione sono la matrice A, il termine noto $\mathbf{b}$, il guess iniziale $\mathbf{x}^{(0)}$, il coefficiente α, la tolleranza per il criterio d'arresto `toll` e il numero massimo di iterazioni ammesse `nmax`. La funzione restituisce la soluzione numerica $\mathbf{x}$ e il numero di iterazioni effettuate k.

   L'intestazione della funzione sarà la seguente:

   $$[\text{x, k}] = \text{richardson(A, b, x0, alpha, toll, nmax)}$$

   Il processo iterativo si arresta quando:

   $$\frac{||\mathbf{r}^{(k)}||}{||\mathbf{b}||} \leq \texttt{toll}$$

   (criterio d'arresto del residuo normalizzato).

```
function [x, k, err] = richardson(A, b, x0, alpha, toll, nmax)

% Metodo di Richardson stazionario
%
% A: matrice del sistema
% b: termine noto
% x0: vettore iniziale
% alpha: coefficiente di Richardson
% toll: tolleranza sul residuo normalizzato
% nmax: massimo numero di iterazioni
%
% x: soluzione ottenuta
% it: numero di iterazioni effettuate

n = size(b,1);
k = 0;

if ((size(A,1) ~= n) || (size(A,2) ~= n) || (size(x0,1) ~= n))
  error('Dimensioni incompatibili')
end

x = x0;
r = b - A*x;
errk = norm(r) / norm(b);
err = errk;

while (errk > toll && k < nmax)
    k = k + 1;
    x = x + alpha*r;
    r = b - A*x;
    errk = norm(r)/norm(b);
    err = [err; errk];
end
```

4. Costruire il termine noto $\mathbf{b}$. Utilizzando la funzione scritta al punto precedente, determinare la soluzione del sistema lineare $A\mathbf{x} = \mathbf{b}$, con A data dal punto 1.

Si ponga $\mathbf{x}^{(0)} = [0, 0, \ldots, 0]^T$, `toll` $= 10^{-6}$ e `nmax` $= 1000$. Si verifichi sperimentalmente che il metodo di Richardson stazionario non precondizionato converge solo se α appartiene all'intervallo trovato nel punto 2. In particolare, si scelga un α al di fuori dell'intervallo e si verifichi che il metodo diverge. Si provi, inoltre, sperimentalmente, che per α_{opt} il metodo converge più velocemente. In particolare, si scelga un α nell'intervallo di convergenza e si osservi il numero di iterazioni per aggiungere a convergenza è maggiore di quello ottenuto utilizzando α_{opt}.

```
% calcolo della transposta
b = transpose([7 4 5 5 5 4 7]);

% si pongono i dati richiesti
x0 = zeros(n,1);
toll = 1e-6;
nmax = 1000;

% si utilizza prima un alpha fuori dal limite del punto 2
alpha = 2;
[xR1,kR1] = richardson(A,b,x0,alpha,toll,nmax)
% xR1 =
%
%   -Inf
%    Inf
%   -Inf
%    Inf
%   -Inf
%    Inf
%   -Inf
%
%
% kR1 =
%
%    208

% adesso ci si avvicina al valore alpha
alpha = 0.11;
[xR2,kR2] = richardson(A,b,x0,alpha,toll,nmax)
% xR2 =
%
%     1.0000
%     1.0000
%     1.0000
%     1.0000
%     1.0000
%     1.0000
%     1.0000
%
%
% kR2 =
%
%     45

% e infine si utilizza alpha opt per verificare
% che sia il migliore
[xR3,kR3] = richardson(A,b,x0,alpha_opt,toll,nmax)
% xR3 =
%
%     1.0000
%     1.0000
%     1.0000
```

```
%      1.0000
%      1.0000
%      1.0000
%      1.0000
%
%
% kR3 =
%
%      22
```

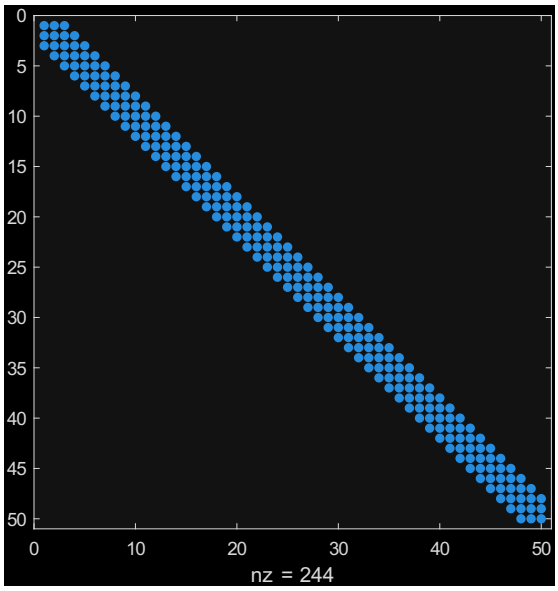
Adesso si passa al gradiente. Si consideri il problema lineare $A\mathbf{x} = \mathbf{b}$, dove la matrice $A \in \mathbb{R}^{n \times n}$ è pentadiagonale:

$$A = \begin{bmatrix} 4 & -1 & -1 & & & & \\ -1 & 4 & -1 & -1 & & & \\ -1 & 4 & -1 & -1 & -1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & -1 & -1 & 4 & -1 & -1 \\ & & & -1 & -1 & 4 & -1 \\ & & & & -1 & -1 & 4 \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} 0.2 \\ 0.2 \\ 0.2 \\ \vdots \\ 0.2 \\ 0.2 \\ 0.2 \end{bmatrix}$$

Si vuole risolvere tale problema con il metodo del gradiente, soddisfando una tolleranza di 10^{-5} per il criterio relativo al residuo normalizzato, a partire dal vettore iniziale $\mathbf{x}_0 = [0, \ldots, 0]^T$.

1. Costruire la matrice A con $n = 50$, il termine noto $\mathbf{b}$ ed il vettore soluzione iniziale $\mathbf{x}_0$. La matrice A è simmetrica e definita positiva? Se ne calcoli il numero di condizionamento, senza utilizzare il comando `cond` di Matlab.

```
n = 50;
A = diag(4*ones(n,1)) + ...
    diag(-ones(n-1,1),-1) + ....
    diag(-ones(n-1,1),1) + ...
    diag(-ones(n-2,1),-2) + ...
    diag(-ones(n-2,1),2);
% si visualizza la sparsity pattern della matrice
spy(A)
% si costruiscono i restanti vettori
b = 0.2*ones(n, 1);
x0 = zeros(n, 1);

% si controlla se e' simmetrica e definita positiva
% sfruttando la velocita' della fattorizzazione di Cholensky
try chol(A);
    disp("La matrice e' simmetrica definita positiva")
catch ME
    disp("La matrice non e' simmetrica definita positiva")
end
eigA = eig(A);
KA = max(eigA)/min(eigA)
% KA =
%
%   336.2412
```



2. Si scriva una funzione che implementi il metodo del gradiente e la si usi per risolvere il sistema lineare $A\mathbf{x} = \mathbf{b}$. L'intestazione della funzione sarà ad esempio:

```
[x, iter, err] = graddyn(A, b, x0, nmax, toll)
```

Dove `err` è il vettore contenente il residuo normalizzato ad ogni iterazione.

```
function [xn, iter, err] = graddyn (A, b, x0, nmax, tol)

% Metodo del gradiente per sistemi lineari
%
% Parametri di ingresso:
%   A       Matrice del sistema
%   b       Termine noto
%   x0      Vettore iniziale
%   nmax    Numero massimo di iterazioni
%   tol     Tolleranza sul test d'arresto
%
% Parametri in uscita
%   xn      Vettore soluzione
%   iter    Iterazioni effettuate
%   err     Vettore contenente gli errori relativi sul residuo

[n, m] = size (A);
if not(n == m)
  error ('matrice non quadrata')
end

iter = 0;
xn = zeros(n,1);

% Iterazioni
bnrm2 = norm (b);
r = b - A * x0;
errk = norm (r) / bnrm2;
err = errk;
```

```
xn = x0;

while (errk > tol) && (iter < nmax)
  z = r;
    transpose_z = transpose(z);
    alpha = transpose_z*r / (transpose_z*A*z);
    xn = xn + alpha * r;
    r = (eye(n)-alpha*A)*r;
    errk = norm (r) / bnrm2;
    err = [err errk];
%     xv = xn;
    iter = iter + 1;
end

if (iter == nmax)
    fprintf('Il metodo graddyn non converge in %d iterazioni \
    n', iter);
end
```

3. Utilizzare la funzione scritta al punto precedente per determinare la soluzione del sistema lineare $A\mathbf{x} = \mathbf{b}$.

```
% si impostano i valori
toll = 1e-5;
nmax = 10000;

% inizia il timer
tic
% si invoca il metodo del gradiente
[xGD, iterGD, errGD] = graddyn(A, b, x0, nmax, toll);
% si salva il tempo finale di esecuzione
timeGD=toc;
```

4. A partire dal `graddyn.m`, si scriva una funzione `gradprec.m` che implementi il metodo del gradiente precondizionato. L'intestazione della funzione sarà ad esempio:

   `[x, iter, err] = gradprec(A, b, x0, nmax, toll)`

   Si risolva il sistema lineare $A\mathbf{x} = \mathbf{b}$ applicando il metodo del gradiente precondizionato con il precondizionatore:

$$P = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}$$

```
function [xn, iter, err] = gradprec (A, b, P, x0, nmax, tol)

% Metodo del gradiente precondizionato
%
% Parametri di ingresso:
%   A       Matrice del sistema
%   b       Termine noto
%   P       Precondizionatore
%   x0      Vettore iniziale
```

```
%   nmax    Numero massimo di iterazioni
%   tol     Tolleranza sul test d'arresto
%
% Parametri in uscita
%   xn      Vettore soluzione
%   iter    Iterazioni effettuate
%   err     Vettore contenente gli errori relativi sul residuo

[n, m] = size (A);
if not(n == m)
  error ('matrice non quadrata')
end

iter = 0;
xn = zeros(n,1);

% Iterazioni
bnrm2 = norm (b);
r = b - A * x0;
errk = norm (r) / bnrm2;
err = errk;
xv = x0;

while (errk > tol) && (iter < nmax)
    z = P\r;
    transpose_z = transpose(z);
    alpha = transpose_z*r / (transpose_z*A*z);
    xn = xv + alpha * z;
    r = r-alpha*A*z;
    errk = norm (r) / bnrm2;
    err = [err errk];
    xv = xn;
    iter = iter + 1;
end

if (iter == nmax)
    fprintf('Il metodo gradprec non converge in %d iterazioni
    \n', iter);
end
```

```
P = diag(2*ones(n,1)) - ...
    diag(ones(n-1,1),-1) - ...
    diag(ones(n-1,1),1);

tic
[xPG, iterPG, errPG] = gradprec(A, b, P, x0, nmax, toll);
timePG=toc;
```

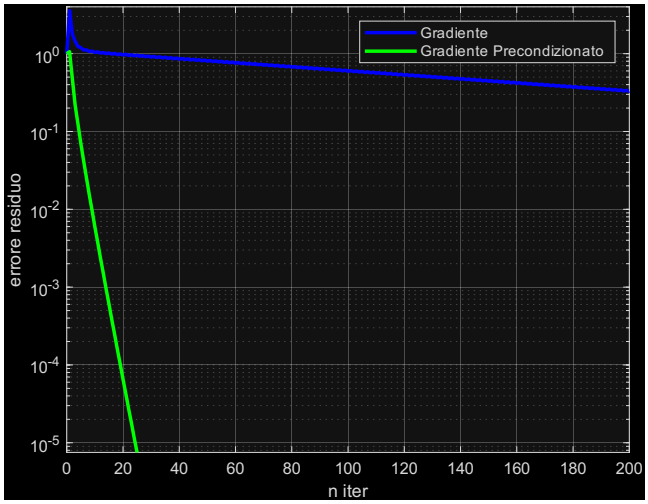
5. Disegnare su un grafico l'andamento del residuo normalizzato:

$$\frac{||\mathbf{r}^{(k)}||}{||\mathbf{b}||}$$

In funzione delle iterazioni k nei due casi (gradiente e gradiente precondizionato), e confrontare le curve ottenute.

```
figure
semilogy(...
    [0:iterGD], ...
    errGD, ...
    'b-', ...
```

```
    [0:iterPG], ...
    errPG, ...
    'g-', ...
    'Linewidth', ...
    2 ...
)
grid on
axis([0 200 0 4])
xlabel('n iter')
ylabel('errore residuo')
legend('Gradiente', 'Gradiente Precondizionato')
```



6. A partire da `richardson.m`, si scriva una funzione `richprec.m` che implementi il metodo di Richardson precondizionato. L'intestazione della funzione sarà ad esempio:

   ```
   [x, iter, err] = richprec(A, b, P, alpha, x0, nmax, toll)
   ```

   Si risolva il sistema lineare $A\mathbf{x} = \mathbf{b}$ applicando il metodo di Richardson precondizionato utilizzando la relativa α_{opt} e il precondizionatore P implementato al punto 4.

```
function [x, it, err] = richprec(A,b,P, alpha, x0,nmax, toll)

% Metodo di Richardson stazionario precondizionato
%
% A: matrice del sistema
% b: termine noto
% P: precondizionatore
% x0: vettore iniziale
% alpha: coefficiente di Richardson
% toll: tolleranza sul residuo normalizzato
% nmax: massimo numero di iterazioni
%
% x: soluzione ottenuta
% it: numero di iterazioni effettuate
% err: vettore contenente gli errori relativi sul residuo


```

```
n = size(b,1);
if ( ...
        not(size(A,1) == size (A,2)) || ...
        not(size(A,1) == n) || ...
        not(size(x0,1) == n) || ...
        not(size (P,2) == n) || ...
        not(size(P,1) == n) ...
)
    error('Dimesioni incompatibili')
end

it = 0;
x = x0;
r = b - A*x;
errk = norm(r)/norm(b);
err = errk;

% loop
while (it < nmax && errk > toll)
    it = it + 1;
    z = P\r;
    x = x + alpha*z;
    r = b - A*x;
    errk = norm(r)/norm(b);
    err = [err; errk];
end
```

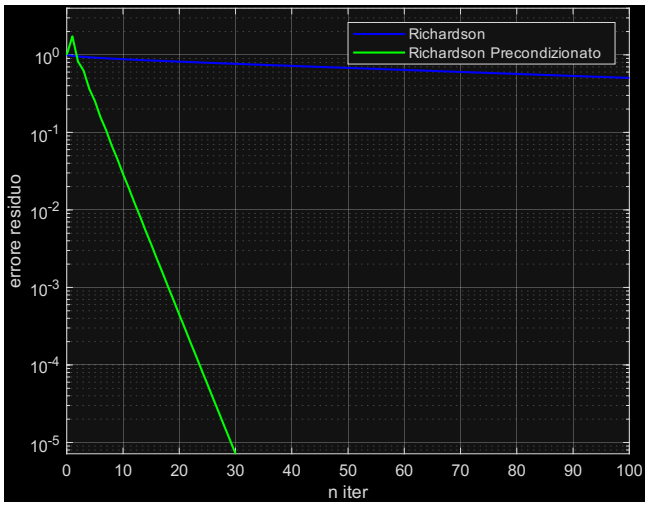
```
lambda_minRP = min(eig(P\A));
lambda_maxRP = max(eig(P\A));
alpha_optRP = 2/(lambda_maxRP + lambda_minRP);
tic
[xRP, iterRP, errRP] = richprec(A, b, P, alpha_optRP, x0, nmax
    , toll);
timeRP=toc;
```

7. Si risolva il sistema lineare $A\mathbf{x} = \mathbf{b}$ con il metodo di Richardson. Disegnare su un grafico l'andamento del residuo normalizzato:

$$\frac{||\mathbf{r}^{(k)}||}{||\mathbf{b}||}$$

In funzione delle iterazioni k nei due casi (Richardson e Richardson precondizionato), e confrontare le curve ottenute.

```
lambda_minR = min(eig(A));
lambda_maxR = max(eig(A));
alpha_optR = 2/(lambda_maxR + lambda_minR);

tic
[xR, iterR, errR] = richardson(A, b, x0, alpha_optR, toll,
    nmax);
timeR=toc;

figure
semilogy([0:iterR], errR, 'b-', [0:iterRP], errRP, 'g-', '
    Linewidth', 1.2)
grid on
axis([0 100 0 4])
xlabel('n iter')
ylabel('errore residuo')
legend('Richardson', 'Richardson Precondizionato')
```


Richardson
Richardson Precondizionato
errore residuo
n iter
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
0
10
20
30
40
50
60
70
80
90
100

6.4 Sistema di equazioni non lineari

6.4.1 Metodo di Newton

Si consideri il problema della ricerca degli zeri del sistema non lineare $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, dove $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$. Definendo:

- $\mathbf{x} = (x_1, \ldots, x_n)^T$
- $\mathbf{f} = (f_1, \ldots, f_n)^T$ con $f_1, \ldots, f_n$ funzioni $\mathbb{R}^n \to \mathbb{R}$

Il problema si può riscrivere nel seguente modo:

$$\begin{cases} f_1(x_1, \ldots, x_n) = 0 \\ f_2(x_1, \ldots, x_n) = 0 \\ \vdots \\ f_n(x_1, \ldots, x_n) = 0 \end{cases}$$

Il metodo di Newton per sistemi non lineari è il seguente. Dato $\mathbf{x}^{(0)} \in \mathbb{R}^n$, per $k \geq 0$:

1. Risolvere il sistema lineare:
$$J\left(\mathbf{x}^k\right) \delta\mathbf{x}^{(k)} = -\mathbf{f}\left(\mathbf{x}^{(k)}\right)$$
2. Ponendo:
$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \delta\mathbf{x}^{(k)}$$

Dove, per un generico punto $\mathbf{y}$, $J(\mathbf{y})$ è la matrice Jacobiana della funzione $\mathbf{f}$ e consiste in una matrice $\mathbb{R}^n \times \mathbb{R}^n$ le cui componenti sono:

$$J_{il}(\mathbf{y}) = \frac{\partial f_i(\mathbf{y})}{\partial y_l} \qquad i, l = 1, \ldots, n$$

Si osservi che:

1. L'applicazione del metodo di Newton richiede ad ogni iterazione la soluzione di un sistema lineare con matrice $A^{(k)} = J\left(\mathbf{x}^{(k)}\right)$.
2. È possibile dimostrare che se:
   (a) $\mathbf{x}^{(0)}$ è "sufficientemente" vicino alla soluzione α
   (b) $J\left(\mathbf{x}^{(k)}\right)$ è una matrice non singolare con $k = 0, 1, \ldots$

   Allora il metodo di Newton converge con ordine 2.

Si consideri il sistema non lineare seguente la cui incognite è il vettore $\mathbf{x} = [x_1, x_2]$:

$$\begin{cases} -x_1 + e^{3x_2} = 1 \\ -x_1 + x_1 x_2^2 = -2 \end{cases}$$

1. Scrivere Il sistema non lineare sotto la forma:

$$\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x})]^T = \mathbf{0}$$

Rappresentare $\mathbf{f}$ mediante una funzione Matlab di tipo *anonymous function* che, ricevuto in input un vettore colonna di 2 elementi, restituisce un vettore colonna di 2 elementi contenente la valutazione di $\mathbf{f}(\mathbf{x})$. Analogamente, calcolare la matrice Jacobiana $J(\mathbf{x})$ di $\mathbf{f}(\mathbf{x})$ e scrivere la *anonymous function* che restituisca tale matrice.

Dunque, il sistema da risolvere è:

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} -x_1 + e^{3x_2} - 1 \\ -x_1 + x_1 x_2^2 + 2 \end{bmatrix} = \mathbf{0}$$

Dove $\mathbf{x} = [x_1, x_2]^T$. La matrice Jacobiana è la seguente:

$$J(\mathbf{x}) = \frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \begin{bmatrix} -1 & 3e^{3x_2} \\ -1 + x_2^2 & 2x_1 x_2 \end{bmatrix}$$

Si introduce il vettore $\mathbf{f}$ e la matrice J utilizzando i seguenti comandi:

```
% il comando @ e' usato per creare funzioni
f = @(x)[-x(1)+exp(3*x(2))-1; -x(1)+x(1)*x(2)^2+2];
J = @(x)[-1 3*exp(3*x(2)); -1+x(2)^2 2*x(1)*x(2)];
```

2. Implementare il metodo di Newton per sistemi. L'intestazione della funzione sarà ad esempio:

```
[xvect, it] = newtonsys(fun, Jf, x0, toll, nmax)
```

Con ovvio significato dei parametri di ingresso e di uscita. Si utilizzi un criterio d'arresto basato sulla norma della differenza tra due iterate successive.

Suggerimento: si utilizzi il comando \ per la risoluzione dei sistemi lineari.

```
function [xvect, it] = newtonsys(fun, Jf, x0, toll, nmax)
%
% [xvect, nit] = newtonsys(fun, Jf, x0, toll, nmax)
%
% Metodo di Newton per sistemi non lineari di dimensione n
%
% Parametri di ingresso:
%
% fun funzione
% Jf matrice Jacobiana
% x0 vett. colonna di dimensione n contenente il dato iniziale
% toll Tolleranza sul test d'arresto
% nmax Numero massimo di iterazioni
%
% Parametri di uscita :
```

```
%
% xvect      Vett. contenente tutte le iterate calcolate
%            (l'ultima componente e' la soluzione)
% it         Iterazioni effettuate

it = 0;
err = toll+1;
xvect = x0;
%
while it < nmax && err >= toll
    f0 = fun(x0);
    Jf0 = Jf(x0);
    dx = -Jf0\f0; % -Jf(x0)\fun(x0)
    x1 = x0 + dx ;
    xvect = [xvect x1];
    it = it + 1;
    err = norm(dx);
    x0 = x1;
end


if it == nmax
    disp('errore - non converge')
end

fprintf('Numero di Iterazioni: %d \n', it);
n = length(x0);
for i = 1:n
    fprintf('x(%d) = %12.8f\n', i , xvect(i,end));
end
```

3. Utilizzare la funzione scritta al punto precedente per calcolare la soluzione del sitema non lineare proposto. Utilizzare una tolleranza di `toll = 1e-6`, un vettore iniziale `x0 = [1;0]`. Fornire il risultato e il numero di iterazioni effettuate.

   Si utilizza la funzione come richiesto.

```
toll=1e-6;
x0=[1;0];
Nmax=1000;

[x,iter] = newtonsys(f,J,x0,toll,Nmax);
% Numero di Iterazioni: 6
% x(1) =   2.39901359
% x(2) =   0.40782842
```

Riferimenti bibliografici

[1] Dede' Luca Fresca Stefania, Botti Michele. Calcolo numerico. Slides from the "Ingegneria Informatica" bachelor's degree course on Politecnico di Milano, 2024.

[2] A. Quarteroni, F. Saleri, and P. Gervasio. *Calcolo Scientifico: Esercizi e problemi risolti con MATLAB e Octave*. UNITEXT. Springer Milan, 2017.

Index

Symbols

A

B

C

D

E

F

T

U