APPUNTI DI ANALISI PER IL CORSO DI MET. ANALIT. E NUMER. PER L'ING. MECC. RISULTATI PRELIMINARI

1. Teorema del valor medio integrale in $\mathbb{R}^n$

I risultati di tutti gli appunti sono validi in $\mathbb{R}^n$ per $n \in \mathbb{N}$, ma in realtà saranno utilizzati per $n = 1, 2, 3$, quindi lo studente potrà limitarsi a considerare questi casi particolari.

Per $r > 0$ numero reale e $\mathbf{x}^0 \in \mathbb{R}^n$ fissati, indicheremo con $B_r(\mathbf{x}^0)$ l'insieme dei punti $\mathbf{x} \in \mathbb{R}^n$ che distano da $\mathbf{x}^0$ meno di r nel senso della norma euclidea (si ricorda che per $\mathbf{x} = (x_1, x_2, ..., x_n) \in \mathbb{R}^n$ la norma euclidea di $\mathbf{x}$, indicata con $|\mathbf{x}|$ o con $\|\mathbf{x}\|$ è data dal numero reale positivo $\left(x_1^2 + x_2^2 + ... + x_n^2\right)^{1/2}$ e la distanza tra due punti $\mathbf{x}$ e $\mathbf{y}$ è $\|\mathbf{x} - \mathbf{y}\|$). In simboli

$$B_r(\mathbf{x}^0) = \left\{ \mathbf{x} \in \mathbb{R}^n \middle| \|\mathbf{x} - \mathbf{x}_0\| < r \right\}$$

In particolare se $n = 1$, $B_r(x^0) = (x^0 - r, x^0 + r)$, se $n = 2$, $B_r(\mathbf{x}^0)$ è il cerchio con centro $\mathbf{x}^0$ e raggio r e se $n = 3$, $B_r(\mathbf{x}^0)$ è la sfera con centro $\mathbf{x}^0$ e raggio r. Indicheremo inoltre con $\bar{B}_r(\mathbf{x}^0)$ la chiususa di $B_r(\mathbf{x}^0)$, cioè $B_r(\mathbf{x}^0)$ unito ai punti di frontiera.

Indicheremo inoltre con $\left|B_r(\mathbf{x}^0)\right|$ la misura di $B_r(\mathbf{x}^0)$, che in particolare per $n = 1$ è $2r$, per $n = 2$ è l'area del cerchio di raggio r e per $n = 3$ è il volume della sfera di raggio r.

Teorema 1.1. *Sia $f : \bar{B}_r(\mathbf{x}^0) \to \mathbb{R}$ continua. Allora esiste $\mathbf{y} \in \bar{B}_r(\mathbf{x}^0)$ tale che*

$$\frac{1}{|B_r(\mathbf{x}^0)|} \int_{B_r(\mathbf{x}^0)} f(\mathbf{x})\, d\mathbf{x} \;=\; f(\mathbf{y})\,. \tag{1.1}$$

La dimostrazione è analoga al caso unidimensionale.

Corollario 1.2. *Nelle stesse ipotesi del Teorema 1.1 si ha che*

$$\lim_{r \to 0} \frac{1}{|B_r(\mathbf{x}^0)|} \int_{B_r(\mathbf{x}^0)} f(\mathbf{x})\, d\mathbf{x} \;=\; f(\mathbf{x}_0)\,. \tag{1.2}$$

La dimostrazione segue immediatamente dal Teorema 1.1 facendo il limite di (1.1) e usando la continuità di f.

2. Scambio tra integrale e derivata

Consideriamo una funzione f definita su un insieme del tipo $A \times B$ con $A \subset \mathbb{R}^m$ e $B \subset \mathbb{R}^n$ insiemi aperti. Scriveremo $f(\mathbf{x}, \mathbf{y})$ dove $\mathbf{x} \in \mathbb{R}^m$ e $\mathbf{y} \in \mathbb{R}^n$. Supponiamo che f sia integrabile in B per ogni $\mathbf{x} \in A$. Quindi $\displaystyle\int_B f(\mathbf{x}, \mathbf{y})\, d\mathbf{y}$ definisce una funzione $F(\mathbf{x})$.

Vale il seguente teorema:

Teorema 2.1. *i) Se f è continua in $A \times B$ allora F è continua in A.*

ii) Se f è continua in $A \times B$ insieme con la derivata parziale $\dfrac{\partial f}{\partial x_i}$ per qualche $1 \leq i \leq n$, allora

$$\frac{\partial F}{\partial x_i}(\mathbf{x}) \;=\; \int_B \frac{\partial f}{\partial x_i}(\mathbf{x}, \mathbf{y})\, d\mathbf{y}\ . \tag{2.1}$$

3. Limiti, derivate e integrali di serie di funzioni

In questa sezione, per semplificare le notazioni, considereremo solo funzioni di una variabile. L'estensione dei risultati a funzioni di più variabili è lasciata allo studente.

Consideriamo una serie di funzioni $\displaystyle\sum_{n=0}^{+\infty} f_n(x)$, con $f_n : A \to \mathbb{R}$. Diremo che la serie soddisfa il **criterio di Weierstrass** in A se esistono numeri reali a_n tali che $|f_n(x)| < a_n$ per $x \in A$ e $n \in \mathbb{N}$, e se $\displaystyle\sum_{n=0}^{+\infty} a_n$ converge. Una serie di funzioni che soddisfa il criterio di Weierstass in A in particolare converge per tutti gli $x \in A$ e quindi definisce in A una funzione somma $S(x)$.

Teorema 3.1. *i) Se $\displaystyle\sum_{n=0}^{+\infty} f_n(x)$ soddisfa il criterio di Weierstrass in A e se esistono finiti i limiti $\lim_{x \to x_0} f_n(x) = l_n$ allora la serie $\displaystyle\sum_{n=0}^{+\infty} l_n$ converge con somma L e risulta*

$$\lim_{x \to x_0} S(x) = L\,.$$

ii) Se $\displaystyle\sum_{n=0}^{+\infty} f_n(x)$ soddisfa il criterio di Weierstrass in A e se tutte le f_n sono continue in A, allora $S(x)$ è continua in A.

Teorema 3.2. *Se tutte le f_n sono derivabili in un intervallo (a, b), $\displaystyle\sum_{n=0}^{+\infty} f_n(x)$ converge per almeno un $x \in (a, b)$ e $\displaystyle\sum_{n=0}^{+\infty} f_n'(x)$ soddisfa il criterio di Weierstrass in (a, b), allora $S(x)$ è derivabile in (a, b) e*

$$S'(x) = \sum_{n=0}^{+\infty} f_n'(x)\,.$$

Si lascia allo studente l'estensione del teorema a derivate di ordine superiore.

Teorema 3.3. *Se tutte le f_n sono integrabili in un intervallo (a, b) e $\displaystyle\sum_{n=0}^{+\infty} f_n(x)$ soddisfa il criterio di Weierstrass in (a, b), allora $S(x)$ è integrabile in (a, b) e*

$$\int_a^b \sum_{n=0}^{+\infty} f_n(x) = \sum_{n=0}^{+\infty} \int_a^b f_n(x)\,.$$

Metodi Analitici e Numerici per l'ingegneria

Cerutti Maria Cristina

A cura di: Andrea Fuso, Gianpiero Gaeta

Indice

Parte I
Richiami di Analisi

1 Funzioni

Teorema 1 (Teorema di Weierstrass). Ogni funzione continua $f : K \to \mathbb{R},\ K \subset \mathbb{R}^n$, insieme chiuso e limitato (compatto), ammette massimo e minimo.

Teorema 2 (Teorema dei valori intermedi o teorema di Darboux). Sia $f : [a,\ b] \to \mathbb{R}$ una funzione continua su un intervallo chiuso e limitato $[a,\ b]$. Detti m e M rispettivamente il minimo e massimo valore di $f(x)$ nell'intervallo $[a,\ b]$, risulta che f assume tutti i valori compresi tra m e M (poichè è una funzione continua).

$$\forall\, c \in [a,\ b]\ \ (\text{Con } m \leq c \leq M)\ \ \exists\, x_0 \in [a,\ b]\ \ \text{t.c. } f(x_0) = c$$

In generale per una funzione continua $f : C \to \mathbb{R},\ C \subseteq \mathbb{R}^n$, insieme connesso, se $\mathbf{x}_1,\ \mathbf{x}_2 \in C$, f assume tutti i valori compresi tra $f(\mathbf{x}_1)$ e $f(\mathbf{x}_2)$.

Teorema 3 (Teorema della media integrale). Sia $f : [a,\ b] \to \mathbb{R}$ una funzione continua (e quindi integrabile), allora esiste $c \in [a,\ b]$ tale che:

$$\frac{1}{b-a}\int_a^b f(x)\,dx = f(c)$$

$$\Rightarrow \int_a^b f(x)\,dx = (b-a)\,f(c)$$

Osservazione 1.1. Essendo l'integrale collegato al calcolo dell'area sottesa dall funzione, il teorema della media integrale afferma che esiste un punto c appartenente all'intervallo, tale per cui l'integrale è calcolabile come l'area di un rettangolo equivalente, avente per base la dimensione dell'intervallo e per altezza l'immagine di c.

Teorema 4 (Teorema fondamentale dell'algebra). Ogni polinomio a coefficienti reali o complessi, di grado maggiore o uguale a 1, ammette almeno una radice complessa. Quindi ogni polinomio a coefficienti reali o complessi di grado $n \geq 1$ ammente nel campo complesso $\mathbb{C}$ esattamente $n$soluzioni (contate con la loro molteplicità).

Definizione 1 (Distanza punto-retta)**.** Si definisce distanza di un punto $P = (x_P,\ y_P)$ da una retta r di equazione $y = mx + q$, la quantità.

$$d(P,\ r) = \frac{|y_P - (mx_P + q)|}{\sqrt{1+m^2}}$$

oppure scrivendo la retta in forma implicita $ax + by + c = 0$ si ha:

$$d(P,\ r) = \frac{|ax_P + by_P + c|}{\sqrt{a^2+b^2}}$$

2 Vettori

Teorema 5 (Disuguaglianza di Schwarz). Per la disuguaglianza di Schwarz il valore assoluto del prodotto scalare di due vettori $\overline{x}$ e $\overline{y}$ è minore o uguale al prodotto delle loro norme:

$$\|\overline{x} \cdot \overline{y}\| \leq \|\overline{x}\| \cdot \|\overline{y}\|$$

3 Matrici

3.1 Algebra delle matrici

Definizione 2 (Matrice diagonale)**.** Una matrice diagonale è una matrice diagonale $D = [d_{ij}]$ che ha tutti gli elementi al di fuori dalla diagonale principale nulli:

$$d_{ij} = 0 \quad \text{se } i \neq j$$

Proposizione 3.1. Il prodotto di due matrici diagonali è ancora una matrice diagonale.

Definizione 3 (Matrice inversa)**.** Siano A e B due matrici quadrate di ordine n. Si dice che B è la matrica inversa di A se:

$$AB = BA = I$$

Dove I è la matrice identità di ordine n. Se esiste una matrice inversa di A, allora si dice che A è invertibile.

Teorema 6 (Condizioni di invertibilità). Per una matrice A quadrata di ordine n le seguenti condizioni sono equivalenti:

1. A ha rango massimo:
$$r(A) = n$$
2. L'unica soluzione di $A\overline{x} = \overline{0}$ è $\overline{x} = \overline{0}$:
$$Ker(A) = \{\overline{0}\}$$
3. A è invertibile.
4. A ha un'inversa sinistra.
5. A ha un'inversa destra.

Definizione 4 (Matrice non singolare)**.** Una matrice A quadrata di ordine n che soddisfi le condizioni equivalenti del teorema 6 si dice **non singolare**. Quindi non singolare è sinonimo di invertibile, e singolare significa non invertibile.

Proposizione 3.2 (Inversa di un prodotto)**.** Siano A e B due matrici quadrate di ordine n. La matrice prodotto AB è invertibile se e solo se A e B sono invertibili. In tal caso:

$$(AB)^{-1} = B^{-1}A^{-1}$$

Definizione 5 (Matrice simmetrica)**.** Una matrice A si dice **simmetrica** se $A^T = A$. Una matrice A si dice **antissimetrica** se $A^T = -A$.

Esempio 1. Per una matrice quadrata di ordine 3:

$$\text{Forma matrice simmetrica} \rightarrow \begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix}$$

$$\text{Forma matrice antissimetrica} \rightarrow \begin{bmatrix} 0 & b & c \\ -b & 0 & e \\ -c & -e & 0 \end{bmatrix}$$

Esempio 2 (Matrice di Hilbert)**.** Una matrice H è una matrice di Hilbert se ha gli elementi h_{ij} pari a:

$$h_{ij} = (i + j - 1)^{-1}$$

$$H = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{n+1} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \cdots & \frac{1}{n+2} \\ \vdots & & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & \frac{1}{n+2} & \cdots & \frac{1}{2n-1} \end{bmatrix}$$

Proposizione 3.3. Una matrice simmetrica è sempre diagonalizzabile con autovalorei reali.

Definizione 6 (Matrice definita positiva)**.** Sia A una matrice simmetrica di ordine n a coefficienti reali. La matrice A si dice definita positiva se per ogni vettore $\overline{x} = [x_1, \ldots, x_n] \in \mathbb{R}^n$, il prodotto $\overline{x}A\overline{x}^T$ è maggiore di zero, ovvero:

$$\text{A definita positiva} \Leftrightarrow \overline{x}A\overline{x}^T > 0 \quad \forall\, \overline{x} \in \mathbb{R}^n \quad \text{Con } \overline{x} \neq \overline{0}$$

Oppure: vedi teorema 7

Osservazione 3.1. Gli autovalori di una matrice definita positiva sono tutti positivi.

Proposizione 3.4 (Proprietà matrice trasposta)**.**

$$(A + B)^T = A^T + B^T$$

$$(tA)^T = tA^T$$

$$\left(A^T\right)^T = A$$

$$(AB)^T = B^T A^T$$

$$\left(A^T\right)^{-1} = \left(A^{-1}\right)^T$$

$$\text{Se } A \text{ è simmetrica} \rightarrow A^T = A$$

$$\text{Se } A \text{ è antisimmetrica} \rightarrow A^T = -A$$

Teorema 7 (Matrice definita positiva)*.* Per una matrice A simmetrica le seguenti condizioni sono equivalenti:

1. A è definita positiva.
2. I minori principali di nord-ovest di A sono positivi.

Definizione 7 (Matrice a dominanza diagonale per righe)**.** Una matrice A si dice che è dominante per righe se è una matrice quadrata di ordine n con gli elementi sulla diagonale principale maggiori o uguali (in valore assoluto) della somma di tutti i restanti elementi della stessa riga (sempre in valore assoluto):

$$\text{Dominanza diagonale in senso debole} \rightarrow |a_{ii}| \geq \sum_{i=1, j \neq i}^{n} |a_{ij}|$$

$$\text{Dominanza diagonale in senso stretto} \rightarrow |a_{ii}| > \sum_{i=1, j \neq i}^{n} |a_{ij}|$$

4 Teorema spettrale

Proposizione 4.1. Sia A una matrice reale simmetrica (Def: 5) di ordine n e sia $q(\overline{x}) = \overline{x}^T A \overline{x}$ allora:

1. A è diagonalizzabile con autovalori reali.
2. Se $A\overline{v} = \lambda \overline{v}$, allora $q(\overline{v}) = \lambda \|\overline{v}\|^2$ (quindi se $\overline{v}$ è autovettore della matrice A).
3. Se λ_{min} e λ_{max} sono il minimo e massimo degli autovalori di A, allora:

$$\lambda_{min} \|\overline{x}\|^2 \leq \overline{x}^T A \overline{x} \leq \lambda_{max} \|\overline{x}\|^2$$

Parte II
Calcolo numerico

5 Risoluzione di sistemi lineari con metodi diretti

5.1 Metodo di Cramer

Teorema 1. ***(Teorema di Cramer)***
Sia $A\overline{x} = \overline{b}$ un sistema lineare quadrato di n equazioni in n incognite. Se $det(A) \neq 0$ il sistema ammette un'unica soluzione $\overline{v} = [x_1, \ldots, x_n]^T$ di componenti:

$$x_i = \frac{det(A_i)}{det(A)}$$

Dove A_i è ottenuta sostituendo la colonna di A con il termine noto $\overline{b}$

Esempio 1.

$$\begin{cases} x + 2y = 3 \\ 3x + 7y = 2 \end{cases} \Rightarrow A = \begin{bmatrix} 1 & 2 \\ 3 & 7 \end{bmatrix} \quad \overline{b} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

$$x = \frac{det(A_x)}{det(A)} = \frac{det\begin{bmatrix} 3 & 2 \\ 2 & 7 \end{bmatrix}}{det\begin{bmatrix} 1 & 2 \\ 3 & 7 \end{bmatrix}} = \frac{det\begin{bmatrix} b_1 & 2 \\ b_2 & 7 \end{bmatrix}}{det\begin{bmatrix} 1 & 2 \\ 3 & 7 \end{bmatrix}} = 17$$

$$y = \frac{det(A_y)}{det(A)} = \frac{det\begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix}}{det\begin{bmatrix} 1 & 2 \\ 3 & 7 \end{bmatrix}} = \frac{det\begin{bmatrix} 1 & b_1 \\ 3 & b_2 \end{bmatrix}}{det\begin{bmatrix} 1 & 2 \\ 3 & 7 \end{bmatrix}} = -7$$

Costo computazionale 1. Il metodo di Cramer richiede un costo computazionale pari a:

$$\#\text{flops} = 2(n+1)!$$

calcolando i determinanti con la regola di Laplace.

5.2 Risoluzione sistemi triangolari

5.2.1 Backward substitution

Un sistema triangolare superiore, è un sistema che in forma matriciale si presenta nel seguente modo:

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,n} \\ 0 & a_{2,2} & \ldots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & a_{n,n} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

Il termine x_n è immediatamente calcolabile come:

$$x_n = \frac{b_n}{a_{n,n}}$$

Per quanto riguarda il termine x_{n-1} l'operazione da eseguire corrisponde a:

$$x_{n-1} = [b_{n-1} - a_{n-1,n}x_n] \frac{1}{a_{n-1,n-1}}$$

Si può quindi estrapolare l'algoritmo per il calcolo del generico x_i:

Proposizione 5.1. Data una matrice triangolare alta, è possibile trovare tutte le soluzioni attraverso il seguente algoritmo:

$$x_n = \frac{b_n}{a_{n,n}}$$

$$x_i = \frac{1}{a_{i,i}} \left[b_i - \sum_{j=i+1}^{n} a_{i,j} x_j \right] \qquad \begin{matrix} i = 1, \dots, n \\ j = i+1, \dots n \end{matrix}$$

Questo algoritmo prende il nome di **sostituzione all'indietro**, conosciuto anche come **backward substitution**.

Costo computazionale 2. Per quanto riguarda il costo computazionale dell'algoritmo si distinguono due casi: il primo passaggio e i successivi. Per il primo passaggio (calcolo dell'elemento x_n) si esegue una sola operazione (la divisione), mentre nel secondo passaggio le operazioni diventano tre: la divisione per $a_{i,i}$, la moltiplicazione e la sottrazione all'interno della parentesi. Quindi il numero di operazioni elementari risulta essere:

$$\#\text{flops} = 1 + (1+2) + \cdots + (1+2i) + \cdots + (1+2(n-1)) = \sum_{i=0}^{n} (1+2i)$$

$$\Downarrow$$

$$\#\text{flops} = n + 2 \sum_{i=1}^{n-1} i = n + 2\frac{n(n-1)}{2} = n^2$$

5.2.2 Forward substitution

Un sistema triangolare inferiore , è un sistema che in forma matriciale si presenta nel seguente modo:

$$\begin{bmatrix} a_{1,1} & 0 & \dots & 0 \\ a_{1,2} & a_{2,2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

Il termine x_1 è direttamente calcolabile come:

$$x_1 = \frac{b_1}{a_{1,1}}$$

Il secondo termine risulta essere pari a:

$$x_2 = \frac{1}{a_{2,2}} \left[b_2 - a_{2,1} x_1 \right]$$

Algoritmo 1. *L'algoritmo per il calcolo della generica x_i risulta quindi essere:*

$$x_i = \frac{1}{a_{i,i}} \left[b_i - \sum_{j=1}^{i-1} a_{i,j} x_j \right] \qquad j = 1, \dots, n$$

Proposizione 5.2. Questo algoritmo per calcolare le x_i incognite prende il nome di **algoritmo di sostituzione in avanti**, conosciuto anche come **forward substitution**.

Osservazione 5.1. Il costo computazionale per il forward substitution è perfettamente analogo a quello del backward substitution, è quindi pari a n^2

Osservazione 5.2. Se A è triangolare non singolare ($detA \neq 0$), gli algoritmi funzionano sempre poichè:

$$detA \neq 0 \quad \Leftrightarrow \quad a_{i,i} \neq 0 \quad \forall \quad i = 1, \dots, n$$

5.3 Metodo di eliminazione di Gauss

Esiste un algoritmo, detto algoritmo di eliminazione di Gauss, che riduce una matrice in una matrice a scala mediante un numero finito di operazioni elementari sulle righe.

Il metodo di eliminazione gaussiana si basa sull'idea di ridurre il sistema $A\overline{x} = \overline{b}$ ad un sistema equivalente (avente cioè la stessa soluzione) della forma $U\overline{x} = \hat{b}$ dove U è triangolare superiore e $\hat{b}$ è un nuovo termine noto. Quest'ultimo sistema potrà essere risolto con il metodo delle sostituzioni all'indietro (*backward substitution* Prop. 5.1).

Funzionamento del metodo

Indicando il sistema originale come $A^{(1)}\overline{x} = \overline{b}^{(1)}$

Si sostituisce la seconda riga con la differenza tra la riga stessa e la prima riga moltiplicata per una costante non nulla, scelto in modo tale da rendere nullo il primo elemento della seconda riga; in questo modo si ottiene un sistema equivalente a quello di partenza → cioè con la stessa soluzione. Tale costante è il motiplicatore $m_{2,1}$:

$$m_{2,1} = \frac{a_{2,1}}{a_{1,1}}$$

$$\overline{b}_2^{(2)} = \overline{b}_2^{(1)} - m_{2,1}\overline{b}_1^{(1)}$$

tale procedimento viene ripetuto per tutte le righe successive.

Facendo questo si ottiene, per la prima colonna, un vettore che ha come unico elemento non nullo quello sulla diagonale.

$$\left[\begin{array}{cccc|c} a_{1,1} & a_{1,2} & \dots & a_{1,n} & b_1 \\ 0 & a_{2,2} & \dots & a_{2,n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & a_{n,2} & \dots & a_{n,n} & b_n \end{array}\right] \Rightarrow \left[\begin{array}{cccc} a_{1,1}^{(1)} & a_{1,2}^{(1)} & \dots & a_{1,n}^{(1)} \\ 0 & a_{2,2}^{(2)} & \dots & a_{2,n}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n,2}^{(2)} & \dots & a_{n,n}^{(2)} \end{array}\right] \left[\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_n \end{array}\right] = \left[\begin{array}{c} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_n^{(2)} \end{array}\right]$$

Dove gli elementi indicati con: $^{(2)}$ corrispondono al risultato delle operazioni svolte sulle righe, dopo l'operazione di MEG sulla prima colonna.

In generale si ha che:

$$m_{i,k} = \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}} \qquad i = k+1, \dots, n$$

Dove gli elementi $a_{k,k}^{(k)}$ sono detti **pivot**.

Quindi nel caso generale, l'operazione eseguita è:

$$a_{i,j}^{(k+1)} = a_{i,j}^{(k)} - m_{i,k}a_{k,j}^{(k)} \quad i,j = k+1, \dots, n$$

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

Osservazione 5.3. L'operazione di MEG sia arresta quando incontra un pivot nullo, per evitare questo si utilizza il pivoting (Thm 5.4.1).

Esempio 3. Si consideri il seguente sistema:

$$\begin{cases} 3x - y + z = 2 \\ 2x + y = 1 \\ -2x - 2y + z = -2 \end{cases} \Longrightarrow \left[\begin{array}{ccc|c} 3 & -1 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ -2 & -2 & 1 & -1 \end{array}\right] = B$$

$$\underbrace{\qquad\qquad}_{A} \quad \underbrace{\quad}_{\vec{b}}$$

Con i passaggi successivi si vuole rendere A (matrice dei coefficienti) triangolare alta, cioè tale da avere i termini sotto la diagonale principale tutti identicamente nulli, per risolvere così un sistema triangolare. Le operazioni da eseguire sono delle particolari combinazioni linari delle righe.

$$A = \begin{bmatrix} \overline{a}_1 \\ \overline{a}_2 \\ \overline{a}_3 \end{bmatrix} \quad \text{Con } \overline{a}_i \text{ righe di } A$$

$$B = \left[\begin{array}{c|c} A & \vec{b} \end{array}\right] = \begin{bmatrix} \overline{b}_1 \\ \overline{b}_2 \\ \overline{b}_3 \end{bmatrix}$$

Si costruiscono quindi successive matrici $B^{(k)}$, equivalenti a B, in cui gli elementi sotto la diagonale principale diventano progressivamente uguali a zero, e sia:

$$B^{(1)} = B$$

$B^{(2)}$ è equivalente a $B^{(1)}$ ed ha gli elementi sottodiagonali della prima colonna nulli; in questo caso si cerca di avere $a_{21}^{(2)} = a_{31}^{(2)} = 0$, quindi:

$$\overline{b}_1^{(2)} = \overline{b}_2^{(2)}$$

$$\overline{b}_2^{(2)} = \overline{b}_2^{(1)} - m_{21}\overline{b}_1^{(1)}$$

m_{21} è quel valore che permette di ottenere $a_{21}^{(2)} = 0$:

$$\begin{bmatrix} 2 & 1 & 0 & 1 \end{bmatrix} - \frac{2}{3}\begin{bmatrix} 3 & -1 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 0 & \frac{5}{3} & -\frac{2}{3} & -\frac{1}{3} \end{bmatrix}$$

$$\Rightarrow m_{21} = \frac{2}{3} = \frac{a_{21}^{(1)}}{a_{11}^{(1)}}$$

pre rendere $a_{31}^{(2)} = 0$:

$$\overline{b}_3^{(2)} = \overline{b}_3^{(1)} - m_{31}\overline{b}_1^{(1)}$$

$$\begin{bmatrix} -2 & -2 & 1 & -2 \end{bmatrix} + \frac{2}{3}\begin{bmatrix} 3 & -1 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 0 & -\frac{8}{3} & \frac{5}{3} & -\frac{2}{3} \end{bmatrix}$$

$$\Rightarrow m_{31} = -\frac{2}{3} = \frac{a_{31}^{(1)}}{a_{11}^{(1)}}$$

La matrice $B^{(2)}$ risulta quindi essere:

$$B^{(2)} = \begin{bmatrix} 3 & -1 & 1 & 2 \\ 0 & \frac{5}{3} & -\frac{2}{3} & -\frac{1}{3} \\ 0 & -\frac{8}{3} & \frac{5}{3} & -\frac{2}{3} \end{bmatrix}$$

Ora si cerca la matrice $B^{(3)}$ tale da avere $a_{32}^{(3)} = 0$, quindi:

$$\overline{b}_1^{(3)} = \overline{b}_1^{(2)}$$

$$\overline{b}_2^{(3)} = \overline{b}_2^{(2)}$$

$$\overline{b}_3^{(3)} = \overline{b}_3^{(2)} - m_{32}\overline{b}_2^{(2)}$$

$$\Rightarrow \begin{bmatrix} 0 & -\frac{8}{3} & \frac{5}{3} & -\frac{2}{3} \end{bmatrix} + \frac{8}{5}\begin{bmatrix} 0 & \frac{5}{3} & -\frac{2}{3} & -\frac{1}{3} \end{bmatrix} = \begin{bmatrix} 0 & 0 & \frac{3}{5} & -\frac{6}{5} \end{bmatrix}$$

$$\Rightarrow m_{32} = -\frac{8}{5} = \frac{a_{32}^{(2)}}{a_{22}^{(2)}}$$

La matrice $B^{(3)}$ risulta essere:

$$B^{(3)} = \begin{bmatrix} 3 & -1 & 1 & 2 \\ 0 & \frac{5}{3} & -\frac{2}{3} & -\frac{1}{3} \\ 0 & 0 & \frac{3}{5} & -\frac{6}{5} \end{bmatrix} \Longrightarrow \begin{cases} 3x - y + z = 2 \\ \frac{5}{3}y - \frac{2}{3}z = -\frac{1}{3} \\ \frac{3}{5}z = -\frac{6}{5} \end{cases} \Rightarrow \overline{x} = \begin{bmatrix} 1 \\ -1 \\ -2 \end{bmatrix}$$

Proposizione 5.3. Valgono le seguenti formule per la somma dei primi N interi, e per la somma dei quadrati dei primi N interi:

$$\sum_{i=1}^{N} i = \frac{N\,(N+1)}{2}$$

$$\sum_{i=1}^{N} i^2 = \frac{(2N+1)\,(N+1)\,N}{6}$$

Costo computazionale 3.

Contributo per il calcolo di m

Essendo il generico $m_{i,k}$ pari a:

$$m_{i,k} = \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}} \qquad i = k+1, \ldots, n$$

Segue che si esegue una sola operazione (la divisione) per $n-(k-1)$ volte, per ciascuna delle righe della matrice successiva alla prima ($k = 2, \ldots, n$).

$$\#\textit{flops} = (n-1) + (n-2) + \cdots + 1 = \sum_{i=1}^{n-1} i = \frac{(n-1)\,(n-1+1)}{2} = \frac{n\,(n-1)}{2}$$

Contributo per il calcolo di $a_{i,j}$

Essendo il generico $a_{i,j}^{(k+1)} = a_{i,j}^{(k)} - m_{i,k}a_{k,j}^{(k)}$, segue che per effettuare tale calcolo si devono eseguire due operazioni elementari: il prodotto $m_{i,k}a_{k,j}^{(k)}$ e la sottrazione. Al primo passaggio bisogna calcolare gli elementi di una matrice $(n-1) \times (n-1)$ (la prima riga della matrice $A^{(2)}$, uguale alla prima riga della matrice $A^{(1)}$, non si calcola così come la prima colonna sottodiagonale di $A^{(2)}$, che ha elementi uguali a zero). Ad ogni passaggio bisogna calcolare gli elementi di una matrice quadrata $(n-(k-1))$. Di conseguenza il numero di operazioni elementari totali (flops) corrisponde a:

$$\#\text{flops} = 2\left[(n-1)^2 + (n-2)^2 + \cdots + 1\right] = 2\sum_{i=1}^{n-1} i^2 = 2\frac{(2n-2+1)\,(n-1+1)\,(n-1)}{6} = \frac{(2n-1)\,(n-1)\,n}{3}$$

Contributo per il calcolo di b

Essendo il generico $b_i^{(k+1)} = b_i^{(k)} - m_{i,k}^{(k)} b_k^{(k)}$, segue che per effettuare tale calcolo si devono eseguire due operazioni elementari: il prodotto $m_{i,k}^{(k)} b_k^{(k)}$ e la sottrazione, effettuate per $n-1$ volte, cioè per gli $n-1$ elementi presenti nel vettore $\overline{b}$ (è il pedice i a scorrere); al passaggio successivo $n-2$ e così via. Di conseguenza il costo computazionale di tale operazione risulta essere:

$$\#\text{flops} = 2\left[(n-1) + (n-2) + \cdots + 1\right] = 2\sum_{i=1}^{n-1} i = 2\frac{n\,(n-1)}{2} = n\,(n-1)$$

Costo computazionale totale

In definitiva il costo computazionale totale è pari a:

$$\#\text{flops} = \frac{n\,(n-1)}{2} + \frac{(2n-1)\,(n-1)\,n}{3} + 2\frac{n\,(n-1)}{2} = \frac{3}{2}n\,(n-1) + \frac{(2n-1)\,(n-1)\,n}{3} \simeq \frac{3}{2}n^2 + \frac{2}{3}n^3 \simeq \frac{2}{3}n^3 \quad (1)$$

Osservazione 1. L'approssimazione è valida per valori di n molto grandi.

Osservazione 2. Il costo computazione per il MEG è molto minore di quello richiesto dal metodo di Cramer:

	MEG	**Cramer**
Costo computazionale	$\frac{2}{3}n^3$	$2\,(n+1)!$

Algoritmo 2. *Sia $k = 1, \ldots, n$ l'indice colonna, l'algoritmo del MEG conta tre passaggi:*

1. *Il calcolo del coefficiente $m_{i,k}$:*
$$m_{i,k} = \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}} \qquad i = 1, \ldots, n-1$$
2. *Il calcolo del generico $a_{i,j}^{(k+1)}$:*
$$a_{i,j}^{(k+1)} = a_{i,j}^{(k)} - m_{i,k} a_{k,j}^{(k)} \qquad i = 1, \ldots, n-1$$
3. *Il calcolo del generico $b_i^{(k+1)}$:*
$$b_i^{(k+1)} = b_i^{(k)} - m_{i,k}^{(k)} b_k^{(k)} \qquad i = 1, \ldots, n-1$$

Osservazione 5.4. È ora possibile calcolare il costo computazionale corretto del MEG, al cui costo precedentemente (Eq: 1) calcolato si deve aggiungere il risultato appena ottenuto:

$$\#\text{flops}_{tot} = \frac{3}{2}n\,(n-1) + \frac{(2n-1)\,(n-1)\,n}{3} + n^2 \simeq \frac{2}{3}n^3$$

Di conseguenza per un numero di passaggi n molto grande, il costo computazionale del MEG dipende sempre dal termine $\frac{2}{3}n^3$ e il contributo dell'algoritmo di backward substitution risulta perciò trascurabile (di un ordine di grandezza inferiore).

5.4 Fattorizzazione LU

Cosa fa 1. Fattorizza la matrice A in modo tale da avere $A = LU$ dove L è una matrice di tipo triangolare bassa, e U una matrice triangolare alta.

Osservazione 5.5. Se esiste una fattorizzazione LU di A, per risolvere il sistema $A\overline{x} = \overline{b}$, si può scrivere $LU\overline{x} = \overline{b}$, che equivale a risolvere i sistemi:

$$\begin{cases} L\overline{y} = \overline{b} \\ U\overline{x} = \overline{y} \end{cases}$$

di cui il primo è triangolare inferiore e il secondo è triangolare superiore (Sezione: 5.2).

Problema 5.1. Come si trovano le matrici L ed U?

Esempio 4. Considerando ad esempio una matrice 2×2 :

$$\begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix} = \begin{bmatrix} l_{1,1} & 0 \\ l_{2,1} & l_{2,2} \end{bmatrix} \begin{bmatrix} u_{1,1} & u_{1,2} \\ 0 & u_{2,2} \end{bmatrix}$$

Il sistema presenta quattro equazioni e sei incognite (gli elementi delle matrici L e U), di conseguenza è un sistema di tipo sottodeterminato. Per risolvere questo problema, in questo caso si possono scegliere arbitrariamente due elementi, ad esempio $l_{1,1}$ e $l_{2,2}$ (gli elementi sulla diagonale della matrice L) e porli uguali a uno, nell'esempio:

$$\begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ l_{2,1} & 1 \end{bmatrix} \begin{bmatrix} u_{1,1} & u_{1,2} \\ 0 & u_{2,2} \end{bmatrix}$$

In questo modo il sistema diventa risolvibile:

$$\begin{cases} u_{1,1} = a_{1,1} \\ u_{1,2} = a_{1,2} \\ l_{2,1} = \frac{a_{2,1}}{a_{1,1}} \ (a_{1,1} \neq 0) \\ u_{2,2} = a_{2,2} - \frac{a_{2,1}}{a_{1,1}} a_{1,2} = \frac{a_{2,2}a_{1,1} - a_{2,1}a_{1,2}}{a_{1,1}} = \frac{detA}{a_{1,1}} \end{cases}$$

Osservazione 5.6. Se si fosse applicato il MEG alla matrice A:

$$m_{2,1} = \frac{a_{2,1}}{a_{1,1}} = l_{1,2}$$

Inoltre la prima riga di U è uguale alla prima riga di A e $u_{2,2} = a_{2,2} - m_{2,1}a_{1,2}$, cioè la stessa operazione che si avrebbe ottenuto applicando il MEG; quindi U è la matrice che si ottiene alla fine del MEG.

In generale per determinare L ed U si devono trovare $n^2 + n$ elementi con n^2 equazioni; si scelgono quindi gli n elementi della diagonale di L uguali ad 1 (**fattorizzazione LU di Gauss**).

Teorema 8. Se tutti i pivots sono diversi da zero, la fattorizzazione LU di Gauss è unica, e si ha che:

$L \rightarrow$ Gli **elementi sottodiagonali sono i moltiplicatori del MEG**

$U \rightarrow$ È la matrice finale del MEG

quindi condizione necessaria e sufficiente per l'esistenza di questa fattorizzazione è che tutti i pivots $a_{k,k}^{(k)}$ del MEG siano diversi da zero (condizione che si verifica quando la matrice non è singolare (Def: 4).

Dimostrazione. Sia $A^{(0)} = A$, si ossercva che $A^{(1)}$, cioè la matrice trasformata dal primo passaggio iterativo del MEG che porta all'azzeramento degli elementi della prima colonna, si può ottenere come:

$$A^{(1)} = M_1 A^{(0)}$$

$$M_1 = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ -m_{2,1} & 1 & 0 & \dots & 0 \\ -m_{3,1} & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -m_{n,1} & 0 & 0 & \dots & 1 \end{bmatrix}$$

- Ora per passare alla matrice $A^{(2)}$ il procedimento è analogo: $A^{(2)} = M_2 A^{(1)} = M_2 M_1 A^{(0)}$:

$$M_2 = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & -m_{3,2} & 1 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & -m_{n,2} & 0 & \dots & 1 \end{bmatrix}$$

- Si procede in questo modo fino ad arrivare alla matrice finale triangolare superiore del MEG, che indichiamo con U. Quindi la generica matrice M_k ha una forma del tipo:

$$M_k = \begin{bmatrix} 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & & 1 & 0 & & 0 \\ 0 & & -m_{k+1,k} & 1 & & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & -m_{n,k} & 0 & \dots & 1 \end{bmatrix} = I_n - \overline{m}_k \overline{e}_k^T$$

  dove $\overline{e}_k$ è il k-esimo vettore della base canonica:

$$\overline{e}_k^T = \begin{bmatrix} 0 & 0 & \cdots & 1 & \cdots & 0 \end{bmatrix}$$

  mentre $\overline{m}_k$ corrisponde ad un vettore del tipo:

$$\overline{m}_k = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ m_{k+1,k} \\ \vdots \\ m_{n,k} \end{bmatrix}$$

  Esempio 5. Prendendo ad esempio una matrice 3×3 , in cui l'elemento $m_{3,2} = 4$, la matrice M_2 risulta essere:

$$M_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -4 & 1 \end{bmatrix} = I_3 - \overline{m}_2 \overline{e}_k^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \left(\begin{bmatrix} 0 \\ 0 \\ 4 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \right) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 4 & 0 \end{bmatrix}$$

- Quindi alla fine del processo si ottiene la matrice A_k che è la matrice ridotta a scala, e che coincide quindi con U :

$$A^{(k)} = U = M_{n-1} M_{n-2} \cdots M_2 M_1 A$$

- Ricordando la proprietà 3.2, è possibile scrivere:

$$A = M_1^{-1} M_2^{-1} \cdots M_{n-2}^{-1} M_{n-1}^{-1} U = LU$$

- Le matrici M_k sono matrici triangolari inferiori con elementi diagonali tutti pari ad uno e con inversa data da

$$M_k^{-1} = I_n + \overline{m}_k \overline{e}_k^T$$

  come si può verificare facilmente essendo $\left(\overline{m}_i \overline{e}_i^T\right) \left(\overline{m}_j \overline{e}_j^T\right)$ uguale alla matrice identicamente nulla per $i \leq j$. Di conseguenza si ha che:

$$A = M_1^{-1} \cdots M_{n-1}^{-1} U = \left(I_n + \overline{m}_1 \overline{e}_1^T\right) \cdots \left(I_n + \overline{m}_{n-1} \overline{e}_{n-1}^T\right) U = \left(I_n + \sum_{i=1}^{n-1} \overline{m}_i \overline{e}_i^T \right) U$$

$$\Rightarrow \quad A = \begin{bmatrix} 1 & 0 & \dots & \dots & 0 \\ m_{21} & 1 & & & \vdots \\ \vdots & m_{32} & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ m_{n1} & m_{n3} & \dots & m_{n,n-1} & 1 \end{bmatrix} U$$

- Di conseguenza risulta essere:

$$M_1^{-1} M_2^{-1} \cdots M_{n-2}^{-1} M_{n-1}^{-1} = L$$

□

Proprietà 1. Sia $A \in \mathbb{R}^{n \times n}$. La fattorizzazione LU di A con $l_{ii} = 1$ per $i = 1, \dots, n$ (fattorizzazione LU di Gauss), esiste ed è unica se e solo se le sottomatrici principali di nord-ovest A_i di A di ordine $i = 1, \dots, n-1$ sono non singolari (Def: 4).

Osservazione 5.7. Nel caso in cui la fattorizzazione LU esiste, il determinante di A si può calcolare come:

$$detA = u_{11} u_{22} \cdots u_{nn} = \prod_{i=1}^{n} u_{ii}$$

Quando si può usare 1.

Condizione necessaria e sufficiente per fattorizzazione LU

Una condizione necessaria e sufficiente affinchè possa esistere la fattorizzazione LU è che i determinanti delle sottomatrici principali di nord-ovest siano $\neq 0$, cioè che non siano singolari (Def: 4).

Condizioni sufficienti per fattorizzazione LU

Condizione sufficiente per la fattorizzazione LU è che:

1. Che la matrice A sia simmetrica (Def: 5) e definita positiva (Def: 6). Siccome A è simmetrica per la Proposizione 4.1 vale:
$$\overline{x}^T A \overline{x} \geq \lambda_{min} \|\overline{x}\|^2$$
2. Che la matrice A sia a dominanza diagonale stretta per le righe (Def: 7).
$$|a_{ii}| > \sum_{i=1, j \neq i}^{n} |a_{ij}|$$

5.4.1 Pivoting

Se si incontra un pivot $a_{k,k}^{(k)} = 0$ il procedimento si arresta. In questo caso siccome A è non singolare, per $j > k$ $\exists\ a_{j,k}^{(k)} \neq 0$. Se così non fosse il determinante della matrice sarebbe pari a zero, e dunque la matrice sarebbe singolare, dunque contraria all'ipotesi di partenza. Scambiando quindi la riga j e la riga k è possibile procedere con la fattorizzazione. Questo procedimento viene detto *pivoting per righe*.

Poiché un valore piccolo (in valore assoluto) del pivot $a_{k,k}^{(k)}$, può amplificare gli eventuali errori di arrotondamento presenti negli elementi $a_{k,j}^{(k)}$ è utile effettuare il pivoting per righe anche se $a_{k,k}^{(k)} \neq 0$ utilizzando come pivot quello maggiore.

Questa operazione viene effettuata tramite la moltiplicazione a sinsitra per la matrice di permutazione P, cioè la matrice ottenuta dalla matrice identità in cui si sono scambiate le righe corrispondenti.

Esempio 6. Supponendo di avere una matrice A della forma:

$$A = \begin{bmatrix} 5 & 0 & 0 \\ 1 & 0 & 0 \\ 3 & 0 & 0 \end{bmatrix}$$

e di voler scambiare le righe 2 e 3 di tale matrice per ottenere la matrice:

$$B = \begin{bmatrix} 5 & 0 & 0 \\ 3 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

moltiplicando A per l'oppurtuna matrice P è posibile ottenere B:

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

$$B = PA = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 5 & 0 & 0 \\ 1 & 0 & 0 \\ 3 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 5 & 0 & 0 \\ 3 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

Se nell'operazione di fattorizzazione avviene il pivoting, alla fine del processo avrò L ed U e la matrice di permutazione P, prodotto di tutte le permutazioni effettuate. In questo caso, per risolvere il sistema, dovrò risolvere:

$$PA\overline{x} = P\overline{b} \quad \Rightarrow \quad LU\overline{y} = P\overline{b} \quad \Rightarrow \begin{cases} L\overline{y} = P\overline{b} \\ U\overline{x} = \overline{y} \end{cases}$$

È anche possibile effettuate il pivoting totale, cioè cercare l'elemento maggiore all'interno della sottomatrice $A_k^{(k)} = \left\{a_{i,j}^{(k)}\right\}_{i,j=k,\ldots,n}$ e scambiare le relative righe e colonne (si veda la Figura 1).

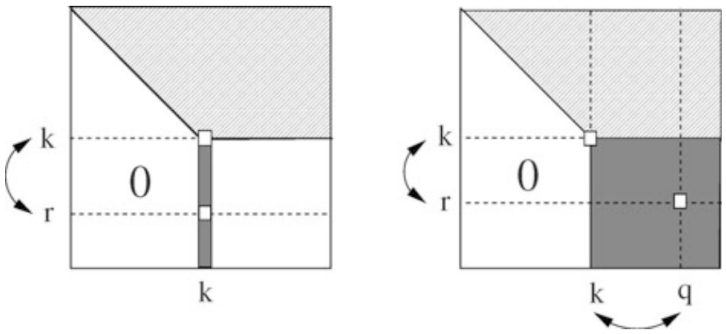


Figura 1: Pivotazione per righe (a sinistra) o totale (a destra). In grigio più scuro, le aree interessate dalla ricerca dell'elemento pivotale

Il **vantaggio** della fattorizzazione LU rispetto al MEG, per la risoluzione di sistemi è che L ed U dipendono dalla sola A e non dal termine noto, la stessa fattorizzazione può essere utilizzata per risolvere diversi sistemi lineari con la stessa matrice A dei coefficienti, ma con termine noto $\overline{b}$ variabile.

Matlab 1. In Matlab il comando per eseguire la fattorizzazione LU su una matrice A è:

```
[L U P]=lu(A)
y=fwsub(L,P*b);
x=bksub(U,y);
```

5.5 Fattorizzazione di Cholesky

Rappresenta un metodo diretto alternativo a LU.

Definizione 8 (Metodo Cholesky)**.** Permette di riscrivere la matrice A come:

$$A = Q^T Q$$

$$Q^T \rightarrow \text{Triangolare inferiore}$$

$$Q \rightarrow \text{Triangolare superiore}$$

Costo computazionale 4. Il costo computazionale associato a questo metodo è la metà di quello di LU:

$$\#\text{flops} = \frac{n^3}{3}$$

5.6 Errore e condizionamento

Definizione 9 (Errore relativo)**.** Sia $\hat{x}$ la soluzione numerica calcolata e sia $\overline{x}$ la soluzione esatta, si definisce errore relativo la quantità:

$$E = \frac{\|\overline{x} - \hat{x}\|}{\|\overline{x}\|}$$

Definizione 10 (Residuo)**.** Sia $\hat{x}$ la soluzione numerica calcolata per il sistema lineare del tipo $A\overline{x} = \overline{b}$, si definisce residuo la quantità:

$$\overline{r} = \overline{b} - A\hat{x} = A\overline{x} - A\hat{x} = A\left(\overline{x} - \hat{x}\right)$$

Allora:

$$(\overline{x} - \hat{x}) = A^{-1}\overline{r}$$

5.6.1 Numero di condizionamento

Definizione 11 (Norma di matrice)**.** La norma di una matrice A è definita come:

$$\|A\| = \underset{\overline{x} \in \mathbb{R}^n}{max} \frac{\|A\overline{x}\|}{\|\overline{x}\|}$$

Poiché $\forall\, \overline{x} \in \mathbb{R}^n$ può essere scritto come $\overline{x} = \|\overline{x}\| \cdot \frac{\overline{x}}{\|\overline{x}\|} = \|\overline{x}\| \cdot \overline{y}$, dove $\overline{y}$ è il versore che indica la direzione di $\overline{x}$, si ha:

$$\frac{\|A\overline{x}\|}{\|\overline{x}\|} = \frac{\|A\left(\|\overline{x}\| \cdot \overline{y}\right)\|}{\|\overline{x}\|} = \|\overline{x}\| \frac{\|A\overline{y}\|}{\|\overline{x}\|}$$

da cui:

$$\|A\| = \underset{\substack{\mathbf{y} \in \mathbb{R}^n \\ \|\mathbf{y}\| = 1}}{max} \|A\mathbf{y}\|$$

Dalla definizione di $\|A\|$ si ha:

$$\|A\| \geq \frac{\|A\overline{x}\|}{\|\overline{x}\|} \quad \Longleftrightarrow \quad \|A\overline{x}\| \leq \|A\| \cdot \|\overline{x}\|$$

Osservazione 5.8. Se A è una matrice simmetrica (Def: 5) e definita positiva (Def: 6) allora $\|A\| = \lambda_{max}$ cioè al massimo degli autovalori.

Dimostrazione. ....

□

Definizione 12 (Numero di condizionamento)**.** Si definisce numero di condizionamento $K(A)$ della matrice A , la quantità:

$$K(A) = \|A\| \cdot \|A^{-1}\|$$

- Se il numero di condizionamento è molto grande, si dice che la matrice è **malcondizionata**.
- Se il numero di condizionamento è piccolo (vicino al valore unitario), si dice che la matrice è **bencondizionata**.

Matlab 2. Il comando Matlab per il numero di condizionamento di una matrice A è:

$$\text{cond}(A)$$

Osservazione 5.9. Nel caso ideale il numero di condizionamento $K(A)$ di una matrice A dovrebbe essere pari a 1 (Def: 3), tuttavia gli errori di approssimazione che si generano in Matlab portano ad avere $K(A) \neq 1$

Proposizione 5.4. Esiste una relazione tra l'errore relativo E e il numero di condizionamento:

$$E \leq K(A) \frac{\|\overline{r}\|}{\|\overline{b}\|}$$

Dimostrazione. Per definizione di errore relativo (Def: 9) si ha che:

$$E = \frac{\|\overline{x} - \hat{x}\|}{\|\overline{x}\|} = \frac{\|A^{-1}\overline{r}\|}{\|\overline{x}\|}$$

- Per la definizione di norma di matrice (Def: 11) :

$$E = \frac{\|A^{-1}\overline{r}\|}{\|\overline{x}\|} \leq \frac{\|A^{-1}\| \cdot \|\overline{r}\|}{\|\overline{x}\|}$$

- Si può inoltre notare (sempre per la definizione di norma) che:

$$\frac{\|A\overline{x}\|}{\|\overline{x}\|} \leq \frac{\|A\| \cdot \|\overline{x}\|}{\|\overline{x}\|} \quad \Rightarrow \quad \overline{x} \geq \frac{\|A\overline{x}\|}{\|A\|} = \frac{\|\overline{b}\|}{\|A\|}$$

- Quindi si ottiene:

$$E \leq \frac{\|A^{-1}\overline{r}\|}{\frac{\|\overline{b}\|}{\|A\|}} = \|A\| \cdot \|A^{-1}\| \frac{\|\overline{r}\|}{\|\overline{b}\|} = K(A) \frac{\|\overline{r}\|}{\|\overline{b}\|} \tag{2}$$

□

Osservazione 5.10. Se A è una matrice simmetrica e definita positiva, l'osservazione 5.8 implica che il numero di condizionamento è:

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

dove λ_{max} e λ_{min} sono rispettivamente il massimo e il minimo degli autovalori di A.

Osservazione 5.11. In genere il residuo $\|\overline{r}\|$ si mantiene piccolo, tuttavia può capitare che $K(A)$ sia tanto grande da non essere compensato dal residuo.

Esempio 7. Prendendo per esempio la matrice di Hilbert (Def: 2) di ordine 4:

$$K(H_4) > 1.5 \times 10^4 \quad e \quad \|\overline{r}\| \simeq 10^{-16}$$

- Passaggi tipici

$$\|A\overline{x}\|^2 = (A\overline{x})^T (A\overline{x}) = \sum_{i=1}^{n} (x_i)^2 (\lambda_i)^2 \leq ...$$

$$\|A\overline{x}\|^2 = (A\overline{x})^T (A\overline{x})$$

Ma ogni vettore $\overline{x}$ in funzione degli autovettori

6 Risoluzione di sistemi lineari con metodi iterativi

I metodi iterativi per risolvere sistemi lineare del tipo $A\overline{x} = \overline{b}$ consistono nel costruire una successione del tipo $\left\{\overline{x}^{(k)}\right\}_{k\in\mathbb{N}}$ tale che $\overline{x}^{(k)}$ converga alla soluzione $\overline{x}$ per k che tende all'infinito, per qualunque vettore iniziale $\overline{x}^{(0)}$, scelto arbitrariamente, cioè:

$$\left\{\overline{x}^{(k)}\right\}_{k\in\mathbb{N}} \quad \text{t.c.} \quad \overline{x}^{(k)} \overset{k\to\infty}{\longrightarrow} \overline{x}$$

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

Questi metodi si basano sostanzialmente su tre passi fondamentali:

1. Determinazione del metodo, cioè come determinare $x^{(k)}$.
2. Dimostrazione della convergenza del metodo.
3. Stima dell'errore per determinare il criterio di arresto.

6.1 Metodo del gradiente

Quando si può usare 2. Quando A è simmetrica e definita positiva, si può usare perchè il metodo del gradiente converge.

A cosa serve 1. Trova una soluzione ad un sistema lineare del tipo Ax=b

Cosa fa 2. Trova il punto di minimo $\overline{x}$ di una funzione Φ partendo da un punto $\overline{x}^{(0)} \in \mathbb{R}^n$.

Vale infatti il seguente:

Teorema 9. *Si consideri* $\Phi(\overline{y}) = \frac{1}{2}\overline{y}^T A\overline{y} - \overline{y}^T\overline{b}$ *con* $\Phi : \mathbb{R}^n \to \mathbb{R}$ *dove* A *e* $\mathbf{b}$ *sono la matrice dei coefficienti e il temine noto del sistema. Allora* $\overline{x}$ *è soluzione di* $A\overline{x} = \overline{b}$ *se e solo se* $\overline{x}$ *è punto di minimo assoluto di* Φ *.*

$$\overline{x} \text{ soluzione di } A\overline{x} = \overline{b} \quad \Longleftrightarrow \quad \nabla\Phi(\overline{x}) = \overline{0}$$

Osservazione 6.1. La funzione $\Phi(\overline{y}) = \frac{1}{2}\overline{y}^T A\overline{y} - \overline{y}^T\overline{b}$ con $\Phi : \mathbb{R}^n \to \mathbb{R}$ nel caso $n = 1$ rappresenta una parabola, e nel caso $n = 2$ un paraboloide. Infatti:

$$n = 1 \Rightarrow \Phi(\overline{y}) = \frac{a}{2}y^2 - by$$

$$n = 2 \Rightarrow \Phi(\overline{y}) = \frac{1}{2}\begin{bmatrix} y_1 & y_2 \end{bmatrix}\begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} - \begin{bmatrix} y_1 & y_2 \end{bmatrix}\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \frac{1}{2}\left(a_{11}y_1^2 + 2a_{12}y_1y_2 + a_{22}y_2^2\right) - y_1b_1 - y_2b_2$$

Dimostrazione. Si dimostrano le due implicazioni:

1. $\Rightarrow$ Si suppone che $\overline{x}$ sia soluzione di $A\overline{x} = \overline{b}$ e si dimostra che sia anche il minimo della funzione Φ :

$$\Phi(\overline{x} + \overline{v}) = \frac{1}{2}(\overline{x} + \overline{v})^T A(\overline{x} + \overline{v}) - (\overline{x} + \overline{v})^T\overline{b} =$$

$$= \overbrace{\frac{1}{2}\overline{x}^T A\overline{x} - \overline{x}^T\overline{b}}^{\Phi(\overline{x})} + \frac{1}{2}\overline{v}^T A\overline{x} + \frac{1}{2}\overline{x}^T A\overline{v} + \frac{1}{2}\overline{v}^T A\overline{v} - \overline{v}^T\overline{b} =$$

$$= \Phi(\overline{x}) + \frac{1}{2}\overline{v}^T\overline{b} + \frac{1}{2}\overline{x}^T A\overline{v} + \frac{1}{2}\overline{v}^T A\overline{v} - \overline{v}^T\overline{b} =$$

$$\text{Ricordando che: } \left(\overline{x}^T A\overline{v}\right) = \left(\overline{x}^T A\overline{v}\right)^T = \overline{v}^T A\overline{x} = \overline{v}^T\overline{b}$$

Poichè A è una matrice simmetrica (Def: 5 , Prop: 3.4):

$$= \Phi(\overline{x}) + \frac{1}{2}\overline{v}^T A\overline{v} \geq \Phi(\overline{x}) + \frac{1}{2}\lambda_{min}\|\overline{v}\|^2 > \Phi(\overline{x}) \qquad \forall\, \mathbf{v} \in \mathbb{R}^n,\ \mathbf{v} \neq \mathbf{0}$$

L'ultima disuguaglianza è valida poichè A è definita positiva (Prop: 4.1).

Quindi $\forall\, \overline{v} \neq 0$ vale che $\Phi(\overline{x} + \overline{v}) > \Phi(\overline{x})$, quindi $\overline{x}$ è il punto di minimo della funzione Φ .

2. $\Leftarrow$ Si suppone che $\overline{x}$ sia punto di minimo della funzione Φ e si dimostra che è soluzione di $A\overline{x} = \overline{b}$:
   Se $\overline{x}$ è punto di minimo della funzione Φ, allora per il teorema di Fermat $\nabla\Phi(\overline{x}) = \overline{0}$

$$\Phi(y_1, \ldots, y_n) = \frac{1}{2}\sum_{i,j=1}^{n} a_{ij}y_iy_j - \sum_{i=1}^{n} b_iy_i$$

$$\frac{\partial\Phi}{\partial y_k} = \overbrace{\frac{1}{2}\sum_{j=1}^{n} a_{kj}y_j}^{\text{Se } i=k} + \overbrace{\frac{1}{2}\sum_{i=1}^{n} a_{ki}y_i}^{\text{Se } j=k} - b_k = \frac{1}{2}\overbrace{\left(\frac{1}{2}\sum_{j=1}^{n} a_{kj}y_j + \frac{1}{2}\sum_{i=1}^{n} a_{ki}y_i\right)}^{\text{Sono uguali perchè } A \text{ è simmetrica}} - b_k =$$

$$= \sum_{i=1}^{n} a_{ki}y_i - b_k = \left(A\overline{y} - \overline{b}\right)_k$$

Ovvero il k-esimo $A\overline{y} - \overline{b}$

$$\Rightarrow \nabla\Phi(\overline{y}) = A\overline{y} - \overline{b} = -\overline{r}$$

$$\Rightarrow \nabla\Phi(\overline{x}) = -\overline{r} = \overline{0} = -\left(A\overline{x} - \overline{b}\right)$$

Si osservi che in realtà per questa seconda parte di dimostrazione non abbiamo utilizzato che A è definita positiva. Quindi l'essere punto di minimo è condizione sufficiente per essere soluzione per una qualsiasi matrice A simmetrica.

Siccome per il teorema di Fermat deve essere uguale a zero, implica che anche il residuo $\overline{r}$ sia pari a zero e che quindi $\overline{x}$ sia soluzione di $A\overline{x} = \overline{b}$

□

Quindi si tratta di costruire un algoritmo per trovare il punto di minimo di $\Phi(\mathbf{y})$. Utilizzeremo il caso $n = 2$ per farci guidare dall'intuizione geometrica.

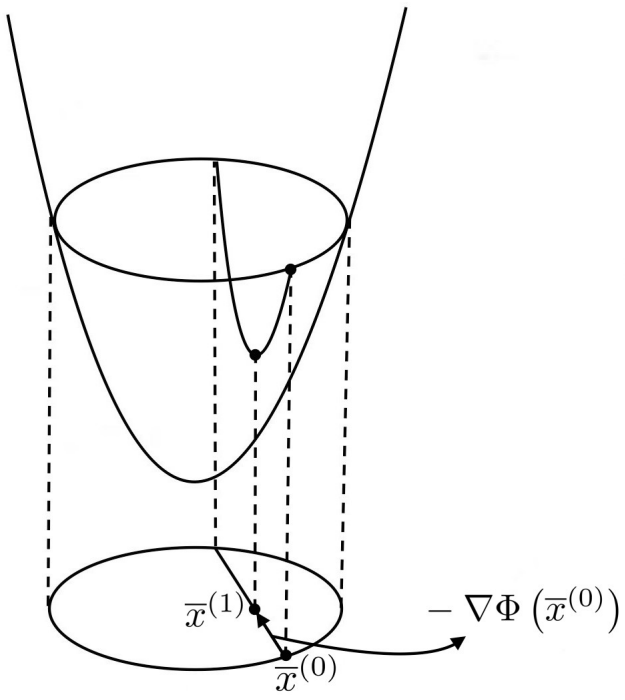


Figura 2: Illustrazione metodo del gradiente.

Per determinare $\overline{x}$ si procede come segue:

1. Si sceglie un punto $\overline{x}^{(0)} \in \mathbb{R}^n$ arbitrario.
2. Si considera l'insieme di livello $\mathscr{L}_{\Phi\left(\overline{x}^{(0)}\right)}$, che in questo caso è un'ellisse.

   L'obiettivo è quello di avvicinarsi al punto di minimo della funzione Φ ; da $\mathbf{x}^{(0)}$ la direzione di massima decrescita è $-\nabla\Phi\left(\overline{x}^{(0)}\right)$.
3. Si considera la retta $\overline{x}\left(\alpha\right) = \overline{x}^{(0)} + \alpha\overline{r}^{(0)}$, da $\overline{x}^{(0)}$ nella direzione opposta al gradiente e valuto Φ lungo questa linea. In generale si ottiene una funzione in una sola variabile, di cui è facile trovare il minimo. Graficamente, nel caso $n = 2$, equivarrebbe a "tagliare" il paraboloide con un piano verticale ottenendo una parabola.
4. Si trova il punto di minimo della funzione $F\left(\alpha\right) = \Phi\left(\mathbf{x}^{(k)} + \alpha\mathbf{r}^{(k)}\right)$ (nel caso $n = 2$ di una parabola, come si vede in Figura 2), che costituisce il punto $\overline{x}^{(1)}$ da cui si parte per l'iterazione successiva.

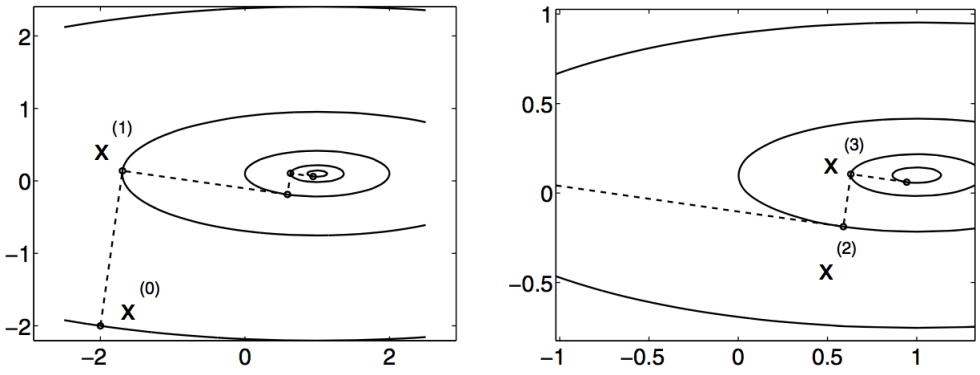


Figura 3: Le prime iterate generate dal metodo del gradiente sulle curve di livello di Φ

- In generale supponendo di avere $\overline{x}^{(k)}$ si cerca $\overline{x}^{(k+1)}$ della forma:

$$\overline{x}^{(k+1)} = \overline{x} - \alpha \overbrace{\nabla\Phi\left(\overline{x}^{(k)}\right)}^{\text{Direzione}} = \overline{x}^{(k)} + \alpha\overline{r}^{(k)}$$

con $\overline{r}^{(k)} = \overline{b} - A\overline{x}^{(k)}$

minimizzando quindi $F(\alpha) = \Phi\left(\mathbf{x}^{(k)} + \alpha\mathbf{r}^{(k)}\right)$. Quindi eguagliando a zero la derivata prima di F otteniamo:

$$F'(\alpha) = \nabla\Phi\left(\overline{x}^{(k+1)}\right) = \nabla\Phi\left(\overline{x}^{(k)} + \alpha\overline{r}^{(k)}\right)\overline{r}^{(k)^T} =$$

per definizione di derivata di una funzione composta in $\mathbb{R}^n$. Ricordando la 2 si ha :

$$= \overline{r}^{(k)^T}\left(A\left(\overline{x}^{(k)} + \alpha\overline{r}^{(k)}\right) - \overline{b}\right) = \overline{r}^{(k)^T}\left(-\overline{r}^{(k)} + \alpha A\overline{r}^{(k)}\right) =$$

$$= -\overline{r}^{(k)^T}\overline{r}^{(k)} + \alpha\overline{r}^{(k)^T}A\overline{r}^{(k)} = 0$$

$$F'(\alpha) = 0 \Longleftrightarrow \alpha = \alpha^{(k)} = \frac{\overline{r}^{(k)^T}\overline{r}^{(k)}}{\overline{r}^{(k)^T}A\overline{r}^{(k)}}$$

$$\alpha^{(k)} = \frac{\overline{r}^{(k)^T}\overline{r}^{(k)}}{\overline{r}^{(k)^T}A\overline{r}^{(k)}} \tag{3}$$

Algoritmo 3.

1. *Si sceglie un $\overline{x}^{(0)}$ arbitrario.*
2. *Si calcola $\overline{r}^{(k)}$ (direzione di discesa):*
$$\overline{r}^{(k)} = \overline{b} - A\overline{x}^{(k)}$$
3. *Si calcola $\alpha^{(k)}$ (lunghezza del passo):*
$$\alpha^{(k)} = \frac{\overline{r}^{(k)^T}\overline{r}^{(k)}}{\overline{r}^{(k)^T}A\overline{r}^{(k)}}$$
4. *Si calcola $\overline{x}^{(k+1)}$:*
$$\overline{x}^{(k+1)} = \overline{x}^{(k)} + \alpha^{(k)}\overline{r}^{(k)}$$

Osservazione 6.2. Il denominatore della 3 è diverso da zero perchè A è definita positiva (Def: 6) e simmetrica (Def: 5), e quando $\overline{r}^{(k)} = \overline{0}$ significa che si è trovata la soluzione al passaggio precedente.

Osservazione 6.3. Il metodo del gradiente fa parte dei metodi di Richardson dinamici, ovvero un metodo per cui l'elemento $\overline{x}^{(k+1)}$ è ottenuto a partire da quello precedente $\overline{x}^{(k)}$, a cui viene sommata un altro elemento (in questo caso pari a $\overline{r}^{(k)}$) moltiplicato per un coefficiente α, detto parametro di rilassamento (o di accelerazione):

- Se α è costante (indipendente dall'iterazione k) si parla di metodo di Richardson stazionario.
- Se $\alpha = \alpha^{(k)}$ (variabile e dipendente dall'iterazione k) si parla di metodo di Richardson dinamico.

6.2 Metodo del gradiente coniugato

Quando si può usare 3. Quando A è simmetrica e definita positiva.

A cosa serve 2. Serve per trovare la soluzione $\overline{x}$ del sistema lineare del tipo $A\overline{x} = \overline{b}$

Cosa fa 3. Si cerca di migliorare la velocità di convergenza (Th: 12) del metodo del gradiente, scegliendo una direzione diversa da quella identificata da $-\nabla\Phi$ (data una generica direzione di discesa $\overline{p}^{(k)}$, troveremo il valore $\alpha^{(k)}$ come quel valore di α che rende minimo $\Phi\left(\overline{x}^{(k)} + \alpha\overline{p}^{(k)}\right)$).

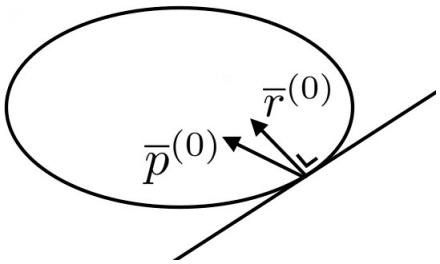


Figura 4: Direzione di $\overline{p}^{(0)}$ rispetto a $\overline{r}^{(0)}$

È possibile trovare delle direzioni $\overline{p}^{(k)}$, lungo cui muoversi, migliori rispetto a $\overline{r}^{(k)}$, rispetto cioè alla direzione definita dal gradiente ($\overline{r}^{(k)} = -\nabla\Phi\left(\overline{x}^{(k)}\right)$); si tratta cioè di una direzione per cui il metodo converge più velocemente alla soluzione $\overline{x}$.

Come si può osservare dalla figura 4, $\overline{p}^{(0)}$ punta direttamente verso la soluzione $\overline{x}$ (che in questo caso coincide con il centro dell'ellisse, si sta sempre facendo riferimento all'esempio del paraboloide), e descrive quindi un angolo rispetto a $\overline{r}^{(0)}$ che per costruzione deve essere minore di $\frac{\pi}{2}$ (deve puntare verso l'interno).

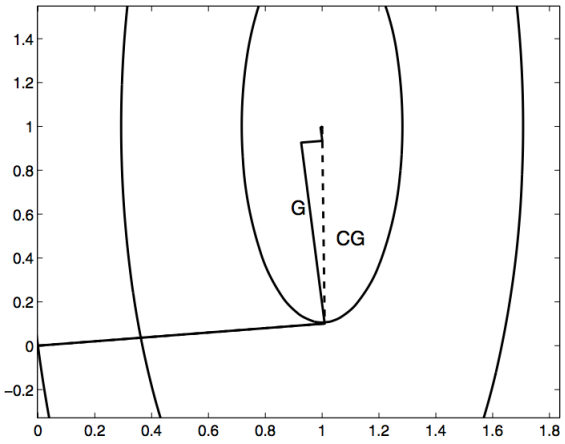


Figura 5: Le direzioni di discesa del gradiente coniugato (in linea tratteggiata e de- notate con CG) e quelle del gradiente (in linea continua ed indicate con G). Si noti come il metodo CG in due iterazioni abbia già raggiunto la soluzione.

Osservazione 6.4. Dato $\mathbf{p}^{(k)}$, l' $\alpha^{(k)}$ che minimizza $\Phi\left(\overline{x}^{(k)} + \alpha^{(k)}\overline{p}^{(k)}\right)$ lo si trova eseguendo gli stessi passaggi che sono stati effettuati per il calcolo nel metodo del gradiente (Eq: 3), ottenendo:

$$\alpha^{(k)} = \frac{\overline{p}^{(k)^T}\overline{r}^{(k)}}{\overline{p}^{(k)^T}A\overline{p}^{(k)}} \tag{4}$$

Osservazione 6.5. Con questa scelta di $\alpha^{(k)}$ si ottiene che il resituo al passaggio $k+1$ è ortogonale a $\overline{p}^{(k)}$:

$$\overline{r}^{(k+1)} \perp \overline{p}^{(k)}$$

Il problema a questo punto è determinare la direzione $\overline{p}^{(k)}$.

Determinazione di $\overline{p}^{(k)}$

A differenza del metodo del gradiente, ogni passo deve essere ottimale tenendo conto anche dei passaggi precedenti: ad ogni passaggio si sceglie una direzione che sia ottimale non solo rispetto all'ultimo, ma a tutti i passaggi precedenti.

1. Si scelgono arbitrariamente $\overline{x}^{(0)}$ (punto di partenza) e $\overline{p}^{(0)}$ (ad esempio uguale a $\overline{r}^{(0)}$, anche se non è necessario).

2. Si trova il punto $\overline{x}^{(1)} = \overline{x}^{(0)} + \alpha^{(0)}\overline{p}^{(0)}$ con $\alpha^{(0)}$ che minimizza $\Phi\left(\overline{x}^{(0)} + \alpha\overline{p}^{(0)}\right)$:

$$\alpha^{(0)} = \frac{\overline{p}^{(0)^T}\overline{r}^{(0)}}{\overline{p}^{(0)^T}A\overline{p}^{(0)}}$$

3. Si trova poi il punto $\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \alpha^{(1)}\mathbf{p}^{(1)}$ nella forma $\mathbf{x}^{(2)} = \mathbf{x}^{(0)} + \alpha^{(0)}\mathbf{p}^{(0)} + \alpha^{(1)}\mathbf{p}^{(1)}$ e si cercano $\alpha^{(0)}$ e $\alpha^{(1)}$ tali che minimizzino $\Phi\left(\overline{x}^{(0)} + \alpha^{(0)}\overline{p}^{(0)} + \alpha^{(1)}\overline{p}^{(1)}\right) = F\left(\alpha^{(0)}, \alpha^{(1)}\right)$, cioè una funzione a due variabili. $\overline{x}^{(2)}$ è il punto di minimo di $\Phi\left(\overline{v}\right)$ tra tutti i $\overline{v}$ della forma $\overline{v} = \overline{x}^{(0)} + \alpha^{(0)}\overline{p}^{(0)} + \alpha^{(1)}\overline{p}^{(1)} \Longrightarrow$ Si sta minimizzando lungo un piano (descritto dai vettori $\alpha^{(0)}\overline{p}^{(0)}$ e $\alpha^{(1)}\overline{p}^{(1)}$).

Osservazione 6.6. Anzichè minimizzare solo l'ultimo passaggio, si scrive il vettore successivo come $\overline{x}^{(0)}$ più una combinazione lineare di tutte le direzioni scelte in precedenza, e si fa in modo che $\overline{x}^{(k+1)}$ sia minimo rispetto a tutte le precedenti direzioni. Cioè si minimizza prima lungo un'unica direzione (quindi lungo una retta), poi lungo due direzioni (quindi rispetto ad un piano) e così via.

Dal teorema di Fermat nel punto di minimo deve essere $\nabla F\left(\alpha^{(0)},\ \alpha^{(1)}\right) = \mathbf{0}$

4. Si calcola quindi $\nabla F\left(\alpha^{(0)}, \alpha^{(1)}\right)$:

$$\frac{\partial F}{\partial \alpha^{(0)}} = \overline{p}^{(0)^T}\nabla\Phi\left(\overline{x}^{(0)} + \alpha^{(0)}\overline{p}^{(0)} + \alpha^{(1)}\overline{p}^{(1)}\right) = \overline{p}^{(0)^T}\left[A\left(\overline{x}^{(0)} + \alpha^{(0)}\overline{p}^{(0)} + \alpha^{(1)}\overline{p}^{(1)}\right) - \overline{b}\right] =$$

$$= \overline{p}^{(0)^T}\left(-\overline{r}^{(0)} + \alpha^{(0)}A\overline{p}^{(0)} + \alpha^{(1)}A\overline{p}^{(1)}\right) = 0$$

$$= -\overline{p}^{(0)^T}\overline{r}^{(0)} + \alpha^{(0)}\overline{p}^{(0)^T}A\overline{p}^{(0)} + \alpha^{(1)}\overline{p}^{(0)^T}A\overline{p}^{(1)} = 0$$

(Per i passaggi vedi quelli effettuati per il calcolo della 3)

$$\frac{\partial F}{\partial \alpha^{(1)}} = \overline{p}^{(1)^T}\nabla\Phi\left(\overline{x}^{(0)} + \alpha^{(0)}\overline{p}^{(0)} + \alpha^{(1)}\overline{p}^{(1)}\right) = \overline{p}^{(1)^T}\left[A\left(\overline{x}^{(0)} + \alpha^{(0)}\overline{p}^{(0)} + \alpha^{(1)}\overline{p}^{(1)}\right) - \overline{b}\right] =$$

$$= \overline{p}^{(1)^T}\left(-\overline{r}^{(0)} + \alpha^{(0)}A\overline{p}^{(0)} + \alpha^{(1)}A\overline{p}^{(1)}\right) =$$

$$= -\overline{p}^{(1)^T}\overline{r}^{(0)} + \alpha^{(0)}\overline{p}^{(1)^T}A\overline{p}^{(0)} + \alpha^{(1)}\overline{p}^{(1)^T}A\overline{p}^{(1)} = 0$$

Osservazione 6.7. Osserviamo che se nella prima equazione si ottiene lo stesso $\alpha^{(0)}$ calcolato al primo passo se $\alpha^{(1)}\overline{p}^{(0)^T}A\overline{p}^{(1)} = 0$

5. Scegliamo $\overline{p}^{(1)}$ tale che $\overline{p}^{(0)^T}A\overline{p}^{(1)} = 0$, si dice che $\overline{p}^{(1)}$ è **A-ortogonale** a $\overline{p}^{(0)}$, in questo modo:

$$\alpha^{(0)} = \frac{\overline{p}^{(0)^T}\overline{r}^{(0)}}{\overline{p}^{(0)^T}A\overline{p}^{(0)}}$$

$$\alpha^{(1)} = \frac{\overline{p}^{(1)^T}\overline{r}^{(1)}}{\overline{p}^{(1)^T}A\overline{p}^{(1)}}$$

Osservazione 6.8. Dalla scelta di $\mathbf{p}^{(1)}$ A-ortogonale a $\mathbf{p}^{(0)}$, posso mettere $\mathbf{r}^{(1)}$ nell'espressione di $\alpha^{(k)}$ al posto di $\mathbf{r}^{(0)}$ perchè $\mathbf{p}^{(1)^T}\mathbf{r}^{(0)} = \mathbf{p}^{(1)^T}\mathbf{r}^{(1)}$

Osservazione 6.9. *A* corregge il vettore per far sì che punti verso il centro.

Quindi nel caso generale si cerca:

$$\overline{x}^{(k+1)} = \overline{x}^{(0)} + \overbrace{\sum_{j=0}^{k} \alpha^{(j)} \overline{p}^{(j)}}^{\text{Combinazione lineare}} \quad \text{t.c.} \quad \overline{x}^{(k+1)} = \overbrace{\text{argmin}\Phi\left(\overline{v}\right)}^{\text{Punto di minimo di } \Phi}$$

$$\text{Tra tutti i } \overline{v} = \overline{x}^{(0)} + \sum_{j=0}^{k} \beta^{(j)} \overline{p}^{(j)}$$

Si tratta cioè del minimo sullo spazio di k dimensioni.

Con conti analoghi a quelli precedenti si trova che la scelta ottimale di $\overline{p}^{(k)}$ è che sia A-ortogonale ai $\overline{p}^{(j)}$ con $j = 1, \ldots, k-1$, tale che sia cioè ortogonali a tutti.

$$\overline{p}^{(k)^T} A \overline{p}^{(k)} = 0$$

$$\overline{p}^{(k)^T} A \left[\overline{r}^{(k+1)} - \beta^{(k)} \overline{p}^{(k)}\right] = \overline{p}^{(k)^T} A \overline{r}^{(k+1)} - \beta^{(k)} \overline{p}^{(k)^T} A \overline{p}^{(k)} = 0$$

$$\beta^{(k)} = \frac{\overline{p}^{(k)^T} A \overline{r}^{(k+1)}}{\overline{p}^{(k)^T} A \overline{p}^{(k)}}$$

$$\alpha^{(k)} = \frac{\overline{p}^{(k)^T} \overline{r}^{(k)}}{\overline{p}^{(k)^T} A \overline{p}^{(k)}}$$

Teorema 10. *Con questa scelta $\overline{p}^{(k+1)}$ è A-ortogonale non solo a $\overline{p}^{(k)}$ ma a tutti i $\overline{p}^{(j)}$ (con $j = 0, \ldots, k$) e siccome la matrice A è simmetrica (Def: 5) e definita positiva (Def: 6), i vettori sono linearmente indipendenti.*

$$\overline{p}^{(k)^T} A \overline{p}^{(k)} = 0$$

Dimostrazione.

L'unica combinazione lineare che possa dare il vettore nullo è quella con tutti i coefficienti nulli, prendendo una combinazione lineare:

$$\overline{v} = \overline{0} = \sum_{j=0}^{k} v^{(j)} \overline{p}^{(j)}$$

$$\overline{p}^{(i)^T} A \overline{v} = \overline{p}^{(i)^T} A \left(\sum_{j=0}^{k} v^{(j)} \overline{p}^{(j)} \right) = \sum_{j=0}^{k} v^{(j)} \overbrace{\overline{p}^{(i)^T} A \overline{p}^{(j)}}^{*} = v^{(i)} \overline{p}^{(i)} A \overline{p}^{(i)} \neq 0$$

$$* \rightarrow \begin{array}{c} \text{È uguale a zero} \\ \text{tranne per } i = j \\ \text{(Per A-ortogonalità)} \end{array}$$

Siccome A è definita positiva (Def: 6):

$$\overline{p}^{(i)^T} A \overline{p}^{(i)} \geq \lambda_{min} \|\overline{p}^{(i)}\|^2 > 0$$

questo per il teorema th: 4.1, $\Rightarrow v^{(j)} = 0 \; \forall \, i$

Tale espressione di $\beta^{(k)}$ deriva dal fatto che $\overline{p}^{(k+1)}$ debba essere A-ortogonale a $\overline{p}^{(k)}$, cioè:

$$\overline{p}^{(k)^T} A \overline{p}^{(k)} = 0$$

$$\overline{p}^{(k)^T} A \left[\overline{r}^{(k+1)} - \beta^{(k)} \overline{p}^{(k)}\right] = \overline{p}^{(k)^T} A \overline{r}^{(k+1)} - \beta^{(k)} \overline{p}^{(k)^T} A \overline{p}^{(k)} = 0$$

Quindi $\overline{v} \in \mathbb{R}^n$ e $\overline{x}^{(n)}$ è il minimo di tutto $\mathbb{R}^n \Rightarrow$ È IL minimo.

□

Corollario 1. *Il metodo è esatto dopo al più n passi*

$$\overline{x}^{(k+1)} = argmin\Phi\left(\overline{v}\right) \quad con \quad \overline{v} = \overline{x}^{(0)} + \sum_{j=0}^{k} \beta^{(j)} \overline{p}^{(j)}$$

$$\beta^{(k)} = \frac{\overline{p}^{(k)^T} A \overline{r}^{(k+1)}}{\overline{p}^{(k)^T} A \overline{p}^{(k)}}$$

Dove $\beta^{(k)}$ è l'entità della correzione (dello spostamento) tra l'elemento $\overline{p}^{(k+1)}$ e quello $\overline{p}^{(k)}$:

$$\overline{p}^{(k+1)} = \overline{r}^{(k+1)} - \beta^{(k)} \overline{p}^{(k)}$$

Osservazione 6.10. Questa conclusione è vera in aritmetica esatta, in Matlab ci sono gli errori di calcolo ed è quindi necessario un criterio di arresto.

Algoritmo 4. *L'algoritmo si articola nei seguenti passi:*

1. *Si scelgono arbitrariamente $\overline{x}^{(0)}$ (punto di partenza) e $\overline{p}^{(0)}$*
2. *Si calcola $\alpha^{(k)}$:*
$$\alpha^{(k)} = \frac{\overline{p}^{(k)^T} \overline{r}^{(k)}}{\overline{p}^{(k)^T} A \overline{p}^{(k)}}$$
3. *Si calcola $\overline{x}^{(k+1)}$:*
$$\overline{x}^{(k+1)} = \overline{x}^{(k)} + \alpha^{(k)} \overline{p}^{(k)}$$
4. *Si calcola $\overline{r}^{(k+1)}$:*
$$\overline{r}^{(k+1)} = \overline{b} - A\overline{x}^{(k+1)} = \overline{r}^{(k)} - \alpha^{(k)} \overline{p}^{(k)}$$
5. *Si calcola $\overline{p}^{(k+1)}$:*
$$\overline{p}^{(k+1)} = \overline{r}^{(k+1)} - \beta^{(k)} \overline{p}^{(k)}$$
6. *Si calcola $\beta^{(k)}$ in modo tale che $\overline{p}^{(k+1)}$ sia A-ortogonale a $\overline{p}^{(k)}$:*
$$\beta^{(k)} = \frac{\overline{p}^{(k)^T} A \overline{r}^{(k+1)}}{\overline{p}^{(k)^T} A \overline{p}^{(k)}}$$

6.3 Velocità di convergenza

Definizione 13. Si definisce norma A (simmetrica e definita positiva) di $\overline{x}$ la quantità:

$$\|\overline{x}\|_A^2 = \overline{x}^T A \overline{x}$$

Osservazione 6.11. Le proprietà di questa norma sono:

- Se A è la matrice identità si ottiene la norma del vettore $\overline{x}$
- È sempre maggiore o uguale di zero.
- È uguale a zero solo se $\overline{x} = \overline{0}$
- Vale omogeneità:
$$\|\alpha \overline{v}\| = |\alpha| \cdot \|\overline{v}\|$$
- Disuguaglianza triangolare:
$$\|\overline{x} + \overline{y}\| \leq \|\overline{x}\| + \|\overline{y}\|$$

Teorema 11 (Velocità di convergenza metodo gradiente)**.** $\forall\, \overline{x}^{(0)} \in \mathbb{R}^n$:

$$\|\overline{x} - \overline{x}^{(k)}\|_A \leq \frac{K(A) - 1}{K(A) + 1} \|\overline{x} - \overline{x}^{(k-1)}\|_A$$

$$\|\overline{x} - \overline{x}^{(k)}\|_A \leq \left(\frac{K(A) - 1}{K(A) + 1}\right)^k \|\overline{x} - \overline{x}^{(0)}\|_A$$

Osservazione 6.12. Se $K(A) \gg 1 \quad \Rightarrow \quad \frac{K(A)-1}{K(A)+1} \simeq 1$ e quindi si ha una convergenza molto lenta.

Teorema 12 (Velocità di convergenza metodo gradiente coniugato)**.** $\forall\, \overline{x}^{(0)} \in \mathbb{R}^n \quad \overline{p}^{(0)} \in \mathbb{R}^n$:

$$\|\overline{x} - \overline{x}^{(k)}\| \leq 2 \left(\frac{\sqrt{K(A)} - 1}{\sqrt{K(A)} + 1}\right)^k \|\overline{x} - \overline{x}^{(0)}\|_A$$

Osservazione 6.13. Per la presenza del termine $\sqrt{K(A)}$ la situazione migliora rispetto al caso del metodo del gradiente, però se $K(A) \gg 1$ si ha che $\left(\frac{\sqrt{K(A)}-1}{\sqrt{K(A)}+1}\right) \simeq 1$ e la convergenza è lenta.

6.4 Criterio di arresto

I criteri di arresto più comuni sono due:

Criterio di arresto sul residuo normalizzato

Sul residuo normalizzato si sceglie una tolleranza $\varepsilon > 0$ e ci si arresta al primo $k = k_{min}$, ovvero il primo valore tale per cui:

$$\frac{\|\overline{r}^{(k_{min})}\|}{\|\overline{b}\|} \leq \varepsilon \qquad \left(\overline{r}^{(k)} = \overline{b} - A\overline{x}^{(k)}\right)$$

Ponendo $\hat{x} = \overline{x}^{(k_{min})}$ e facendo gli stessi conti effettuati per il caso dei metodi diretti (Eq: 2) si ottiene:

$$\frac{\|\overline{x} - \hat{x}\|}{\|\overline{x}\|} \leq K(A) \frac{\|\overline{r}^{(k_{min})}\|}{\|\overline{b}\|} \leq \varepsilon K(A)$$

Osservazione 6.14. Siccome di solito il numero di condizionamento $K(A)$ è un numero molto grande, per avere una buona precisione è necessario scegliere un numero ε molto piccolo. Di conseguenza il numero di condizionamento influenza anche il numero di iterazioni (quello minimo per avere una soluzione con una precisione accettabile).

Criterio di arresto sull'incremento

Si basa sulla differenza tra una iterazione e la successiva:

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

Quindi il criterio d'arresto, preso un $\varepsilon > 0$, corrisponde al primo $k = k_{min}$ tale per cui:

$$\|\overline{\delta}^{(k_{min})}\| \leq \varepsilon \|\overline{b}\|$$

6.5 Precondizionamento

Idea: anzichè risolvere il sistema $A\overline{x} = \overline{b}$ possiamo risolvere:

$$P^{-1}A\overline{x} = P^{-1}\overline{b}$$

e scegliere una matrice P abbastanza vicina ad A tale che:

$$P \approx A \xRightarrow{\text{t.c.}} P^{-1}A \approx I \quad \text{(Matrice identità)} \Rightarrow K\left(P^{-1}A\right) \approx 1$$

in modo tale da abbattere il numero di condizionamento, per velocizzare il metodo.

Idealmente, se si prendesse $P = A$ si avrebbe $K\left(P^{-1}A\right) = K\left(A^{-1}A\right) = K\left(I\right) = 1$. Tuttavia calolare A^{-1} equivale a risolvere il sistema $A\overline{x} = \overline{b}$, di conseguenza è necessario trovare una matrice P che sia facile da invertire da un lato, e che tale da rendere $K\left(P^{-1}A\right)$ vicino a 1.

Procederemo in realtà in questo modo: supponendo che P sia una matrice simmetrica e definita positiva, allora:

$$\exists\ P^{\frac{1}{2}} \text{ t.c. } \ P^{\frac{1}{2}}P^{\frac{1}{2}} = P$$

Infatti una matrice simmetrica e definita positiva è diagonalizzabile con tutti gli autovalori positivi e con matrice V degli autovettori ortogonale, e può quindi essere scritta come:

$$P = VDV^T$$

$$V^TV = I \rightarrow \text{Perchè } V \text{ è una matrice ortogonale}$$

$$D \rightarrow \text{Matrice diagonale con gli autovalori sulla diagonale principale}$$

Di coseguenza risulta essere:

$$P^{\frac{1}{2}} = VD^{\frac{1}{2}}V^T$$

dove $D^{\frac{1}{2}}$ è la matrice diagonale che sulla diagonale principale ha le radici quadrate degli autovalori $\sqrt{\lambda_i}$ (ciò è possibile perchè A è definita positiva).

Infatti:

$$P^{\frac{1}{2}}P^{\frac{1}{2}} = \left(VD^{\frac{1}{2}}V^T\right)\left(VD^{\frac{1}{2}}V^T\right) \overset{*}{=} VDV^T = P$$

Quindi si può scrivere:

$$A\overline{x} = \overline{b} \Longleftrightarrow P^{-\frac{1}{2}}A\overbrace{P^{-\frac{1}{2}}\overline{y}}^{\overline{x}} = P^{-\frac{1}{2}}\overline{b}$$

dove:

$$P^{-\frac{1}{2}} = \left(P^{\frac{1}{2}}\right)^{-1} = \left(VD^{\frac{1}{2}}V^T\right)^{-1} \overset{*}{=} \left(VD^{-\frac{1}{2}}V^T\right) \tag{5}$$

l'uguaglianza * è possibile grazie alle proprietà della matrice trasposta (Prop: 3.4). $D^{-\frac{1}{2}}$ è una matrice diagonale che sulla diagonale principale ha i reciproci delle radici degli autovalori;

si trova $\overline{y}$ e poi si calcola $\overline{x}$come $\overline{x} = P^{-\frac{1}{2}}\overline{y}$

$$P^{-\frac{1}{2}}AP^{-\frac{1}{2}} \rightarrow \text{Simmetrica e definita positiva}$$

- Per verificare la simmetria:

$$\left(P^{-\frac{1}{2}}AP^{-\frac{1}{2}}\right)^T = \left(P^{-\frac{1}{2}}\right)^T A\left(P^{-\frac{1}{2}}\right)^T$$

essendo A simmetrica per ipotesi (ricordando che per applicare il metodo del gradiente A deve essere simmetrica definita positiva), quindi $A^T = A$

Inoltre (considerando l'Eq: 5):

$$\left(P^{-\frac{1}{2}}\right)^T = \left(VD^{-\frac{1}{2}}V^T\right)^T = VD^{-\frac{1}{2}}V^T = P^{-\frac{1}{2}}$$

$$\Rightarrow \left(P^{-\frac{1}{2}}AP^{-\frac{1}{2}}\right)^T = P^{-\frac{1}{2}}AP^{-\frac{1}{2}} = A_P$$

- Per verificare che è definita positiva:

$$\overline{y}^TA_P\overline{y} = \overbrace{\overline{y}^TP^{-\frac{1}{2}}}^{\overline{x}^T}A\overbrace{P^{-\frac{1}{2}}\overline{y}}^{\overline{x}} = \overline{x}^TA\overline{x} > 0 \quad \text{tranne per } \overline{x} = \overline{0} \Leftrightarrow \overline{y} = \overline{0} \quad \left(\begin{array}{c}\text{Essendo } A\\ \text{definita positiva}\end{array}\right)$$

Osservazione 6.15. Si possono quindi applicare il metodo del gradiente o il metodo del gradiente coniugato a $P^{-\frac{1}{2}}AP^{-\frac{1}{2}}\overline{y} = P^{-\frac{1}{2}}\overline{b}$, con $\overline{x} = P^{-\frac{1}{2}}\overline{y}$

Esiste una vasta letteratura sulle possibili P facili da invertire e tali che $K\left(P^{-\frac{1}{2}}A\right) \approx 1$, un esempio è scegliere P come matrice diagonale avente gli elementi diagonali di A sulla diagonale:

$$\{a_{ii}\}_{i=1}^{n}$$

Quando si può usare 4. Si usa nei metodi iterativi (ad esempio nel metodo del gradiente).

A cosa serve 3. Serve a ridurre il numero di condizionamento della matrice.

Algoritmo 5. *Precondizionatori diagonali: in generale la scelta di P come la diagonale di A si può rivelare efficace nel caso in cui A sia una matrice simmetrica definita positiva. In generale la scelta di P come la diagonale di A si può rivelare efficace nel caso in cui A sia una matrice simmetrica definita positiva.*

Matlab 3 (Metodo del gradiente precondizionato)**.**

```
X=pcg(A,b);
```

Metodi Analitici e Numerici per l'ingegneria

Cerutti Maria Cristina

A cura di: Andrea Fuso, Gianpiero Gaeta

Indice

7 Approssimazione di autovalori e autovettori

7.1 Metodo delle potenze

Quando si può usare 5. Si usa quando A è diagonalizzabile, che significa che la molteplicità geometrica di tutti gli autovalori è uguale a quella algebrica, o alternativamente quando esiste una base di autovettori.

A cosa serve 4. Il metodo delle potenze serve per il calcolo dell'autovalore di modulo massimo, λ_1 , sotto opportune ipotesi.

Funzionamento del metodo

Euristicamente procediamo in questo modo:

- Dato un vettore iniziale $\mathbf{x}^{(0)}$, il metodo consente di calcolare la successione di vettori, il cui $k-esimo$ è:

$$\mathbf{x}^{(k)} = A^{(k)}\mathbf{x}^{(0)}$$

- Supponendo che A sia una matrice diagonalizzabile (prima ipotesi), esiste una base di autovettori, di conseguenza il vettore iniziale può essere scritto come una loro combinazione lineare:

$$\mathbf{x}^{(0)} = \sum_{i=1}^{n} \alpha^{(i)}\mathbf{v}^{(i)}$$

- Ricordando poi che $A\mathbf{v}^{(i)} = \lambda_i\mathbf{v}^{(i)}$ (con $i = 1, \dots, n$) e sfruttando la precedente relazione si ottiene:

$$\mathbf{x}^{(k)} = A^{(k)}\mathbf{x}^{(0)} = \sum_{i=1}^{n} \alpha^{(i)}A^{(k)}\mathbf{v}^{(i)} = \sum_{i=1}^{n} \alpha^{(i)}\lambda^{k}\mathbf{v}^{(i)}$$

- Come seconda ipotesi si assume $\alpha^{(1)} \neq 0$ e si ha:

$$\mathbf{x}^{(k)} = \alpha^{(k)}\lambda_1^k\mathbf{v}^{(1)} + \sum_{i=2}^{n} \alpha^{(i)}\lambda_i^k\mathbf{v}^{(i)} = \alpha^{(1)}\lambda_1^k \left(\mathbf{v}^{(1)} + \sum_{i=2}^{n} \left(\frac{\alpha^{(i)}}{\alpha^{(1)}}\right)\left(\frac{\lambda_i}{\lambda_1}\right)^k \mathbf{v}^{(i)}\right)$$

Osservazione 7.1. Il vettore iniziale $\mathbf{x}^{(0)}$ ha una componente non nulla lungo l'autovettore $\mathbf{v}^{(1)}$, corrispondente all'autovalore λ_1.

- Supponendo che esiste un autovalore λ_1 strettamente maggiore degli altri: $|\lambda_1| > |\lambda_2| \geq \cdots \geq \cdots > |\lambda_n|$ (terza ipotesi), i rapporti $\left|\frac{\lambda_i}{\lambda_1}\right| < 1$ e quindi $\left(\frac{\lambda_i}{\lambda_1}\right)^k$ convergono a zero per k che tende a $+\infty$, quindi si potrebbe dire che il vettore $\mathbf{x}^{(k)}$ tende ad essere parallelo all'autovettore $\mathbf{v}^{(1)}$.

Tale procedimento in realtà non converge ad un vettore in quanto:

$$\text{Underflow} \rightarrow \text{Quando } |\lambda_1| < 1,\ \lambda_1^k \text{ converge a zero}$$

$$\text{Overflow} \rightarrow \text{Quando } |\lambda_1| > 1,\ \lambda_1^k \text{ diverge a } +\infty$$

Per evitare questo tipo di problemi si normalizza il vettore $\overline{x}^{(k)}$ ad ogni iterazione, per evitare che la sua norma cresca o decresca eccessivamente:

$$\mathbf{y}^{(k)} = \frac{\mathbf{x}^{(k)}}{\|\mathbf{x}^{(k)}\|}$$

$$\mathbf{x}^{(k)} = A\mathbf{y}^{(k-1)}$$

$$\Rightarrow \quad \beta(k) = \frac{1}{\prod_{i=0}^{n} \|\mathbf{x}^{(k)}\|}$$

$$\Rightarrow \mathbf{x}^{(k)} = \beta(k)\,\lambda_1^k \left(\alpha^{(1)}\mathbf{v}^{(1)} + \sum_{i=2}^{n} \alpha^{(i)} \left(\frac{\lambda_i}{\lambda_1}\right)^k \mathbf{v}^{(i)} \right)$$

In questo modo $\beta(k)$ compensa λ_1^k :

Non diverge se $|\lambda_1| \gg 1$

Non converge a zero se $|\lambda_1| < 1$

Infatti:

$$\beta(k) = \frac{1}{\|A^{(k)}\mathbf{x}^{(0)}\|} = \frac{1}{\|\lambda_1^k \left(\alpha^{(1)}\mathbf{v}^{(1)} + \mathbf{q}^{(k)}\right)\|}$$

$$\mathbf{q}^{(k)} = \sum_{i=2}^{n} \alpha^{(i)} \left(\frac{\lambda_i}{\lambda_1}\right)^k \mathbf{v}^{(i)} \longrightarrow \mathbf{0} \quad (\text{Per } k \to +\infty)$$

$$\Rightarrow \beta(k)\,|\lambda_1|^k \longrightarrow \frac{1}{|\alpha^{(1)}| \cdot \|\mathbf{v}^{(1)}\|} \quad (\text{Per } k \to +\infty)$$

$$\Rightarrow \mathbf{x}^{(k)} \longrightarrow \frac{1}{|\alpha^{(1)}| \cdot \|\mathbf{v}^{(1)}\|} \alpha^{(1)}\mathbf{v}^{(1)} \quad (\text{Per } k \to +\infty)$$

Teorema 13. *Sia $A \in \mathbb{C}^{n\times n}$ una matrice diagonalizzabile i cui autovalori soddisfano la seguente relazione: $|\lambda_1| > |\lambda_2| \geq \cdots \geq \cdots > |\lambda_n|$, con λ_1 autovalore di modulo massimo. Assumendo $\alpha^{(1)} \neq 0$, esiste $C > 0$ tale che:*

$$\|\tilde{\mathbf{x}} - \mathbf{v}^{(1)}\| = C \left| \begin{array}{c} \lambda_2 \\ \lambda_1 \end{array} \right|^k + o\left(\left| \begin{array}{c} \lambda_2 \\ \lambda_1 \end{array} \right|^k \right) \qquad k \geq 1$$

dove:

$$\tilde{\mathbf{x}} = \frac{\mathbf{x}^{(k)}\|A^{(k)}\mathbf{x}^{(0)}\|}{\alpha^{(1)}} = \mathbf{v}^{(1)} + \sum_{i=1}^{n} \frac{\alpha^{(i)}}{\alpha^{(1)}} \left(\frac{\lambda_i}{\lambda_1}\right)^k \mathbf{v}^{(i)} \qquad k = 1, 2, \ldots$$

Osservazione 7.2. Questo metodo funziona se $\alpha^{(1)} \neq 0$, cioè se si è scelto un vettore $\mathbf{x}^{(0)}$ con componente lungo $\mathbf{v}^{(1)}$ diversa da zero. In questo caso però, nell'implementazione in Matlab, non bisogna preoccuparsi di tale problematica, poichè in questo caso gli errori di approssimazione giocano a proprio favore: se si avesse scelto $\mathbf{x}^{(0)}$ con $\alpha^{(0)} = 0$ le iterazioni successive avrebbero fatto comparire uan componente parallela a $\mathbf{v}^{(1)}$.

Osservazione 7.3. La velocità di convergenza di tale metodo è proporzionale a:

$$\text{Velocità di convergenza} \simeq \left|\frac{\lambda_2}{\lambda_1}\right|^k$$

cioè a quel termine che converge a zero più lentamente di tutti gli altri.

7.2 Metodo delle potenze generalizzato

Considerando l'inversa di A: A^{-1} . Se A è diagonalizzabile, lo è anche l'inversa, quindi se $|\lambda_{n-1}| > |\lambda_n|$ (autovalore di A), allora è possibile applicare il metodo delle potenze a A^{-1} per trovare λ_n, reciproco dell'autovalore di modulo massimo per A^{-1}.

Algoritmo 6 (Metodo delle potenze inverso)**.** *L'algoritmo si articola nei seguenti passi:*

1. *Si sceglie un* $\mathbf{x}^{(0)}$ *arbitrario (punto di partenza).*
2. *Si calcola* $\mathbf{y}^{(0)}$ *come:*
$$\mathbf{y}^{(0)} = \frac{\mathbf{x}^{(0)}}{\|\mathbf{x}^{(0)}\|}$$
3. *L'elemento* $\mathbf{x}^{(k+1)}$ *è poi calcolato come:*
$$\mathbf{x}^{(k+1)} = A^{-1}\mathbf{y}^{(k)}$$
Osservazione 7.4. $\mathbf{y}^{(k)}$ è conosciuto al passaggio precedente.
4. *Una volta calcolato* $\mathbf{x}^{(k+1)}$ *è possibile calcolare* $\mathbf{y}^{(k+1)}$ *:*
$$\mathbf{y}^{(k+1)} = \frac{\mathbf{x}^{(k+1)}}{\|\mathbf{x}^{(k+1)}\|}$$

Osservazione 7.5. Ad ogni passo si ha un sistema lineare di cui cambia il termine noto (in questo caso $\mathbf{y}^{(k)}$), di conseguenza viene molto comodo usare il metodo LU (Sezione: 5.4).

È possibile un'ulteriore generalizzazione: se esiste un autovalore di A vicino ad un certo $\mu \in \mathbb{C}$ allora è possibile applicare il metodo delle potenze inverse per determinare l'autovalore λ, che ha distanza minima da μ, applicando alla matrice $A - \mu I$

Criterio di arresto

$$\|\mathbf{y}^{(k+1)} - \mathbf{y}\| \leq \varepsilon$$

8 Equazioni e sistemi non lineari

8.1 Metodo di bisezione

Teorema 14 (Teorema degli zeri)**.** *Sia $f : [a, b] \to \mathbb{R}$ una funzione continua e sia il prodotto $f(a) f(b) < 0$ allora:*

$$\exists\, \alpha \ \in \ (a, b) \ \ t.c.\ f(\alpha) = 0$$

Dimostrazione. La dimostrazione è effettuata mediant eil metodo di bisezione:

- Si suppone (ad esempio) che $f(a) < 0$ e che $f(b) > 0$
- Si prende in considerazione punto medio c preso a metà dell'intervallo (a, b) :

$$a^{(0)} = a \quad ; \quad b^{(0)} = b \quad ; \quad c^{(0)} = \frac{a^{(0)} + b^{(0)}}{2}$$

- A questo punto si calcola $f\left(c^{(0)}\right)$ e si hanno tre possibilità:
  1. $f\left(c^{(0)}\right) = 0 \ \Rightarrow$ il metodo si arresta perchè è stato trovato quel valore di α tale che $f(\alpha) = 0$
  2. $f\left(c^{(0)}\right) < 0 \ \Rightarrow$ si considera il nuovo intervallo $\left[a^{(1)},\ b^{(1)}\right] = \left[c^{(0)},\ b^{(0)}\right]$
  3. $f\left(c^{(0)}\right) > 0 \ \Rightarrow$ si considera il nuovo intervallo $\left[a^{(1)},\ b^{(1)}\right] = \left[a^{(0)},\ c^{(0)}\right]$
- Il nuovo intervallo $\left[a^{(1)},\ b^{(1)}\right]$ è incluso nel precedente ($\left[a^{(1)},\ b^{(1)}\right] \subset \left[a^{(0)},\ b^{(0)}\right]$) e di ampiezza:

$$b^{(1)} - a^{(1)} = \frac{b^{(0)} - a^{(0)}}{2} = \frac{b - a}{2}$$

- Quindi generalizzando al k-esimo passaggio si ha:

$$\left[a^{(k)},\ b^{(k)}\right] \quad ; \quad c^{(k)} = \frac{a^{(k)} + b^{(k)}}{2} = \frac{b - a}{2^k}$$

- Si valuta il generico $c^{(k)}$ nella funzione f:
  1. Se $f\left(c^{(k)}\right) = 0 \ \Rightarrow$ il metodo si arresta perchè è stato trovato quel valore di α tale che $f(\alpha) = 0$
  2. Se $f\left(c^{(k)}\right) < 0 \ \Rightarrow$ si considera il nuovo intervallo $\left[a^{(k+1)},\ b^{(k+1)}\right] = \left[c^{(k)},\ b^{(k)}\right]$
  3. Se $f\left(c^{(k)}\right) > 0 \ \Rightarrow$ si considera il nuovo intervallo $\left[a^{(k+1)},\ b^{(k+1)}\right] = \left[a^{(k)},\ c^{(k)}\right]$
- Si ottiene quindi l'intervallo $\left[a^{(k+1)},\ b^{(k+1)}\right]$ tale che $f\left(a^{(k+1)}\right) < 0$ e $f\left(b^{(k+1)}\right) > 0$ e ampiezza pari a :

$$b^{(k+1)} - a^{(k+1)} = \frac{b^{(k)} - a^{(k)}}{2} = \frac{b^{(0)} - a^{(0)}}{2^{k+1}} = \frac{b - a}{2^{k+1}}$$

- Si costruiscono ora le successioni $\left\{a^{(k)}\right\}_{k\in\mathbb{N}}$ e $\left\{b^{(k)}\right\}_{k\in\mathbb{N}}$ tali che:

$$f\left(a^{(k)}\right) < 0 \quad ; \quad f\left(b^{(k)}\right) > 0$$

$$b^{(k)} - a^{(k)} = \frac{b - a}{2^k} \longrightarrow 0 \quad (\text{Per } k \to +\infty)$$

$$\left.\begin{matrix} \left\{a^{(k)}\right\} \text{ è non decrescente} \\ \left\{b^{(k)}\right\} \text{ è non crescente} \end{matrix}\right] \to \text{Sono limitate} \Rightarrow \quad \begin{matrix} \text{Teorema esistenza del limite} \\ \text{per le successioni monotone} \end{matrix}$$

Quindi $a^{(k)} \nearrow \overline{a}$ ($a^{(k)}$ tende per difetto a $\overline{a}$) e $b^{(k)} \searrow \overline{b}$ ($b^{(k)}$ tende per eccesso a $\overline{b}$), ma

$$\overline{b} - \overline{a} = \lim_{k\to+\infty} b^{(k)} - a^{(k)} = \lim_{k\to+\infty} \frac{b - a}{2^k} = 0 \quad \Rightarrow \quad \overline{b} = \overline{a} = \alpha \ (\text{Candidata radice dell'equazione})$$

$$\left.\begin{array}{ccccc} f(\alpha) & = & \lim\limits_{k\to+\infty} f\left(a^{(k)}\right) & \leq & 0 \\ & \uparrow & & \uparrow & \\ & \text{Per continuità} & & \text{Teorema permanenza segno} & \\ & \downarrow & & \downarrow & \\ f(\alpha) & = & \lim\limits_{k\to+\infty} f\left(b^{(k)}\right) & \geq & 0 \end{array}\right] \Rightarrow f(\alpha) = 0$$

□

Implementazione metodo di bisezione

Algoritmo 7. *L'algoritmo per il metodo di bisezione si articola nei seguenti passi:*

1. *Scelta degli estremi dell'intervallo $a^{(0)} = a$ e $b^{(0)} = b$ in modo tale che $f(a)f(b) < 0$:*

$$\Rightarrow x^{(0)} = \frac{b-a}{2}$$

   dove $x^{(0)}$ è l'ampiezza dell'intervallo iniziale.

2. *Per i passaggi $k \geq 1$:*
   (a) *Se $f\left(x^{(k-1)}\right) = 0 \Rightarrow \alpha = x^{(k-1)}$*
   (b) *Se $f\left(x^{(k-1)}\right) < 0 \Rightarrow a^{(k)} = x^{(k-1)}$ e $b^{(k)} = b^{(k-1)}$*
   (c) *Se $f\left(x^{(k-1)}\right) > 0 \Rightarrow a^{(k)} = a^{(k-1)}$ e $b^{(k-1)} = x^{(k-1)}$*

3. *Il nuovo valore:*

$$x^{(k)} = \frac{a^{(k-1)} + b^{(k-1)}}{2}$$

4. *Si calcola l'errore:*

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

   dove $\left|I^{(k)}\right|$ è la misura dell'intervallo $I^{(k)} = \left[a^{(k)},\, b^{(k)}\right]$

5. *Si determina il criterio di arresto, ci si arresta a k_{min} , primo k per cui:*

$$\left(\frac{1}{2}\right)^{k+1}(b-a) < \varepsilon$$

$$\left(\frac{1}{2}\right)^{k+1} < \frac{\varepsilon}{b-a} \quad \Longleftrightarrow \quad 2^{k+1} > \frac{b-a}{\varepsilon}$$

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

Osservazione 8.1. È un criterio molto semplice ma converge mlto lentamente e non tiene conto di come è fatta la funzione.

8.2 Metodo di Newton

Sia $f\ : I = [a,\ b] \to \mathbb{R}$ tale da essere $f \in C^1\left(I\right)$ ([1]), con $f'\left(\alpha\right) \neq 0$, allora il metodo di Newton permette di trovare quel valore α tale per cui $f\left(\alpha\right) = 0$ attraverso i seguenti passaggi:

1. Si parte da un $x^{(0)}$ arbitrario e si scrive l'equazione della retta tangente al grafico di f nel punto $\left(x^{(0)},\ f\left(x^{(0)}\right)\right)$:

$$y = f\left(x^{(0)}\right) + f'\left(x^{(0)}\right) \cdot \left(x - x^{(0)}\right)$$

   Osservazione 8.2. Corrisponde al polinomio di Taylor della funzione f del primo ordine, che determina appunto la retta (quella tangente) che meglio approssima la funzione f nell'intorno di $x^{(0)}$

2. Si trova il punto di intersezione $x^{(1)}$ della retta y con l'asse delle ascisse, si trova cioè la radice della retta tangente:

$$f\left(x^{(0)}\right) + f'\left(x^{(0)}\right) \cdot \left(x^{(1)} - x^{(0)}\right) = 0$$

$$\Rightarrow x^{(1)} = x^{(0)} - \frac{f\left(x^{(0)}\right)}{f'\left(x^{(0)}\right)}$$

3. In generale alla k-esima iterazione la retta tangente al punto $x^{(k)}$ risulta essere:

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

   cioè quella retta tangente al punto $\left(x^{(k)},\ f\left(x^{(k)}\right)\right)$.

4. Si trova il punto $x^{(k+1)}$ come l'intersezione della retta y precedentemente calcolata e l'asse delle x :

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

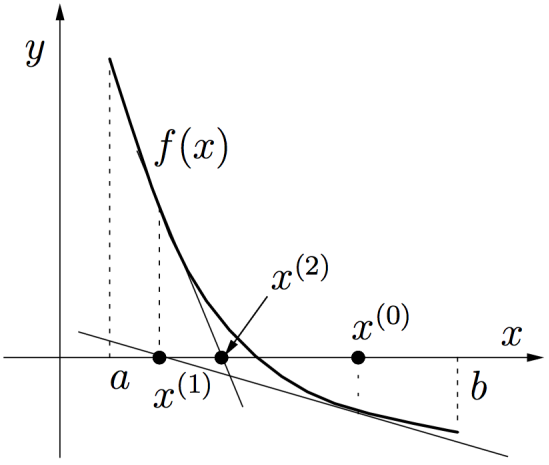


Figura 6: I primi due passi del metodo di Newton per la funzione $f(x)$.

Osservazione 8.3. Come si può notare dalla forma di $x^{(k+1)}$, è necessario (come stabilito nelle ipotesi iniziali) che $f'\left(x^{(k)}\right) \neq 0$, tuttavia nell'implementazione del metodo in Matlab si può trascuare questa eventualità, poichè per gli errori di approssimazione è difficile che si verifichi $f\left(x^{(k)}\right) = 0$

[1] $f \in C^1\left(I\right)$ significa che per la funzione f si può definire la derivata prima f' e che questa è continua sull'intervallo I

Osservazione 8.4. Dallo sviluppo di Taylor nel punto $x^{(k)}$ si ha:

$$f\left(x^{(k+1)}\right) = f\left(x^{(k)}\right) + f'\left(x^{(k)}\right) \cdot \left(x^{(k+1)} - x^{(k)}\right) + o\left(x^{(k+1)} - x^{(k)}\right)$$

Inoltre se $f \in C^2(I)$ è possibile usare il resto secondo Lagrange:

$$\exists\, c^{(k)} \; tra \; x^{(k)} \; e \; x^{(k+1)} \text{ tale che:}$$

$$o\left(x^{(k+1)} - x^{(k)}\right) = \frac{f''\left(c^{(k)}\right)}{2}\left(x^{(k+1)} - x^{(k)}\right)^2$$

Osservazione 8.5. L'incremento $\delta^{(k)}$ tra due iterazioni successive risulta essere:

$$\delta^{(k)} = x^{(k+1)} - x^{(k)} = \frac{f\left(x^{(k+1)}\right) - f\left(x^{(k)}\right) + o\left(x^{(k+1)} - x^{(k)}\right)}{f'\left(x^{(k)}\right)}$$

8.2.1 Problemi di punto fisso

Data una funzione $\Phi(x) : I \to \mathbb{R}$, continua sull'intervallo I, si definisce punto fisso della funzione f quel valore di $\alpha \in I$ tale che:

$$\Phi(\alpha) = \alpha$$

in sostanza è quel punto α di intersezione tra la funzione Φ e la bisettrice del primo e terzo quadrante.

Osservazione 8.6. Il metodo di Newton può essere visto come un problema di individuazione del punto fisso di una funzione $\Phi(x)$ della forma:

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

$$\Phi(x) = x \iff f(\alpha) = 0$$

Supponendo $|\Phi'(\alpha)| < 1$ allora, preso un valore arbitrario di partenza $x^{(0)}$, le iterazioni di punto fisso risultano essere:

$$x^{(k+1)} = \Phi\left(x^{(k)}\right)$$

se esiste il limite $\lim_{k \to +\infty} x^{(k)} = \alpha$, allora α è punto fisso, infatti:

$$x^{(k+1)} \longrightarrow \alpha \quad \text{e} \quad \Phi\left(x^{(k)}\right) \longrightarrow \Phi(\alpha)$$

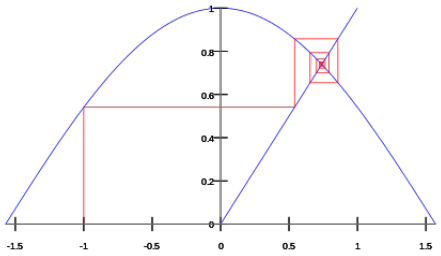


Figura 7: Esempio iterazioni per l'individuazione del punto fisso

Esempio 8.1. Nella figura 7 è rappresentata una porzione della funzione $y = \cos(x)$, per la determinazione del punto fisso:

1. Si sceglie per esempio:
$$x^{(0)} = -1$$
2. Si trova quel punto $x^{(1)}$ per cui $\Phi\left(x^{(0)}\right) = x^{(1)}$:
$$\begin{cases} y = \cos\left(x^{(0)}\right) \\ y = x \end{cases} \Rightarrow x = x^{(1)} = \cos\left(x^{(0)}\right) = 0,54$$

3. A questo punto al passaggio successivo si ottiene:

$$\begin{cases} y = \cos\left(x^{(1)}\right) \\ y = x \end{cases} \quad \Rightarrow x = x^{(2)} = \cos\left(x^{(1)}\right) = 0,86$$

Teorema 15 (Convergenza del metodo del punto fisso)**.** *Se:*

$$\Phi'(\alpha) = \cdots = \Phi^{(p)}(\alpha) = 0 \quad ; \quad \Phi^{(p+1)}(\alpha) \neq 0$$

allora:

$$\lim_{k\to+\infty} \frac{x^{(k+1)} - \alpha}{\left(x^{(k)} - \alpha\right)^{p+1}} = \frac{\Phi^{(p+1)}(\alpha)}{(p+1)!}$$

Dimostrazione. Dallo sviluppo in serie di Taylor calcolato in α si ha:

$$x^{(k+1)} = \Phi\left(x^{(k)}\right) = \underbrace{\Phi(\alpha)}_{=\alpha} + \cdots + \frac{\Phi^{(p+1)}\left(\eta^{(k)}\right)}{(p+1)!}\left(x^{(k)} - \alpha\right)^{p+1} \qquad \left(\text{Con } \eta^{(k)} \text{compreso tra } x^{(k)} \text{e } \alpha\right)$$

$$\Rightarrow x^{(k+1)} - \alpha = \frac{\Phi^{(p+1)}\left(\eta^{(k)}\right)}{(p+1)!}\left(x^{(k)} - \alpha\right)^{p+1}$$

siccome $x^{(k)}$ tende ad α, allora anche $\eta^{(k)}$ (che è compreso tra $x^{(k)}$ e α) deve tendere ad α, quindi:

$$\frac{x^{(k+1)} - \alpha}{\left(x^{(k)} - \alpha\right)^{p+1}} = \frac{\Phi^{(p+1)}\left(\eta^{(k)}\right)}{(p+1)!} \underset{k\to+\infty}{\longrightarrow} \frac{\Phi^{(p+1)}(\alpha)}{(p+1)!}$$

□

Definizione 14. Nel caso di equazione non lineare, α ha molteplicità m se $f(\alpha) = f'(\alpha) = \cdots = f^{(m-1)}(\alpha) = 0$ e $f^{(m)}(\alpha)$ ([2]). Equivalentemente se $f(x)$ può essere scritta come:

$$f(x) = (x-\alpha)^m g(x) \quad \text{t.c.} \quad g(\alpha) \neq 0$$

Osservazione 8.7. Se $f'(\alpha) = 0$ (con α radice di molteplicità maggiore a 1) il metodo di Newton converge semplicemente alla soluzione; è tuttavia possibile recuperare la convergenza quadratica con una modifica, per fare ciò è necessario affrontare il problema di punto fisso.

Tornando al metodo di Newton, nel caso di $f'(\alpha) \neq 0$ (ipotesi di Newton) applicando il teorema per il problema di punto fisso:

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

$$\Phi'(x) = 1 - \frac{\left[f'(x)\right]^2 - f(x) f''(x)}{\left[f'(x)\right]^2}$$

$$\Phi'(\alpha) = 1 - \frac{\left[f'(\alpha)\right]^2}{\left[f'(\alpha)\right]^2} = 0 \qquad (\text{Ricordando che: } f(\alpha) = 0)$$

$$\Phi''(x) = \frac{\left[f'(x) f''(x) - f(x) f'''(x)\right]\left[f'(x)\right]^2 - 2f'(x)\left[f''(x)\right]^2 f(x)}{\left[f'(x)\right]^4}$$

$$\Phi''(\alpha) = \frac{\left[f'(\alpha)\right]^3 f''(\alpha)}{\left[f'(\alpha)\right]^4} = \frac{f''(\alpha)}{f'(\alpha)} \neq 0 \qquad \text{Se: } f''(\alpha) \neq 0$$

[2]Dove $f^{(n)}(\alpha)$ rappresenta la derivata di ordine n della funzione f valutata in α.

vale il teorema della convergenza del metodo del punto fisso (Th: 15) con $p = 1$ e si ha quindi:

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

cioè convergenza quadratica.

Supponendo ora che α sia radice con molteplicità $m > 1$ per $f(x)$:

$$f(\alpha) = f'(\alpha) = \cdots = f^{(m-1)}(\alpha) = 0 \quad e \quad f^{(m)}(\alpha) \neq 0 \Leftrightarrow f(x) = (x-\alpha)^m g(\alpha) \quad (\text{Con } g(\alpha) \neq 0)$$

(per l'implicazione $\Leftrightarrow$ vedi definizione 14)

$\Phi(x)$ può essere scrita come:

$$\Phi(x) = x - \frac{f(x)}{f'(x)} = x - \frac{(x-\alpha)^m g(x)}{m(x-\alpha)^{m-1} g(x) + (x-\alpha)^m g'(x)} =$$

$$= x - \frac{(x-\alpha)^m g(x)}{(x-\alpha)^m \left[\frac{m}{(x-\alpha)} g(x) + g'(x)\right]} = x - \frac{(x-\alpha) g(x)}{m \cdot g(x) + (x-\alpha) g'(x)}$$

$$\Phi'(x) = 1 - \frac{\left[(x-\alpha) g'(x) + g(x)\right] m \cdot g(x) - \left[m \cdot g'(x) + (x-\alpha) g'(x) + (x-\alpha) g(x)\right]}{\left[m \cdot g(x) + (x-\alpha) g'(x)\right]^2}$$

$$\Phi'(\alpha) = 1 - \frac{m\left[g(\alpha)\right]^2}{m^2\left[g(\alpha)\right]^2} = 1 - \frac{1}{m} \neq 0 \Leftrightarrow m > 1$$

Il teorema vale solo con $p = 0$, cioè si ha convergenza semplice.

Osservazione 8.8. Se davanti al termine $\frac{1}{m}$ si avesse come coefficiente m si avrebbe che $\Phi'(\alpha) = 0$

8.2.2 Formula di Newton modificata

Se α ha moltiplicità $m > 1$, si considera:

$$x^{(k+1)} = x^{(k)} - m\frac{f\left(x^{(k)}\right)}{f'\left(x^{(k)}\right)}$$

questo coefficiente m permette di avere:

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

In questo modo si ha che $\Phi'_m(\alpha) = 0$, vale quindi il teorema della convergenza del metodo del punto fisso (Th: 15) con $p = 1$ e si ha quindi convergenza quadratica.

Osservazione 8.9. Se la funzione e la sua derivata intersecano l'asse delle ascisse nello stesso punto, cioè se $f(\alpha) = f'(\alpha) = 0$, allora α è una radice non semplice.

8.3 Criteri di arresto

Definizione 15. Possibili criteri di arresto (si sta definendo $f(x) = 0$:

- Incremento normalizzato, ci si arresta per $k = k_{min}$, al primo k t.c. $\frac{|\overline{x}^{(k_{min})} - x^{(k_{min}-1)}|}{|x^{(k_{min})}|} < \varepsilon$
- Sul residuo $r^{(k)} = f(x_k)$

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

  - Buon criterio se $f'(x) \simeq 1$ non si discosti troppo da uno, altrimenti DISEGNI
    - Disegno 1 derivata $\gg$1 quindi grafico molto pendente, sovrstima l'errore perchè $f(x_k)$ è molto maggiore rispetto alla distanza tra $x^{(k)}$ eα; $\Rightarrow$ sovrastima dell'errore
    - Disegno 2 derivata $\ll$1, uindi grafico molto piatto, la stima dell'errore di $f(x_k)$ è molto minore rispetto alla distanza tra $x^{(k)}$ e $\alpha \Rightarrow$ sottostima dell'errore.

$$|f\left(x^{(k)}\right)| \ll |\alpha - x^{(k)}|$$

8.4 Soluzione di sistemi di equazioni non lineari

Si hanno n equazioni in n incognite, quindi:

$$\begin{cases} f_1(x_1, \ldots, x_n) = 0 \\ f_2(x_1, \ldots, x_n) = 0 \\ \vdots \\ f_n(x_1, \ldots, x_n) = 0 \end{cases} \tag{6}$$

$$\mathrm{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$$

$$\mathbf{F} : A \subseteq \mathbb{R}^n \to \mathbb{R}^n$$

$\mathbf{F}$ è la funzione che ad un certo vettore $\overline{x}$ associa a su volta un vettore di $\mathbb{R}^n$ scritto come $f_1, \ldots, f_n$:

$$\mathbf{F} = (f_1, \ldots f_n)$$

Il sistema (6) può essere scritto come un'equazione vettoriale non lineare nel seguente modo:

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

Osservazione 8.10. Nel caso di una equazione, anzichè risolvere $f(x) = 0$, si sceglie un x_0 arbitrario e si calcola la tangente ad f nel punto x_0 :

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

ponendo $g(x) = 0$, ci si propone ora di eseguire gli stessi passi per una funzione $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n$. Quindi l'idea è di sostituire la funzione $\mathbf{F}$ con una sua approssimazione lineare $\mathbf{G}$ nel punto $\mathbf{x}_0$, che deve quindi avere lo stesso punto di partenza.

In questo caso ci sono n variabili e n funzioni, di conseguenza le possibili derivate parziali sono $n \times n$, e costituiscono quindi una matrice, detta matrice Jacobiana $J_F(\mathbf{x}_0)$:

$$\mathbf{G}(\mathbf{x}) = \mathbf{F}(\mathbf{x}_0) + J_F(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$$

Osservazione 8.11. La matrice Jacobiana di $\mathbf{F}$ è la matrice che contiene tutte le derivate prime tale che l'elemeno ij :

$$[J_F(\mathbf{x}_0)]_{ij} = \frac{\partial f_i}{\partial x_j}(\mathbf{x}_0) = \begin{bmatrix} \frac{\partial f_1(\mathbf{x}_0)}{\partial x_1} & \frac{\partial f_1(\mathbf{x}_0)}{\partial x_2} & \cdots & \frac{\partial f_1(\mathbf{x}_0)}{\partial x_n} \\ \frac{\partial f_2(\mathbf{x}_0)}{\partial x_1} & \frac{\partial f_2(\mathbf{x}_0)}{\partial x_2} & \cdots & \frac{\partial f_2(\mathbf{x}_0)}{\partial x_n} \\ \vdots & & & \vdots \\ \frac{\partial f_n(\mathbf{x}_0)}{\partial x_1} & \frac{\partial f_n(\mathbf{x}_0)}{\partial x_2} & \cdots & \frac{\partial f_n(\mathbf{x}_0)}{\partial x_n} \end{bmatrix}$$

Per trovare $\mathbf{x}_1$ risolvo:

$$\mathbf{G}(\mathbf{x}_1) = \mathbf{0}$$

Cioè $\mathbf{F}(\mathbf{x}_0) + J_F(\mathbf{x}_0)(\mathbf{x}_1 - \mathbf{x}_0) = \mathbf{0}$ cioè:

$$J_F(\mathbf{x}_0)(\mathbf{x}_1 - \mathbf{x}_0) = -\mathbf{F}(\mathbf{x}_0)$$

In questo caso si parla di inversa
Ipotesi: $J_F(\mathbf{x}_0)$ è invertibile ($detJ_F(\mathbf{x}_0) \neq 0$), quindi la formula diventa:

$$\mathbf{x}_1 = \mathbf{x}_0 - J_F^{-1}(\mathbf{x}_0)\mathbf{F}(\mathbf{x}_0)$$

In generale però $\mathbf{x}_{k+1} = \mathbf{x}_k - J_F^{-1}(\mathbf{x}_k)\mathbf{F}(\mathbf{x}_k)$, il calcolo della matrice inversa non è conveniente.

Osservazione 8.12. Per l'implementazione numerica si considera l'incremento:

$$\delta \mathbf{x}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$$

e nella formula $\mathbf{x}_{k+1} = \mathbf{x}_k - J_f^{-1}(\mathbf{x}_k)\,\mathbf{F}(\mathbf{x}_k)$ (metodo di Newton per i sistemi lineari), $\delta \mathbf{x}_k$ è la soluzione del sistema lineare:

$$J_F(\mathbf{x}_k)\,\delta \mathbf{x}_k = -\mathbf{F}(\mathbf{x}_k)$$

Quindi, ricapitolando:

- Si sceglie un $\mathbf{x}_0$ aribitrariamente
- Definito l'incremento $\delta \mathbf{x}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$, è possibile risolvere il sistema lineare:

$$J_F(\mathbf{x}_k)\,\delta \mathbf{x}_k = -\mathbf{F}(\mathbf{x}_k)$$

- Infine si calcola $\mathbf{x}_{k+1}$ come:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \delta \mathbf{x}_k$$

Esempio 8.2. Due equazioni in due incognite

$$\begin{cases} x_1^2 + x_2^2 - 3 = 0 & f_1 \\ 2x_1x_2 + 3x_1^2 = 0 & f_2 \end{cases}$$

Calcolo la Jacobiana:

$$J_F(\overline{x}) = \begin{bmatrix} 2x_1 & 2x_2 \\ 2x_2 + 6x_1 & 2x_1 \end{bmatrix}$$

Sulla prima riga si hanno le derivate parziali di f_1 rispetto a x_1 (posto 11) e a x_2 (posto 12) e lo stesso vale per f_2 nella seconda riga

Se ad esempio $\overline{x}_0 = \begin{bmatrix} 1 & 1 \end{bmatrix}$, la matrice Jacobiana sarebbe:

$$J_F(\overline{x}_0) = \begin{bmatrix} 2 & 2 \\ 8 & 2 \end{bmatrix}$$

Politecnico di Milano

Corso di Metodi Analitici e Numerici per l'Ingegneria

Note sulle equazioni alle derivate parziali: *diffusione*

Cristina Cerutti

Equazione di diffusione

$$\boxed{\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2}}$$

1. Problema di Cauchy-Dirichlet: impostazione del problema

Si vuole trovare la soluzione $u(x,t)$ del seguente problema[1]:

$$u_t = k u_{xx} \quad , \quad 0 \le x \le \pi \ , \ t \ge 0 \tag{1}$$

$$u(0,t) = a \quad , \quad u(\pi,t) = b \tag{2}$$

$$u(x,0) = g(x) \ , \tag{3}$$

dove a e b sono due costanti assegnate e $g(x)$ è una funzione assegnata sull'intervallo $[0,\pi]$. Si noti preliminarmente che affinché il *dato iniziale* (3) sia compatibile con le *condizioni al contorno* (2) deve aversi $f(0) = a$ ed $f(\pi) = b$; se cosí non fosse non potremo sicuramente richiedere che la soluzione sia continua[2].

Si definisca una nuova funzione $v(x,t)$ tale che

$$v(x,t) \equiv u(x,t) - r(x) \quad , \quad r(x) = \frac{b-a}{\pi}\, x + a \ ; \tag{4}$$

risulta immediato verificare che se $u(x,t)$ è soluzione dell'equazione differenziale (1) soddisfacente le condizioni (2) e (3) allora $v(x,t)$ deve essere soluzione del seguente problema

$$v_t = v_{xx} \quad , \quad 0 \le x \le \pi \ , \ t \ge 0 \tag{5}$$

$$v(0,t) = v(\pi,t) = 0 \tag{6}$$

$$v(x,0) = h(x) \quad , \quad h(x) \equiv g(x) - r(x) \ . \tag{7}$$

---

[1] Si ricorre nel seguito alla notazione usuale per cui $u_t = \partial u/\partial t$, $u_x = k\partial u/\partial x$ e cosí via.

[2] In realtà il metodo di risoluzione che svilupperemo ci permetterà di risolvere anche il problema con dati al bordo che non si saldano con continuità (che ha comunque importanti applicazioni fisiche) solo rilassando leggermente la nozione di soluzione (v. S.V.Z.Z.). Questo fatto notevole si esprime dicendo che l'equazione di diffusione ha effetto regolarizzante.

Si osservi che la funzione lineare $r(x)$ definita in (4) è una soluzione dell'equazione (1) e soddisfa le condizioni al contorno (2). Nel prossimo paragrafo vedremo il significato fisico di questa funzione, che rappresenta *la soluzione a regime* o *stazionaria*. In alcuni contesti è detta *rilevamento al bordo*.

Fin qui siamo solo passati dal problema (1)-(2)-(3) per u al problema (5)-(6)-(7) per $v = u - r$. Vedremo subito per quale motivo ciò risulti conveniente. Per il momento si osservi che v deve soddisfare l'equazione di diffusione[3] con condizioni al contorno nulle ($v = 0$ agli estremi dell'intervallo $[0, \pi]$) e dato iniziale $h(x) = g(x) - r(x)$ (non $g(x)$!).

2. Separazione di variabili

Cerchiamo ora una soluzione dell'equazione (5), *non* identicamente nulla, che soddisfi le condizioni al contorno (6) e della forma

$$v(x,t) = X(x)\, T(t) \quad . \tag{8}$$

Si sono introdotte due funzioni incognite X e T, la prima dipendente dalla sola variabile x, la seconda dipendente dalla sola variabile t. Inserendo la (8) nell'equazione (5) ed eseguendo le derivate si ottiene[4] $X(x)T'(t) = kX''(x)T(t)$, ovvero

$$\frac{T'(t)}{k\, T(t)} = \frac{X''(x)}{X(x)} \quad . \tag{9}$$

Dato che x e t sono variabili indipendenti, la precedente equazione puó essere soddisfatta se e solo se[5]

$$T'(t)/T(t) = k\lambda \tag{10}$$

$$X''(x)/X(x) = \lambda \quad , \tag{11}$$

dove λ è una costante (indipendente da x e t) da determinarsi. Il valore (gli infiniti valori, come vedremo) della costante λ si determina risolvendo le equazioni differenziali ordinarie (10) e (11), facendo uso della condizione al contorno (6): $v(0,t) = v(\pi,t) = 0$ per ogni $t \geq 0$.

---

[3]Tale equazione è anche nota come "'equazione del calore"' o "'equazione di Fourier"'.

[4]Gli apici indicano derivazione rispetto all'argomento della funzione su cui compaiono: $X'(x) = dX/dx$, $T'(t) = dT/dt$, ecc...

[5]Derivando la (9) rispetto a t si ottiene $d/dt(T'/T) = d/dt(X''/X) = 0$, dato che X dipende da x e non da t; quindi $T'/T = costante \equiv \lambda$. Si ottiene lo stesso risultato derivando rispetto a x.

Per una soluzione v della forma (8) tale condizione diventa $X(0)T(t) = X(\pi)T(t) = 0$ per ogni $t \geq 0$, ovvero

$$X(0) = X(\pi) = 0 \quad . \tag{12}$$

La soluzione dell'equazione (10), ovvero $T'(t) - k\lambda T(t) = 0$, è data da[6]

$$T(t) = \gamma e^{k\lambda t} \quad , \tag{13}$$

essendo γ una costante arbitraria. Si osservi che dalla forma della soluzione (13) si intuisce che deve aversi necessariamente $\lambda \leq 0$ se si vuole una soluzione ben definita (limitata) per "'tempi"' positivi.

Passiamo ora a discutere la soluzione dell'equazione (11) soggetta alla condizione ai limiti (12). Si osservi che si tratta di risolvere un'equazione differenziale del secondo ordine a coefficienti costanti, $X'' - \lambda X = 0$, con assegnato "'dato iniziale"' e "'finale"', cioè con la X assegnata per due valori di x. A differenza del problema di Cauchy usuale, in cui si assegna il valore di X e della sua derivata X' in uno stesso punto x_0, nel presente problema *ai limiti* ci si aspetta che la soluzione dell'equazione che soddisfi le condizioni $X(0) = X(\pi) = 0$ esista soltanto per determinati valori del parametro λ che vi compare. Mostriamo questo esplicitamente, analizzando i vari casi.

- $\boxed{\lambda > 0}$ In questo caso la soluzione generale dell'equazione (11) $X'' + \lambda X = 0$ è $X(x) = C_1\ e^{\sqrt{\lambda}\ x} + C_2\ e^{-\sqrt{\lambda}\ x}$. La condizione ai limiti implica $C_1 = 0$ e $C_2 = 0$, ovvero $X(x) \equiv 0$. Dunque per $\lambda > 0$ si ha la sola soluzione banale $v \equiv 0$.
- $\boxed{\lambda = 0}$ In questo caso l'equazione (11) diventa $X''(x) = 0$, con soluzione $X(x) = C_1 x + b$, essendo C_1 e C_2 due costanti da determinare. Le condizioni ai limiti $X(0) = X(\pi) = 0$ implicano $C_2 = 0$ e $C_1\pi + C_2 = 0$, ovvero $C_1 = C_2 = 0$. Ne segue $X(x) \equiv 0$. Anche in questo caso si ha la sola soluzione banale $v \equiv 0$.
- $\boxed{\lambda < 0 = -\omega^2}$ In questo caso la soluzione generale della (11) è $X(x) = C_1 \cos(\omega\ x) + C_2 \sin(\omega\ x)$. Le condizioni ai limiti implicano $C_1 = 0$ e $C_2\ \sin(\omega\ \pi) = 0$; la seconda è identicamente soddisfatta se $\omega = n \in \mathbb{N}$, ovvero per valori interi positivi di ω, dato che $\sin(n\pi) = 0\ \forall n$.

Dunque le sole soluzioni non banali (non identicamente nulle) per la X sono date da

$$X_n(x) = c_n \sin(nx) \quad , \quad n = 1, 2, 3 \ldots \tag{14}$$

[6]Si scriva l'equazione in forma differenziale: $dT/T = \lambda dt$. Essendo $dT/T = d \ln G$, si ha $d(\ln T - \lambda t) = 0$, ovvero $\ln T - \lambda t = costante$, da cui segue la (13)

Poiché abbiamo trovato che gli unici valori permessi per λ sono quelli corrispondenti a $\lambda = n^2$, $n \in \mathbb{N}$, le $T(t)$ della forma (13) permesse sono date da

$$T_n(t) = \gamma_n \, e^{-k\,n^2 t} \quad , \quad n = 1, 2, 3 \ldots \tag{15}$$

Ne segue che esistono infinite soluzioni dell'equazione (5), ovvero tutte quelle date dai prodotti $X_n T_n$; precisamente:

$$\boxed{v_n(x,t) = b_n \, e^{-k\,n^2 t} \sin(nx) \quad , \quad n \in \mathbb{N}} \, , \tag{16}$$

dove abbiamo posto $b_n = c_n \gamma_n$.

3. Soluzione del problema di Cauchy

La soluzione piú generale dell'equazione di diffusione (5) si ottiene sommando le soluzioni (16), dato che l'equazione di diffusione è lineare e omogenea e vale quindi il principio di sovrapposizione. Consideriamo, per cominciare, la somma di un numero N finito di tali soluzioni:

$$v^{(N)}(x,t) \equiv \sum_{n=1}^{N} b_n \, e^{-n^2 t} \sin(nx) \, . \tag{17}$$

Osserviamo che la $v^{(N)}(x,t)$ è soluzione dell'equazione di diffusione, ovvero $v_t^{(N)} = v_{xx}^{(N)}$, e soddisfa le condizioni al contorno (6), $v^{(N)}(0,t) = v^{(N)}(\pi,t) = 0$. Questo vale indipendentemente dalla scelta dei coefficienti c_n.

Fino ad ora non abbiamo fatto uso della condizione (7), ovvero non abbiamo tenuto conto del dato iniziale cui deve soddisfare la soluzione del problema. Come vedremo, sará proprio il dato iniziale a determinare completamente i coefficienti b_n che compaiono nella somma (17).

Cominciamo con l'osservare che

$$v^{(N)}(x,0) \equiv \sum_{n=1}^{N} b_n \sin(nx) \, . \tag{18}$$

Supponiamo per il momento di aver fissato un certo valore di N, e che la (17) sia effettivamente la soluzione cercata del problema (5)-(6)-(7). La condizione iniziale (7), in particolare, richiede $v^{(N)}(x,0) = h(x)$, ovvero

$$\sum_{n=1}^{N} b_n \sin(nx) = h(x) \quad (?) \, . \tag{19}$$

Il punto interrogativo a lato dell'uguaglianza indica che, almeno per il momento, ci stiamo chiedendo se esiste un determinato insieme di coefficienti $b_1, \ldots, b_N$ per cui l'uguaglianza sia verificata. Come vedremo, per una $h(x)$ generica e per N finito, l'uguaglianza (19) è verificata solo approssimativamente, con grado di approssimazione tanto migliore quanto piú grande è N, cioè quanto piú alto si sceglie il numero di funzioni seno (o *armoniche*) nella somma.

Per determinare i coefficienti b_n definiamo la quantitá seguente:

$$D_N^2(b_1, \ldots, b_N) \equiv \int_0^\pi \left(\sum_{n=1}^N b_n \sin(nx) - h(x) \right)^2 \, dx \ , \tag{20}$$

lo scarto quadratico medio di $v^{(N)}(x, 0)$ da $h(x)$. Si vede quindi che se D_N^2, vista come funzione dei b_n, è piccola, la nostra soluzione di prova $v^{(N)}$ al tempo $t = 0$ e il dato iniziale h risultano vicini (nel senso della norma L_2)[7]. Cerchiamo allora quei particolari valori dei b_n in corrispondenza dei quali la funzione D_N^2 risulta avere valore minimo.

Il problema che si pone è il seguente: determinare $b_1, \ldots, b_N$ tali che

(1) il gradiente di D_N^2 sia nullo, cioè

$$\frac{\partial D_N^2}{\partial b_s} = 0 \quad , \quad s = 1, \ldots, N \ ;$$

(2) la matrice hessiana (matrice delle derivate seconde) associata a D_N^2 sia definita positiva, cioè l'equazione caratteristica

$$\det \left(\mathbb{I} - \lambda \left[\frac{\partial^2 D_N^2}{\partial b_s \partial b_p} \right] \right) = 0$$

abbia tutte le N radici $\lambda_1, \ldots, \lambda_N$ positive[8].

Derivando la (20) rispetto a c_s e uguagliando a zero si ottiene

$$\frac{\partial D_N^2}{\partial b_s} = 2 \int_0^\pi \left(\sum_{n=1}^N b_n \sin(nx) - h(x) \right) \sin(sx) \, dx = 0$$

ovvero

$$\sum_{n=1}^N b_n \int_0^\pi \sin(nx) \sin(sx) \, dx = \int_0^\pi h(x) \sin(sx) \, dx \ ,$$

[7]La (20) si puó riscrivere in modo compatto $D_N^2 = \left\| v^{(N)}\big|_{t=0} - h \right\|^2$; dunque D_N è proprio la distanza di $v^{(N)}$, calcolata al tempo $t = 0$, da h, il dato iniziale, in $L_2[0, \pi]$.

[8]Il simbolo $\mathbb{I}$ indica la matrice identitá $N \times N$; ricordiamo che le radici dell'equazione caratteristica di una matrice non sono altro che i suoi autovalori.

che, data l'identitá notevole

$$\frac{2}{\pi}\int_0^\pi \sin(nx)\sin(sx) = \delta_{ns} \equiv \begin{cases} 1\ , & s = n \\ 0\ , & s \neq n \end{cases}\ , \tag{21}$$

diviene

$$\sum_{n=1}^{N} b_n \frac{\pi}{2}\ \delta_{ns} = \int_0^\pi h(x)\sin(sx)\ dx\ .$$

Sfruttando la proprietá della delta di Kronecker δ_{ns} definita in (21) si ottiene infine l'espressione del generico coefficiente b_s che soddisfa la proprietá (1) sopra:

$$\boxed{b_s = \frac{2}{\pi}\int_0^\pi h(x)\sin(sx)\ dx\ \ ,\ \ s = 1,\dots,N}\ . \tag{22}$$

Per quanto riguarda la condizione (2), osserviamo che

$$\frac{\partial^2 D_N^2}{\partial b_s \partial b_p} = 2\int_0^\pi \sin(px)\sin(sx)\ dx = \pi\delta_{sp}\ ,$$

cioè la matrice hessiana associata a D_N^2 è diagonale, con elementi diagonali tutti uguali e pari a π. Le radici dell'equazione caratteristica associata a D_N^2 (gli autovalori di D_N^2) valgono tutte π e sono quindi strettamente positive. Ne segue che la funzione $D_N^2(b_1,\dots,b_N)$ ha un minimo assoluto in corrispondenza dei valori dei coefficienti dati dalla formula (22). Tali coefficienti, determinati dal dato iniziale $h(x)$, prendono il nome di *coefficienti di Fourier*, e la somma (17) corrispondente (cioè quella che si ottiene sostituendo ai c_n i valori calcolati secondo la (22)) si chiama *somma di Fourier* della soluzione. Quest'ultima rappresenta un'approssimazione della soluzione vera del problema dell'equazione di diffusione. Si potrebbe far vedere, ma non lo faremo qui, che per $N \to \infty$, il valore minimo di D_N^2, quello corrispondente al valore (22) dei b_n, tende a zero. In altre parole, prendendo in considerazione un numero N molto grande di armoniche nella somma (18) e scegliendo i coefficienti b_n secondo la (22) si approssima molto bene il dato iniziale $h(x)$, tanto meglio quanto piú grande è N; nel limite $N \to \infty$ si ricostruisce esattamente il dato iniziale, precisamente:

$$\lim_{N\to\infty} D_N^2 = 0 \ \Leftrightarrow\ \lim_{N\to\infty} v^{(N)}(x,0) \overset{L_2}{=} h(x)\ \ .$$

Corrispondentemente, e anche questo si potrebbe dimostrare,

$$\lim_{N\to\infty} v^{(N)}(x,t) \overset{L_2}{=} v(x,t)\ ,$$

cioè la soluzione cercata del problema (5)-(6)-(7) ammette la rappresentazione in *serie di Fourier*

$$\boxed{v(x,t) \stackrel{L_2}{=} \sum_{n=1}^{\infty} b_n e^{-n^2 t} \sin(nx)} \tag{23}$$

con i c_n dati dalla (22). La scritta "'L_2"' posta sopra i segni di uguaglianza sopra e in particolare nella (23) significa che la differenza tra lato sinistro (la soluzione del problema) e lato destro (la serie di Fourier) dell'uguaglianza è nulla in norma L_2.

4. Comportamento asintotico ($t \to \infty$) della soluzione

Si moltiplichino entrambi i membri dell'equazione (5) per $2v$ e si integri a sinistra e a destra sull'intervallo $[0, \pi]$; si ottiene

$$\int_0^\pi 2vv_t \, dx = \int_0^\pi 2vv_{xx} \, dx \quad ,$$

ovvero, portando fuori dal segno di integrale la derivata rispetto al tempo a sinistra e integrando per parti a destra

$$\frac{d}{dt}\int_0^\pi v^2 \, dx = 2(vv_x)\Big|_0^\pi - 2\int_0^\pi (v_x)^2 \, dx \quad .$$

Facendo ora uso delle condizioni al contorno (6) si vede subito che il primo termine a destra nell'ultima equazione scritta si annulla ($v = 0$ per $x = 0$ e $x = \pi$). Si ottiene quindi

$$\frac{d}{dt}\int_0^\pi v^2 \, dx = -2\int_0^\pi (v_x)^2 \, dx$$

che può essere riscritta in forma compatta introducendo la notazione standard[9] per la norma L_2 in $[0, \pi]$:

$$\frac{d}{dt}\|v\|^2 = -2\|v_x\|^2 \, . \tag{24}$$

Ora, si supponga che al tempo t sia $\|v_x\|^2 > 0$; la (24) implica allora $d\|v\|^2/dt < 0$ e cioè $\|v\|$ decresce nel tempo. Questo processo si arresta solo se da un certo istante in poi $\|v_x\|^2 = 0$, ovvero $v = cost.$; dato che $v = 0$ in 0 e in π, l'unica possibiltà in questo caso è $v = cost. \equiv 0$[10]. In definitiva la (24) implica

$$\lim_{t\to\infty} v(x,t) = 0 \quad , \tag{25}$$

[9]Precisamente $\|f\|^2 \equiv \int_0^\pi f^2 \, dx$ per ogni f a quadrato integrabile in $[0, \pi]$.

[10]Attenzione: si ricorda che $\|f\| = 0$ implica $f = 0$ *quasi ovunque*, ovvero tranne al piú su un insieme di misura nulla nel senso di Lebesgue.

almeno quasi ovunque. Vedremo nella prossima sezione che in realtà, per dati iniziali sufficientemente "buoni", v tende a zero uniformemente, con velocità esponenziale ($|v| \leq cost.\ e^{-t}$).

L'equazione di diffusione è dissipativa. Infatti la relazione (24), ovvero $d\|v\|^2/dt \leq 0$, puó essere interpretata fisicamente pensando a $\|v\|^2$ come all'energia associata al sistema: l'energia decresce nel tempo, ovvero viene dissipata. Per analogia, si pensi all'equazione differenziale ordinaria $\dot{x} = -x$. Se $x(t)$ è soluzione di tale equazione, allora vale la relazione $d(x^2)/dt = -2x^2 \leq 0$.

Ovviamente, dato che $u = v + r$, dalla (25) segue anche

$$\lim_{t\to\infty} u(x,t) = r(x) = \frac{b-a}{\pi}\ x + a \quad , \tag{26}$$

e quindi, basandosi su una semplice analisi preliminare, si è ottenuta una informazione di importanza fondamentale: la soluzione asintotica del problema (1)-(2)-(3) è data dalla $r(x)$ introdotta in (4) che, si osservi, dipende dalle condizioni al contorno ma non dal dato iniziale.

5. Alcune osservazioni

Vediamo alcune proprietá generali dei coefficienti della serie di Fourier (22). Per prima cosa, tenendo presente che $|\sin(nx)| \leq 1$, si ha

$$\begin{aligned} |b_n| = \quad & \left|\frac{2}{\pi}\int_0^\pi h(x)\sin(nx)\ dx\right| \leq \\ & \leq \frac{2}{\pi}\int_0^\pi |h(x)|\ |\sin(nx)|\ dx \leq \frac{2}{\pi}\int_0^\pi |h(x)|\ dx\ , \end{aligned}$$

ovvero, se il dato iniziale $h(x)$ è modulo-integrabile (appartiene ad $L_1[0,\pi]$) i coefficienti di Fourier sono limitati superiormente. Possiamo scrivere in forma compatta[11]

$$|b_n| \leq \frac{2}{\pi}\|h\|_1\ .$$

Dalla (23) segue allora

$$\begin{aligned} |v(x,t)| \leq \quad & \sum_{n=1}^\infty |b_n| e^{-n^2 t} \leq \frac{2}{\pi}\|h\|_1 \sum_{n=1}^\infty e^{-n^2 t} \leq \\ & \leq \frac{2}{\pi}\|h\|_1 \sum_{n=1}^\infty e^{-nt} = \frac{2}{\pi}\|h\|_1\ \frac{1}{e^t - 1} \quad . \end{aligned}$$

[11]Ricordiamo la definizione di norma L_p: $\|f\|_p \equiv \left(\int |f|^p\ dx\right)^{1/p}$, dove p è un numero positivo.

Questo mostra che la soluzione $v(x,t)$, o meglio, la sua rappresntazione in serie di Fourier, tende a zero per $t \to \infty$, uniformemente in $x \in [0,\pi]$.

Un'altra considerazione. Supponiamo che il dato iniziale $h(x)$ sia derivabile in $(0,\pi)$ (cioè esista continua $h'(x)$). Con una integrazione per parti nella (22) e tenendo conto del fatto che $h(0) = h(\pi) = 0$, otteniamo

$$b_n = \frac{2}{n\pi}\left[\int_0^\pi h'(x)\cos(nx)\ dx\right]$$

da cui segue che $\lim_{n\to\infty} c_n = 0$.

Nel caso in cui la $h'(x)$ ammetta ad esempio un punto di discontinuitá a salto finito in $x_0 \in (0,\pi)$, si "'spacca"' l'integrale in due e *poi* si integra per parti:

$$\begin{aligned} b_n = \quad & \frac{2}{\pi}\left[\int_0^{x_0} h(x)\sin(nx)\ dx + \int_{x_0}^\pi h(x)\sin(nx)\ dx\right] = \\ & = \frac{2}{n\pi}\left[-h'(x)\cos(nx)|_0^{x_0} + \int_0^{x_0} h'(x)\cos(nx)\ dx + \right. \\ & \left. -h'(x)\cos(nx)|_{x_0}^{\pi} + \int_{x_0}^\pi h'(x)\cos(nx)\ dx\right] . \end{aligned} \tag{27}$$

Anche in questo caso si vede che $\lim_{n\to\infty} b_n = 0$.

6. Suggerimenti per lo svolgimento degli esercizi

Lo schema per lo svolgimento degli esercizi relativi all'equazione di diffusione è il seguente. Si chiede di risolvere il problema seguente per la $u(x,t)$:

$$\begin{aligned} & u_t = u_{xx} \quad , \quad 0 \le x \le \pi \ , \ t \ge 0 \\ & u(0,t) = a \quad , \quad u(\pi,t) = b \\ & u(x,0) = g(x) \ . \end{aligned}$$

Si introducono la funzione lineare (*rilevamento al bordo*)

$$r(x) \equiv \frac{b-a}{\pi}\ x + a$$

e la funzione

$$v(x,t) \equiv u(x,t) - r(x) \ ,$$

soluzione del problema con condizioni al bordo nulle e dato iniziale modificato $h(x)$:

$$\begin{aligned} & v_t = v_{xx} \quad , \quad 0 \le x \le \pi \ , \ t \ge 0 \\ & v(0,t) = v(\pi,t) = 0 \\ & v(x,0) = h(x) \equiv g(x) - r(x) \quad (!!!) \ . \end{aligned}$$

La soluzione di questo nuovo problema è data in serie di Fourier

$$v(x,t) = \sum_{n=1}^{\infty} b_n e^{-n^2 t} \sin(nx)$$

dove i coefficienti b_n che vi compaiono, *devono* essere calcolati esplicitamente tramite la formula

$$\boxed{b_n = \frac{2}{\pi} \int_0^{\pi} \left(g(x) - \frac{b-a}{\pi}\ x - a \right) \sin(nx)\ dx} \quad .$$

La funzione in parentesi è proprio $h = f - r$, ovvero il dato iniziale per la v. I numeri a e b e la funzione $f(x)$ sono forniti dal problema per la u. La determinazione dei c_n è in sostanza l'unico calcolo esplicito da fare. Si esegue tipicamente integrando per parti, ed integrando la funzione trigonometrica che compare nell'integrando (al primo passaggio $\sin(nx)$, al secondo $\cos(nx)$ ecc...), in modo tale che compaiano successivamente le derivate del dato iniziale h. Se questo ha un'espressione polinomiale, tale procedura si conclude in un numero finito di passi. Il generico coefficiente di Fourier b_n è una funzione dell'intero n, non puó dipendere da x, e deve decrescere al crescere di n.

La soluzione del problema per la $u(x,t)$ si scrive alla fine:

$$u(x,t) = v(x,t) + r(x) = \sum_{n=1}^{\infty} c_n e^{-n^2 t} \sin(nx) + \frac{b-a}{\pi}\ x + a\ .$$

Valgono sempre i limiti

$$\lim_{t\to\infty} v(x,t) = 0 \quad ; \quad \lim_{t\to\infty} u(x,t) = r(x)\ .$$

7. Esercizi proposti: problema di Cauchy-Dirichlet

Si chiede di trovare la soluzione $u(x,t)$ del problema

$$\begin{aligned} &u_t = u_{xx} \quad , \quad 0 \le x \le \pi\ ,\ t \ge 0 \\ &u(0,t) = a \quad , \quad u(\pi,t) = b \\ &u(x,0) = g(x)\ , \end{aligned}$$

calcolandone esplicitamente la serie di Fourier, nei casi seguenti.

(1) $a = b = 0$; $g(x) = \sin^3(x)$.

(2) $a = b = 0$; $g(x) = x(\pi - x)$.

(3) $a = 0$, $b = \pi^3$; $g(x) = x^3$.

(4) $a = 1$, $b = 0$; $g(x) = \begin{cases} 1\ ,\ 0 \le x \le 1 \\ \frac{\pi - x}{\pi - 1}\ ,\ 1 \le x \le \pi \end{cases}$.

(5) $a = 0$, $b = 0$; $g(x) \begin{cases} x\left(\frac{\pi}{2} - x\right)\ ,\ 0 \le x \le \pi/2 \\ 0\ ,\ \pi/2 \le x \le \pi \end{cases}$.

(6) $a = b = 0$; $g(x) = \frac{\pi}{2} - \left|\frac{\pi}{2} - x\right|$.

Si raccomanda di tracciare il grafico del dato iniziale $g(x)$ e del rilevamento $r(x)$ come prima cosa.

Risposte

Vengono fornite risposte parziali (riguardanti i coefficienti della serie di Fourier, con qualche suggerimento) relative ai problemi precedenti.

(1) $b_1 = 3/4$; $b_2 = 0$; $b_3 = -1/4$; $b_n = 0 \ \ \forall\ n \ge 4$.

♠ Sugg.: $\sin^3(x) = \frac{3}{4}\sin(x) - \frac{1}{4}\sin(3x)$.

(2) $b_n = \begin{cases} \frac{8}{\pi n^3}\ ,\ n\ dispari \\ 0\ ,\ n\ pari \end{cases}$.

(3) $b_n = \frac{12(-1)^n}{n^3}$.

(4) $b_n = \frac{2\sin(n)}{n^2\pi(\pi-1)}$.

(5) $b_{2k+1} = -\frac{(-1)^k}{(2k+1)^2} + \frac{4}{\pi(2k+1)^3}\ \ ,\ \ k \ge 0$;
$b_{2k} = -\frac{(-1)^k - 1}{2\pi k^3}\ \ ,\ \ k \ge 1$.

♠ Sugg.: $\sin(n\pi/2) = 0$ per n pari; $\cos(n\pi/2) = 0$ per n dispari. Inoltre $\sin((2k+1)\pi/2) = (-1)^k$; $\cos((2k)\pi/2) = (-1)^k$. Ricordare che nella serie di Fourier si parte da $n = 1$.

(6) $b_n = 0$ per n pari; $b_{2k+1} = \frac{4(-1)^k}{\pi(2k+1)^2}$, $k \ge 0$.

8. Problema di Cauchy-Neumann

Si vuole trovare la soluzione $u(x,t)$ del seguente problema di Cauchy-Neumann:

$$u_t = k u_{xx}\ \ ,\ \ 0 \le x \le \pi\ ,\ t \ge 0 \tag{28}$$

$$u_x(0,t) = c\ \ ,\ \ u_x(\pi,t) = d \tag{29}$$

$$u(x,0) = g(x)\ , \tag{30}$$

dove c e d sono due costanti assegnate e $g(x)$ è una funzione assegnata sull'intervallo $[0, \pi]$. Per trovare la soluzione, in modo analogo a quanto abbiamo fatto per il problema di Cauchy-Dirichlet, vogliamo prima ridurci al problema omogeneo, cio'e trovare la soluzione $v(x,t)$ del problema

$$v_t = v_{xx} \quad , \quad 0 \leq x \leq \pi \ , \ t \geq 0 \tag{31}$$

$$v_x(0,t) = v_x(\pi,t) = 0 \tag{32}$$

$$v(x,0) = \tilde{g}(x) \quad , \ . \tag{33}$$

dove $\tilde{g}(x) = g(x) - r_N(x)$ è una funzione scelta opportunamente (analoga del rilevamento al bordo per il problema di Dirichlet) in modo tale che v soddisfi (28) e (29). Si osserva che in questo caso non sarà più sufficiente una funzione della sola x (in quanto per poter assegnare dati di Neumann diversi ai due estremi non basta un polimonio di primo grado ma è necessario un polinomio quadratico, che però non soddisfa (28) e quindi ....). Si lascia allo studente la verifica che la più semplice funzione $r_N(x,t)$ che soddisfa queste proprietà è

$$r_N(x) = \frac{d-c}{2\pi}\ x^2 + cx + \frac{d-c}{\pi}\ t \tag{34}$$

e quindi

$$\tilde{g}(x) = g(x) - \frac{d-c}{2\pi}\ x^2 - cx \ . \tag{35}$$

In modo del tutto analogo a quanto fatto per il problema di Cauchy-Dirichlet, si separano le variabili e questa volta si trova (verificare!!)

$$v(x,t) = \alpha_0 + \sum_{n=1}^{\infty} \alpha_n e^{-n^2 t} \cos(nx)$$

dove i coefficienti c_n che vi compaiono, *devono* essere calcolati esplicitamente tramite la formula

$$\boxed{\alpha_0 = \frac{1}{\pi} \int_0^{\pi} \left(g(x) - \frac{d-c}{2\pi}\ x^2 - cx \right)\ dx}$$

e

$$\boxed{\alpha_n = \frac{2}{\pi} \int_0^{\pi} \left(f(x) - \frac{d-c}{2\pi}\ x^2 - cx \right) \cos(nx)\ dx} \quad .$$

La soluzione del problema per la $u(x,t)$ si scrive alla fine:

$$u(x,t) = v(x,t) + r_N(x) = \alpha_0 + \sum_{n=1}^{\infty} \alpha_n e^{-n^2 t} \cos(nx) + \frac{d-c}{2\pi}\, x^2 + cx \quad .$$

9. Esercizi proposti: problema di Cauchy-Neumann

Si chiede di trovare la soluzione $u(x,t)$ del problema

$$\begin{aligned} &u_t = u_{xx} \quad , \quad 0 \le x \le \pi \ , \ t \ge 0 \\ &u_x(0,t) = c \quad , \quad u_x(\pi,t) = d \\ &u(x,0) = g(x) \ , \end{aligned}$$

calcolandone esplicitamente la serie di Fourier, nei casi seguenti.

(1) $c = d = 0$; $g(x) = 4\sin^3(x)$.

(2) $c = 0$, $d = 2\pi$; $g(x) = 2\cos^2(x) - 3 + x^2$.

(3) $c = \pi$, $d = \pi$; $g(x) = 0$.

Si raccomanda di tracciare il grafico del dato iniziale $g(x)$ e del rilevamento $r_N(x)$ come prima cosa.

Equazione di Laplace

Operatore di Laplace in coordinate polari (in $\mathbb{R}^2$)

È utile, quando si ha un dominio a simmetria radiale, scrivere Δu in coordinate polari.

Passaggio in coordinate polari:

$$\begin{cases} x(\rho,\theta) = \rho\cos(\theta) & \rho \in (0,+\infty) \\ y(\rho,\theta) = \rho\sin(\theta) & \theta \in [-\pi,\pi) \end{cases}$$

posto $v(\rho,\theta) = u(\rho\cos(\theta), \rho\sin(\theta))$, dove:

$$\rho = \sqrt{x^2+y^2} \qquad ; \qquad \theta = \begin{cases} \arctan\frac{y}{x} & I \text{ e } IV \text{ quadrante} \\ \arctan\frac{y}{x} + \pi & II \text{ e } IV \text{ quadrante} \end{cases}$$

$$u_x = \frac{\partial u}{\partial x} = \frac{\partial v}{\partial \rho}\frac{\partial \rho}{\partial x} + \frac{\partial v}{\partial \theta}\frac{\partial \theta}{\partial x}$$

$$u_y = \frac{\partial u}{\partial y} = \frac{\partial v}{\partial \rho}\frac{\partial \rho}{\partial y} + \frac{\partial v}{\partial \theta}\frac{\partial \theta}{\partial y}$$

$$\frac{\partial \rho}{\partial x} = \frac{x}{\sqrt{x^2+y^2}} = \cos\theta$$

$$\frac{\partial \rho}{\partial \theta} = \sin\theta$$

$$\frac{\partial \theta}{\partial x} = \frac{-\frac{y}{x^2}}{1+\frac{y^2}{x^2}} = -\frac{y}{x^2+y^2} = -\frac{\sin\theta}{\rho}$$

$$\frac{\partial \theta}{\partial y} = \frac{\frac{1}{x}}{1+\frac{y^2}{x^2}} = \frac{x}{x^2+y^2} = \frac{\cos\theta}{\rho}$$

$$\Delta v = v_{xx} + v_{yy}$$

Calcoliamo ora u_{xx}:

$$u_x = v_\rho\cos\theta - v_\theta\frac{\sin\theta}{\rho}$$

$$u_{xx} = \frac{\partial v_x}{\partial \rho}\frac{\partial \rho}{\partial x} + \frac{\partial u_x}{\partial \theta}\frac{\partial \theta}{\partial x}$$

$$\frac{\partial v_x}{\partial \rho} = v_{\rho\rho}\cos\theta - v_{\rho\theta}\frac{\sin\theta}{\rho} + v_\theta\frac{\sin\theta}{\rho^2}$$

$$\frac{\partial v_x}{\partial \theta} = v_{\rho\theta}\cos\theta - v_\rho\sin\theta - v_{\theta\theta}\frac{\sin\theta}{\rho} - u_\theta\frac{\cos\theta}{\rho}$$

quindi:

$$u_{xx} = v_{\rho\rho}\cos^2\theta - 2v_{\rho\theta}\frac{\sin\theta\cos\theta}{\rho} + v_\rho\frac{\sin^2\theta}{\rho} + v_{\theta\theta}\frac{\sin^2\theta}{\rho^2} + 2v_\theta\frac{\sin\theta\cos\theta}{\rho^2}$$

Calcoliamo ora u_{yy}:

$$u_y = v_\rho\sin\theta + v_\theta\frac{\cos\theta}{\rho}$$

$$u_{yy} = \frac{\partial v_y}{\partial\rho}\frac{\partial\rho}{\partial y} + \frac{\partial v_y}{\partial\theta}\frac{\partial\theta}{\partial y}$$

$$\frac{\partial v_y}{\partial\rho} = v_{\rho\rho}\sin\theta - v_\theta\frac{\cos\theta}{\rho^2} + v_{\rho\theta}\frac{\cos\theta}{\rho}$$

$$\frac{\partial v_y}{\partial\theta} = v_{\rho\theta}\sin\theta + v_\rho\cos\theta + v_{\theta\theta}\frac{\cos\theta}{\rho} - v_\theta\frac{\sin\theta}{\rho}$$

quindi:

$$u_{yy} = v_{\rho\rho}\sin^2\theta + 2v_{\rho\theta}\frac{\sin\theta\cos\theta}{\rho} + v_\rho\frac{\cos^2\theta}{\rho} + v_{\theta\theta}\frac{\cos^2\theta}{\rho^2} - 2v_\theta\frac{\sin\theta\cos\theta}{\rho^2}$$

Quindi risulta essere:

$$u_{xx} + u_{yy} = v_{\rho\rho} + \frac{1}{\rho}v_\rho + \frac{1}{\rho^2}v_{\theta\theta}$$

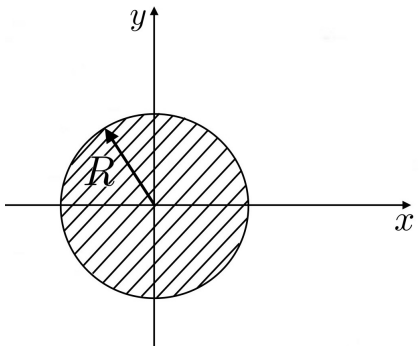
Da ora in avanti scriveremo $u(\rho,\theta)$ per indicare la $v(\rho,\theta)$.

Metodo di separazione delle variabili

Prendiamo per esempio il problema di Dirichlet sul cerchio di raggio R e centrato in $(0,0)$.

Il sistema:

$$\begin{cases}\Delta u(x,y) = 0 & (x,y)\in B_R(0) = \{(x,y)\,|x^2+y^2<R^2\}\\ u(x,y) = f(x,y) & (x,y)\in\partial B_R(0) = \{(x,y)\,|x^2+y^2=R^2\}\end{cases}$$



posto $u(\rho,\theta) = u(\rho\cos(\theta), \rho\sin(\theta))$, si ha:

$$\Delta u(\rho,\theta) = u_{xx}(\rho\cos(\theta),\rho\sin(\theta)) + u_{yy}(\rho\cos(\theta),\rho\sin(\theta)) = u_{\rho\rho}(\rho,\theta) + \frac{1}{\rho}u_{\rho}(\rho,\theta) + \frac{1}{\rho^2}u_{\theta\theta}(\rho,\theta)$$

per tanto l'equazione

$$\begin{cases} u_{xx} + u_{yy} = 0 & B_R(0) \\ u = f & \partial B_R(0) \end{cases}$$

in coordinate polari diventa:

$$\begin{cases} u_{\rho\rho} + \frac{1}{\rho}u_{\rho} + \frac{1}{\theta}u_{\theta\theta} = 0 & (\rho,\theta) \in [0,R) \times [-\pi,\pi) \\ u(R,\theta) = f(\theta) := f[R\cos(\theta), R\sin(\theta)) \end{cases}$$

1° passo

Si cercano tutte le possibili soluzioni $U(\rho,\theta)$ di

$$u_{\rho\rho} + \frac{1}{\rho}u_{\rho} + \frac{1}{\rho^2}u_{\theta\theta} = 0 \qquad (\rho,\theta) \in [0,R) \times [-\pi,\pi)$$

con le seguenti proprietà:

1. $U(\rho,\theta) = h(\rho)\,g(\theta)$, con $g(\theta)$ periodica di periodo 2π
2. $U(\rho,\theta)$ limitata su $B_R(0)$

2° passo

Tra tutte queste possibili soluzioni si sceglie quella che soddisfa

$$\tilde{U}(R,\theta) = \tilde{f}(\theta)$$

Primo passo

Supponiamo che $U(\rho,\theta) = h(\rho)\,g(\theta)$ sia soluzione di

$$u_{\rho\rho} + \frac{1}{\rho}u_{\rho} + \frac{1}{\rho^2}u_{\theta\theta} = 0 \qquad (\rho,\theta) \in [0,R) \times [-\pi,\pi)$$

sostituendo nell'equazione si trova

$$h''(\rho)\,g(\theta) + \frac{1}{\rho}h'(\rho)\,g(\theta) + \frac{1}{\rho^2}h(\rho)\,g''(\theta) = 0$$

$$g(\theta)\left[h''(\rho) + \frac{1}{\rho}h'(\rho)\right] = -g''(\theta)\left[\frac{1}{\rho^2}h(\rho)\right]$$

$$\frac{g''(\theta)}{g(\theta)} = -\frac{h''(\rho) + \frac{1}{\rho}h'(\rho)}{\frac{1}{\rho^2}h(\rho)} \qquad (\rho, \theta) \in [0, R) \times [-\pi, \pi)$$

Affichè una funzione della sola θ e della sola ρ siano uguali devono essere costanti. Quindi si ha:

$$\begin{cases} \frac{g''(\theta)}{g(\theta)} = \lambda \\ \\ -\frac{h''(\rho)+\frac{1}{\rho}h'(\rho)}{\frac{1}{\rho^2}h(\rho)} = \lambda \end{cases} \qquad \begin{cases} g''(\theta) - \lambda g(\theta) = 0 & (1) \\ \\ \rho^2 h''(\rho) + \rho h'(\rho) + \lambda h(\rho) = 0 & (2) \end{cases}$$

Risolviamo (1): l'equazione caratteristica è $\eta^2 - \lambda = 0 \Rightarrow \eta^2 = \lambda$, in base al segno di λ le soluzioni variano:

$$\lambda > 0 \longrightarrow g(\theta) = c_1 e^{\sqrt{\lambda}\theta} + c_2 e^{-\sqrt{\lambda}\theta}$$

$$\lambda = 0 \longrightarrow g(\theta) = c_1 + c_2\theta$$

$$\lambda < 0 \longrightarrow c_1 \sin\left(\sqrt{|\lambda|}\theta\right) + c_2 \cos\left(\sqrt{|\lambda|}\theta\right)$$

noi vogliamo che g sia $2\pi -$ periodica

- Se $\lambda > 0$, g non è mai periodica
- Se $\lambda = 0$, $g(\theta) = c_1$ è l'unica soluzione $2\pi -$ periodica
- Se $\lambda < 0$, $g(\theta) = c_1 \sin(n\theta) + c_2 \cos(n\theta)$, con $n = 1, 2, 3, \ldots$, solo le possibili soluzioni $2\pi -$ periodiche; in particolare deve essere

$$\sqrt{|\lambda|} = n \qquad |\lambda| = n^2 \qquad \lambda = -n^2 \qquad n = 1, 2, 3, \ldots$$

Adesso risolviamo la (2) con $\lambda = 0$ e $\lambda = -n^2 \quad (n = 1, 2, 3 \ldots)$.

- $\lambda = 0$

  L'equazione (2) è $\rho^2 h''(\rho) + \rho h'(\rho) = 0$

$$\frac{h''(\rho)}{h(\rho)} = -\frac{1}{\rho} \qquad \int \frac{h''(\rho)}{h(\rho)} d\rho = -\int \frac{1}{\rho} d\rho$$

$$\ln(h'(\rho)) = -\ln(\rho) + K_1$$

$$e^{\ln(h'(\rho))} = \frac{1}{e^{\ln(\rho)}} e^{K_1}$$

$$h'(\rho) = \frac{1}{\rho} K_2$$

  la soluzione è del tipo

$$h(\rho) = K_2 \ln(\rho) + K_3$$

- $\lambda = -n^2$

  L'equazione (2) è $\rho^2 h''(\rho) + \rho h'(\rho) - n^2 h(\rho) = 0$, è un'equazione di Eulero omogenea. Si cercano soluzioni del tipo $K\rho^\alpha$

$$\frac{\partial}{\partial \rho} K\rho^\alpha = K\alpha\rho^{\alpha-1} \qquad \frac{\partial^2}{\partial \rho^2} K\rho^\alpha = K\alpha(\alpha-1)\rho^{\alpha-2}$$

  sostituendo nell'equazione si trova

$$K\left[\rho^2\alpha(\alpha-1)\rho^{\alpha-2} + \rho\alpha\rho^{\alpha-1} - n^2\rho^\alpha\right] = 0$$

$$K\rho^\alpha\left[\alpha^2 + \cancel{\alpha} - \cancel{\alpha} - n^2\right] = 0$$

$$\text{Da cui: } \alpha^2 - n^2 = 0 \qquad \alpha = \pm n$$

  La soluzione ottenuta è pertanto di tipo

$$h(\rho) = K_1\rho^n + K_2\rho^{-n}$$

In maniera equivalente, per risolvere (2) avremmo potuto operare il seguente cambio di variabili:

$$\rho = e^s \qquad s = \ln(\rho) \qquad \frac{ds}{d\rho} = \frac{1}{\rho}$$

così da avere:

$$\frac{dh}{d\rho} = \frac{dh}{ds}\frac{ds}{d\rho} = \frac{dh}{ds}\frac{1}{\rho}$$

$$\frac{d^2h}{d\rho^2} = \frac{d}{d\rho}\left(\frac{dh}{d\rho}\right) = \frac{d}{d\rho}\left(\frac{dh}{ds}\right)\frac{1}{\rho} - \frac{dh}{ds}\frac{1}{\rho^2} =$$

$$= \frac{d}{ds}\left(\frac{dh}{ds}\right)\frac{ds}{d\rho}\frac{1}{\rho} - \frac{dh}{ds}\frac{1}{\rho^2} = \frac{d^2h}{ds^2}\frac{1}{\rho^2} - \frac{dh}{ds}\frac{1}{\rho^2}$$

L'equazione (2) diventa quindi:

$$\rho^2\frac{d^2h}{d\rho^2} + \rho\frac{dh}{d\rho} + \lambda h = 0$$

ovvero:

$$\boxed{\frac{d^2h}{ds^2} + \lambda h = 0}$$

- Per $\lambda = 0$ rimane $\frac{d^2h}{ds^2} = 0$ ovvero $h(s) = c_1 + c_2 s$

$$\Rightarrow \boxed{h(\rho) = c_1 + c_2\ln(\rho)}$$

- Per $\lambda = -n^2$ si ha $\frac{d^2h}{ds^2} - n^2 h = 0$ le cui soluzioni sono

$$h(s) = c_1 e^{-ns} + c_2 e^{ns}$$
$$\Rightarrow \boxed{h(\rho) = c_1 \rho^{-n} + c_2 \rho^n}$$

In conclusione tutte le funzioni della forma:

$$c_0,\ \overline{c}_0 \ln(\rho),\ c_n \rho^n \cos(n\theta),\ \overline{c}_n \rho^{-n} \cos(n\theta),$$
$$d_n \rho^n \sin(n\theta),\ \overline{d}_n \rho^{-n} \sin(n\theta)$$

e, per la linearità di Δ, anche tutte le loro combinazioni lineari, soddisfano

$$\Delta u(\rho, \theta) = 0 \qquad (\rho, \theta) \in (0, R] \times [-\pi, \pi)$$

Secondo passo

In particolare, noi vogliamo soluzioni di

$$(*) \begin{cases} \Delta u(\rho, \theta) = 0 & (\rho, \theta) \in (0, R] \times [-\pi, \pi) \\ u(R, \theta) = f(\theta) & \theta \in [-\pi, \pi) \end{cases}$$

che siano limitate in $B_R(0)$.

La condizione sulla limitatezza si traduce nel fatto che i termini ρ^{-n}, $\ln(\rho)$ non compaiono, cioè

$$\overline{c}_0 = \overline{c}_n = \overline{d}_n = 0 \qquad \forall n = 1, 2, 3 \ldots$$

Quindi una possibile soluzione di $(*)$ avrà la forma

$$U(\rho, \theta) = c_0 + \sum_{n=1}^{+\infty} c_n \rho^n \cos(n\theta) + \sum_{n=1}^{+\infty} d_n \rho^n \sin(n\theta) \qquad (\bullet)$$

con le costanti $c_0,\ c_n,\ d_n$ da determinare in base alla f.

Se $f \in \mathscr{L}^2([-\pi, \pi))$ possiamo scrivere[1]

$$f(\theta) \cong \frac{a_0}{2} + \sum_{n=1}^{+\infty} a_n \cos(n\theta) + \sum_{n=1}^{+\infty} b_n \sin(n\theta) \qquad (\bullet\bullet)$$

Imponendo $U(R, \theta) = f(\theta)$ e uguagliando le due espressioni $(\bullet)$ e $(\bullet\bullet)$ termine a termine si ha

$$c_0 = \frac{a_0}{2} \qquad c_n R^n = a_n \qquad d_n R^n = b_n$$

$$\text{da cui: } c_0 = \frac{a_0}{2} \qquad c_n = \frac{a_n}{R^n} \qquad d_n = \frac{b_n}{R^n} \qquad n = 1, 2, 3 \ldots$$

La soluzione di $(*)$ sarà dunque:

$$U(\rho, \theta) = \frac{a_0}{2} + \sum_{n=1}^{+\infty} \left(\frac{a_n}{R^n}\right) \rho^n \cos(n\theta) + \sum_{n=1}^{+\infty} \left(\frac{b_n}{R^n}\right) \rho^n \sin(n\theta)$$

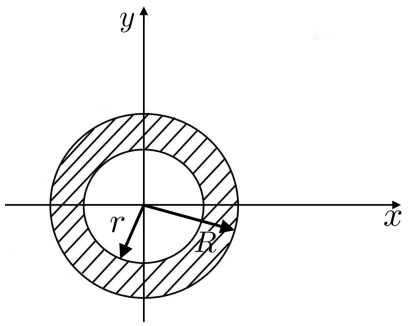
dove $a_0,\ a_n,\ b_n$ sono i coefficienti di Fourier di f.

[1] Se f è continua in θ questa è una vera uguaglianza.

Problema di Dirichlet sulla corona circolare

Il problema in coordinate polari è formulato in questo modo:

$$(**)\begin{cases}\Delta u(\rho,\theta)=0 & (\rho,\theta)\in[r,R]\times[-\pi,\pi)\\ u(r,\theta)=g(\theta) & \theta\in[-\pi,\pi)\\ u(R,\theta)=f(\theta) & \theta\in[-\pi,\pi)\end{cases}$$



Anche in questo caso si usa il metodo di separazione delle veriabili.

Primo passo

Il primo passo è virtualmente identico al caso del cerchio, con la differenza che ora i termini che contengono $\ln(\rho)$, ρ^{-n} sono limitate sulla corona circolare (che non contiene 0). Quindi in questo caso possiamo scegliere ogni combinazione lineare di

$$c_0,\ \overline{c}_0\ln(\rho),\ c_n\rho^n\cos(n\theta),\ \overline{c}_n\rho^{-n}\cos(n\theta),$$
$$d_n\rho^n\sin(n\theta),\ \overline{d}_n\rho^{-n}\sin(n\theta)$$

per determinare una soluzione di $(**)$.

Secondo passo

Scriviamo la soluzione di $(**)$ come

$$U(\rho,\theta)=c_0+\overline{c}_0\ln(\rho)+\sum_{n=1}^{+\infty}\left(c_n\rho^n+\overline{c}_n\rho^{-n}\right)\cos(n\theta)+\sum_{n=1}^{+\infty}\left(d_n\rho^n+\overline{d}_n\rho^{-n}\right)\sin(n\theta)\qquad(\bullet)$$

con i coefficienti $c_0,\ \overline{c}_0,\ c_n,\ \overline{c}_n,\ d_n,\ \overline{d}_n$ da determinare in base alle funzioni alle funzioni f e g.

Come nel caso precedente scriviamo

$$f(\theta)\cong\frac{a_0}{2}+\sum_{n=1}^{+\infty}a_n\cos(n\theta)+\sum_{n=1}^{+\infty}b_n\sin(n\theta)$$

(••)

$$g(\theta) \cong \frac{\alpha_0}{2} + \sum_{n=1}^{+\infty} \alpha_n \cos(n\theta) + \sum_{n=1}^{+\infty} \beta_n \sin(n\theta)$$

dove

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} f(\theta)\, d\theta \qquad a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(\theta)\cos(n\theta)\, d\theta \qquad b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(\theta)\sin(n\theta)\, d\theta$$

$$\alpha_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} g(\theta)\, d\theta \qquad \alpha_n = \frac{1}{\pi}\int_{-\pi}^{\pi} g(\theta)\cos(n\theta)\, d\theta \qquad \beta_n = \frac{1}{\pi}\int_{-\pi}^{\pi} g(\theta)\sin(n\theta)\, d\theta$$

Imponendo nelle espressioni (•) e (••)

$$\begin{cases} U(r,\theta) = f(\theta) \\ U(R,\theta) = g(\theta) \end{cases}$$

e uguagliando termine a termine, si ottengono i sistemi

$$\begin{cases} c_0 + \overline{c}_0 \ln(R) = \frac{a_0}{2} \\ c_0 + \overline{c}_0 \ln(r) = \frac{\alpha_0}{2} \end{cases}$$

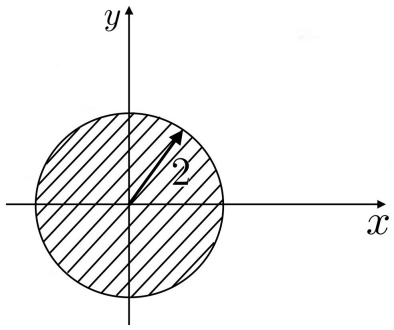
$$\begin{cases} c_n R^n + \overline{c}_n R^{-n} = a_n \\ c_n r^n + \overline{c}_n r^{-n} = \alpha_n \end{cases}$$

$$\begin{cases} d_n R^n + \overline{d}_n R^{-n} = b_n \\ d_n r^n + \overline{d}_n r^{-n} = \beta_n \end{cases}$$

da cui si ricavano i coefficienti $c_0,\ \overline{c}_0,\ c_n,\ \overline{c}_n,\ d_n,\ \overline{d}_n$ della soluzione del problema $(**)$.

Esercizio 1.

$$\begin{cases} \Delta u = 0 & \text{su } B_2(0) = \{(x,y)\,|x^2+y^2<4\} \\ u = x^4 & \text{su } \partial B_2(0) = \{(x,y)\,|x^2+y^2=4\} \end{cases}$$



In coordinate polari

$$\begin{cases} \Delta u(\rho,\theta) = u_{\rho\rho} + \frac{1}{\rho}u_\rho + \frac{1}{\rho^2}u_{\theta\theta} = 0 & (\rho,\theta) \in [0,2] \times [-\pi,\pi) \\ u(2,\theta) = (2\cos(\theta))^4 & \theta \in [-\pi,\pi) \end{cases}$$

Usando l'uguaglianza

$$(\cos(\theta))^2 = \left[\frac{1}{2}\cos(2\theta) + \frac{1}{2}\right]$$

calcoliamo la serie di Fourier di $f(\theta) = 16(\cos(\theta))^4$

$$(\cos(\theta))^4 = \left[(\cos(\theta))^2\right]^2 = \left[\frac{1}{2}\cos(2\theta) + \frac{1}{2}\right]^2 = \frac{1}{4}(\cos(2\theta))^2 + \frac{1}{4} + \frac{1}{2}\cos(2\theta) =$$

$$= \frac{1}{4}\left[\frac{1}{2}\cos(4\theta) + \frac{1}{2}\right] + \frac{1}{4} + \frac{1}{2}\cos(2\theta) =$$

$$= \frac{1}{8}\cos(4\theta) + \frac{1}{2}\cos(2\theta) + \frac{3}{8}$$

pertanto

$$f(\theta) = 16(\cos(\theta))^4 = 2\cos(4\theta) + 8\cos(2\theta) + 6$$

$$\frac{a_0}{2} = 6 \qquad a_2 = 8 \qquad a_4 = 2$$

$$a_n = 0 \quad \forall\, n \neq 0, 2, 4$$

$$b_n = 0 \quad \forall\, n = 1, 2, 3\ldots$$

scrviamo

$$U(\rho,\theta) = c_0 + \sum_{n=1}^{+\infty} c_n \rho^n \cos(n\theta) + \sum_{n=1}^{+\infty} d_n \rho^n \sin(n\theta)$$

e imponiamo $U(2,\theta) = f(\theta)$ termine a termine, cioè

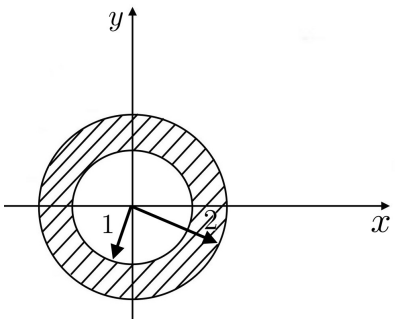
$$\begin{cases} c_0 = 6 \\ c_2\left(2^1\right) = 8 \\ c_4\left(2^4\right) = 2 \end{cases} \qquad \begin{cases} c_0 = 6 \\ c_2 = 2 \\ c_4 = \frac{1}{8} \end{cases}$$

La soluzione cercata è dunque

$$\tilde{U}(\rho,\theta) = 6 + 2\rho^2\cos(2\theta) + \frac{1}{8}\rho^4\cos(4\theta)$$

Esercizio 2. Risolvere

$$\begin{cases} \Delta u = 0 & (\rho,\theta) \in [1,2] \times [-\pi,\pi) \\ u(1,\theta) = (\cos(\theta))^2 & \theta \in [-\pi,\pi) \\ u(2,\theta) = 1 + \sin(\theta) & \theta \in [-\pi,\pi) \end{cases}$$



Nella corona circolare cerchiamo soluzioni del tipo

$$U(\rho,\theta) = c_0 + \overline{c}_0 \ln(\rho) + \sum_{n=1}^{+\infty} \left(c_n \rho^n + \overline{c}_n \rho^{-n}\right) \cos(n\theta) + \sum_{n=1}^{+\infty} \left(d_n \rho^n + \overline{d}_n \rho^{-n}\right) \sin(n\theta)$$

In questo caso dobbiamo imporre

$$U(1,\theta) = g(\theta) = (\cos(\theta))^2 = \frac{1}{2}\left[\cos(2\theta) + 1\right] = \underbrace{\frac{1}{2}}_{\frac{\alpha_0}{2}} + \underbrace{\frac{1}{2}}_{\alpha_2} \cos(2\theta)$$

$$U(2,\theta) = f(\theta) = \underbrace{1}_{\frac{a_0}{2}} + \underbrace{1}_{b_n} \sin(\theta)$$

Uguagliando termine a termine otteniamo

$$\begin{cases} c_0 + \overline{c}_0 \ln(1) = \frac{\alpha_0}{2} = \frac{1}{2} \\ c_0 + \overline{c}_0 \ln(2) = \frac{a_0}{2} = 1 \end{cases} \qquad \begin{cases} c_0 = \frac{1}{2} \\ \frac{1}{2} + \overline{c}_0 \ln(2) = 1 \end{cases} \qquad \begin{cases} c_0 = \frac{1}{2} \\ \overline{c}_0 = \frac{1}{2\ln(2)} \end{cases}$$

$$\begin{cases} c_2 1^1 + \overline{c}_2 1^{-1} = \alpha_2 = \frac{1}{2} \\ c_2 2^2 + \overline{c}_2 2^{-2} = a_2 = 0 \end{cases} \qquad \begin{cases} c_2 + \overline{c}_2 = \frac{1}{2} \\ 4c_2 + \frac{\overline{c}_2}{4} = 0 \end{cases} \qquad \begin{cases} c_2 = -\frac{1}{30} \\ \overline{c}_2 = \frac{8}{15} \end{cases}$$

$$\begin{cases} d_1 1^1 + \overline{d}_1 1^{-1} = \beta_2 = 0 \\ d_1 2^2 + \overline{d}_1 2^{-2} = b_2 = 1 \end{cases} \qquad \begin{cases} d_1 + \overline{d}_1 = 0 \\ 2d_1 + \frac{\overline{d}_1}{2} = 1 \end{cases} \qquad \begin{cases} d_1 = \frac{2}{3} \\ \overline{d}_1 = -\frac{2}{3} \end{cases}$$

da cui la soluzione cercata è:

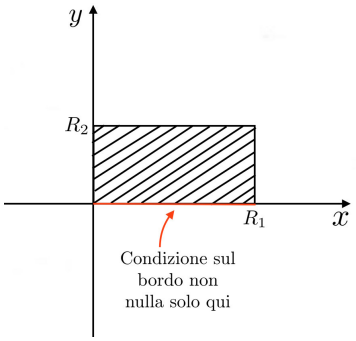
$$\tilde{U}(\rho,\theta) = \frac{1}{2} + \frac{1}{2\ln(2)} \ln(\rho) - \frac{1}{30}\rho^2 \cos(2\theta) + \frac{8}{15}\rho^{-2}\cos(2\theta) + \frac{2}{3}\rho \sin(\theta) - \frac{2}{3}\rho^{-1}\sin(\theta)$$

Osservazione 1. Si possono risolvere con la stessa tecninca, il problema di Neumann sul cerchio e sulla corona circolare, il problema misto sulla corona circplare in cui per uno dei due cerchi si utilizza Dirichlet e per l'altro Neumann.

Problema di Dirichlet nel rettangolo in $\mathbb{R}^2$

Risolvere:

$$(***)\begin{cases} \Delta u = 0 & (x,y) \in [0,R_1] \times [0,R_2] \\ u(0,y) = 0 & y \in [0,R_2] \\ u(R_1,y) = 0 & y \in [0,R_2] \\ u(x,0) = f(x) & x \in [0,R_1] \\ u(x,R_2) = 0 & x \in [0,R_1] \end{cases}$$



Anche questo caso si svolge con il metodo della separazione delle variabili.

Primo passo

Cerchiamo soluzioni del tipo

$$\Delta u = 0 \qquad (x,y) \in [0,R_1] \times [0,R_2]$$

che sono della forma

$$u(x,y) = h(x)\, g(y)$$

sostituendo nell'equazione si trova

$$u_{xx} + u_{yy} = h''(x)\, g(y) + h(x)\, g''(y) = 0$$

cioè

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

Essendo il primo membro solo funzione della sola xe il secondo della sola y, l'uguaglianza è soddisfatta se e solo se

$$\frac{h''(x)}{h(x)} = \lambda = -\frac{g''(y)}{g(y)}$$

cioè

$$\begin{cases} h''(x) - \lambda h(x) = 0 & (1) \\ g''(y) + \lambda g(y) = 0 & (2) \end{cases}$$

con le condizioni iniziali

$$(1)\begin{cases} h(0)=0 \\ h(R_1)=0 \end{cases} \quad \left(\text{da} \quad \begin{array}{c} u(0,y)=h(0)\,g(y)=0 \\ u(R_1,y)=h(R_1)\,g(y)=0 \\ \forall\, y\in[0,R_2] \end{array}\right)$$

$$(2) \quad g(R_2)=0 \qquad \left(\text{da} \quad \begin{array}{c} u(x,R_2)=h(x)\,g(R_2)=0 \\ \forall\, x\in[0,R_1] \end{array}\right)$$

- Risolviamo l'equazione (1)

  L'equazione caratteristica è $\eta^2-\lambda=0$ che porta a soluzioni di tipo

$$\begin{array}{cc} h(x)=c_1e^{\sqrt{\lambda}x}+c_2e^{-\sqrt{\lambda}x} & \lambda>0 \\ h(x)=c_1+c_2x & \lambda=0 \\ h(x)=c_1\cos\left(\sqrt{|\lambda|}x\right)+c_2\sin\left(\sqrt{|\lambda|}x\right) & \lambda<0 \end{array}$$

  ora sostituiamo le condizioni iniziali nelle tre tipologie e vediamo cosa si salva:

  1. Prima tipologia ($\lambda>0$)

$$h(0)=c_1+c_2=0 \implies c_1=-c_2$$

$$h(R_1)=c_1e^{\sqrt{\lambda}R_1}+c_2e^{-\sqrt{\lambda}R_1}=c_1\left(e^{\sqrt{\lambda}R_1}+e^{-\sqrt{\lambda}R_1}\right) \iff \begin{array}{c} \cancel{\lambda=0} \\ e \\ c_1=0 \end{array} \quad \Rightarrow c_1=c_2$$

     non si salva nessuna soluzione (tranne quella nulla).

  2. Seconda tipologia ($\lambda=0$)

$$h(0)=c_1=0$$

$$h(R_1)=c_1+c_2R_1=c_2R_1=0 \implies c_2=0=c_1$$

     anche qui non si salva nessuna soluzione.

  3. Terza tipologia ($\lambda<0$)

$$h(0)=c_1=0$$

$$h(R_1)=\cancel{c_1\cos\left(\sqrt{|\lambda|}R_1\right)}+c_2\sin\left(\sqrt{|\lambda|}R_1\right)=0$$

$$\iff \sqrt{|\lambda|}R_1=n\pi$$

$$\underset{\lambda<0}{\iff} \lambda=-\frac{n^2\pi^2}{R_1^2} \qquad n=1,2,3\ldots$$

     cioè si salvano le soluzioni del tipo

$$h(x)=c\sin\left(\frac{n\pi}{R_1}x\right) \qquad n=1,2,3\ldots$$

- Ora risolviamo l'equazione (2) con $\lambda = -\frac{n^2\pi^2}{R_1^2}$ $\qquad n = 1, 2, 3 \ldots$

$$g''(y) - \frac{n^2\pi^2}{R_1^2} g(y) = 0 \qquad n = 1, 2, 3 \ldots$$

L'equazione caratteristica è

$$\eta^2 - \frac{n^2\pi^2}{R_1^2} = 0 \qquad \eta = \pm\frac{n\pi}{R_1}$$

che porta a soluzioni del tipo

$$g(y) = c_1 e^{\frac{n\pi}{R_1} y} + c_2 e^{-\frac{n\pi}{R_1} y}$$

sostituendo la condizione iniziale

$$g(R_2) = c_1 e^{\frac{n\pi}{R_1} R_2} + c_2 e^{-\frac{n\pi}{R_1} R_2} = 0$$

cioè

$$c_2 = -c_1 e^{2n\pi \frac{R_2}{R_1}}$$

Otteniamo pertanto soluzioni del tipo

$$g(y) = c_1 \left(e^{\frac{n\pi}{R_1} y} - e^{-\frac{n\pi}{R_1} y} e^{2n\pi \frac{R_2}{R_1}} \right) =$$

$$\stackrel{*}{=} -c_1 2 e^{n\pi \frac{R_2}{R_1}} \left(\sinh\left(\frac{n\pi}{R_1} (R_2 - y) \right) \right) =$$

$$= k \sinh\left(\frac{n\pi}{R_1} (R_2 - y) \right) \qquad n = 1, 2, 3 \ldots, \qquad k = \text{costante}$$

Excursus su sinh **e** cosh

$$\sinh(z) \stackrel{\text{def}}{=} \frac{e^z - e^{-z}}{2} \qquad \cosh(z) \stackrel{\text{def}}{=} \frac{e^z + e^{-z}}{2}$$

$$\sinh(-z) = -\sinh(z) \qquad \cosh(-z) = \cosh(z)$$

$$\sinh(z_1 + z_2) = \sinh(z_1)\cosh(z_2) + \sinh(z_2)\cosh(z_1)$$

$$\cosh(z_1 + z_2) = \cosh(z_1)\cosh(z_2) + \sinh(z_1)\sinh(z_2)$$

$$[\cosh(z)]^2 - [\sinh(z)]^2 = 1$$

$$e^{\alpha} - e^{-\alpha} e^{2\beta} = e^{\beta}\left(e^{\alpha-\beta} - e^{\beta-\alpha}\right) = -2e^{\beta}\left(\frac{e^{\beta-\alpha} - e^{\alpha-\beta}}{2}\right) = -2e^{\beta}\sinh(\beta - \alpha)$$

questa è la formula usata in $*$, con $\alpha = \frac{n\pi}{R_1} y$ e $\beta = \frac{n\pi}{R_1} R_2$

conclusione

Tutte le soluzioni della forma

$$c_n \sin\left(\frac{n\pi}{R_1}x\right)\sinh\left(\frac{n\pi}{R_1}(R_2-y)\right) \qquad n=1,2,3\ldots \tag{1}$$

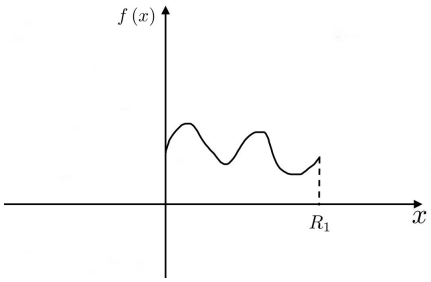
e, per la linearità di Δ, anche tutte le loro combinazioni lineari, soddisfano

$$\Delta u(x,y)=0 \qquad (x,y)\in[0,R_1]\times[0,R_2]$$

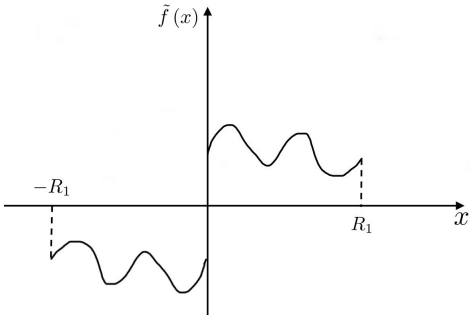
Secondo passo

Dobbiamo trovare una particolare combinazione lineare $\tilde{U}$ del tipo (1) per cui sia $\tilde{U}(x,0)=f(x) \quad x\in[0,R_1]$

Abbiamo



Possiamo pensare di prolungare f per disparità su $[-R_1,R_1]$



Allora

$$\tilde{f}(x)\cong\sum_{n=1}^{+\infty}b_n\sin\left(\frac{n\pi}{R_1}x\right) \qquad x\in[-R_1,R_1]$$

$$b_n=\frac{1}{R_1}\int_{-R_1}^{R_1}\sin\left(\frac{n\pi}{R_1}x\right)\tilde{f}(x)\,dx \quad \underset{\text{f dispari}}{\overset{=}{\downarrow}} \quad \frac{2}{R_1}\int_0^{R_1}\sin\left(\frac{n\pi}{R_1}\right)f(x)\,dx$$

Abbiamo dunque per la soluzione di $(***)$ una scrittura del tipo

$$U(x,y) = \sum_{n=1}^{+\infty} c_n \sinh\left(\frac{n\pi}{R_1}(R_2 - y)\right) \sin\left(\frac{n\pi}{R_1}x\right)$$

mentre possiamo scrivere per il lato f

$$\tilde{f}(x) \cong \sum_{n=1}^{+\infty} b_n \sin\left(\frac{n\pi}{R_1}x\right) \qquad \text{dove } b_n = \frac{2}{R_1}\int_0^{R_1} f(x) \sin\left(\frac{n\pi}{R_1}x\right) dx$$

Imponendo $U(x,0) = f(x)$ ed uguagliando ciascun coeffciente si ottiene:

$$c_n \sinh\left(\frac{n\pi}{R_1}R_2\right) = b_n \qquad n = 1, 2, 3 \ldots$$

da cui si riava

$$c_n = \frac{b_n}{\sinh\left(\frac{n\pi}{R_1}R_2\right)}$$

La soluzione di $(***)$ cercata è dunque

$$\tilde{U}(x,y) = \sum_{n=1}^{+\infty} \frac{b_n}{\sinh\left(\frac{n\pi}{R_1}R_2\right)} \sinh\left(\frac{n\pi}{R_1}(R_2 - y)\right) \sin\left(\frac{n\pi}{R_1}x\right)$$

Osservazione 1. Quando il problema ha condizioni al contorno non nulle su tutti e quattro i lati

$$\begin{cases} \Delta u = 0 & (x,y) \in [0,R_1] \times [0,R_2] \\ u(0,y) = f_1(y) & y \in [0,R_2] \\ u(R_1,y) = f_2(y) & y \in [0,R_2] \\ u(x,0) = f_3(x) & x \in [0,R_1] \\ u(x,R_2) = f_4(x) & x \in [0,R_1] \end{cases}$$

si cercano separatamente le quattro soluzioni $u_1,\ u_2,\ u_3,\ u_4$ dei problemi con condizioni al contorno tutte nulle tranne che una. Poi per la linearità di Δ si ha che la soluzione cercata è $u_1 + u_2 + u_3 + u_4$

Risolvere il problema

$$\begin{cases} \Delta u = 0 & (x,y) \in [0,1] \times [0,2] \\ u(0,y) = y & y \in [0,2] \\ u(1,y) = 0 & y \in [0,2] \\ u(x,0) = 0 & x \in [0,1] \\ u(x,2) = 0 & x \in [0,1] \end{cases}$$

metodo della separazione delle variabili:

cerchiamo soluzioni del tipo

$$u(x,y) = h(x)\, g(y)$$

sostituendo nell'equazione

$$h''(x)\, g(y) + h(x)\, g''(y) = 0$$

da cui

$$-\frac{g''(y)}{g(y)} = \frac{h''(x)}{h(x)} = \lambda$$

$$\Longrightarrow \begin{cases} g''(y) + \lambda g(y) = 0 & (1) \\ h''(x) - \lambda h(x) = 0 & (2) \end{cases} \qquad \text{Con } \begin{cases} g(0) = 0 \\ g(2) = 0 \end{cases} \quad e \qquad h(1) = 0$$

Equazione (1):

$$\eta^2 + \lambda = 0 \qquad \eta^2 = -\lambda$$

$$\begin{array}{ll} \lambda < 0 & g(y) = c_1 e^{\sqrt{|\lambda|}y} + c_2 e^{-\sqrt{|\lambda|}y} \\ \lambda = 0 & c_1 + c_2 y \\ \lambda > 0 & c_1 \cos\left(\sqrt{\lambda} y\right) + c_2 \sin\left(\sqrt{\lambda} y\right) \end{array}$$

Sostituendo le condizioni iniziali si trovano solo soluzioni del terzo tipo con

$$c_1 = 1 \qquad \sqrt{\lambda} = \frac{\pi n}{2} \qquad \Rightarrow \lambda = \frac{\pi^2 n^2}{4}$$

scriviamo dunque

$$g_n(y) = c \sin\left(\frac{\pi n}{2} y\right) \qquad n = 1, 2, 3 \ldots$$

e troviamo le rispettive soluzioni dell'equazione (2).

$$h''(x) - \frac{n^2\pi^2}{4} h(x) = 0$$

equazione caratteristica:

$$\eta^2 - \frac{n^2\pi^2}{4} = 0 \qquad \text{da cui } \eta = \pm\frac{n\pi}{2}$$

eoluzioni possibili:

$$h(x) = c_1 e^{\frac{n\pi}{2}x} + c_2 e^{-\frac{n\pi}{2}x}$$

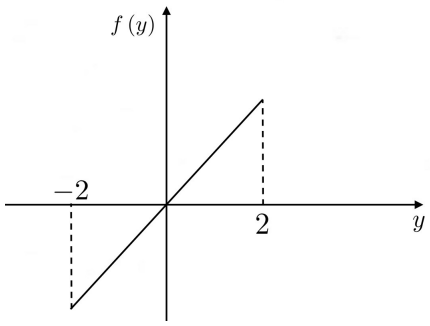
sostituiamo la condizione al contorno:

$$h_n(x) = c\left(e^{\frac{n\pi}{2}x} - e^{-\frac{n\pi}{2}x} e^{nx}\right) = -2ce^{\frac{n\pi}{2}} \sinh\left(\frac{n\pi}{2}(1-x)\right) = \tilde{c} \sinh\left(\frac{n\pi}{2}(1-x)\right)$$

Otteniamo dunque soluzioni di $\Delta u = 0$ su $[0,1] \times [0,2]$ della forma

$$U(x,y) = \sum_{n=1}^{+\infty} c_n \sinh\left(\frac{n\pi}{2}(1-x)\right) \sin\left(\frac{\pi n}{2} y\right)$$

Per quanto riguarda il dato al contorno, abbiamo



$$f(y) = \sum_{n=1}^{+\infty} b_n \sin\left(\frac{n\pi}{2} y\right)$$

$$\text{dove } b_n = \frac{1}{2}\int_{-2}^{2} y \sin\left(\frac{\pi n}{2} y\right) dy = \frac{2}{2}\int_{0}^{2} y \sin\left(\frac{\pi n}{2} y\right) dy =$$

$$= \left[y\left(-\frac{2}{\pi n} \cos\left(\frac{\pi n}{2} y\right)\right)\right]_0^2 + \cancel{\int_0^2 \frac{2}{\pi n} \cos\left(\frac{\pi n}{2} y\right) dy} =$$

$$= -\frac{4}{\pi n}(-1)^n = \frac{4}{\pi n}(-1)^{n+1}$$

Imponendo $U(0, y) = f(y)$ ed uguagliando termine a termine si ha

$$c_n \sinh\left(\frac{n\pi}{2}\right) = \frac{4}{n\pi}(-1)^{n+1} \qquad \text{da cui } c_n = \frac{\frac{4}{n\pi}}{\sinh\left(\frac{n\pi}{2}\right)}(-1)^{n+1} \qquad n = 1, 2, 3 \ldots$$

La soluzione cercata è dunque

$$\tilde{U}(x, y) = \sum_{n=1}^{+\infty} \frac{\frac{4}{n\pi}}{\sinh\left(\frac{n\pi}{2}\right)}(-1)^{n+1} \sinh\left(\frac{n\pi}{2}(1-x)\right) \sin\left(\frac{n\pi}{2} y\right)$$

ESERCIZI SULL'EQUAZIONE DI LAPLACE.

N.B. I problemi indicati con $(*)$ sono più difficili e non sono considerati essenziali per il superamento della prova parziale. Gli altri sono invece considerati requisiti minimi per il superamento della prova. Un eventuale problema del tipo di quelli con $(*)$ nel compito, se risolto correttamente, darà un bonus di punteggio per chi aspira ad un voto alto (≥ 25).

1. Si consideri il problema di Dirichlet per l'equazione di Laplace nel rettangolo $[0, L] \times [0, M]$ sul piano xy

$$(PD) \begin{cases} u_{xx} + u_{yy} = 0 \quad , & 0 \leq x \leq L \quad , \quad 0 \leq y \leq M \\ u(x, 0) = f_1(x) \quad , & 0 \leq x \leq L \\ u(x, M) = f_2(x) \quad , & 0 \leq x \leq L \\ u(0, y) = g_1(y) \quad , & 0 \leq y \leq M \\ u(L, y) = g_2(y) \quad , & 0 \leq y \leq M \end{cases}$$

dove si assume che il dato sul bordo del rettangolo sia continuo. Grazie al principio di sovrapposizione possiamo scrivere la soluzione $u(x, y)$ del problema (PD) come somma delle soluzioni dei quattro sottoproblemi di Dirichlet seguenti, ciascuno con condizioni omogenee su tre dei quattro lati del rettangolo:

$$(PD1) \begin{cases} u^1_{xx} + u^1_{yy} = 0 & 0 \leq x \leq L \quad , \quad 0 \leq y \leq M \\ u^1(x, 0) = f_1(x) & 0 \leq x \leq L \\ u^1(x, M) = 0 & 0 \leq x \leq L \\ u^1(0, y) = u^1(L, y) = 0 \quad , & 0 \leq y \leq M \end{cases}$$

$$(PD2) \begin{cases} u^2_{xx} + u^2_{yy} = 0 & 0 \leq x \leq L \quad , \quad 0 \leq y \leq M \\ u^2(x, 0) = 0 & 0 \leq x \leq L \\ u^2(x, M) = f_2(x) & 0 \leq x \leq L \\ u^2(0, y) = u^2(L, y) = 0 \quad , & 0 \leq y \leq M \end{cases}$$

$$(PD3) \begin{cases} u^3_{xx} + u^3_{yy} = 0 & 0 \leq x \leq L \quad , \quad 0 \leq y \leq M \\ u^3(x, 0) = u^3(x, M) = 0 \quad , & 0 \leq x \leq L \\ u^3(0, y) = g_1(y) & 0 \leq y \leq M \\ u^3(L, y) = 0 & 0 \leq y \leq M \end{cases}$$

$$(PD4) \begin{cases} u^4_{xx} + u^4_{yy} = 0 & 0 \leq x \leq L \quad , \quad 0 \leq y \leq M \\ u^4(x, 0) = u^4(x, M) = 0 \quad , & 0 \leq x \leq L \\ u^4(0, y) = 0 & 0 \leq y \leq M \\ u^4(L, y) = g_2(y) & 0 \leq y \leq M \end{cases}$$

La soluzione di (PD) sarà allora $u(x, y) = u_1(x, y) + u_2(x, y) + u_3(x, y) + u_4(x, y)$ (DOMANDA: Dove esattamente?) Trovare le soluzioni dei quattro problemi di Dirichlet $(PD1)$, $(PD2)$, $(PD3)$, $(PD4)$ con il metodo di separazione delle variabili; verificare in particolare che

$$u^1(x, y) = \sum_{n=1}^{+\infty} A_n \sinh \frac{n\pi(M - y)}{L} \sin \frac{n\pi x}{L}$$

dove

$$A_n = \frac{2}{L \sinh(n\pi M/L)} \int_0^L f_1(x) \sin \frac{n\pi x}{L} dx = \frac{b_n}{\sinh(n\pi M/L)}$$

con b_n coefficienti di Fourier di f_1 sviluppata in serie di soli seni. Trovare le formule analoghe per gli altri problemi.

2. Applicare il metodo descritto sopra per trovare la soluzione al problema di Dirichlet come (PD) con $L = \pi/2,\ M = \pi,\ f_1 = \sin x,\ f_2 = \cos 2x,\ g_1 = \sin \frac{y}{2}$ e $g_2 = \cos y$.

3. Trovare la soluzione $u(x, y)$ dei seguenti problemi misti con il metodo di separazione delle variabili

(1)

$$\begin{cases} u_{xx} + u_{yy} = 0 & 0 \le x \le L \quad , \quad 0 \le y \le M \\ u(x, 0) = 0 & 0 \le x \le L \\ u(x, M) = f(x) & 0 \le x \le L \\ u_x(0, y) = u_x(L, y) = 0 \quad , & 0 \le y \le M \end{cases}$$

(2)

$$\begin{cases} u_{xx} + u_{yy} = 0 & 0 \le x \le L \quad , \quad 0 \le y \le M \\ u(x, 0) = u(x, M) = 0 \quad , & 0 \le x \le L \\ u_x(0, y) = g(y) & 0 \le y \le M \\ u_x(L, y) = 0 & 0 \le y \le M \end{cases}$$

Per ciascuno dei problemi sopra elencati descrivere una situazione fisica di cui ne sia un modello.

(∗)**4.** Sia $u(x, y)$ la soluzione del seguente problema di Neumann per l'equazione di Laplace

$$\begin{cases} u_{xx} + u_{yy} = 0 & 0 \le x \le L \quad , \quad 0 \le y \le M \\ u_y(x, 0) = 0 & 0 \le x \le L \\ u_y(x, M) = f(x) & 0 \le x \le L \\ u_x(0, y) = u_x(L, y) = 0 \quad , & 0 \le y \le M \end{cases}$$

(a) Si spieghi, senza risolvere il problema, sotto quale condizione fisica questo problema può essere risolto.

(b) Si risolva il problema con il metodo di separazione delle variabili e si mostri che il metodo funziona solo se vale la condizione trovata in **(a)**.

(c) Si determini la costante arbitraria, nella soluzione di **(b)** considerando la soluzione dell'equazione di diffusione indipendente dal tempo con dato iniziale $u(x, y, 0) = g(x, y)$.

5. Trovare tutti i polinomi omogenei di grado 3 e 4 che sono armonici in tutto $\mathbb{R}^2$.

6. Risolvere il problema di Dirichlet per l'equazione di Laplace in un disco di raggio R con dato al bordo $f(\theta)$ (con f periodica di periodo 2π), dopo aver scritto l'operatore di Laplace in coordinate polari. Risolvere poi il problema con $f(\theta) = -2\cos 3\theta + \sin 2\theta$.

7. Sia $u(x, y)$ la soluzione del problema di Dirichlet per l'equazione di Laplace nel triangolo di vertici l'origine O, il punto $A(0, 2)$ ed il punto $B(1, 1)$ con dato $2(2 - x)(1 - x)$ sul lato AB e nullo sugli altri due. Mostrare che

$$y(y - x) \leq u(x, y) \leq 0$$

sul suddetto triangolo.

8. Si risolva, con il metodo di separazione delle variabili, l'equazione di Laplace nel quarto di un cerchio di raggio R con le condizioni (date in coordinate polari)

(1) $u_\theta(r, 0) = u\left(r, \frac{\pi}{2}\right) = 0 \qquad u(1, \theta) = f(\theta)$

(2) $u_\theta(r, 0) = u_\theta\left(r, \frac{\pi}{2}\right) = 0 \qquad u(1, \theta) = f(\theta)$

Appunti di analisi funzionale

1 Soluzioni deboli 1-d

1.1 Motivazione dell'introduzione di concetto soluzione debole

La formulazione classica per un'equazione differenziale ed il concetto di soluzione classica (cioè una funzione che soddisfa l'equazione punto per punto e che ammette tutte le derivate coinvolte nell'equazione continue) in molti casi, spesso derivanti da applicazioni fisiche, è inadeguata. Vediamo alcuni esempi.

Cominciamo a considerare il semplice problema di Dirichlet 1-d

$$\begin{cases} -u''(x) = f(x) & 0 < x < 1, \\ u(0) = 0 & u(1) = 0 \end{cases} \tag{1.1}$$

che governa ad esempio la configurazione di equilibrio di un filo elastico, con tensione pari ad uno e fissato agli estremi, in regime di piccoli spostamenti e soggetto ad una forza verticale di intensità f. Il significato fisico di $f(x)$ è che la forza complessiva agente sul tratto $(0, x)$ è

$$F(x) = \int_0^x f(t)\, dt \tag{1.2}$$

La soluzione u rappresenta lo spostamento verticale del filo rispetto alla posizione di riposo $u \equiv 0$. Per vedere che la formulazione classica (1.1), che richiede l'esistenza di una soluzione $u \in C^2\big((0,1)\big) \cap C^0\big([0,1]\big)$ che soddisfi l'equazione puntualmente, non è in generale adeguata si consideri ad esempio il caso di un filo soggetto a carichi diversi.

Esempio 1.1. *Come primo esempio consideriamo un filo con estremi fissati, soggetto a carico unitario concentrato in $x = \frac{1}{2}$ trasversale al filo e rivolto verso il basso (in questo caso f è rappresentabile con una delta di Dirac[1] in $x = \frac{1}{2}$,*

[1] La delta di Dirac $\delta_0(x)$ è una una funzione generalizzata che può essere definita come " $\lim_{n\to\infty}$ " $f_n(x)$ dove

$$f_n(x) = \begin{cases} 0 & x < 0 \ \wedge \ x > \frac{1}{n} \\ n & 0 \le x \le \frac{1}{n} \end{cases} \tag{1.3}$$

Questo limite vale zero se $x \neq 0$ e $+\infty$ se $x = 0$, quindi non é una funzione come quelle che abbiamo considerato finora. Poiché $\int_{-\infty}^{+\infty} f_n(x)\, dx = 1$ (area rettangolo di base $\frac{1}{n}$ e altezza n) diciamo che anche

$$\int_{-\infty}^{+\infty} \delta_0(x)\, dx = 1 \tag{1.4}$$

indicata con $\delta_{\frac{1}{2}}(x)$). Dall'interpretazione che abbiamo dato di $F(x)$ chiaramente si ha che

$$F(x) = \begin{cases} 0 & 0 < x \le \frac{1}{2} \\ -1 & \frac{1}{2} \le x \le 1 \end{cases} \tag{1.7}$$

L'intuizione ci dice che la soluzione fisica esiste ed è continua. Proviamo a calcolarla.

Dalla prima riga della (1.1) *'integrado' e usando* (1.7) *otteniamo*

$$-u' = F(x) + c = \begin{cases} c & 0 < x \le \frac{1}{2} \\ c - 1 & \frac{1}{2} \le x \le 1 \end{cases} \tag{1.8}$$

con c costante arbitraria da determinare. Avremo allora che

$$u(x) = \int_0^x u'(t)\, dt = \int_0^x (-F(t) - c)\, dt\,. \tag{1.9}$$

Si osservi che la $u(x)$ così definita soddisfa già $u(0) = 0$ ed è continua su tutto l'intervallo $[0, 1]$. Da (1.8) *e da* (1.7) *otteniamo che quando $0 \le x < \frac{1}{2}$*

$$u(x) = -\int_0^x c\, dt = -c\, x \tag{1.10}$$

mentre quando $\frac{1}{2} \le x \le 1$

$$u(x) = -\int_0^{\frac{1}{2}} c\, dt + \int_{\frac{1}{2}}^x (1 - c)\, dt = -\frac{c}{2} + (1 - c)\left(x - \frac{1}{2}\right). \tag{1.11}$$

Resta ora da determinare c in modo che sia soddisfatta la condizione $u(1) = 0$. Sostituendo in (1.11) *si ottiene la condizione*

$$u(1) = -\frac{c}{2} + (1 - c)\frac{1}{2} = -\,c + \frac{1}{2} = 0 \tag{1.12}$$

da cui $c = \frac{1}{2}$. Quindi la soluzione di (1.1) *calcolata in questo modo é*

$$u(x) = \begin{cases} -\frac{1}{2}\, x & 0 \le x < \frac{1}{2} \\ \frac{1}{2}\,(x - 1) & \frac{1}{2} \le x \le 1 \end{cases} \tag{1.13}$$

Quindi la $\delta_0(x)$ rappresenta una massa unitaria concentrata nell'origine. La $\delta_0(x)$ puó essere vista anche come la "derivata" della funzione di Heavyside

$$H(x) = \begin{cases} 0 & x < 0 \\ 1 & x \ge 0 \end{cases} \tag{1.5}$$

Possiamo poi definire $\delta_a(x) = \delta_0(x - a)$ che concentra massa in un arbitrario punto a. La (1.4) implica che se $g(x)$ é una funzione continua

$$\int_{-\infty}^{+\infty} g(x)\delta_a(x)\, dx = g(a). \tag{1.6}$$

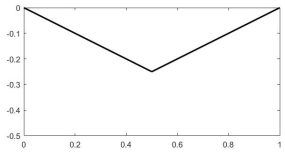


Fig. 1.a

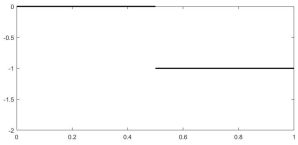


Fig. 1.b

Chiaramente questa soluzione non è di classe $C^2(0,1)$ come richiede la definizione di soluzione classica di un'equazione differenziale del secondo ordine, anzi addirittura non è neanche di classe $C^1(0,1)$ non essendo derivabile in $x = \frac{1}{2}$. Si potrebbe pensare che allora basti indebolire leggermente la definizione chiedendo in questo caso che l'equazione sia soddisfatta punto per punto tranne in $x = \frac{1}{2}$ ma in questo modo si perderebbero tutte le informazioni su cosa succede in quel punto che nel nostro esempio è proprio quello più importante, dove succede 'tutto'. Inoltre se ci accontentassimo di una soluzione che soddisfa l'equazione in tutti i punti tranne in $x = \frac{1}{2}$, anche richiedendo la continuità avremmo che tutte le funzioni del tipo

$$g(x) = \begin{cases} -k\,x & 0 \le x < \frac{1}{2} \\ k\,(x-1) & \frac{1}{2} \le x \le 1 \end{cases} \tag{1.14}$$

con $k > 0$ potrebbero essere soluzioni, perché $g''(x) = 0$ per $x \neq \frac{1}{2}$. Inoltre g'' é "infinita" in $x \neq \frac{1}{2}$ ma se anche queste fossero soluzione sarebbe contraddetta l'evidenza fisica che la soluzione (configuazione della corda) è unica.

Nell'ottica dell'introduzione di un diverso concetto di soluzione che possa includere anche esempi di questo tipo, osserviamo che se moltiplichiamo l'equazione in (1.1) con $f(x) = -\delta_{\frac{1}{2}}(x)$ per una qualsiasi funzione $v \in C^1(0,1)$ con $v(0) = v(1) = 0$ (indicheremo l'insieme di queste funzioni con $C^1_0(0,1)$, dove il pedice indica che sono funzioni zero al bordo) e integriamo tra 0 e 1, otteniamo

$$-\int_0^1 u''(x)\,v(x)\,dx \;=\; -\int_0^1 \delta_{\frac{1}{2}}(x)\,v(x)\,dx \tag{1.15}$$

Abbiamo visto nella nota sulla δ di Dirac che l'integrale a destra è uguale a $-v(\frac{1}{2})$. A sinistra integrando per parti, usando l'annullamento di $v(x)$ al bordo e la (1.13), ed il teorema fondamentale del calcolo integrale, otteniamo

(1.16)
$$-\int_0^1 u''(x)v(x)\,dx = -u'(x)v(x)\Big]_0^1 + \int_0^1 u'(x)v'(x)\,dx = \int_0^1 u'(x)\,v'(x)\,dx$$

$$= \int_0^{\frac{1}{2}} \left(-\frac{1}{2}\right) v'(x)\,dx + \int_{\frac{1}{2}}^1 \left(\frac{1}{2}\right) v'(x)\,dx = -\frac{1}{2}\,v(x)\Big]_0^{\frac{1}{2}} + \frac{1}{2}\,v(x)\Big]_{\frac{1}{2}}^1 \tag{1.17}$$

$$= -\frac{1}{2}\,v\left(\frac{1}{2}\right) - \frac{1}{2}\,v\left(\frac{1}{2}\right) = -v\left(\frac{1}{2}\right). \tag{1.18}$$

Abbiamo quindi mostrato che la nostra soluzione (1.13) è una funzione $u \in C_0(0,1)$ (cioè continua in $(0,1)$, che vale 0 ai due estremi dell'intervallo) che soddisfa la seguente relazione

$$\int_0^1 u'(x)\, v'(x)\, dx = \int_0^1 f(x)\, v(x)\, dx \tag{1.19}$$

per ogni funzione $v \in C_0^1(0,1)$.

Mostrare per esercizio che è l'unica funzione tra quelle del tipo (1.14) che soddisfa questa relazione (1.19) (cioè la relazione è verificata solo se $k = \frac{1}{2}$).

Esercizio 1.1. *Considerare il caso di due carichi unitari concentrati rispettivamente in $x = 2/5$ e $x = 3/5$.*

Esempio 1.2. *Consideriamo ora il caso di un carico di intensità unitaria, distribuito uniformente sull'intervallo $[2/5, 3/5]$, cioè nel problema (1.1) abbiamo*

$$f(x) = \begin{cases} 0 & x \in (0, 2/5) \\ -1 & x \in [2/5, 3/5] \\ 0 & x \in (3/5, 1) \end{cases} \tag{1.20}$$

quindi una funzione "più ragionevole" dal punto di vista matematico.

Soluzione. Come nell'esempio 1.1, integrando una prima volta, otteniamo

$$u'(x) = -\int_0^x f(x)\, dx + k = \begin{cases} k & x \in (0, 2/5) \\ x - 2/5 + k & x \in [2/5, 3/5] \\ 1/5 + k & x \in (3/5, 1) \end{cases} \tag{1.21}$$

e integrando una seconda volta (usando la condizione $y(0) = 0$)

$$u(x) = \int_0^x u'(x)\, dx = \begin{cases} k\, x & x \in (0, 2/5) \\ x^2/2 + (k - 2/5)x + 2/25 & x \in [2/5, 3/5] \\ (1/5 + k)x - 1/10 & x \in (3/5, 1) \end{cases} . \tag{1.22}$$

sostituendo ancora la condizione $y(1) = 0$, otteniamo $k = -1/10$ e quindi

$$u(x) = \int_0^x u'(x)\, dx = \begin{cases} -1/10\, x & x \in (0, 2/5) \\ x^2/2 - 1/2\, x + 2/25 & x \in [2/5, 3/5] \\ 1/10\ x - 1/10 & x \in (3/5, 1) \end{cases} . \tag{1.23}$$

Di nuovo abbiamo ottenuto una soluzione per il problema ragionevole dal punto di vista fisico, ma che non è una soluzione matematica nel senso classico, in quanto ammette derivata prima continua ma non ammette derivata seconda in tutto $(0,1)$. Quindi ancora una volta dobbiamo dare una nozione di soluzione più generale di quella classica.

Si consideri più in generale il seguente problema di Dirichlet omogeneo per una **equazioni ellittica in forma di divergenza**

$$\begin{cases} -\big(A(x)u'(x)\big)' + B(x)u'(x) + C(x)u(x) = f(x) & x \in (\alpha, \beta) \\ u(\alpha) = u(\beta) = 0 \end{cases} , \tag{1.24}$$

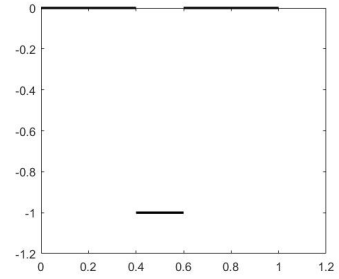


Fig. 2.a

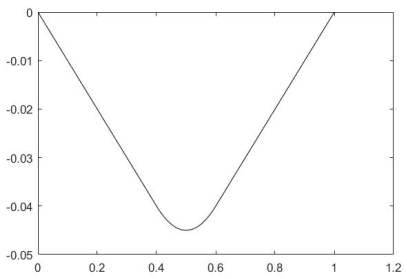


Fig. 2.b

Se $A(x) \geq \lambda > 0$ in (α, β) l'equazione si dice ellittica. Se le funzioni $A', B, C, f \in C^0(\alpha, \beta)$ il problema (1.24) ammette una soluzione classica $u \in C^2\big((\alpha, \beta)\big) \cap C^0\big([\alpha, \beta]\big)$. Se, come in (1.15), moltiplichiamo l'equazione in (1.24) per una qualsiasi funzione $v \in C_0^1(0,1)$ e integriamo tra α e β abbiamo

(1.25)
$$\int_\alpha^\beta \Big[-\big(A(x)u'(x)\big)' + B(x)u'(x) + C(x)u(x)\Big]\, v(x)\, dx = \int_\alpha^\beta f(x)\, v(x)\, dx\,.$$

Integrando poi per parti il primo termine a sinistra e usando le condizioni di Dirichlet otteniamo

$$-\int_\alpha^\beta \big(A(x)u'(x)\big)'\, v(x)\, dx \;=\; -A(x)\, u'(x)\, v(x)\Big]_\alpha^\beta + \int_\alpha^\beta A(x)\, u'(x)\, v(x)'\, dx \;=$$

$$= \int_\alpha^\beta A(x)\, u'(x)\, v(x)'\, dx \tag{1.26}$$

sostituendo ora (1.26) in (1.25) otteniamo (omettendo la dipendenza da x per alleggerire la scrittura)

$$\int_\alpha^\beta \left[A\, u'\, v' + B\, u'\, v \;+\; C\, u\, v\right]\, dx = \int_\alpha^\beta f\, v\, dx\,. \tag{1.27}$$

Osserviamo che la formula (1.27) ha significato anche se la soluzione u non ha 2 derivate continue perché la derivata seconda non compare più, ed inoltre anche la derivata prima compare solo integrata, quindi potrebbe anche non essere continua. Allo stesso tempo le funzioni A, B, C, f così come v' non debbono essere necessariamente continue. Da queste osservazioni può nascere appunto la nostra 'candidata definizione di soluzione'. Il nostro obiettivo sarà quindi quello di utilizzare la formula (1.27) per dare una definizione di soluzione più generale (o più *debole*) di quella classica, che potrà consistere nel cercare una funzione u in un insieme (anzi spazio vettoriale) opportuno, tale che (1.27) sia valida per tutte le funzioni v anch'esse in uno spazio da definire. Dovremo quindi identificare lo spazio in cui cercare la u soluzione (ma vedremo anche sarà lo

stesso in cui prendere anche le funzioni v che saranno chiamate *funzioni test*) in modo tale che la formula (1.27) abbia senso, ma anche in modo tale che il problemo risulti ben posto.

D'altra parte dobbiamo assicurarci che nel caso in cui però una soluzione classica esista la nostra definizione riproduca proprio quella soluzione e non un'altra. Mettiamo l'accento sul fatto che stiamo cercando una definizione di soluzione che sia più debole nel senso che ci permetta di risolvere problemi più generali, tipo quelli degli esempi considerati prima (anzi di trovare proprio la soluzione che ha significato fisico in quegli esempi) ma che non aumenti le soluzioni nel caso in cui una soluzione classica esista. A questo proposito mostriamo che se $u \in C^2\big((\alpha,\beta)\big) \cap C^0_0\big([\alpha,\beta]\big)$ soddisfa (1.27) per ogni funzione $v \in C^1_0(0,1)$ allora la u soddisfa puntualmente l'equazione nel problema (1.24) (le condizioni di Dirichlet omogenee sono soddisfatte dal fatto che si prende u in C^0_0). Sotto queste ipotesi integrando per parti il primo termine dell'integrale di sinistra in (1.27) e utilizzando il fatto che la funzione test v è zero agli estremi, abbiamo

$$\int_\alpha^\beta [A\,u'\,v' + B\,u'\,v \,+\, C\,u\,v]\;dx \;=$$

$$=\; A\,u'(x)\,v(x)\Big]_\alpha^\beta - \int_\alpha^\beta \Big[\big(A\,u'\big)'\,v\Big]\;dx \,+\, \int_\alpha^\beta [B\,u'\,v \,+\, C\,u\,v]\;dx \;=$$

$$=\; -\int_\alpha^\beta \Big[\big(A\,u'\big)'\,v \,+\, B\,u'\,v \,+\, C\,u\,v\Big]\;dx \tag{1.28}$$

cioè

$$\int_\alpha^\beta \Big[-\big(A\,u'\big)' + B\,u' \,+\, C\,u \,-\, f\Big]\;v(x)\;dx \;=\; 0. \tag{1.29}$$

Vogliamo concludere che la validità di (1.29) per ogni funzione test v implica che la funzione in parentesi è nulla per ogni $x \in (\alpha,\beta)$ e quindi soddisfa l'equazione in (1.24) punto per punto ed è quindi soluzione classica. Si osservi che per le ipotesi fatte la funzione in parentesi quadra è continua, quindi se fosse diversa da zero (diciamo per fissare le idee > 0) in un punto $x_0 \in (\alpha,\beta)$ per il teorema di permannenza del segno sarebbe > 0 in $(x_0-\epsilon, x_0+\epsilon)$. Si potrebbe allora scegliere una test $v \geq 0$ tale che $v(x) = 0$ per $x \notin (x_0-\epsilon, x_0+\epsilon)$ e $v(x_0) > 0$. É facile concludere a questo punto che l'integrale (1.29) sarebbe positivo e non nullo.

Osserviamo a questo punto che il termine di sinistra di (1.27) associa ad (u,v) un numero reale e dipende linearmente (cioè conserva somma e prodotto per uno scalare) sia da u che da v: è detto per questo *forma bilineare* e sarà indicato con $a(u,v)$. Analogamente il termine di sinistra associa a v un numero reale e dipende linearmente da v: sarà chiamato *funzionale lineare* ed indicato con $F(v)$. Le definizioni precise di a e di F saranno date più avanti. Possiamo quindi dire che la **formulazione debole** del problema di Dirichlet (1.24) ha la forma

Trovare una funzione u tale che, per ogni funzione v si abbia

$$a(u,v) = F(v) \tag{1.30}$$

con u e v appartenenti ad opportuni spazi di funzioni.

Vedremo che l'ambiente naturale in cui cercare la soluzione del problema qui sopra è quello degli spazi di Hilbert e che u e v dovranno appartenere allo stesso spazio (di Hilbert) e poi vedremo quali spazi di Hilbert sono adatti a risolvere i diversi problemi di questo tipo derivanti da equazioni differenziali.

1.2 Cenni di analisi funzionale: spazi normati, con prodotto interno e spazi di Hilbert

Questa parte verrà completata successivamente. Al momento si può fare riferimento al libro di testo S. Salsa, F. Vegni, A. Zaretti e P. Zunino, *Invito alle equazioni a derivate parziali*, Springer (2008).

Considereremo sempre insiemi che sono spazi vettoriali, cioè chiusi rispetto alla somma ed al prodotto per uno scalare.

Definizione 1.1. ***Norma e distanza o metrica*** *si veda libro a pag. 264*

Esempi di spazi vettoriali normati:

Esempio 1.3. *Insieme delle funzioni continue su un intervallo $[\alpha, \beta]$, indicato con $C^0\big([\alpha, \beta]\big)$ con la norma $\|u\|_C = max_{x\in[\alpha,\beta]}\ |u(x)|$*

Esempio 1.4. *Insieme delle funzioni che elevate al quadrato sono integrabili*[2] *su un intervallo (α, β), indicato con $L^2\big(\alpha, \beta\big)$ con la norma $\|u\|_2 = \left(\int_\alpha^\beta\ |u(x)|^2\ dx\right)^{1/2}$*

Esempio 1.5. *Insieme delle funzioni che sono integrabili in modulo su un intervallo (α, β), indicato con $L^1\big(\alpha, \beta\big)$ con la norma $\|u\|_1 = \int_\alpha^\beta\ |u(x)|\ dx$*

Esempio 1.6. *Nello spazio $C^0\big([\alpha, \beta]\big)$ dell'esempio 1.3 possiamo anche considerare le norme $\|u\|_2$ o $\|u\|_2$ e come vedremo avrà proprietà differenti.*

Esempio 1.7. *Anche l'insieme $C_0^0\big([\alpha, \beta]\big)$ delle funzioni continue che sono nulle agli estremi dell'intervallo è uno spazio vettoriale ed in esso possiamo anche considerare si la norma $\|u\|_C$ che le norme $\|u\|_1$ o $\|u\|_2$. Anche le funzioni*

[2]Il concetto corretto di integrazione in cui ambientare questi discorsi e dimostrare in modo rigoroso i fatti di teoria, è quello dell'integrale secondo Lebesgue. Non è scopo di questo corso entrare negli aspetti tecnici dell'integrale di Lebesgue non essendo necessario per le nostre finalità. L'idea è quella di pensare a successioni di funzioni di classe C^1 a pezzi e di definire l'integrale del limite come limite degli integrali della successione. Dal punto di vista pratico ci interessano solo i seguenti tre fatti:

- due funzioni che differiscono in un insieme di misura nulla (per noi sarà in un numero finito o un'infinità numerabile di punti) sono indistinguibili dal punto di vista degli integrali (per la precisione si dirà che le due funzioni sono "uguali quasi ovunque", q. o.) e per i nostri scopi saranno considerate uguali. Quindi gli elementi degli spazi di funzioni integrabili sono in realtà *classi di equivalenza* di funzioni uguali q.o.
- una funzione è integrabile se e solo se lo è il suo modulo
- le funzioni che in modulo sono integrabili in senso generalizzato (secondo Riemann) sono **integrabili** secondo Lebesgue. (Quindi ad esempio $f(x) = x^\alpha$ con $-1 < \gamma < 0$ sono integrabili secondo Lebesgue in $[0, 1]$)

continue che sono nulle in uno degli estremi costituiscono uno spazio vettoriale, mentre ovviamente le funzioni che hanno tutte uno stesso valore diverso da zero non costituiscono uno spazio vettoriale.

Definizione 1.2. ***Spazio completo e Spazio di Banach*** *si veda libro a pag. 265*

Esempio 1.8. *Lo spazio $C^0\big([\alpha,\beta]\big)$ dell'esempio 1.3 con la norma $\|u\|_C$ è completo. Per la dimostrazione si veda un libro di analisi 2 che parli di convergenza uniforme.*

Esempio 1.9. *Lo spazio $C^0\big([\alpha,\beta]\big)$ con la norma $\|u\|_2$* ***non*** *è completo. Si veda libro a pag. 266 dove si mostra che la successione $\{f_n(t)\}$ a metà pagina converga in norma $\|u\|_2$ alla funzione di Heaviside*

$$H(t) = \begin{cases} 0 & -1 \le t < 0, \\ 1 & 0 \le t \le 1 \end{cases}$$

che non è continua cioè non appartiene a $C^0[-1,1]$.

Esempio 1.10. *Lo spazio $L^2\big(\alpha,\beta\big)$ con la norma $\|u\|_2$ è completo così come lo spazio $L^1\big(\alpha,\beta\big)$ con la norma $\|u\|_1$ (vedi esempi 1.4 e 1.5).*

In qquello che segue introdurremo il concetto di prodotto interno (o scalare) e vedremo che uno spazio con prodotto interno (e completo) avrà tutte le proprietà di R^N, teorema parallelogramma, Pitagora, ortogonalità, proiezioni etc..

Definizione 1.3. ***Prodotto interno, norma indotta dal prodotto interno e spazio di Hilbert*** *si veda libro a pag. 268, pag. 269 e pag. 270.*

Proprietà 1.1.

- *Disuguaglianza di Schwartz - pag. 269*
- *Legge del parallelogramma - pag. 269*

Esempio 1.11. *Lo spazio $L^2\big(\alpha,\beta\big)$ è uno spazio di Hilbert. Si veda libro a pag. 270 es. 7.3.*

Definizione 1.4. ***Ortogonalità e sottospazio chiuso*** *si veda libro a pag. 271 e 272.*

Proprietà 1.2.

- *Teorema di Pitagora - pag. 272*
- *Teorema della proiezione - pag. 273*

Definizione 1.5. ***Basi ortonormali e coefficienti di Fourier generalizzati*** *si veda libro a pag. 275.*

Definizione 1.6. ***Operatori e funzionali lineari e limitati*** *si veda libro a pag. 278 e 279.*

Teorema 1.3. ***Teorema di rappresentazione di Riesz*** *si veda libro a pag. 281.*

Definizione 1.7. ***Forme bilineari*** *si veda libro a pag. 283.*

Calcolo numerico dei coefficienti di Fourier

Sia $f : [-\pi, \pi] \to \mathbb{R}$ una funzione dispari che ammette serie di Fourier convergente

$$f(x) \simeq \sum_{k=1}^{\infty} b_k \sin(kx)\,.$$

Sappiamo che

$$b_k = \frac{2}{\pi} \int_0^{\pi} f(x) \sin(kx)\, dx \tag{1}$$

e vogliamo un algoritmo numerico efficiente per calcolare i primi N coefficienti b_k. Potremmo utilizzare una formula di quadratura per approssimare l'integrale (1). Poiché $\sin(0) = \sin(k\pi) = 0$, usando la formula del trapezio abbiamo

$$b_k = \frac{2}{\pi} \int_0^{\pi} f(x) \sin(kx)\, dx \simeq \frac{2}{N} \sum_{j=1}^{N-1} f\left(\frac{\pi j}{N}\right) \sin\left(\frac{\pi k j}{N}\right)\,. \tag{2}$$

Si osservi che in questo modo si calcolano in realtà in modo approssimato i primi $N-1$ coefficienti perché la formula (2) da sempre zero per $k = N$.

Analogamente, se $f : [-\pi, \pi] \to \mathbb{R}$ è una funzione pari che ammette serie di Fourier convergente

$$f(x) \simeq \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos(kx)$$

allora

$$a_k = \frac{2}{\pi} \int_0^{\pi} f(x) \cos(kx)\, dx \tag{3}$$

e, utilizzando la formula del punto medio,

$$a_k = \frac{2}{\pi} \int_0^{\pi} f(x) \cos(kx)\, dx \simeq \frac{2}{N} \sum_{j=1}^{N} f\left(\frac{\pi(2j-1)}{2N}\right) \cos\left(\frac{\pi k(2j-1)}{2N}\right)\,. \tag{4}$$

Per effettuare in modo efficiente i calcoli (2)-(4) è conveniente definire la trasformata discreta di Fourier.

La trasformata discreta di Fourier

Sia $N = 2M$ e consideriamo lo spazio vettoriale $\mathbb{C}^N$ con prodotto scalare definito da

$$u \cdot v = \sum_{k=0}^{N-1} u_k \overline{v}_k\,.$$

Sia $\{\mathbf{e}_j\}$, $j = 0, \ldots, N-1$, la base canonica, definita da $(\mathbf{e}_j)_k = \delta_{jk}$. Ricordando la formula di Eulero

$$e^{ix} = \cos x + i \sin x\,,$$

definiamo la base di Fourier $\{\mathbf{w}_j\}$ come segue:

$$\{\mathbf{w}_j\}_k = \frac{1}{\sqrt{N}} e^{\frac{2\pi ijk}{N}} .$$

Poiché la funzione e^{ix} è periodica di periodo 2π, è utile considerare i numeri interi modulo N, ossia identificare un intero k con $k+N$. Si verifica con un semplice calcolo che $\{\mathbf{w}_j\}$ è una base ortonormale, infatti se $j \neq l$

$$\mathbf{w}_j \cdot \mathbf{w}_l = \frac{1}{N} \sum_{k=0}^{N-1} e^{\frac{2\pi ijk}{N}} e^{-\frac{2\pi ilk}{N}} = \frac{1}{N} \sum_{k=0}^{N-1} e^{\frac{2\pi i(j-l)k}{N}} = \frac{1}{N} \frac{1 - e^{\frac{2\pi i(j-l)N}{N}}}{1 - e^{\frac{2\pi i(j-l)}{N}}} = 0 ,$$

mentre

$$\mathbf{w}_j \cdot \mathbf{w}_j = \frac{1}{N} \sum_{k=0}^{N-1} e^{\frac{2\pi ijk}{N}} e^{-\frac{2\pi ilk}{N}} = \frac{1}{N} \sum_{k=0}^{N-1} 1 = 1 .$$

Quindi ogni vettore $\mathbf{x} \in \mathbb{C}^N$ può essere rappresentato come combinazione lineare dei vettori $\{\mathbf{w}_j\}$, cioè

$$\mathbf{x} = \sum_{k=0}^{N-1} \hat{x}_k \mathbf{w}_k \qquad \text{ossia} \qquad x_j = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \hat{x}_k e^{\frac{2\pi ijk}{N}} . \tag{5}$$

Definiamo la sua *trasformata discreta di Fourier* (DFT) come il vettore $\hat{\mathbf{x}} \in \mathbb{C}^N$ le cui componenti sono i coefficienti $\hat{x}_k$ di x in (5).

Poiché $\{\mathbf{w}_j\}$ è una base ortonormale, ne segue immediatamente che

$$\hat{x}_k = (x, \mathbf{w}_k) = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} x_j e^{-\frac{2\pi ijk}{N}} . \tag{6}$$

Se $\mathbf{x} \in \mathbb{R}^N$ (ossia se le parti immaginarie di x sono nulle), allora $\hat{x}_j = \overline{\hat{x}_{N-j}}$ per ogni $j = 1, \ldots, N/2$ e in particolare $\mathrm{Im}(\hat{x}_0) = \mathrm{Im}(\hat{x}_{N/2}) = 0$. Quindi N coefficienti reali bastano per descrivere completamente $\hat{x}$.

MATLAB utilizza una definizione alternativa, ma ovviamente equivalente (si ricordi che gli indici dei vettori di MATLAB iniziano sempre da 1):

$$\hat{x}_k = \sum_{j=1}^{N} x_j e^{-\frac{2\pi i(j-1)(k-1)}{N}} \qquad \text{e} \qquad x_j = \frac{1}{N} \sum_{k=0}^{N-1} \hat{x}_k e^{\frac{2\pi i(j-1)(k-1)}{N}} .$$

La funzione usata da MATLAB è `fft`, perché per motivi di efficienza computazionale non usa direttamente la formula (6), ma un algoritmo chiamato Fast Fourier Transform.

DST. Si consideri $\mathbf{x}$ reale e dispari, nel senso che $x_j = -x_{N-j}$ per ogni j (sempre modulo N, cosa che in particolare implica che $x_0 = x_{N/2} = 0$). Allora $\mathrm{Re}(\hat{x}_j) = 0$ per ogni j, $\mathrm{Im}(\hat{x}_j) = -\mathrm{Im}(\hat{x}_{N-j})$ e $\hat{x}_0 = \hat{x}_{N/2} = 0$, quindi $\hat{x}$ è

determinato da $N/2-1$ coefficienti reali. Possiamo definire una trasformazione lineare $DST : \mathbb{R}^{N/2-1} \to \mathbb{R}^{N/2-1}$ come la funzione che trasforma le $N/2-1$ componenti che indentificano un vettore reale e dispari nelle $N/2-1$ componenti della sua DFT. La funzione usata da matlab è `dst` ed è definita da

$$\hat{x}_k = \sum_{j=1}^{M} x_j \sin\left(\frac{\pi jk}{M+1}\right) \qquad \text{e} \qquad x_j = \frac{2}{M+1}\sum_{k=1}^{M} \hat{x}_k \sin\left(\frac{\pi jk}{M+1}\right) \tag{7}$$

(dove rispetto alle formule precedenti $M = N/2-1$). La DST ha il notevole vantaggio sulla DFT di trasformare vettori reali in vettori reali, tuttavia non è efficiente calcolare la DST con la formula (7), quindi qualsiasi algoritmo per il calcolo della DST (tra cui quello adottato da MATLAB) sfrutta l'algoritmo FFT. In MATLAB la DST è calcolata con il comando `dst` e la trasformata inversa si ottiene con il comando `idst`.

DCT. Se `x` è reale e pari, ossia $x_j = x_{N-j}$, allora $\hat{x}$ è reale e quindi $\hat{x}$ è determinato da $N/2+1$ coefficienti reali. Possiamo allora definire una trasformazione lineare $DCT : \mathbb{R}^{N/2+1} \to \mathbb{R}^{N/2+1}$ come la funzione che trasforma le $N/2+1$ componenti che identificano un vettore reale e dispari nelle $N/2+1$ componenti della sua DFT. La funzione usata da matlab è `dct` ed è definita da

$$\hat{x}_1 = \sqrt{\frac{1}{M}}\sum_{j=1}^{M} x_k \tag{8}$$

e

$$\hat{x}_k = \sqrt{\frac{2}{M}}\sum_{j=1}^{M} x_j \cos\left(\frac{\pi(k-1)(2j-1)}{2M}\right) \quad k = 2,\ldots,M\,. \tag{9}$$

tali che

$$x_j = \sqrt{\frac{1}{M}}\,\hat{x}_1 + \sqrt{\frac{2}{M}}\sum_{k=2}^{M} \hat{x}_k \cos\left(\frac{\pi(k-1)(2j-1)}{2M}\right) \quad j = 1,\ldots,M\,,$$

Il calcolo dei coefficienti di Fourier con MATLAB

Dalle formule (2) e (7) segue che il vettore $\hat{x}$ dei primi N coefficienti di Fourier di una funzione dispari si ottiene calcolando la `dst` del vettore x di componenti $x_j = f\left(\frac{\pi j}{N+1}\right)$, $j = 1,\ldots,N$, e moltiplicandolo per $\dfrac{2}{N+1}$.

Analogamente dalle formule (4), (8) e (9) il vettore $\hat{x_c}$ dei primi N coefficienti di Fourier di una funzione pari si ottiene calcolando la `dct` del vettore x di componenti $x_j = f\left(\frac{\pi(2j-1)}{2N}\right)$, $j = 1,\ldots,N$ e moltiplicando la prima componente per $\dfrac{2}{\sqrt{N}}$ e le altre componenti per $\sqrt{\dfrac{2}{N}}$.

Se si vogliono calcolare i coefficienti di Fourier di una funzione qualsiasi $f(x)$, è conveniente scomporla nella somma di una funzione pari e di una dispari:

$$f(x) = p(x) + d(x) \quad \text{con} \quad p(x) = \frac{f(x) + f(-x)}{2} \quad \text{e} \quad d(x) = \frac{f(x) - f(-x)}{2}$$

per poi applicare le formule (2) e (4) rispettivamente a $d(x)$ e $p(x)$. Se $d_j = d\left(\frac{\pi j}{N+1}\right)$ e $p_j = p\left(\frac{\pi(2j-1)}{2N}\right)$, (con $j = 1, \ldots, N$), $\tilde{d} = \texttt{dst}(d)$ e $\tilde{p} = \texttt{dct}(p)$, segue che

$$a_0 = \frac{2}{\sqrt{N}}\tilde{p}_1\,, \quad a_k = \sqrt{\frac{2}{N}}\widetilde{p}_{k+1}\,, \quad b_k = \frac{2}{N+1}\widetilde{d}_k\,.$$

Esercizio

Si ricavi la matrice di rigidezza degli elementi finiti lineari per il problema:

$$\begin{cases} -au'' + bu' + cu = f \quad 0 < x < L \\ u(0) = u(L) = 0 \end{cases}$$

$$\text{Spazio}: \quad V = H_0^1([0,L]) = \left\{ v \in H^1([0,L]) \ \textit{t.c.}\ v(0) = v(L) = 0 \right\}$$

Formulazione debole:
Trovare $u \in H_0^1([0,L])$ tale che

$$\underbrace{\int_0^L au'v'\,dx + \int_0^L bu'v\,dx + \int_0^L cuv\,dx}_{\mathcal{A}(u,v)} = \underbrace{\int_0^L fv\,dx}_{F(v)}$$

Sotto opportune ipotesi vale Lax-Milgram, quindi il problema è ben posto (esistenza ed unicità della soluzione).

Metodo di Galerkin

Ambientare il problema in uno spazio $V_h \subset V$ con V_h finito dimensionale: $\dim V_h \equiv N_h < \infty$

Problema di Galerkin
Trovare $u_h \in V_h$ tale che $\mathcal{A}(u_k, v_k) = F(v_h)\ \forall v_h \in V_h$

Indicando con $\{\varphi_j : \ j = 1, 2, \ldots, N_h\}$ una base di V_h, testare che $\forall\, v_h \in V_h$ equivale a farlo per le sole funzioni di base:

$$\text{trovare } u_h \in V_h \ \textit{t.c.}\ \mathcal{A}(u_h, \varphi_i) = F(\varphi_i) \quad \text{per } i = 1, \ldots, N_h$$

Inoltre, u_h può essere scritta come

$$u_h = \sum_{j=1}^{N_h} u_j \varphi_j(x)$$

dove u_j, $j = 1, \ldots, N_h$, sono coefficienti incogniti.
Allora Galerkin diventa: trovare u_j, $j = 1, \ldots, N_h$, tali che

$$\mathcal{A}\left(\sum_{j=1}^{N_h} u_j \varphi_j(x), \varphi_i(x) \right) = F(\varphi_i(x)) \qquad \text{per } i = 1, \ldots, N_h$$

dalla linearità di $\mathcal{A}(\cdot,\cdot)$:

$$\sum_{j=1}^{N_h} u_j \mathcal{A}(\varphi_j(x), \varphi_i(x)) = F(\varphi_i(x)) \qquad \forall \varphi_i,\ i = 1, \ldots, N_h$$

Sia allora $\mathbf{u} \in \mathbb{R}^{N_h},\ \mathbf{A} \in \mathbb{R}^{N_h \times N_h},\ \mathbf{f} \in \mathbb{R}^{N_h}$

$$[\mathbf{u}]_i = u_i \qquad [\mathbf{A}]_{ij} = \mathcal{A}(\varphi_j, \varphi_i) \qquad [\mathbf{f}]_i = F(\varphi_i)$$

il problema equivale a risolvere il sistema lineare:

$$\mathbf{A}\mathbf{u} = \mathbf{f}$$

Svolgimento

Scelta spazio:

$$[0, L] = \overset{k}{\underset{i=1}{U}} [x_{i-1}, x_i] \quad x_i = hi \quad h = \frac{L}{k}$$

Funzioni lineari a tratti $: X_h^1 = \left\{ v_h \in C^0([0,L]) :\ v_h|_{(x_{i-1}, x_i)} \in \mathbb{P}^1(x_{i-1}, x_i) \right\}$

lo spazio sarà allora

$$V_h = \left\{ v_h \in X_h^1 :\ v_h(0) = v_h(1) = 0 \right\} \subset V$$

Le funzioni di base sono allora:

$$\varphi_i(x) = \begin{cases} \frac{x - x_{i-1}}{h} & x \in [x_{i-1}, x_i] \\ \frac{x_i - x}{h} + 1 & x \in [x_i, x_{i+1}] \\ 0 & \text{altrove} \end{cases} \Longrightarrow \varphi_i'(x) = \begin{cases} 1/h & x \in [x_{i-1}, x_i] \\ -1/h & x \in [x_i, x_{i+1}] \\ 0 & \text{altrove} \end{cases}$$

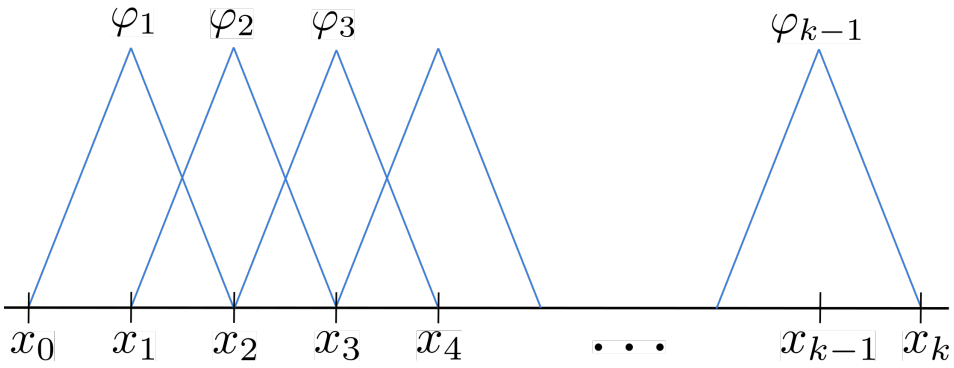


Figura 1: Illustrazione delle funzioni di base.

con $N_h = k - 1$

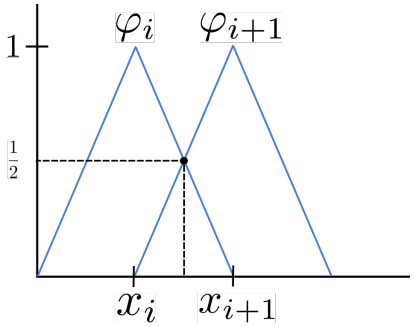
Calcoliamo ora gli elementi della matrice $\mathbf{A}$:

$$[\mathbf{A}]_{ij} = \mathcal{A}(\varphi_j, \varphi_i) = \int_0^L a\varphi_j'\varphi_i'\,dx + \int_0^L b\varphi_j'\varphi_i\,dx + \int_0^L c\varphi_j\varphi_i\,dx$$

Poichè gli integrali sono non nulli solo dove le funzioni di base si sovrappongono, la matrice $\mathbf{A}$ è una **matrice tridiagonale**.

- Elementi sovradiagonali:

$$[\mathbf{A}]_{i,i+1} = \mathcal{A}(\varphi_{i+1}, \varphi_i)$$



Utilizzeremo la regola di Simpson per integrare esattamente il terzo termine (che è un termine di secondo grado)

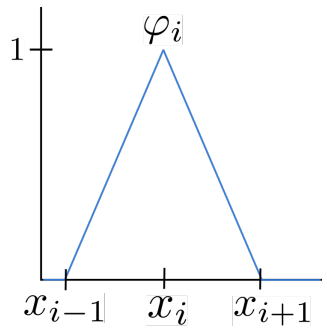
$$\text{Regola di Simpson: } \int_a^b f(x)\ dx = \frac{h}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]$$

$$[\mathbf{A}]_{i,i+1} = \mathcal{A}(\varphi_{i+1}, \varphi_i) = \int_{x_i}^{x_{i+1}} a\left(\frac{1}{h}\right)\left(-\frac{1}{h}\right)dx + \int_{x_i}^{x_{i+1}} b\left(\frac{1}{h}\right)\varphi_i dx + \int_{x_i}^{x_{i+1}} c\varphi_{i+1}\varphi_i dx =$$

$$= a\left(-\frac{1}{h^2}\right)h + \frac{b}{h}\frac{h}{2} + c\frac{h}{6}\left[0 + 4\frac{1}{2}\frac{1}{2} + 0+\right] = -\frac{a}{h} + \frac{b}{2} + \frac{ch}{6}$$

- Elementi diagonali:

$$[\mathbf{A}]_{i,i} = \mathcal{A}(\varphi_i, \varphi_i)$$



$$[\mathbf{A}]_{i,i} = \mathcal{A}(\varphi_i, \varphi_i) = \int_{x_{i-1}}^{x_{i+1}} \left(a(\varphi_i')^2 + b\varphi_i'\varphi_i + c\varphi_i\varphi_i \right) dx =$$

spezzando l'integrale:

$$= \int_{x_{i-1}}^{x_i} \left(a(\varphi_i')^2 + b\varphi_i'\varphi_i + c\varphi_i\varphi_i \right) dx + \int_{x_i}^{x_{i+1}} \left(a(\varphi_i')^2 + b\varphi_i'\varphi_i + c\varphi_i\varphi_i \right) dx =$$

$$= a\frac{1}{h}\frac{1}{h}h + a\left(-\frac{1}{h}\right)\left(-\frac{1}{h}\right)h + \frac{1}{h}\frac{h}{2} + b\left(-\frac{1}{h}\right)\frac{h}{2} + c\frac{h}{6}\left[0 + 4\frac{1}{4} + 1\right] + c\frac{h}{6}\left[1 + 4\frac{1}{4} + 0\right] =$$

$$= \frac{2a}{h} + \frac{2}{3}ch$$

- Elementi sottodiagonali:

Con calcoli analoghi a quelli svolti per gli elementi sovradiagonali si ottiene:

$$[\mathbf{A}]_{i+1,i} = -\frac{a}{h} - \frac{b}{2} + \frac{ch}{6}$$

In definitiva:

$$\mathbf{A} = \frac{a}{h}\begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & \ddots \\ & & \ddots & \ddots \end{bmatrix} + b\begin{bmatrix} 0 & 1/2 & & \\ -1/2 & \ddots & \ddots & \\ & \ddots & \ddots & \ddots \\ & & \ddots & \ddots \end{bmatrix} + ch\begin{bmatrix} 2/3 & 1/6 & & \\ 1/6 & \ddots & \ddots & \\ & \ddots & \ddots & \ddots \\ & & \ddots & \ddots \end{bmatrix}$$

come già sottolineato, $\mathbf{A}$ è tridiagonale, è inoltre simmetrica se $b = 0$.