

# **Appunti del corso di *Metodi numerici e ottimizzazione***

L<sup>A</sup>T<sub>E</sub>X Ninjas

Andrea Cimino

Marco Cornolti

Emanuel Marzini

Davide Mascitti

Lorenzo Muti

Marco Stronati

`{cimino,cornolti,marzini,mascitti,muti,stronati}@cli.di.unipi.it` \*

set 2010-set 2011



# Indice

<b>1</b>	<b>Richiami di Algebra Lineare</b>	<b>9</b>
1.1	Richiami su matrici . . . . .	9
1.1.1	Proprietà importanti sulle matrici . . . . .	9
1.1.2	Alcune classi importanti di matrici . . . . .	10
1.1.2.1	Matrici triangolari . . . . .	10
1.1.2.2	Matrici simmetriche . . . . .	10
1.1.2.3	Matrici unitarie . . . . .	12
1.1.3	Prodotto scalare . . . . .	13
1.1.3.1	Matrici normali . . . . .	14
1.1.3.2	Matrici definite positive . . . . .	14
1.1.3.3	Matrice partizionata a blocchi . . . . .	15
1.1.3.4	Matrici riducibili . . . . .	15
1.2	Autovalori ed autovettori . . . . .	16
1.3	Forme canoniche . . . . .	21
1.3.1	Forma canonica di Jordan . . . . .	21
1.3.2	Forma canonica di Schur . . . . .	24
1.4	Alcune proprietà delle matrici definite positive . . . . .	27
1.5	Localizzazione degli autovalori . . . . .	28
1.6	Predominanza diagonale . . . . .	28
1.7	Norme . . . . .	29
1.7.1	Norme matriciali . . . . .	30
<b>2</b>	<b>Richiami di Analisi e Ottimizzazione</b>	<b>33</b>
2.1	Richiami di analisi in $\mathbb{R}^n$ . . . . .	33
2.1.1	Funzioni di più variabili a valori reali . . . . .	36
2.1.2	Derivate . . . . .	38
2.1.2.1	Derivate seconde . . . . .	44
2.1.3	Funzioni di più variabili a valori vettoriali . . . . .	44
2.1.4	Derivate di ordine superiore . . . . .	44
<b>3</b>	<b>Ottimizzazione: regione ammissibile e condizioni di ottimalità</b>	<b>49</b>
3.0.5	Risultati importanti su insiemi convessi e funzione convessa . . . . .	52
3.1	Ottimizzazione non vincolata . . . . .	57
3.1.1	Condizioni di ottimalità . . . . .	57
<b>4</b>	<b>Risoluzione di Sistemi Lineari</b>	<b>63</b>
4.1	Propagazione dell'errore . . . . .	63
4.2	Fattorizzazione . . . . .	64
4.3	Metodo di Gauss . . . . .	67
4.4	Matrici Elementari di Gauss . . . . .	69
4.5	Matrici elementari di Householder . . . . .	71
4.5.1	Massimo Pivot . . . . .	74
4.5.2	Complessità . . . . .	75
4.6	Fattorizzazione di Cholesky . . . . .	75
4.6.1	Complessità . . . . .	75

4.6.2	Stabilità . . . . .	76
4.7	Complessità sui sistemi lineari . . . . .	76
4.7.1	Tabella riassuntiva . . . . .	76
4.8	Sistemi Lineari: metodi iterativi . . . . .	76
4.8.1	Convergenza . . . . .	76
4.8.2	Richiami: decomposizione additiva . . . . .	77
4.9	Particolari decomposizioni additive . . . . .	79
4.9.1	Metodi iterativi di Jordan e Gauss-Seidel . . . . .	79
4.9.2	Metodo di Jacobi . . . . .	79
4.9.3	Metodo di Gauss-Seidel . . . . .	79
4.9.4	Condizioni sufficienti per convergenza Jacobi e Gauss-Seidel . . . . .	80
4.10	Matrici tridiagonali . . . . .	83
<b>5</b>	<b>Metodi di risoluzione per sistemi non lineari</b>	<b>85</b>
5.1	Generalità sui metodi iterativi per sistemi non lineari . . . . .	85
5.2	Metodo di Newton-Raphson (delle tangenti) . . . . .	89
5.2.1	Newton-Raphson su funzioni convesse . . . . .	90
<b>6</b>	<b>Ottimizzazione non vincolata</b>	<b>95</b>
6.1	Metodi risolutivi per i problemi di ottimizzazione non vincolata . . . . .	97
6.1.1	Ricerca monodimensionale . . . . .	99
6.1.2	La massima decrescita . . . . .	100
6.2	Metodi del gradiente . . . . .	101
6.2.1	Proprietà del Metodo del Gradiente Esatto . . . . .	101
6.2.2	Critica al metodo del gradiente esatto . . . . .	104
6.2.3	Metodi del gradiente con ricerca inesatta . . . . .	108
6.2.4	Metodo di Netwon-Raphson . . . . .	113
<b>7</b>	<b>Metodo del gradiente coniugato</b>	<b>119</b>
7.1	Motivazioni e strumenti . . . . .	119
7.2	Definizione dei parametri . . . . .	120
7.2.1	Passo di discesa . . . . .	120
7.2.2	Aggiornamento del residuo . . . . .	121
7.2.3	Scelta della direzione di discesa . . . . .	121
7.3	Definizione algoritmica del metodo . . . . .	123
7.4	Convergenza del metodo . . . . .	123
7.4.1	Direzioni A-coniugate e residui ortogonali . . . . .	123
7.4.2	Misura dell'errore . . . . .	124
7.5	Metodo del gradiente coniugato preconditionato . . . . .	125
<b>8</b>	<b>Metodo del Gradiente coniugato non lineare</b>	<b>127</b>
8.1	Considerazioni preliminari . . . . .	127
8.2	Approssimazione del passo . . . . .	127
8.3	Convergenza del metodo . . . . .	128
8.4	Variazioni del metodo . . . . .	130
<b>9</b>	<b>Metodi per l'ottimizzazione non vincolata senza derivate</b>	<b>131</b>
9.1	Metodo di ricerca diretta a compasso . . . . .	131
9.1.1	Descrizione dell'algoritmo . . . . .	132
9.1.2	Teorema di convergenza . . . . .	132
<b>10</b>	<b>Il problema lineare dei minimi quadrati</b>	<b>135</b>
10.1	Metodo delle equazioni normali . . . . .	135

10.2	Metodo QR . . . . .	137
10.2.1	Casi e Costi . . . . .	139
10.3	Norme di matrici non quadrate . . . . .	140
10.4	SVD: Decomposizione ai valori singolari di una matrice . . . . .	140
10.4.1	Calcolo dei valori e vettori singolari . . . . .	144
10.5	Risoluzione del problema dei minimi quadrati con i valori singolari . . . . .	145
10.6	Pseudoinversa di Moore-Penrose . . . . .	146
10.7	Calcolo della soluzione di minima norma con il metodo del gradiente coniugato . . . . .	147
10.7.1	Tabella riassuntiva . . . . .	148
10.8	SVD troncata . . . . .	148
<b>11</b>	<b>Il problema nonlineare dei minimi quadrati</b>	<b>151</b>
11.1	Minimi quadrati lineari . . . . .	151
11.2	Problema dei minimi quadrati non lineare . . . . .	152
11.3	Metodo di Gauss-Newton . . . . .	153
11.4	Regressioni (non) lineari - data/curve fitting . . . . .	156
<b>12</b>	<b>Ottimizzazione vincolata con regione ammissibile convessa</b>	<b>159</b>
12.1	Condizioni di ottimalità con regione ammissibile convessa . . . . .	160
12.2	Metodi risolutivi . . . . .	164
12.2.1	Metodi delle direzioni ammissibili . . . . .	164
12.2.2	Algoritmo di Frank-Wolfe (Gradiente condizionato) . . . . .	167
12.2.3	Metodi del gradiente proiettato . . . . .	168
12.2.4	Metodo del gradiente proiettato . . . . .	173
<b>13</b>	<b>Ottimizzazione vincolata con regione ammissibile non necessariamente convessa</b>	<b>175</b>
13.1	Condizioni di ottimalità . . . . .	175
13.1.1	Eliminazione della condizione di convessità . . . . .	175
13.1.1.1	Caso con soli vincoli di disuguaglianza . . . . .	177
13.1.1.2	Caso con introduzione di vincoli di uguaglianza . . . . .	183
13.2	Metodi per la risoluzione . . . . .	185
13.2.1	Alcuni approcci alla risoluzione del problema . . . . .	185
13.2.2	Excursus di metodi sull'ottimizzazione vincolata . . . . .	187
13.2.2.1	Trasformazione di problemi in forma non vincolata: penalizzazione esterna	187
13.2.2.2	Metodo dei moltiplicatori . . . . .	190
13.2.2.3	Penalizzazione interna: metodi barriera . . . . .	193
<b>14</b>	<b>Calcolo di autovalori e autovettori</b>	<b>201</b>
14.1	Condizionamento del problema . . . . .	201
14.2	Metodo delle potenze . . . . .	203
14.2.1	Underflow/Overflow e normalizzazione . . . . .	205
14.2.2	Rilassamento delle ipotesi dell'algoritmo . . . . .	206
14.2.3	Varianti del metodo delle potenze . . . . .	207
14.2.3.1	Variante di Wielandt (metodo delle potenze inverse) . . . . .	207
14.2.3.2	Variante della deflazione . . . . .	208
14.3	Riduzione di una matrice hermitiana in forma tridiagonale: il metodo di Householder . . . . .	208
14.3.1	Trasformazioni . . . . .	208
14.4	Il metodo di Householder . . . . .	209
14.5	Calcolo degli autovalori delle matrici tridiagonali hermitiane con la successione di Sturm . . . . .	211
14.6	Metodo QR . . . . .	213
14.6.1	Metodo QR con trasformazioni di Householder . . . . .	214
14.6.2	Convergenza . . . . .	214
14.6.3	Indebolimento delle condizioni del teorema . . . . .	218

---

14.6.4	Tecnica di translazione: QR con shift . . . . .	218
14.6.5	Matrice di Sylvester . . . . .	218
<b>15</b>	<b>DFT: Trasformata discreta di Fourier</b>	<b>219</b>
<b>16</b>	<b>Richiami da Wikipedia</b>	<b>225</b>
16.0.6	Teorema di Lagrange (o valor medio) . . . . .	225
16.0.7	Disuguaglianza triangolare . . . . .	225

# Prologo

Queste dispense del corso di Metodi Numerici ed Ottimizzazione sono state elaborate durante il corso dell'A.A. 2010/2011, tenuto dal prof. Giancarlo Bigi e dal prof. Roberto Bevilacqua, nel primo anno in cui il corso è stato attivato nel Corso di Laurea Magistrale in Informatica dell'Università di Pisa.

*Sono state preparate da alcuni studenti soprattutto per avere una base per studiare il sostanzioso programma, quindi mancano di organicità e di un linguaggio uniforme. Inoltre nella versione attuale vi sono probabilmente molti errori.* La speranza è che i professori (con l'aiuto degli studenti dei prossimi anni) possano farlo diventare, se non un libro di testo, almeno una dispensa ben fatta.

Il lavoro principale è stato quello di prendere gli appunti delle lezioni in  $\LaTeX$ , fatto da Andrea Cimino. Gli appunti sono poi stati sistemati, rielaborati ed integrati da Andrea Cimino, Marco Cornolti, Emmanuel Marzini, Davide Mascitti, Lorenzo Muti, Marco Stronati.

Spesso abbiamo consultato il materiale on-line offerto dai prof. Bevilacqua (parte sui metodi numerici) e Bigi (parte sull'ottimizzazione).

Si ringraziano gli sviluppatori degli strumenti open-source che abbiamo utilizzato per scrivere questa dispensa, in particolare:  $\LaTeX$ , Texmaker, Inkscape, Subversion.

Il lavoro – sia nella versione compilata che nei sorgenti  $\LaTeX$  – è pubblicato sotto licenza Creative Commons Attribuzione–Non commerciale–Condividi allo stesso modo 3.0 Italia (CC BY-NC-SA 3.0).

Chiunque è libero di riprodurre, distribuire, comunicare al pubblico, esporre in pubblico, rappresentare, eseguire e recitare quest'opera.

Chiunque è libero di modificare quest'opera e ridistribuirla, sotto le seguenti condizioni:

**Attribuzione** l'opera derivata deve citare gli autori originali.

**Non Commerciale** non si può usare quest'opera né un'opera derivata per fini commerciali.

**Condividi allo stesso modo** l'opera derivata deve essere rilasciata sotto una licenza uguale o compatibile con questa.

Maggiori informazioni su <http://creativecommons.org/licenses/by-nc-sa/3.0/it/deed.it>

Pisa, 28 settembre 2011





# 1 Richiami di Algebra Lineare

## 1.1 Richiami su matrici

Le matrici alle quali ci riferiremo saranno definite su campo dei complessi  $\mathbb{C}$ . Lavoreremo quindi con dei vettori su  $\mathbb{C}$ :  $v \in \mathbb{C}^h$ ,  $A \in \mathbb{C}^{m \times n}$ .

### 1.1.1 Proprietà importanti sulle matrici

#### **Definizione 1.1 (Sottoinsieme chiuso rispetto alla moltiplicazione)**

Un sottoinsieme di  $\mathbb{C}^{n \times n}$  si dice chiuso rispetto all'operazione di moltiplicazione, se date due matrici  $A$  e  $B$  appartenenti al sottoinsieme, anche il prodotto  $AB$  appartiene al sottoinsieme.

I seguenti sottoinsiemi di  $\mathbb{C}^{n \times n}$  sono chiusi rispetto all'operazione di moltiplicazione:

- matrici triangolari superiori (inferiori),
- matrici triangolari superiori (inferiori) in senso stretto
- matrici unitarie

La moltiplicazione fra matrici gode della proprietà associativa, di quella distributiva rispetto all'addizione, ma non di quella commutativa.

#### **Proprietà 1.1 (Proprietà sul prodotto di matrici)**

- $A + 0 = 0 + A = A$  (la matrice nulla è l'elemento neutro della somma)
- $A + (-A) = 0$  (esistenza di un elemento opposto per la somma)
- $(A + B) + C = A + (B + C)$  (proprietà associativa della somma)
- $A + B = B + A$  (proprietà commutativa della somma)
- $(AB)C = A(BC)$  (proprietà associativa del prodotto)
- $(A + B)C = AC + BC$  (proprietà distributiva)
- $C(A + B) = CA + CB$  (proprietà distributiva)

#### **Proprietà 1.2 (Sull'inversione di matrici)**

Vale la seguente proprietà:

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

Vale inoltre la seguente proprietà

 **Proprietà 1.3 (Reverse order law)**

$$(AB)^H = B^H A^H \quad (1.1)$$

dove  $H$  è l'operatore di trasposizione coniugata.

**Domanda aperta**

Quali condizioni sono necessarie affinché valga la reverse order law? Una dimostrazione di tale proprietà: si ponga

$$M = AB$$

Allora valgono le seguenti implicazioni

$$A^{-1}M = A^{-1}AB \Rightarrow B^{-1}A^{-1}M = B^{-1}B \Rightarrow B^{-1}A^{-1}M = I$$

Ma allora  $M$  deve essere l'inversa di  $B^{-1}A^{-1}$ , ma  $M = AB$ .

Il punto di questa dimostrazione che non mi è chiaro è: quali condizione (ad esempio sull'invertibilità di  $A$  e  $B$ ) sono necessarie?

Ciuffo: Semplicemente che  $A$  e  $B$  siano invertibili, altrimenti l'operatore di inversa non dà risultati.

## 1.1.2 Alcune classi importanti di matrici

### 1.1.2.1 Matrici triangolari

**Definizione 1.2 (Matrice triangolare)**

Le matrici triangolari inferiori (superiori) sono matrici quadrate che hanno tutti gli elementi al di sopra (sotto) della diagonale principale nulli.

Il determinante di una matrice triangolare  $A$  può essere calcolato come il prodotto dei suoi elementi principali.

$$\det(A) = \prod_{i=1}^n a_{ii}$$

**Definizione 1.3 (Matrice Hessemberg)**

Una matrice di Hessemberg è una matrice "quasi" triangolare. In particolare è detta superiore se ha valori pari a zero sotto la prima sottodiagonale, viceversa nel caso inferiore.

**Esempio 1.4 (Hessemberg superiore e inferiore)**

$$\begin{pmatrix} 1 & 4 & 2 & 3 \\ 3 & 4 & 1 & 7 \\ 0 & 2 & 3 & 4 \\ 0 & 0 & 1 & 3 \end{pmatrix} \quad \begin{pmatrix} 1 & 2 & 0 & 0 \\ 5 & 2 & 3 & 0 \\ 3 & 4 & 3 & 7 \\ 5 & 6 & 1 & 1 \end{pmatrix}$$

### 1.1.2.2 Matrici simmetriche

**Definizione 1.5 (Matrice simmetrica)**

Una matrice simmetrica è una matrice quadrata che ha la proprietà di essere uguale alla sua trasposta. Quindi  $A$  è simmetrica sse:

$$A = A^T \quad \text{cioè} \quad a_{ij} = a_{ji}, \quad \forall i, \forall j. \quad a_{ij} \in A$$

La classe delle matrici simmetriche è un sottoinsieme delle matrici Hermitiane.

**Definizione 1.6 (Matrice trasposta coniugata)**

Data una matrice  $A \in C^{m \times n}$ , si definisce matrice trasposta coniugata di  $A$  la matrice  $B \in C^{n \times m}$  tale che

$$b_{ij} = \overline{a_{ji}}$$

dove  $\overline{a_{ji}}$  è il coniugato del numero complesso  $a_{ji}$ , e si indica

$$B = A^H$$

**Definizione 1.7 (Matrice Hermitiana)**

Una matrice Hermitiana è una matrice a valori complessi che coincide con la propria trasposta coniugata (o matrice aggiunta).

$$A = A^H$$

Proprietà importanti di tali matrici:

- sulla diagonale principale *devono* essere presenti solamente numeri reali.
- gli autovalori sono reali.
- se  $A$  è reale hermitiana, allora risulta  $A^T = A$  ed è detta *simmetrica*.

**Proprietà 1.4**

Se  $A \in C_{n \times n}$  è una matrice hermitiana, cioè  $A = A^H$ , e  $\mathbf{x} \in C_n$ , il numero

$$\alpha = \mathbf{x}^H A \mathbf{x}$$

è reale. Infatti, poiché  $A$  è hermitiana, si ha:

$$\overline{\alpha} = \overline{\mathbf{x}^H A \mathbf{x}} = (\mathbf{x}^H A \mathbf{x})^H = ((\mathbf{x}^H A) \mathbf{x})^H \stackrel{1.1)}{=} \mathbf{x}^H (\mathbf{x}^H A)^H = \mathbf{x}^H (A^H \mathbf{x}) = \mathbf{x}^H A^H \mathbf{x} = \mathbf{x}^H A \mathbf{x} = \alpha.$$

**Esempio 1.8 (Matrice Hermitiana)**

$$\begin{pmatrix} 2 & 3+i \\ 3-i & 4 \end{pmatrix}$$

### 1.1.2.3 Matrici unitarie

#### Definizione 1.9 (Matrice unitaria)

Una matrice  $U$  è detta unitaria se soddisfa la condizione:

$$U^H U = U U^H = I$$

dove  $I$  è la matrice identità.

#### Proprietà

- Dal punto di vista della complessità ottenere l'inversa di una matrice, che in generale ha costo cubico in  $n$ , per le unitarie è immediato, dato che  $U^{-1} = U^H$ .
- Le matrici unitarie hanno autovalori di modulo 1.
- $U^{-1}$  è ancora una matrice unitaria.
- Le matrici unitarie sono delle isometrie, o rispettano la norma 2, ossia

$$\|Ux\|_2 = \|x\|_2$$

Infatti  $\|Ux\|_2 = \sqrt{x^H \underbrace{U^H U}_I x} = \|x\|_2$

- Le matrici unitarie danno garanzia di stabilità in tutti i calcoli in cui intervengono.
- Se  $A$  è reale unitaria, allora  $A^T A = A A^T = I$  ed è detta *ortogonale*.

Inoltre le matrici unitarie su  $\mathbb{R}^n$  sono tali che le loro colonne formano una base ortonormale di  $\mathbb{R}^n$ , cioè per ogni coppia di vettori della base il loro prodotto scalare è zero.

#### Esempio 1.10

Ad esempio:

$$\begin{bmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{bmatrix}$$

è una base ortonormale di  $\mathbb{R}^2$ .

#### Domanda aperta

Queste considerazioni valgono anche per  $\mathbb{R}$ ?

Ciuffo: certo che valgono anche per  $\mathbb{R}$ : l'unica matrice unitaria per  $\mathbb{R}$  è  $[1]$ , quindi la proprietà è banalmente dimostrata dato che non ci sono le coppie di colonne diverse della matrice per cui dovrebbe valere la proprietà.

#### Matrici unitarie importanti

**Matrici di rotazione piana (2x2)** Ruota un vettore di un angolo  $\theta$

$$G = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$$

**Matrice di rotazione piana (n x n)**

$$G = \begin{bmatrix} 1 & 0 & 0 & \dots & \dots & \dots & 0 & 0 & 0 \\ 0 & 1 & 0 & \dots & \dots & \dots & 0 & 0 & 0 \\ 0 & 0 & \dots & \dots & \dots & \dots & \dots & 0 & 0 \\ 0 & 0 & \dots & \cos(\theta) & \dots & -\sin(\theta) & \dots & 0 & 0 \\ 0 & 0 & \dots & \dots & 0 & \dots & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & \dots & 0 & \dots & \dots & 0 & 0 \\ 0 & 0 & \dots & \sin(\theta) & \dots & -\cos(\theta) & \dots & 0 & 0 \\ 0 & 0 & \dots & \dots & \dots & \dots & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & \dots & \dots & 0 & 1 & 0 \\ 0 & 0 & 0 & \dots & \dots & \dots & 0 & 0 & 1 \end{bmatrix}$$

**Matrice di permutazione** Una matrice di permutazione  $P_\pi$ , per la permutazione  $\pi$  si ottiene da  $I$  permutandone le righe.

**Esempio 1.11 (Matrice di permutazione)**

$$P_\pi = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Dato che le matrici di permutazione sono matrici ortogonali, cioè  $P_\pi P_\pi^T = I$ :

- sono matrici unitarie
- l'inversa esiste e si scrive  $P_\pi^{-1} = P_{\pi^{-1}} = P_\pi^T$

Nel caso di un vettore colonna  $x$ ,  $P_\pi x$  ne permuta le righe.

**1.1.3 Prodotto scalare**

su  $\mathbb{R}^n$

$$\langle u, v \rangle = \sum_{i=1}^n u_i v_i = u^T v$$

Si può inoltre definire la norma euclidea di un vettore tramite il prodotto scalare in  $\mathbb{R}^n$ :

$$\|u\| = \sqrt{\langle u, u \rangle} = \sqrt{\sum_{i=1}^n u_i^2}$$

su  $\mathbb{C}^n$

$$\langle u, v \rangle = \sum_{i=1}^n \bar{u}_i v_i = u^H v$$

Inoltre  $u^H u = \sum_{i=1}^n \bar{u}_i u_i = \sum_{i=1}^n |u_i|^2$  e quindi anche in questo caso basta fare la radice quadrata per ottenere la norma euclidea.

 **Proprietà 1.5**

Il prodotto scalare  $f(u, v) = \langle u, v \rangle$  su  $\mathbb{C}^n$  gode delle seguenti proprietà:

- $f(au + bv, w) = af(u, w) + bf(v, w)$
- $f(u, av + bw) = af(u, w)$
- $f(u, v) = \overline{f(v, u)}$
- $f(u, u) \geq 0$  e  $f(u, u) = 0 \iff u = 0$

Nota: il prodotto di un complesso per il suo coniugato ne dà il modulo al quadrato:

$$(Re(\lambda) + i Im(\lambda)) (Re(\lambda) - i Im(\lambda)) = Re^2(\lambda) + Im^2(\lambda)$$

### 1.1.3.1 Matrici normali

**Definizione 1.12 (Matrice normale)**

Una matrice quadrata  $A$  è normale se

$$A^H A = A A^H$$

Questa classe contiene anche le hermitiane e le unitarie, infatti:

- se una matrice è hermitiana allora è normale
- se una matrice è unitaria allora è normale

 **Proprietà 1.6**

Una matrice normale può essere diagonalizzata per mezzo di matrici unitarie.

### 1.1.3.2 Matrici definite positive

**Definizione 1.13 (Matrice definita positiva)**

Sia  $A \in \mathbb{C}_{n \times n}$  una matrice hermitiana e sia  $x \in \mathbb{C}_n, x \neq 0$ . Allora se il numero reale  $x^H A x > 0$  si dice che la matrice  $A$  è definita positiva. Analogamente:

- se  $x^H A x \geq 0$ ,  $A$  è semidefinita positiva
- se  $x^H A x \leq 0$ ,  $A$  è semidefinita negativa
- se  $x^H A x < 0$ ,  $A$  è definita negativa

Sono una sottoclasse delle hermitiane.

 **Proprietà 1.7**

Una matrice  $A^H A$  è semi-definita positiva.

Infatti  $x^H A^H A x = (Ax)^H A x = \|Ax\|^2 \geq 0$  Se  $A$  rango massimo allora  $A^H A$  è definita positiva.

### 1.1.3.3 Matrice partizionata a blocchi

A volte è possibile dividere una matrice in sottomatrici o *blocchi*, e questo può agevolare delle operazioni come il prodotto, che può essere ridefinito in termini del prodotto dei blocchi.

#### Esempio 1.14 (4 blocchi)

$$A = \left[ \begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array} \right] \quad B = \left[ \begin{array}{c|c} B_{11} & B_{12} \\ \hline B_{21} & B_{22} \end{array} \right]$$

Prodotto:

$$AB = \left[ \begin{array}{c|c} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ \hline A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{array} \right]$$

Attenzione! I blocchi non sono commutativi!

### 1.1.3.4 Matrici riducibili

#### Definizione 1.15 (Matrice riducibile)

Una matrice  $A$  si dice *riducibile* se esiste una matrice di permutazione  $\Pi$  tale che

$$B = \Pi \cdot A \cdot \Pi^T = \left[ \begin{array}{c|c} B_{11} & B_{12} \\ \hline O & B_{22} \end{array} \right]$$

cioè con i blocchi diagonali  $B_{11}$  e  $B_{22}$  quadrati e  $B_{21} = 0$ .

Questa proprietà è legata alla presenza di elementi nulli nella matrice di partenza  $A$ . Pretendere che gli zeri siano sotto o sopra è arbitrario.

#### Matrici riducibili nella risoluzione di sistemi lineari

Supponiamo  $A$  riducibile: vogliamo risolvere il sistema di equazioni

$$Ax = b$$

Allora

$$\begin{aligned} \Pi Ax &= \Pi b && \text{applichiamo permutazione} \\ \underbrace{\Pi A \Pi^T}_B \Pi x &= \Pi b && \Pi \text{ è ortogonale, cioè } \Pi \Pi^T = I \\ B \Pi x &= \Pi b && \text{ponendo } \Pi x = y \quad \Pi b = c \\ B y &= c \end{aligned}$$

Riscrivendo in forma matriciale, enfatizzando il fatto che si lavora con matrici a blocchi:

$$\left[ \begin{array}{c|c} B_{11} & B_{12} \\ \hline 0 & B_{22} \end{array} \right] \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$$

abbiamo triangolarizzato a blocchi la matrice senza fare calcoli.

Ora possiamo risolvere il seguente sistema lineare:

$$\begin{cases} B_{11}y_1 + B_{12}y_2 = c_1 \\ B_{22}y_2 = c_2 \end{cases}$$

Per riavere la soluzione originale basta riapplicare la permutazione  $x = \Pi^T y$  (costo zero).

Il costo totale di risolvere il nuovo sistema è  $\frac{n^3}{12}$  contro il costo del normale metodo di Gauss  $\frac{n^3}{3}$ .

**Teorema 1.16**

Una matrice è riducibile sse il grafo associato non è fortemente connesso.

Un grafo è fortemente connesso se da ogni nodo è possibile arrivare ad ogni altro nodo.

**1.2 Autovalori ed autovettori****Definizione 1.17 (Autovalore/Autovettore)**

Siano  $A \in \mathbb{C}^{n \times n}$ ,  $x \in \mathbb{C}^n$ ,  $x \neq 0$  e  $\lambda \in \mathbb{C}$ . Se vale

$$Ax = \lambda x$$

allora  $x$  e  $\lambda$  sono detti rispettivamente autovettore e autovalore di  $A$ .

**Definizione 1.18 (Spettro di A)**

Lo spettro di una matrice  $A$  è costituito dall'insieme degli autovalori di  $A$

**Definizione 1.19 (Raggio spettrale di A)**

Siano  $\lambda_1, \dots, \lambda_n \in \mathbb{C}$  gli autovalori di  $A$ .

Il raggio spettrale di  $A$ , denotato con  $\rho(A)$ , è definito come

$$\rho(A) = \max_{1, \dots, n} |\lambda_i|$$

**Definizione 1.20 (Traccia di A)**

Sia  $A \in \mathbb{C}^{n \times n}$ . Si definisce traccia di  $A$  la quantità

$$\text{tr}(A) = \sum_{i=1}^n a_{ii}$$

Valgono inoltre le seguenti proprietà

**Proprietà 1.8**

$$\text{tr}A = \sum_{i=1}^n \lambda_i \quad \det A = \prod_{i=1}^n \lambda_i$$

Gli autovalori sono le soluzioni dell'equazione caratteristica  $p(\lambda) = 0$ , dove

$$p(\lambda) = \det(A - \lambda I) = a_0 \lambda^n + a_1 \lambda^{n-1} + \dots + a_n \quad (1.2)$$

In particolare vale la proprietà



 **Proprietà 1.9**

$$a_1 = -(-1)^{n-1} \text{tr}A \quad a_n = \det A$$

Quindi il tutto può essere riscritto come

$$p(\lambda) = (-1)^n \lambda^n + (-1)^{n-1} \text{tr}(A) \lambda^{n-1} + \dots + \det A$$

è il polinomio caratteristico di grado  $n$ .

Il problema degli autovalori è un problema non lineare e il grado del polinomio cresce col crescere della dimensione della matrice.

Enunciamo inoltre un importante teorema, omettendone la dimostrazione. Ci sarà utile in alcune dimostrazioni

 **Teorema 1.21**

Una matrice  $A$  di ordine  $n$  è diagonalizzabile se e solo se ha  $n$  autovettori linearmente indipendenti. Inoltre le colonne della matrice  $S$ , per cui  $S^{-1}AS$  è diagonale, sono gli autovettori di  $A$

**Definizione 1.22 (Sviluppo Laplace)**

Lo sviluppo di Laplace è un metodo di calcolo del determinante, che risulta efficiente per matrici non troppo grandi e sparse. Si procede scegliendo una riga, la  $i$ -esima, tramite la formula:

$$\det(A) = \sum_{j=1}^n (-1)^{i+j} a_{i,j} \det(A_{ij})$$

con  $A_{ij}$  la matrice ottenuta da  $A$  eliminando la riga  $i$ -esima e la colonna  $j$ -esima. Esiste uno sviluppo analogo anche lungo la  $j$ -esima colonna.

**Matrici Hermitiane**
 **Proprietà 1.10**

Se  $A$  Hermitiana allora gli autovalori sono reali ( $\lambda \in \mathbb{R}$ ).

*Dimostrazione.* Usando la definizione di autovalore

$$Ax = \lambda x \quad \text{con } x \neq 0$$

Calcolo la trasposta coniugata in entrambi i membri

$$x^H A^H = x^H \lambda^H = x^H \bar{\lambda} = \bar{\lambda} x^H$$

Moltiplico entrambi i membri per  $x$  (diverso da 0 per ipotesi)

$$x^H A^H x = \bar{\lambda} x^H x \quad (1)$$

Prendo la definizione di autovalore e moltiplico entrambi i membri per  $x^H$

$$x^H Ax = x^H \lambda x = \lambda x^H x \quad (2)$$

Dato che  $A$  è hermitiana (1) e (2) sono uguali, da cui

$$\lambda x^H x = \bar{\lambda} x^H x$$

Dato che  $x^H x \neq 0$  ( $\sum |x_i|^2 \neq 0$ ) allora posso semplificare, ottenendo

$$\lambda = \bar{\lambda} \Rightarrow \lambda \in \mathbb{R}$$

□

## Matrici Unitarie

Verifichiamo gli autovalori per le matrici unitarie:

### Proprietà 1.11

Sia  $U$  unitaria allora i suoi autovalori hanno modulo 1.

$$|\lambda| = 1$$

*Dimostrazione.* Sia  $\lambda$  un autovalore. Poiché per le matrici unitarie vale la proprietà

$$|Ux| = |x|$$

inoltre vale per definizione

$$U^H U = I$$

Moltiplicando ciascun membro di  $Ux = \lambda x$  per se stesso trasposto coniugato, usando la reverse order law (1.1), otteniamo

$$x^H \underbrace{U^H U}_I x = x^H x \bar{\lambda} \lambda \Rightarrow \underbrace{x^H x}_{\text{scalare}} = |\lambda|^2 \underbrace{x^H x}_{\text{scalare}} \Rightarrow 1 = |\lambda|^2$$

□

## Autovalori per matrici definite positive

### Proprietà 1.12

$A$  è definita positiva se e solo se i suoi autovalori sono reali positivi.

*Dimostrazione.* Al solito prendiamo la definizione di autovalore

$$Ax = \lambda x$$

Moltiplicando i due membri per  $x^H$  a sinistra

$$x^H Ax = x^H \lambda x$$

che è equivalente a

$$x^H Ax = \lambda x^H x$$

Utilizzando per ipotesi

- $x^H x > 0$  per una delle (1.5)
- $x^H Ax > 0$  essendo  $A$  definita positiva

possiamo scrivere

$$\underbrace{x^H Ax}_{>0} = \lambda \underbrace{x^H x}_{>0}$$

Ma allora  $\lambda$  deve essere necessariamente positivo  $\square$

Quindi definita positiva implica tutti gli autovalori positivi e vale anche il viceversa anche se dobbiamo dimostrarlo.



### Teorema 1.23

?? Una matrice hermitiana  $A$  è definita positiva se e solo se i determinanti di tutte le sottomatrici principali di testa di  $A$  (e quindi anche il determinante di  $A$ ) sono positivi.

### Definizione 1.24

Sottomatrici principali sono quelle matrici create scegliendo colonne e righe con gli stessi indici.

### Esempio 1.25

Prendendo gli indici 1 e 3 da una matrice  $4 \times 4$  otteniamo una sottomatrice principale  $2 \times 2$  con la diagonale sovrapposta a quella della matrice di origine.



### Teorema 1.26 (Teorema di Cayley - Hamilton)

Sia  $p$  il polinomio caratteristico di  $A$  allora

$$p(A) = 0$$

*Dimostrazione.* (Rimandiamo la dimostrazione, quando si parlerà di diagonalizzazione)  $\square$

Vediamo per  $n = 2$

Il polinomio caratteristico è dato da

$$p(\lambda) = (-1)^2 \lambda^2 - \text{tr}(A) \cdot \lambda + \det A$$

Allora il teorema afferma che per ogni matrice  $A$  vale

$$p(A) = A^2 - \text{tr}(A) \cdot A + \det A \cdot I = 0$$

Non è ovvio che quello sia 0, è quindi un teorema importante.

Vediamo nel caso  $A \quad m \times n$

$$0 = a_0 A^n + a_1 A^{n-1} \dots a_n I = p(A) \tag{1.3}$$

Se  $A$  è invertibile, moltiplicando per  $A^{-1}$  otteniamo

$$0 = a_0 A^{n-1} + a_1 A^{n-2} \dots \underline{a_n} A^{-1}$$

$a_n \neq 0$  perchè sappiamo che è il determinante, quindi

$$A^{-1} = \frac{1}{a_n} \cdot (\text{polinomio in } A \text{ di grado } n-1)$$

**Domanda aperta**

Come è stata fatta questo accenno alla matrice inversa? Cosa si voleva fare vedere? Si voleva fare vedere per caso che  $1/a_n$  è un autovalore di  $A^{-1}$

Non serve considerare polinomi in una matrice di grado troppo elevato (maggiore o uguale a  $n$ )  
Sia  $S$  un polinomio di grado  $m \geq n$  dove  $n$  è l'ordine della matrice. Allora

$$S(x) = \underbrace{p(x)}_m \underbrace{q(x)}_n + \underbrace{r(x)}_{m-n} \quad \text{divisione fra polinomi}$$

$$S(A) = \underbrace{p(A)q(A)}_{=0 \text{ (1.26)}} + r(A) = 0$$

Quindi abbiamo un polinomio di grado  $< n$

C'è un polinomio (polinomio minimo): È possibile che esistano altri polinomi  $S(\lambda)$  con  $S(A) = 0$  e grado minore di  $n$

**Definizione 1.27 (Polinomio minimo)**

Si definisce polinomio minimo  $\psi(\lambda)$  il polinomio monico (ossia con primo coefficiente uguale a 1) di grado minimo tale che  $\psi(A) = 0$  (matrice nulla)

From Wikipedia

Data una matrice quadrata  $A$  a valori in un certo campo  $K$ , si considera l'insieme

$$I = \{p \in K[x] \mid p(A) = 0\}$$

di tutti i polinomi che annullano  $A$ . Questo insieme risulta essere un ideale nell'anello  $K[x]$  di tutti i polinomi con coefficienti in  $K$ .

L'anello  $K[x]$  è un anello euclideo: è infatti possibile fare una divisione fra polinomi con resto. Conseguentemente, è un anello ad ideali principali: ogni ideale è generato da un unico elemento. In particolare,  $I = (m(x))$  è generato da un elemento  $m(x)$ . Tale elemento è unico solo a meno di moltiplicazione per una costante non nulla: è quindi unico se lo si suppone "monico" (cioè con coefficiente 1 nel termine  $x^k$  più grande). Si definisce quindi il "polinomio minimo" di  $A$  tale polinomio  $m(x)$ .

**Esempio 1.28**

$$A = I_n$$

$$P(\lambda) = (1 - \lambda)^n \quad \det(I - \lambda I) = \begin{bmatrix} 1 - \lambda & 0 \\ 0 & 1 - \lambda \end{bmatrix}$$

$$\psi(\lambda) = -(1 - \lambda) = \lambda - 1$$

$$\psi(I) = I - I = 0$$

**Esempio 1.29**

From Wikipedia

Consideriamo per esempio la matrice

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

Il suo polinomio caratteristico è dato da

$$p(\lambda) = \det \begin{bmatrix} 1 - \lambda & 2 \\ 3 & 4 - \lambda \end{bmatrix} = (1 - \lambda)(4 - \lambda) - 2 \cdot 3 = \lambda^2 - 5\lambda - 2.$$

Il teorema di Cayley–Hamilton sostiene che:

$$A^2 - 5A - 2I_2 = 0$$

il che si può facilmente verificare.

## 1.3 Forme canoniche

### Definizione 1.30 (Matrici simili)

Due matrici quadrate  $A$  e  $B$  si dicono simili se esiste una matrice  $M$  invertibile per cui vale

$$A = M^{-1}BM$$

è facile verificare quindi che la similitudine fra matrici è una relazione di equivalenza nell'insieme delle matrici  $M_{n \times n}$

### Proprietà 1.13

Matrici simili godono delle seguenti proprietà:

- hanno stesso rango;
- hanno stessa traccia;
- hanno stesso determinante;
- hanno stesso polinomio caratteristico: questo implica che due matrici simili hanno stessi autovalori.

### Definizione 1.31 (Molteplicità algebrica e geometrica)

Sia  $\lambda_i$  un autovalore di una matrice  $A$ .

Diremo molteplicità algebrica di  $\lambda_i$  la sua molteplicità nel polinomio caratteristico.

Diremo inoltre molteplicità geometrica di  $\lambda_i$  la dimensione del relativo autospazio  $V_{\lambda_i}$

Da ricordarsi che una matrice è diagonalizzabile se e solo se molteplicità algebrica e geometrica coincidono. Se non fosse così, ricordando che la molteplicità algebrica è sempre maggiore o uguale a quella geometrica, possiamo cercare di ricondurre una data matrice, ad una matrice che abbia una struttura simile a quella di una matrice diagonale, ed è detta forma canonica di Jordan. Ogni matrice quadrata è riconducibile a forma di Jordan.

### 1.3.1 Forma canonica di Jordan

La forma canonica di Jordan di una matrice quadrata  $A$  definisce una matrice triangolare  $J$  simile ad  $A$  che ha una struttura il più possibile vicina ad una matrice diagonale. La matrice è diagonale se e solo se  $A$  è diagonalizzabile, altrimenti è divisa in blocchi detti blocchi di Jordan.

La forma canonica caratterizza univocamente la classe di similitudine di una matrice, cioè due matrici sono simili se e solo se hanno la stessa forma di Jordan (a meno di permutazione dei blocchi).

### Blocco di Jordan

Un *blocco di Jordan* di ordine  $k$  è una matrice triangolare superiore con  $k$  righe costituita nel seguente modo:

$$\begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \cdots & \cdots & 0 & \lambda \end{pmatrix}$$

in cui ogni elemento della diagonale è uguale a  $\lambda$  ed in ogni posizione (" $i$ ", " $i$ " + 1) si trova un 1. Il suo polinomio caratteristico è  $(x - \lambda)^k$ , e quindi ha  $\lambda$  come unico autovalore con molteplicità algebrica  $k$ . D'altra parte, l'autospazio relativo a  $\lambda$  è:

$$\ker(J_k(\lambda) - \lambda I) = \ker \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \cdots & \cdots & 0 & 0 \end{pmatrix} = \text{Span} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix}$$

avente, quindi, dimensione 1. Dal teorema di diagonalizzabilità segue che se  $k > 1$  il blocco di Jordan non è diagonalizzabile.

### Matrice di Jordan

Una matrice di Jordan è una matrice a blocchi del tipo

$$J = \begin{pmatrix} J_1 & & 0 \\ & \ddots & \\ 0 & & J_k \end{pmatrix}$$

dove  $J_i$  è un blocco di Jordan con autovalore  $\lambda_i$ . Ogni blocco di Jordan contribuisce con un autospazio unidimensionale relativo a  $\lambda_i$ .

- La molteplicità geometrica di  $\lambda_i$ , definita come la dimensione del relativo autospazio, è pari al numero di blocchi con autovalore  $\lambda_i$ .
- La molteplicità algebrica di  $\lambda_i$ , definita come la molteplicità della radice  $\lambda_i$  nel polinomio caratteristico di  $J$ , è pari alla somma degli ordini di tutti i blocchi con autovalore  $\lambda_i$ .

Il teorema di diagonalizzabilità asserisce che  $J$  è diagonalizzabile se e solo se le molteplicità algebriche e geometriche coincidono, ovvero se e solo se i blocchi hanno tutti ordine pari ad 1: in altre parole,  $J$  è diagonalizzabile se e solo se è già diagonale.

#### Esempio 1.32 (Esempio da Wikipedia)

Calcoliamo la forma canonica di Jordan della matrice

$$A = \begin{pmatrix} 5 & 4 & 2 & 1 \\ 0 & 1 & -1 & -1 \\ -1 & -1 & 3 & 0 \\ 1 & 1 & -1 & 2 \end{pmatrix}$$

Il suo polinomio caratteristico è

$$(x - 4)^2(x - 2)(x - 1)$$

, quindi i suoi autovalori sono 4, 4, 2 e 1. Ricordiamo che, se indichiamo con  $m_{alg}(\lambda)$  e  $m_{geo}(\lambda)$  le molteplicità algebrica e geometrica di un autovalore  $\lambda$  valgono sempre le seguenti disuguaglianze:

$$1 \leq m_{geo}(\lambda) \leq m_{alg}(\lambda)$$

Quindi in questo caso le molteplicità algebriche e geometriche degli autovalori 2 e 1 sono tutte 1, e l'unica grandezza da trovare è la molteplicità geometrica di 4, che può essere 1 o 2. La molteplicità geometrica di un autovalore indica il numero di blocchi di Jordan presenti relativi a quell'autovalore. Vediamo che

$$\dim \ker(A - 4I) = 1.$$

Segue quindi che "A" non è diagonalizzabile, e l'autovalore 4 ha un solo blocco di Jordan. I dati che abbiamo sono sufficienti a determinare la matrice di Jordan, che è la seguente:

$$J = \begin{pmatrix} 4 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Se  $B$  è  $\begin{cases} \text{diagonale (a blocchi)} \\ \text{triangolare (a blocchi)} \end{cases}$

Forma canonica di Jordan: Qualunque matrice  $A$  è riconducibile a forma canonica di Jordan

$$S^{-1}AS = J$$

### Esempio 1.33 (Esempio visto a lezione)

Prendiamo il caso  $n = 9$ , con i seguenti autovalori:

$$\begin{cases} \lambda_1 = 2 & \text{con molteplicità algebrica 5, geometrica 2} \\ \lambda_2 = 3 & \text{con molteplicità algebrica 2, geometrica 1} \\ \lambda_3 = 1 & \text{con molteplicità algebrica 2, geometrica 2} \end{cases}$$

Allora la forma canonica prende gli autovalori distinti. La matrice di Jordan sarà composta dai seguenti blocchi

$$J = \begin{bmatrix} B_{5 \times 5} & 0 & 0 \\ 0 & B_{2 \times 2} & 0 \\ 0 & 0 & B_{2 \times 2} \end{bmatrix}$$

$$J = \begin{bmatrix} 2 & 1 & & & & & & & \\ & 2 & 1 & & & & & & \\ & & 2 & 1 & & & & & \\ & & & 2 & & & & & \\ & & & & 2 & & & & \\ & & & & & 2 & & & \\ & & & & & & 3 & 1 & \\ & & & & & & & 3 & \\ & & & & & & & & 1 \\ & & & & & & & & & 1 \end{bmatrix}$$

Polinomio caratteristico :

$$p(\lambda) = \underbrace{(\lambda - 2)^5(\lambda - 3)^2(\lambda - 1)^2}_{\text{checkme!!}}$$

Questo non dice come spaccare il blocco più grande i blocchi più piccoli. (3,2) (4,1) etc ...

Adesso è facile verificare il teorema di Caley Hamilton perché

$$p(A) = p\left(\underbrace{SJS^{-1}}_{\text{diagonale a blocchi}}\right)$$

$$S^{-1}AS = J$$

$$a_0(SJS^{-1})^2 + a_1(SJS^{-1}) + a_2I$$

$$SJ^2S^{-1}a_1(SJS^{-1}) + a_2SS^{-1} = Sp(j)S^{-1}$$

I polinomi dei singoli blocchi fanno zero.

Prendiamo il primo blocco e lo chiamiamo C.

$$p(C) = (-1)^9$$

$$(C - 2I)^5(C - 3I)^2(C - I)^2$$



**Nota**

**CHECKME**

Se svolgiamo

$$(C - 2I)^5 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}^5 = 0$$

$\sigma(\lambda_i) \geq$  dimensione del blocco

Basta che si annulli una delle componenti di sopra anche il prodotto totale sia 0.



### Teorema 1.34

La forma canonica di Jordan è diagonale (a elementi) (A è diagonalizzabile) se e solo se esistono n autovettori linearmente indipendenti.

Caso particolare: poiché ad autovalori distinti corrispondono autovettori linearmente indipendenti, quindi se A possiede n autovalori distinti allora è diagonalizzabile.

## 1.3.2 Forma canonica di Schur



### Teorema 1.35 (Forma canonica di Schur)

Sia  $A \in \mathbb{C}^{n \times n}$  e siano  $\lambda_1, \dots, \lambda_n$  i suoi autovalori. Allora esiste una matrice unitaria U e una matrice triangolare superiore T i cui elementi principali sono i  $\lambda_i$ , tali che

$$A = UTU^H$$



$T = U^H A U$  è detta *forma di Schur* di  $A$ .

Dato che  $T$

- è simile ad  $A$ , ha gli stessi autovalori
- è triangolare, ha gli autovalori lungo la diagonale.

Inoltre ricordiamo che essendo  $U$  unitaria,  $U^H = U^{-1}$ .



### Nota

La dimostrazione è copiata pari pari dal libro del professore.

*Dimostrazione.* Si procede per induzione. Per  $n = 1$  la tesi vale con  $U = [1] = a_{11}$ .

per  $n > 1$ , sia  $x_1$  l'autovettore normalizzato corrispondente all'autovalore  $\lambda_1$  e sia  $S$  lo spazio generato da  $x_1$ . Indicata con  $y_2, \dots, y_n$  una base ortonormale dello spazio  $S^\perp$ , la matrice

$$Q = [x_1 | y_2 | \dots | y_n]$$

è unitaria e  $Q^H x_1 = e_1$ . Si considera la matrice

$$B = Q^H A Q$$

la cui prima colonna è

$$B e_1 = Q^H A Q e_1 = Q^H A x_1 = Q^H \lambda_1 x_1 = \lambda_1 Q^H x_1 = \lambda_1 e_1$$

e quindi  $B$  può essere partizionata nel seguente modo:

$$B = \begin{bmatrix} \lambda_1 & c^H \\ 0 & A_1 \end{bmatrix}$$

dove  $c \in \mathbb{C}^{n-1}$  e  $A_1 \in \mathbb{C}^{(n-1) \times (n-1)}$ . Per l'ipotesi induttiva esiste una matrice unitaria  $U_1 \in \mathbb{C}^{(n-1) \times (n-1)}$ , tale che

$$A_1 = U_1 A_2 U_1^H$$

dove  $A_2 \in \mathbb{C}^{(n-1) \times (n-1)}$  tale che

$$A_1 = U_1 A_2 U_1^H$$

dove  $A_2 \in \mathbb{C}^{(n-1) \times (n-1)}$  è triangolare superiore. Allora risulta

$$A = Q B Q^H = Q \begin{bmatrix} \lambda_1 & c^H \\ 0 & A_1 \end{bmatrix} Q^H = Q \begin{bmatrix} \lambda_1 & c^H \\ 0 & U_1 A_2 U_1^H \end{bmatrix} Q^H$$

Indicando con  $U^2 \in \mathbb{C}^{m \times n}$  la matrice unitaria

$$U_2 = \begin{bmatrix} 1 & 0^H \\ 0 & U_1 \end{bmatrix}$$

si ha:

$$A = Q U_2 \begin{bmatrix} \lambda_1 & c^H U_1 \\ 0 & A_2 \end{bmatrix} U_2^H Q^H$$

Poiché la matrice  $U = Q U_2$  è ancora unitaria in quanto prodotto di matrici unitarie, risulta

$$A = U \begin{bmatrix} \lambda_1 & c^H U_1 \\ 0 & A_2 \end{bmatrix} U^H$$

da cui la tesi, essendo  $A_2$  matrice triangolare superiore.  $\square$

**Nota**

Il professore ha enunciato la seguente proprietà:

Se  $A$  è Hermitiana e applichiamo il Teorema di Schur,  $T$  risulta diagonale.

La dimostrazione è stata abbozzata, e qui la ripropongo in versione completa, come dal libro

**Teorema 1.36**

Sia  $A$  una matrice hermitiana di ordine  $n$ , cioè  $A = A^H$  e siano  $\lambda_1, \dots, \lambda_n$  i suoi autovalori. Allora esiste una matrice unitaria  $U$  tale che

$$A = U \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} U^H$$

*Dimostrazione.* Per il teorema 1.35 si ha  $T = U^H A U$ , dove  $T$  è una matrice triangolare superiore e  $U$  è unitaria. Poiché  $A = A^H$ , si ha

$$T^H = (U^H A U)^H = U^H A^H U = U^H A U = T$$

cioè la matrice triangolare  $T$  risulta essere una matrice diagonale con gli elementi principali reali e per il teorema 1.21 le colonne di  $U$ , che sono ortonormali perché  $U$  è unitaria, risultano essere gli autovettori di  $A$   $\square$

**Nota**

Anche il seguente teorema e dimostrazione è copiato pari pari dal libro, poichè i passaggi sono molto più chiari.

**Teorema 1.37**

Una matrice  $A \in \mathbb{C}^{n \times n}$  è normale, cioè  $A^H A = A A^H$ , se e solo esiste una matrice unitaria  $U$  tale che

$$A = U \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} U^H$$

Le matrici normali ( $A^H A = A A^H$ ) sono tutte e sole quelle diagonalizzabili con  $U$  unitaria

$$A = U D U^H$$

*Dimostrazione.*  $\implies$ : Si supponga dapprima che  $A$  sia normale. Per il teorema 1.35 esiste una matrice  $U$  unitaria tale che

$$T = U^H A U$$

con  $T$  matrice triangolare superiore e si ha:

$$T^H T = U^H A^H U U^H A U = U^H A^H A U \quad T T^H = U^H A U U^H A^H U = U^H A A^H U$$

Poiché  $A$  è normale, ne segue che

$$T^H T = T T^H \tag{1.4}$$

e quindi anche  $T$  è normale. Si dimostra per induzione su  $n$  che  $T$  è diagonale. Se  $n = 1$  questo è ovvio. Se  $n > 1$ , poiché  $T$  è triangolare superiore, per l'elemento  $p_{11}$  della matrice  $P = T^H T = T T^H$ , si ha

$$p_{11} = \overline{t_{11}} t_{11} = |\lambda_1|^2 \quad \text{e} \quad p_{11} = \sum_{j=1}^n t_{1j} \overline{t_{1j}} = |\lambda_1|^2 + \sum_{j=2}^n |t_{1j}|^2$$

da cui

$$t_{1j} = 0, \quad \text{per } j = 2, \dots, n$$

cioè la prima riga di  $T$  ha tutti gli elementi nulli eccetto quello principale. Indicata con  $T_{n-1}$  la sottomatrice ottenuta da  $T$  cancellando la prima riga e la prima colonna dalla (1.4) segue che

$$T_{n-1}^H T_{n-1} = T_{n-1} T_{n-1}^H$$

Per l'ipotesi induttiva  $T_{n-1}$  è diagonale, e quindi  $T$  risulta diagonale.  $\square$

## 1.4 Alcune proprietà delle matrici definite positive

$$x^H A x > 0 \quad \text{per } x \neq 0$$

Ricordiamo le proprietà:

- $A$  definita positiva  $\iff$  tutti gli autovalori di  $A$  sono reali positivi.
- $A$  definita positiva  $\iff$  le sottomatrici principali di testa hanno determinante positivo (ricordiamo che ci sono le  $a_{ii}$ )



### Nota

Qui il professore enunciato il seguente teorema, dimostrandone a lezione solo un verso.



### Teorema 1.38

Sia  $A$  una matrice hermitiana di ordine  $n$  e siano  $\lambda_1, \dots, \lambda_n$  i suoi autovalori. Allora  $A$  è definita positiva se e solo se  $\lambda_i > 0, i = 1, \dots, n$ .

*Dimostrazione.* Vediamo l'implicazione contraria ossia autovalori reali positivi sono quelle definite positive e questo lo vediamo con Schur. (la freccia opposta del primo punto).

Poiché  $A$  è hermitiana, risulta

$$A = U D U^H$$

con  $U$  matrice unitaria e  $D$  matrice diagonale avente come elementi principali gli autovalori  $\lambda_i, i = 1, \dots, n$  di  $A$ . Se  $x \in \mathbb{C}^n, x \neq 0$ , si ha

$$x^H A x = x^H U D U^H x = y^H D y \quad (1.5)$$

dove il vettore  $y = U^H x$  non può essere uguale a 0, perché  $U$  non è singolare. Dalla (1.5) si ha:

$$\begin{aligned} x^H A x &= (\bar{y}_1, \dots, \bar{y}_n) \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \bar{y}_1 \lambda_1 y_1 + \bar{y}_2 \lambda_2 y_2 + \dots + \bar{y}_n \lambda_n y_n = \\ &= \sum_{i=1}^n \underbrace{|y_i|^2}_{\geq 0} \underbrace{\lambda_i}_{> 0} > 0 \end{aligned}$$

poiché gli autovalori  $\lambda_i$  sono tutti positivi e gli  $|y_i|$  sono tutti non nulli.



### Nota

Non ho segnato e capito perchè l'ultimo passaggetto è  $> 0$

Ciuffo: dai miei appunti ho che: il caso  $= 0$  si ha quando ogni  $y_i = 0$ , quindi l'intero vettore  $y$  dev'essere 0, ma dato che  $y = U^H x$  e che per ipotesi di  $A$  definita positiva abbiamo  $x \neq 0$ , e dato che  $U$  è unitaria, allora esiste un  $y_i > 0$  che rende la sommatoria  $> 0$ . Di questa spiegazione va inserita la versione in italiano

$\square$

## 1.5 Localizzazione degli autovalori

### Definizione 1.39 (Cerchi di Gerschgorin)

Sia  $A \in \mathbb{C}^{m \times n}$ . I cerchi del piano complesso

$$K_i = \{z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j=1, j \neq i}^n |a_{ij}|\}, i = 1, 2, \dots, n$$

di centro  $a_{ii}$  e raggio  $r = \sum_{j=1, j \neq i}^n |a_{ij}|$  sono detti cerchi di Gershgorin.



### Teorema 1.40 (Primo teorema di Gershgorin)

Gli autovalori della matrice  $A$  di ordine  $n$  sono tutti contenuti in

$$\bigcup_{i=1, \dots, n} K_i$$



### Teorema 1.41 (Terzo teorema di Gershgorin)

Se la matrice  $A$  di ordine  $n$  è irriducibile, ogni autovalore  $\lambda$ , che sta sulla frontiera dei cerchi di Gershgorin a cui appartiene, sta sulla frontiera di tutti i cerchi di Gershgorin. In particolare questo vale per gli autovalori che appartengono alla frontiera dell'unione dei cerchi di Gershgorin.

### Esempio 1.42



#### Nota

Ricopiare l'esempio

$$\begin{array}{ccc} 2 & 1 & 0 \\ 1 & 2 & 1 \\ & & 1 \\ 0 & 1 & 2 \end{array}$$

Terzo teorema: Grafo fortemente connesso

$0$  è autovalore??

$$0 \leq \lambda_i \leq 4$$

No perchè non appartiene alla frontiera  $F$  di  $F(k_1)$  e  $F(k_n)$ .

## 1.6 Predominanza diagonale

**Definizione 1.43 (Predominanza diagonale (per righe))**

Una matrice  $A \in \mathbb{C}^{n \times n}$  si dice a predominanza diagonale (debole) se per ogni  $i = 1, \dots, n$  risulta

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

ed esiste almeno un indice  $s$  per cui

$$|a_{ss}| > \sum_{j=1, j \neq s}^n |a_{sj}|$$

Una matrice  $A \in \mathbb{C}^{n \times n}$  si dice a predominanza diagonale (forte) se per ogni  $i = 1, \dots, n$  risulta

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

Con la forte acchiappiamo anche le matrici irriducibili. Si dimostra che con la predominanza diagonale debole e irriducibilita allora:

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

**Nota**

Andrea: enuncio solamente il prossimo teorema perché ne viene fatto uso in seguito

**Teorema 1.44 (equivalenza delle norme)**

Siano  $\|\cdot\|'$  e  $\|\cdot\|''$  due norme vettoriali. Allora le due norme sono topologicamente equivalenti, nel senso che esistono due costanti  $\alpha$  e  $\beta \in \mathbb{R}$ ,  $0 < \alpha \leq \beta$ , tali che per ogni  $x \in \mathbb{C}^n$

$$\alpha \|x\|'' \leq \|x\|' \leq \beta \|x\|''$$

## 1.7 Norme

**Definizione 1.45 (Norma)**

Una funzione  $\mathbb{C}^n \rightarrow \mathbb{R}$

$$x \rightarrow \|x\|$$

che verifica le seguenti proprietà

1.  $\|x\| \geq 0$        $\|x\| = 0$  se e solo se  $x = 0$
2.  $|\alpha x| = |\alpha| \|x\|$
3.  $\|x + y\| \leq \|x\| + \|y\|$

**Nota**

A lezione è stata enunciata un'ulteriore proprietà (Opzionale?)  $f(xy) \leq f(x) \cdot f(y)$ , perchè?

Tipi di norme

**Definizione 1.46**

Sia  $x \in \mathbb{C}^m$ , si definiscono

$$\|x\|_1 = \sum_{i=1}^n |x_i| \quad \text{norma 1}$$

$$\|x\|_2 = \sqrt{\sum_{i=1}^n |x_i|^2} = \sqrt{x^H x} \quad \text{norma 2}$$

$$\|x\|_\infty = \max_{i=1, \dots, n} |x_i| \quad \text{norma } \infty$$

La norma 2 è quella che corrisponde alla lunghezza euclidea del vettore  $x$

**1.7.1 Norme matriciali****Nota**

Il professore è stato molto rapido nel parlare di norme matriciali, cerchiamo col libro di definire per benino le cose.

**Definizione 1.47 (Norma matriciale)**

Una funzione  $\mathbb{C}^{n \times n} \rightarrow \mathbb{R}$

$$A \rightarrow \|A\|$$

che verifica le seguenti proprietà:

1.  $\|A\| \geq 0$  e  $\|A\| = 0$  se e solo se  $A = 0$
2.  $\|\alpha A\| = |\alpha| \|A\|$  per ogni  $\alpha \in \mathbb{C}$
3.  $\|A + B\| \leq \|A\| + \|B\|$  per ogni  $B \in \mathbb{C}^{n \times n}$
4.  $\|AB\| \leq \|A\| \|B\|$  per ogni  $B \in \mathbb{C}^{n \times n}$

è detta norma matriciale

Mostriamo che sia possibile associare ad una norma vettoriale una corrispondente norma matriciale. Si osservi che, poiché la norma vettoriale è una funzione continua, l'insieme

$$\{x \in \mathbb{C}^n : \|x\| = 1\}$$

è chiuso; inoltre, poiché per il teorema (1.44) esiste  $\alpha$  tale che  $\|x\|_\infty \leq \alpha \|x\|$ , ossia  $\max_{i=1, \dots, n} |x_i| \leq \alpha \|x\|$ , l'insieme è anche limitato. Poiché una funzione continua assume su un sottoinsieme chiuso e limitato di  $\mathbb{C}^n$  massimo e minimo si ha che esiste

$$\max_{\|x\|=1} \|Ax\|$$

**Nota**

Andrea: lascio commentata la parte presa e lezione, si accennava a delle proprietà submoltiplicative...

**Definizione 1.48 (Norma matriciale indotta)**

La norma definita da

$$\|A\| = \max_{\|x\|=1} \|Ax\|$$

viene detta norma matriciale indotta dalla norma vettoriale  $\| \cdot \|$



### Teorema 1.49

$\rho$  raggio spettrale

Dalle tre norme vettoriali definite precedentemente, si ottengono le corrispondenti norme matriciali indotte

$$\begin{aligned} \|A\|_1 &= \max_{j=1,\dots,n} \sum_{i=1}^n |a_{ij}| && \text{norma 1} \\ \|A\|_2 &= \sqrt{\rho(A^H A)} && \text{norma 2} \\ \|A\|_\infty &= \max_{i=1,\dots,n} \sum_{j=1}^n |a_{ij}| && \text{norma } \infty \end{aligned}$$



### Nota

A lezione è stata vista solo la dimostrazione per la norma 2. Rispetto alla dimostrazione del libro, è stato fatto l'uso dei quadrati per evitare di portarsi in continuazione delle radici. Scrivo la dimostrazione del libro, e commento quella scritta a lezione ( $A^H A$  è Hermitiana)

*Dimostrazione.*

$$\max_{\|x\|=1} \|Ax\|^2 = N^2 \quad N^2 = \rho(A^H A)$$

Bisogna fare vedere due cose:

- $\forall x, \|x\|_1 = 1 \quad \|Ax\|^2 \leq N^2$
- $\exists x_1$  tale che  $\|Ax_1\|^2 = N^2$

Per il primo punto: poiché la matrice  $A^H A$  è hermitiana, per il teorema (1.36) risulta

$$A^H A = U D U^H$$

, dove  $U$  è unitaria e  $D$  diagonale con gli autovalori di  $A^H A$  come elementi principale. Se  $A = 0$ , allora  $\rho(A^H A) = 0$ , e inversamente, se  $\rho(A^H A) = 0$ , risulta  $D = 0$  e quindi  $A = 0$ . Se  $A \neq 0$ , si ha

$$x^H A^H A x \geq 0 \quad \text{per } x \neq 0$$

Procedendo in modo analogo a (1.38), risulta che gli autovalori di  $A^H A$  sono non negativi e per almeno un di essi, corrispondente al raggio spettrale di  $A^H A$ , si ha

$$\lambda_1 = \rho(A^H A) > 0$$

Sia  $x$  tale che  $\|x\|_2 = 1$  e  $y = U^H x$ , poiché  $U$  è unitaria, poiché per ogni matrice unitaria risulta

$$\|Ax\|_2 = \|x\|_2 \quad \forall x \in \mathbb{C}^n$$

risulta  $\|y\|_2 = 1$  e quindi

$$\begin{aligned} \max_{\|x\|_2=1} \|Ax\|_2 &= \max_{\|x\|_2=1} \sqrt{x^H A^H A x} = \max_{\|y\|_2=1} \sqrt{y^H D y} = \max_{\|y\|_2=1} \sqrt{\sum_{i=1}^n \lambda_i |y_i|^2} \\ &\leq \sqrt{\sum_{i=1}^n \lambda_i |y_i|^2} = \sqrt{\lambda_1} = \sqrt{\rho(A^H A)} \end{aligned}$$

Per il secondo punto: dobbiamo verificare che esiste un vettore  $x$ ,  $\|x\|_2 = 1$ , per cui

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

Questo vettore è  $x_1$ , autovettore di  $A^H A$  relativo all'autovalore  $\lambda_1$  normalizzato in modo che  $\|x_1\|_2 = 1$ . Infatti risulta

$$x_1^H A^H A x_1 = \lambda_1 x_1^H x_1 = \lambda_1 = \rho(A^H A)$$

□





## 2 Richiami di Analisi e Ottimizzazione

Vedremo Problemi di ottimizzazione:

$$f : \mathbb{R}^n \rightarrow \mathbb{R}, D \subseteq \mathbb{R}^n$$
$$\min\{f(x) : x \in D\}$$

Trovare  $\bar{x} \in \mathbb{R}^n$  tale che

- $\bar{x} \in D$  (ammissibile)
- $f(\bar{x}) \leq f(x)$  per ogni  $x \in D$  (punto di minimo)

Successione:

$$\{x_k\}_{k \in \mathbb{N}} \subseteq \mathbb{R}^n \quad x_1, x_2, \dots$$

Useremo la norma 2 su  $\mathbb{R}^n$  :  $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$

### 2.1 Richiami di analisi in $\mathbb{R}^n$

#### Definizione 2.1 (Limite di una successione)

$\bar{x} \in \mathbb{R}^n$  si dice limite di  $\{x_k\}_{k \in \mathbb{N}}$  se

$$\forall \varepsilon > 0 \exists \bar{k} \quad t.c. \quad \|x_k - \bar{x}\|_2 \leq \varepsilon \quad \forall k \geq \bar{k}$$

#### Esempio 2.2

$x_k = (1/k, 1/k)$  : il limite è il vettore  $(0, 0)$

Non tutte le successioni convergono

#### Esempio 2.3

$x_k = ((1/k), (-1)^k)$ : in questo caso il limite non esiste

#### Definizione 2.4 (Sottosuccessione)

Selezione degli elementi  $k_j$  tale che l'indice vada a  $+\infty$ .

#### Esempio 2.5

Sottosuccessione  $k_j = 2j - 1$  ( $k_j$  assume valori dispari)

Sottosuccessione  $k_j = 2j$  ( $k_j$  assume valori pari)

Nel caso dell'esempio, le singole sottosuccessioni definite precedente le successioni convergono a

- $(0, 1)$
- $(0, -1)$

**Definizione 2.6 (Punti di accumulazione della successione)**

$\bar{x}$  è un punto di accumulazione di una successione  $\{x_k\}$  se esiste una sottosequenza infinita di indici  $k_1, k_2, k_3, \dots$  tale che

$$\bar{x} = \lim_{j \rightarrow +\infty} x_{k_j}$$

**Teorema 2.7 (Bolzano-Weierstrass)**

Sia  $\{x_k\}$  una successione tale che esista una costante  $M > 0$  per cui  $\|x_k\| < M \forall k$ . Allora  $\{x_k\}$  ammette una sottosuccessione convergente.

**Definizione 2.8 (Insieme di vicini di  $x$ )**

Dato un punto  $x \in \mathbb{R}^n$ , chiamiamo  $\mathcal{N} \in \mathbb{R}^n$  insieme di vicini di  $x$  se è un insieme aperto che contiene  $x$ . Un insieme di vicini di  $x$  molto utile è il seguente

**Definizione 2.9 (Palla aperta di raggio  $\epsilon$  attorno a  $x$ )**

Una palla aperta di raggio  $\epsilon$  attorno a  $x$  è definita come

$$B(x, \epsilon) = \{y \mid \|y - x\| < \epsilon\}$$

In un piano bidimensionale abbiamo un cerchio.

**Definizione 2.10**

$A \subseteq \mathbb{R}^n$  si dice aperto se

$$\forall x \in A \exists \epsilon > 0 \text{ t.c. } B(x, \epsilon) \subseteq A$$

**Esempio 2.11 (Esempi di insiemi aperti)**

- $] - 1, 1[$
- $\mathbb{R}^n$
- $\emptyset$

**Proprietà 2.1**

unione di aperti è aperto  $A_\alpha \text{ aperti} \rightarrow \cup_\alpha A_\alpha \text{ è aperto}$

 **Proprietà 2.2**

- l'intersezione di un insieme finito di aperti è un aperto
- l'unione infinita di aperti è un aperto

L'intersezione infinita di  $B(0, \frac{1}{k}) = \{0\}$  non è un aperto

**Definizione 2.12 (Punto interno)**

Sia  $A \subseteq \mathbb{R}^n$ ,  $x$  si dice punto interno di  $A$  se  $\exists \varepsilon > 0$  t.c.  $B(x, \varepsilon) \subseteq A$

 **Proprietà 2.3**

$A$  aperto  $\iff A = \{\text{punti interni di } A\}$

**Definizione 2.13 (Chiuso)**

$A \subseteq \mathbb{R}^n$  si dice chiuso se  $\mathbb{R}^n \setminus A$  (complementare) è aperto

**Esempio 2.14 (Insiemi chiusi)**

- $\overline{B(x, \varepsilon)} = \{y \in \mathbb{R}^n \mid \|y - x\|_2 \leq \varepsilon\}$
- $\emptyset$
- $[-1, 1]$

**Definizione 2.15 (Punto di chiusura)**

Punto  $x \in \mathbb{R}^n$  si dice punto di chiusura se

$$\forall \varepsilon > 0 : B(x, \varepsilon) \cap A \neq \emptyset$$

 **Proprietà 2.4**

$A$  è chiuso  $\iff A = \{\text{punti di chiusura}\}$

Notazione:  $(\overline{A}, cl(A))$

 **Proprietà 2.5**

- $A_\alpha$  chiusi  $\Rightarrow \bigcap_\alpha A_\alpha$  è chiuso. L'intersezione infinita di chiusi è un chiuso.

- $A_i$  chiusi  $i = 1 \dots k \Rightarrow \bigcup_{i=1}^k A_i$  è chiuso. L'unione finita di chiusi è un chiuso.

Ci sono però degli elementi che possono essere sia aperti che chiusi e altri che non sono né aperti né chiusi  
Gli insiemi vuoti e  $\mathbb{R}^n$  sono sia aperti che chiusi.

**Esempio 2.16 (Insieme di elementi che non sono né aperti né chiusi)**

Prendiamo un rettangolo ed una palla (figura!)

$$([-1, 0] \times [-1, 1]) \cup B(0, 1) = A$$

- $(-1 - \epsilon, 0) \in B((-1, 0), \epsilon) \quad \forall \epsilon : A$  non è aperto
- $x_k = (1 - \frac{1}{k}, 0) \in A \quad x_k \rightarrow (1, 0) \notin A : A$  non è chiuso

 **Proprietà 2.6**

$A \subseteq \mathbb{R}^n$  è chiuso  $\iff \forall \{x_k\} \subseteq A$  per cui  $x_k \rightarrow \bar{x}$  risulta  $\bar{x} \in A$

**Definizione 2.17 (Insieme limitato)**

$A \subseteq \mathbb{R}^n$  si dice limitato se  $\exists M > 0$  per cui  $\|x\|_2 \leq M \quad \forall x \in A$

**Definizione 2.18 (Insieme compatto)**

$A \subseteq \mathbb{R}^n$  si dice compatto se  $A$  è limitato e chiuso.

Gli insiemi compatti sono interessanti per il seguente motivo

 **Teorema 2.19 (Bolzano-Weirstrass)**

Sia  $A \subseteq \mathbb{R}^n$  un insieme compatto.

Ogni successione  $\{x_k\}_k \subseteq A$  ammette una sottosuccessione convergente. Ossia

$$\exists \{x_{k_j}\} \subseteq A, \bar{x} \in A \text{ t.c. } x_{k_j} \rightarrow_{j \rightarrow +\infty} \bar{x}$$

**2.1.1 Funzioni di più variabili a valori reali****Definizione 2.20 (Funzione continua)**

$f$  si dice continua in  $\bar{x} \in \mathbb{R}^n$  se

$$\forall \epsilon > 0 \quad \exists \delta > 0 \quad \text{t.c.} \quad \forall x \in \mathbb{R}^n : \|x - \bar{x}\|_2 \leq \delta \Rightarrow |f(x) - f(\bar{x})| \leq \epsilon$$

 **Proprietà 2.7**

$f$  è continua in  $\bar{x} \in \mathbb{R}^n \iff \forall \{x_k\}$  t.c.  $x_k \rightarrow \bar{x}$  risulta  $\lim_{k \rightarrow +\infty} f(x_k) = f(\bar{x})$

**Esempio 2.21 (Esempi di funzioni continue)**

- $f(x) = \|x\|_2$
- $f(x_1, x_2) = \sin(\pi x_1 x_2)$

**Definizione 2.22 (Lipschitziana)**

Una funzione lipschitziana è una funzione che ha una crescita limitata, nel senso che il rapporto tra variazione di ordinata e variazione di ascissa non può superare un valore fissato, detto costante di Lipschitz. È una condizione più forte della continuità.

$$\frac{\|f(x_2) - f(x_1)\|}{\|x_2 - x_1\|} \leq L$$

**Definizione 2.23 (Estremo superiore)**

Sia  $(X, \leq)$  un insieme totalmente ordinato,  $E \subseteq X$ . Se esiste un elemento  $y \in X$  tale che:

- $y$  è un maggiorante di  $E$
- se  $z < y$  allora  $z$  non è un maggiorante di  $E$

diciamo che  $y$  è estremo superiore di  $E$ , in simboli  $y = \sup E$  e diciamo che  $E$  è limitato superiormente.

La definizione di estremo inferiore è simmetrica.

**Definizione 2.24 (Massimo)**

Si definisce elemento massimo di  $S$  un  $M \in S$  tale che

$$\forall a \in S, a \leq M$$

La definizione di minimo è simmetrica.

**Teorema 2.25 (Weirstrass per funzioni continue)**

Sia  $A \subseteq \mathbb{R}^n$  compatto e  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  continua su  $A$ .

Allora  $f$  ammette massimo e minimo su  $A$ .

*Dimostrazione.* Poniamo  $l = \inf\{f(x) : x \in A\} \in [-\infty, +\infty]$ . Sicuramente esiste successione  $f(x_k) \rightarrow l$  con  $x_k \in A$ . (per definizione di estremo inferiore). Ma

- Per la (2.19),  $A$  compatto  $\Rightarrow \exists x_{k_j} \rightarrow \bar{x}$
- Per la (2.7)  $f$  continua  $\Rightarrow f(x_{k_j}) \rightarrow f(\bar{x})$

Allora, con l'unicità del limite,  $f(\bar{x}) \Rightarrow l \neq -\infty \Rightarrow \min\{f(x) : x \in A\}$

□

**Esempio 2.26**

$n = 1$   $f(t) = e^{-t}$ ,  $A = \mathbb{R}_+$

$$\inf\{f(t) : t \in A\} = 0$$

Ma il minimo non esiste! Quindi l'insieme non è compatto.

### 2.1.2 Derivate

Per le funzioni  $f : \mathbb{R} \rightarrow \mathbb{R}$  abbiamo definito la derivata come

$$\lim_{t \rightarrow 0} \frac{f(\bar{x} + t) - f(\bar{x})}{t}$$

Nel caso più generale di funzioni  $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$  consideriamo il punto  $\bar{x}$  e l'insieme di vettori

$$\{\bar{x} + tv \mid t \geq 0\} \quad v \in \mathbb{R}^n, \|v\|_2 = 1 \quad v : \text{direzione}$$

#### Definizione 2.27 (Derivata direzionale)

$f$  si dice derivabile in  $x \in \mathbb{R}^n$  nella direzione  $v$  se esiste

$$D_v f(\bar{x}) = \lim_{t \rightarrow 0} \frac{f(\bar{x} + tv) - f(\bar{x})}{t}$$

Le derivate direzionali sono una generalizzazione delle derivate parziali, nelle quali si scelgono come vettori direzionali quelli del tipo

$$e_i = (0, 0, 0, \dots, 0, \underset{\text{i-esima}}{1}, 0, \dots, \dots, 0)$$

ecco una definizione più formale

#### Definizione 2.28 (Derivata parziale)

Si definisce derivata parziale di  $f$  in  $\vec{x}$  rispetto alla variabile  $k$ -esima  $x_k$  il limite, se esiste finito

$$\frac{\partial f}{\partial x_k}(\vec{x}) = \lim_{h \rightarrow 0} \frac{f(x_1, x_2, \dots, x_k + h, \dots, x_n) - f(x_1, x_2, \dots, x_n)}{h}$$

#### Esempio 2.29 (Derivata parziale di $\sin(\pi x_1 x_2)$ )

$$f(x_1, x_2) = \sin(\pi x_1 x_2)$$

Le 2 derivate parziali sono

$$\frac{\partial f}{\partial x_1}(x) = \pi x_2 \cos(\pi x_1 x_2) \quad \frac{\partial f}{\partial x_2}(x) = \pi x_1 \cos(\pi x_1 x_2)$$

#### Definizione 2.30 (Gradiente)

$\nabla f(\bar{x}) = \left( \frac{\partial f}{\partial x_1}(\bar{x}), \frac{\partial f}{\partial x_2}(\bar{x}), \dots, \frac{\partial f}{\partial x_n}(\bar{x}) \right)$  è detto gradiente di  $f$  in  $\bar{x}$

Informalmente, il gradiente indica la direzione di massima pendenza.

Nell'esempio (2.1), il vettore

$$\nabla f(x) = \begin{pmatrix} \pi x_2 \cos(\pi x_1 x_2) \\ \pi x_1 \cos(\pi x_1 x_2) \end{pmatrix}$$

,i cui elementi sono le derivate parziali precedentemente calcolate, è il *gradiente di  $f$  in  $\bar{x}$* .

Prendiamo  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $v \in \mathbb{R}^n$ ,  $\|v\| = 1$

**Definizione 2.31 (Two-sided derivate)**

$$\frac{\partial f}{\partial v}(\bar{x}) = \lim_{t \rightarrow 0} \frac{[f(\bar{x} + t \cdot v) - f(\bar{x})]}{t}$$

**Definizione 2.32 (One-sided directional derivate)**

$$f'(\bar{x}, v) = \lim_{t \rightarrow 0^+} \frac{[f(\bar{x} + t \cdot v) - f(\bar{x})]}{t}$$

Il limite invece che andare a 0 da entrambi i lati, ci va solo da  $0^+$ .

**Nota**

A volte esiste una e non l'altra.

**Esempio 2.33 (Derivate della norma)**

$$f(x) = \|x\|_2 = \left( \sum_{i=1}^n x_i^2 \right)^{1/2} \quad v \in \mathbb{R}^n \quad \|v\|_2 = 1 \quad , \bar{x} = 0$$

$$\frac{f(\bar{x} + tv) - f(\bar{x})}{t} = \frac{f(tv)}{t} = \frac{\left( \sum_{i=1}^n t^2 v_i^2 \right)^{1/2}}{t} = \frac{|t| \left( \sum_{i=1}^n v_i^2 \right)^{1/2}}{t} = \operatorname{sgn}(t) \|v\|_2 = \operatorname{sgn}(t)$$

Quindi  $f'(\bar{x}, v) = 1$  mentre  $\frac{\partial f}{\partial v}(\bar{x})$  non esiste.

**Esempio 2.34 (Esempi di calcolo di derivata in  $\mathbb{R}^n$ )**

Prendiamo ad esempio la funzione  $f$  definita nel seguente modo

$$f(x_1, x_2) = \begin{cases} \left( \frac{x_1^2 x_2}{x_1^2 + x_2^2} \right)^2 & \text{se } (x_1, x_2) \neq (0, 0) \\ \text{indef.} & \text{se } (x_1, x_2) = (0, 0) \end{cases} \quad (2.1)$$

Ponendo  $x_2 = \alpha x_1^2$ , risulta

$$f(x_1, \alpha x_1^2) = \left[ \alpha x_1^4 \cdot (x_1^4 + \alpha^2 x_1^4) \right]^2 = \left[ \alpha / (1 + \alpha^2) \right]^2$$

**Domanda aperta**

Perchè ha voluto introdurre questo  $\alpha$ ? Voleva mostrare qualcosa in particolare?

$f$  risulta non continua in  $\bar{x} = (0, 0)$ . Consideriamo infatti le seguenti successioni:

- $f\left(\frac{1}{k}, \frac{1}{k^2}\right)$ . Sostituendo i parametri in (2.1) otteniamo il valore costante  $1/4$ . Per  $k \rightarrow +\infty$  la successione converge a  $(0, 0)$ , ma risulta

$$\lim_{k \rightarrow \infty} f\left(\frac{1}{k}, \frac{1}{k^2}\right) = 1/4 \quad \neq \quad ? = f(\bar{x})$$

- $f(\frac{1}{k}, \frac{2}{k^2})$ : discorso analogo al caso precedente. Risulta infatti  $f(\frac{1}{k}, \frac{2}{k^2}) = 4/25$ , ma

$$\lim_{k \rightarrow \infty} f\left(\frac{1}{k}, \frac{2}{k^2}\right) = 4/25 \neq ? = f(\bar{x})$$

### Domanda aperta

Perchè ha mostrato due esempi? Non ne bastava solo uno per provare la non continuità?

Facciamo vedere che esistono le derivate direzionali di  $f$  in  $\bar{x} = (0, 0)$  in tutte le direzioni. Ricordiamo che  $x$  e  $v$  sono vettori e  $t$  è uno scalare. Poniamo  $f(\bar{x}) = 0$ .

$$\begin{aligned} \frac{\partial f}{\partial v}(\bar{x}) &= \frac{\partial f}{\partial v}((0, 0)) = \lim_{t \rightarrow 0} \frac{f((0, 0) + t(v_1, v_2)) - \overbrace{f((0, 0))}^0}{t} = \\ \lim_{t \rightarrow 0} \frac{f((tv_1, tv_2))}{t} &= \lim_{t \rightarrow 0} \frac{[t^3 v_1^2 v_2 / (t^4 v_1^4 + t^2 v_2^2)]^2}{t} = \lim_{t \rightarrow 0} t v_1^4 v_2^2 / (t^2 v_1^4 + v_2^2)^2 \xrightarrow{t \rightarrow 0} 0 \end{aligned}$$

Dall'esempio precedente abbiamo stabilito quindi che la proprietà:

$$\varphi : \mathbb{R} \rightarrow \mathbb{R}, \quad \varphi \text{ derivabile in } \bar{x} \implies \varphi \text{ continua in } \bar{x}$$

non vale in generale in  $\mathbb{R}^n$ .



### Nota

Le derivate in tutte le direzioni esistono, quindi non è come nelle equazioni ad una variabile: se  $f$  è derivabile, non è necessariamente continua.

### Definizione 2.35 (Funzione lineare)

Una funzione  $L$  è lineare se e solo se

$$x, y \in \mathbb{R}^n \quad \alpha, \beta \in \mathbb{R} \quad L(\alpha x + \beta y) = \alpha \cdot L(x) + \beta \cdot L(y)$$

ovvero:

$$L(x) = l^T x = \sum_{i=1}^n l_i x_i$$

### Definizione 2.36 (Funzione differenziabile)

$f : \mathbb{R}^n \rightarrow \mathbb{R}$  si dice differenziabile in  $\bar{x} \in \mathbb{R}^n$  se  $\exists L : \mathbb{R}^n \rightarrow \mathbb{R}$  lineare tale che

$$\forall h \in \mathbb{R}^n \quad f(\bar{x} + h) = f(\bar{x}) + L(h) + r_{\bar{x}}(h)$$

con  $r_{\bar{x}}(h)$  un resto tale che  $\frac{r_{\bar{x}}(h)}{\|h\|_2} \rightarrow 0$  per  $\|h\|_2 \rightarrow 0$

### Domanda aperta

Mi sembra di aver capito che una funzione differenziabile implica che tutte le derivate, da tutte le direzioni, sono esprimibili come combinazione lineare delle derivate della base canonica. La cosa verrà esplicitata nell'esercizio (2.1)



In altri termini, la differenza tra la funzione in un punto e la funzione calcolata in  $\bar{x}$  è approssimata da una funzione lineare.

 **Proprietà 2.8**

Sia  $f : \mathbb{R}^n \rightarrow \mathbb{R}$

$f$  differenziabile in  $\bar{x} \implies f$  continua in  $\bar{x}$



**Nota**

(Quest'ultima proprietà è simile a quella delle funzioni a una variabile)

 **Proprietà 2.9**

Sia  $f : \mathbb{R}^n \rightarrow \mathbb{R}$

$f$  differenziabile in  $\bar{x} \implies f$  ammette derivate in  $\bar{x}$  in ogni direzione  $v$ .

*Dimostrazione.*

$$\frac{\partial f}{\partial v}(\bar{x}) = \lim_{t \rightarrow 0} \frac{f(\bar{x} + tv) - f(\bar{x})}{t} = \lim_{t \rightarrow 0} \frac{L(tv) + r_{\bar{x}}(tv)}{t} = \lim_{t \rightarrow 0} \frac{tL(v) + r_{\bar{x}}(tv)}{t} = L(v) + \lim_{t \rightarrow 0} \frac{r_{\bar{x}}(tv)}{t} = L(v)$$

$$\|tv\|_2 = t\|v\|_2 = t$$

(Nota che  $\|v\|_2 = 1$ )  $\square$

**Domanda aperta**

Da quello che ho capito dalle note una volta stabilito che

$f$  differenziabile in  $\bar{x} \implies f$  ammette derivate in  $\bar{x}$  in ogni direzione  $v$ .

Si vuole avere un metodo esplicito per calcolare la derivata direzionale. Chiedere lumi. A occhio mi viene da dire che, una volta calcolate le derivate parziali della base canonica, possiamo ricavare le derivate delle direzioni facendo un semplice prodotto riga per colonna. La condizione che deve essere rispettata è che  $f$  deve essere differenziabile.



**Osservazione 2.37 (Calcolo esplicito della derivata direzionale)**

Sfruttiamo le seguenti ipotesi:

$$v \in \mathbb{R}^n \quad \|v\|_2 = 1 \quad v = \sum_{i=1}^n v_i e_i \quad \underbrace{f \text{ differenziabile}}_{\text{CHECKME}}$$

calcoliamo quindi la derivata direzionale

$$\frac{\partial f}{\partial v}(\bar{x}) = L(v) = L\left(\sum_{i=1}^n v_i e_i\right) \stackrel{1)}{=} \sum_{i=1}^n v_i \cdot L(e_i) = \sum_{i=1}^n v_i \frac{\partial f}{\partial x_i}(\bar{x}) = \nabla f(\bar{x})^T v$$

1.  $L$  è lineare

**Esercizio 2.1**

Esercizio per casa: studia

$$f(x_1, x_2) = \begin{cases} x_1^2 x_2 / (x_1^2 + x_2^2) & \text{se } (x_1, x_2) \neq (0, 0) \\ 0 & \text{se } (x_1, x_2) = (0, 0) \end{cases}$$

**Svolgimento**

Vediamo che questa funzione non è differenziabile in  $\bar{x} = (0, 0)$

$$\frac{\partial f}{\partial v}(\bar{x}) = \lim_{t \rightarrow 0} \frac{t^3 v_1^2 v_2 / t^2 \sqrt{v_1^2 + v_2^2}}{t} = v_1^2 v_2 \quad (\text{poichè } \|v\|_2^2 = v_1^2 + v_2^2 = 1)$$

In particolare abbiamo

$$\frac{\partial f}{\partial x_1}(\bar{x}) = \frac{\partial f}{\partial x_2}(\bar{x}) = 0$$

Per farlo vedere, specialmente in questa fase iniziale del corso, esplicitiamo i conti:

$$\frac{\partial f}{\partial x_1}(\bar{x}) = 1^2 \cdot 0 = 0$$

Nell'altro caso

$$\frac{\partial f}{\partial x_2}(\bar{x}) = 0^2 \cdot 1 = 0$$

ma  $\frac{\partial f}{\partial v}(\bar{x}) \neq 0$  per ogni altra direzione. Quindi  $\frac{\partial f}{\partial v}(\bar{x}) \neq \nabla f(x)^T v = 0$

**Domanda aperta**

È vero che il motivo profondo della disuguaglianza è che la derivata in ogni direzione non è esprimibile come combinazione lineare delle derivate parziali sulle direzioni della base canonica?

**Teorema 2.38 (Differenziale totale)**

Sia  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , tale da ammettere derivate parziali in ogni  $x \in B(\bar{x}, \epsilon)$  per qualche  $\epsilon > 0$ . Allora

$$\frac{\partial f}{\partial x_i} : B(\bar{x}, \epsilon) \rightarrow \mathbb{R} \text{ continua in } \bar{x} \implies f \text{ differenziabile in } \bar{x}$$

$$f \rightarrow \frac{\partial f}{\partial x_i}(\bar{x})$$

**Esempio 2.39**

Vediamo un'applicazione pratica del teorema (2.38) appena enunciato. Dalla tabella di verità delle implicazioni possiamo stabilire che vale la proprietà

$$(a \implies b) \iff (\neg b \implies \neg a)$$

Facciamo vedere quindi che non derivabilità implica non continuità.

Riprendiamo l'esercizio (2.1):

$$\frac{\partial f}{\partial x_1}(x) = \underbrace{D(x_1^2 x_2 / (x_1^2 + x_2^2))}_{\text{rispetto a } x_1} \stackrel{\square}{=} \frac{D(x_1^2 x_2)(x_1^2 + x_2^2) - x_1^2 x_2 D(x_1^2 + x_2^2)}{(x_1^2 + x_2^2)^2} =$$

$$\frac{2x_1 x_2^3 + 2x_2^3 x_1 - 2x_1 x_2^3}{(x_1^2 + x_2^2)^2} = \frac{2x_2^3 x_1}{(x_1^2 + x_2^2)^2}$$

1. Derivata del rapporto

Ponendo  $x_1 = x_2 = k$ ,  $k \neq 0$  otteniamo

$$\frac{\partial f}{\partial x_1}(x_1, x_2) = \frac{2k^4}{(2k^2)^2} = \frac{1}{2}$$

Mentre, come visto precedentemente nell'esempio

$$\frac{\partial f}{\partial x_1}(0, 0) = 0$$

### Proprietà 2.10 (Funzioni composte)

Sia  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  differenziabile in  $\bar{x} \in \mathbb{R}^n$ ,  $\Phi : \mathbb{R} \rightarrow \mathbb{R}$  derivabile in  $f(\bar{x}) \in \mathbb{R}^n$ .

Allora  $\Phi \circ f : \mathbb{R}^n \rightarrow \mathbb{R}$  è differenziabile in  $\bar{x}$  con

$$\nabla(\Phi \circ f)(\bar{x}) = \Phi'(f(\bar{x})) \nabla f(\bar{x})$$



### Teorema 2.40 (Valor medio)

$f : \mathbb{R}^n \rightarrow \mathbb{R}$  è differenziabile (su tutto  $\mathbb{R}^n$ ) con continuità (cioè  $x \mapsto \nabla f(x)$  è continua) su  $\mathbb{R}^n$ .

Dati  $\bar{x}, h \in \mathbb{R}^n$ , esiste  $t \in (0, 1)$  t.c.

$$f(\bar{x} + h) = f(\bar{x}) + \nabla f(\bar{x} + th)^T h$$



### Nota

Notare che  $x + th$  appartiene al segmento di estremi  $x$  e  $x + h$ . Ricordare inoltre il caso

$$\Phi : \mathbb{R} \rightarrow \mathbb{R} : \Phi(t_1) = \Phi(t_0) + \Phi'(\xi)(t_1 - t_0) \text{ per un opportuno } \xi \in (t_1, t_0)(t_0, t_1)$$

### Definizione 2.41 (Formula di Taylor del 1° Ordine)

Sia  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  differenziabile, allora la formula di Taylor del primo ordine risulta

$$f(\bar{x} + h) = f(\bar{x}) + \nabla f(\bar{x})^T h + r(h) \quad \text{con} \quad \frac{r(h)}{\|h\|_2} \xrightarrow{\|h\|_2 \rightarrow 0} 0$$

Quest'ultima è una semplice riscrittura della definizione alla luce della proprietà (2.9)

**Definizione 2.42 (Iperpiano tangente)**

$$h = x - \bar{x}, h \approx 0 \rightarrow f(x) \approx f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x})$$

Equazione iperpiano tangente al grafico di  $f$  nel punto  $(\bar{x}, f(\bar{x}))$

$$\{(x, f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x})) \mid x \in \mathbb{R}^n\}$$

**2.1.2.1 Derivate seconde**

$$\begin{aligned} f : \mathbb{R}^n &\rightarrow \mathbb{R} \text{ differenziabile} \\ x &\rightarrow \frac{\partial f}{\partial v}(x) \quad \frac{\partial f}{\partial v} : \mathbb{R}^n \rightarrow \mathbb{R} \\ \bar{x} \in \mathbb{R}^n, \quad v \in \mathbb{R}^n, \quad \|v\|_2 = 1, \quad w \in \mathbb{R}^n, \quad \|w\|_2 = 1 \\ \frac{\partial}{\partial w} \left( \frac{\partial f}{\partial v} \right) (\bar{x}) &= \lim_{t \rightarrow 0} \frac{\frac{\partial f}{\partial v}(\bar{x} + tw) - \frac{\partial f}{\partial v}(\bar{x})}{t} \\ w = e_1, \quad v = e_j & \\ \frac{\partial}{\partial x_1} \left( \frac{\partial f}{\partial x_2} \right) &\rightsquigarrow \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ &\text{si scrive} \end{aligned}$$

Esercizio:  $f(x_1, x_2) = \sin(\pi x_1 x_2)$  calcolare le 4 derivate parziali.

**2.1.3 Funzioni di più variabili a valori vettoriali****Definizione 2.43 (Matrice Jacobiana)**

$$F : \mathbb{R}^n \rightarrow \mathbb{R}^n \quad F = (f_1, \dots, f_n)$$

Matrice dei gradienti: righe sono i gradienti.

$$J_F(x) = \begin{pmatrix} \nabla f_1(x)^T \\ \nabla f_2(x)^T \\ \dots \\ \nabla f_n(x)^T \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1(x)}{\partial(x_1)}(x) & \dots & \frac{\partial f_1(x)}{\partial(x_n)}(x) \\ \dots & \dots & \dots \\ \frac{\partial f_n(x)}{\partial(x_1)}(x) & \dots & \frac{\partial f_n(x)}{\partial(x_n)}(x) \end{pmatrix}$$

Il determinante della matrice Jacobiana è detto Jacobiano.

**2.1.4 Derivate di ordine superiore**

$f$  differenziabile su  $\mathbb{R}^n \Rightarrow \exists \frac{\partial f}{\partial v}(x)$  per ogni  $x \in \mathbb{R}^n, v \in \mathbb{R}^n$  con  $\|v\|_2 = 1$ .

$\frac{\partial f}{\partial v} : \mathbb{R}^n \rightarrow \mathbb{R}$  è una funzione  $\rightarrow$  ammette derivate nelle varie direzioni?

Limitiamoci alle derivate parziali:  $w = e_i, v = e_j \rightarrow \frac{\partial}{\partial x_j} \rightsquigarrow \frac{\partial^2 f}{\partial x_i \partial x_j} \quad i = j \rightsquigarrow \frac{\partial^2 f}{\partial x_i^2}$

Funzioni  $\mathbb{R}^n \rightarrow \mathbb{R}$  differenziabile:

$$\frac{\partial f}{\partial(x_j)} : \mathbb{R}^n \rightarrow \mathbb{R}$$

**Esempio 2.44**

Definiamo  $f$  come  $f(x_1, x_2) = \sin(\pi x_1 x_2)$

$$\begin{aligned} \frac{\partial}{\partial(x_1)} f(x) &= \pi x_2 \cos(\pi x_1 x_2) \\ \frac{\partial}{\partial(x_2)} f(x) &= \pi x_1 \cos(\pi x_1 x_2) \\ \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) &= \pi \cos(\pi x_1 x_2) - \pi^2 x_2 x_1 \sin(\pi x_1 x_2) \\ \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) &= \pi \cos(\pi x_1 x_2) - \pi^2 x_1 x_2 \sin(\pi x_1 x_2) \end{aligned}$$

Le due derivate parziali sono uguali. Scambiando l'ordine di derivazione abbiamo ottenuto lo stesso risultato: non è un caso, deriva dal seguente teorema.



**Teorema 2.45 (Schwarz/inversione ordine di derivazione)**

Sia  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  tale che per ogni  $i, j = 1 \dots n$  tali che  $\frac{\partial^2 f}{\partial x_j \partial x_i}$ ,  $\frac{\partial^2 f}{\partial x_i \partial x_j}$  esistano in  $B(\bar{x}, \varepsilon)$  per  $\bar{x} \in \mathbb{R}^n$ ,  $\varepsilon > 0$  e siano continue in  $\bar{x}$ . Allora:

$$\frac{\partial^2 f}{\partial x_j \partial x_i} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

**Definizione 2.46 (Matrice Hessiana di  $f$  in  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  in  $\bar{x} \in \mathbb{R}^n$ )**

$$\nabla^2 f(\bar{x}) = \left\{ \frac{\partial^2 f}{\partial x_i \partial x_j} \right\}_{\substack{i=1 \dots n \\ j=1 \dots n}} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(\bar{x}) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(\bar{x}) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\bar{x}) \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\bar{x}) & \dots & \dots & \frac{\partial^2 f}{\partial x_n \partial x_n}(\bar{x}) \end{pmatrix}$$

Differenziabile due volte  $\frac{\partial^2 f}{\partial x_i \partial x_j}(x)$  esista per ogni  $x \in \mathbb{R}^n$  e  $\frac{\partial^2 f}{\partial x_i \partial x_j} : \mathbb{R}^n \rightarrow \mathbb{R}$  sia continua su  $\mathbb{R}^n$ . Questa è una matrice *simmetrica*

**Definizione 2.47 (Formula di Taylor con resto)**

$$f(\bar{\mathbf{x}} + \mathbf{h}) = f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^T \mathbf{h} + \frac{1}{2} \mathbf{h}^T \nabla^2 f(\bar{\mathbf{x}} + t\mathbf{h}) \mathbf{h} \quad \text{con } t \in (0, 1) \text{ opportuno}$$

**Definizione 2.48 (Formula di Taylor secondo ordine)**

$$f(\bar{\mathbf{x}} + \mathbf{h}) = f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^T \mathbf{h} + \frac{1}{2} \mathbf{h}^T \nabla^2 f(\bar{\mathbf{x}}) \mathbf{h} + r_{\bar{\mathbf{x}}}(\mathbf{h}) \quad \text{con } \frac{r_{\bar{\mathbf{x}}}(\mathbf{h})}{\|\mathbf{h}\|_2} \xrightarrow{\|\mathbf{h}\|_2 \rightarrow 0} 0$$

$$\mathbf{h}^T \nabla^2 f(\bar{\mathbf{x}}) = \sum_{i=1}^n \sum_{j=1}^n = \frac{\partial^2 f}{\partial x_i \partial x_j}(\bar{\mathbf{x}}) h_i h_j$$

Questa formula è interessante quando  $h$  è piccolo, infatti

$$\mathbf{h} = \mathbf{x} - \bar{\mathbf{x}} \approx 0 : f(\mathbf{x}) \approx f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^T (\mathbf{x} - \bar{\mathbf{x}}) + \underbrace{\frac{1}{2} (\mathbf{x} - \bar{\mathbf{x}})^T \nabla^2 f(\bar{\mathbf{x}}) (\mathbf{x} - \bar{\mathbf{x}})}_{\text{funzione non lineare}}$$

approssimazione quadratica di  $f$  vicino a  $\mathbf{x}$

Funzioni alle quali saremo molto interessanti sono le funzioni quadratiche

**Definizione 2.49 (Funzione Quadratica)**

Siano  $Q \in \mathbb{R}^{n \times n}$ ,  $b \in \mathbb{R}^n$ ,  $c \in \mathbb{R}$ . La funzione:  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  è detta funzione quadratica se

$$f(x) = \frac{1}{2} x^T Q x + b^T x + c = \frac{1}{2} \sum_{k=1}^n \sum_{l=1}^n q_{kl} x_k x_l + \sum_{k=1}^n b_k x_k + c$$

 **Nota**

Una forma quadratica è una funzione quadratica in cui sono presenti solo i termini di secondo grado, ossia

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

 **Osservazione 2.50 (Q simmetrica in una funzione quadratica)**

Cosa accade se  $Q$  è simmetrica in una funzione lineare? Valgono le seguenti proprietà:

1. (Gradiente):  $\nabla f(x) = Qx + b$
2. (Matrice Hessiana):  $\nabla^2 f(x) = Q$

Dimostriamo 1. Vediamo cosa vale la derivata parziale in una generica componente  $j$

$$\begin{aligned} \frac{\partial f}{\partial x_j}(x) &= \frac{1}{2} \left[ \sum_{l=1}^n q_{jl} x_l + \sum_{k=1}^n q_{kj} x_k \right] + b_j = \frac{1}{2} \left[ \sum_{l=1}^n q_{lj} x_l + \sum_{k=1}^n q_{kj} x_k \right] + b_j \quad \stackrel{\text{simmetrica}}{=} \\ &= \sum_{l=1}^n q_{jl} x_l + b_j = (Qx)_j + b_j \quad \Rightarrow \quad \nabla f(x) = Qx + b \end{aligned}$$

Discorso analogo per il punto 2.

$$\frac{\partial f}{\partial x_i \partial x_j}(x) = \frac{\partial f}{\partial x_i} \left( \frac{\partial f}{\partial x_j \partial x_j}(x) \right) = \frac{\partial f}{\partial x_i} \left( \sum_{l=1}^n q_{jl} + b_j \right) = q_{ji} = q_{ij} \quad \Rightarrow \quad \nabla^2 f(x) = Q$$

**Esercizio 2.2**

Per esercizio: trovare l'approssimazione di Taylor del secondo ordine di  $f(x_1, x_2) = -x_1^4 - x_2^2$  centrata in  $\bar{x} = (0, 2/5)$

**Svolgimento**

$$\nabla f(x) = \begin{pmatrix} -4x_1^3 \\ -2x_2 \end{pmatrix} \quad \nabla^2 f(x) = \begin{pmatrix} -12x_1^2 & 0 \\ 0 & -2 \end{pmatrix}$$

$$\bar{x} = (0, 2/5) \quad \nabla f(\bar{x}) = \begin{pmatrix} 0 \\ -4/5 \end{pmatrix} \quad \nabla^2 f(\bar{x}) = \begin{pmatrix} 0 & 0 \\ 0 & -2 \end{pmatrix}$$

$$\begin{aligned} f(\bar{x} + h) &= f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) + \frac{1}{2} (x - \bar{x})^T \nabla^2 f(\bar{x}) (x - \bar{x}) = \\ &= -\frac{4}{25} - \frac{4}{5} \left(x_2 + \frac{2}{5}\right) - 2 \left(x_2 + \frac{2}{5}\right)^2 = \\ &= -\frac{12}{25} - \frac{4}{5} x_2 - 2x_2^2 - \frac{8}{25} - \frac{8}{5} x_2^2 = \\ &= \frac{20}{25} - \frac{12}{5} x_2 - 2x_2^2 \end{aligned}$$

**TODO**

*Magari mettere un bel grafico 3d per evidenziare il comportamento di Taylor.*





# 3 Ottimizzazione: regione ammissibile e condizioni di ottimalità

## TODO

Dare un nome più sensato al capitolo ed alle sezioni.

$$f : \mathbb{R}^n \rightarrow \mathbb{R}, D \subseteq \mathbb{R}^n$$

$$(P) \quad \min\{f(x) : x \in D\}$$

Problema: trovare  $\bar{x} \in \mathbb{R}^n$  tale che:

- $\bar{x} \in D$
- $f(\bar{x}) \leq f(x) \quad \forall x \in D$

### Classificazione della regione ammissibile

$$\begin{cases} D = \mathbb{R}^n & \text{Ottimizzazione non vincolata} \\ D \subset \mathbb{R}^n & \text{Ottimizzazione vincolata} \end{cases}$$

### Classificazione della funzione obiettivo

$$f \text{ obiettivo} : \begin{cases} \text{lineare} \\ \text{non lineare} \end{cases}$$

### Definizione 3.1 (Punto di minimo globale)

$\bar{x} \in \mathbb{R}^n$  si dice punto di minimo globale di (P) se

- $x \in D$
- $f(\bar{x}) \leq f(x) \quad \forall x \in D$

In questo caso  $f(\bar{x})$  si dice *valore ottimo* di (P)

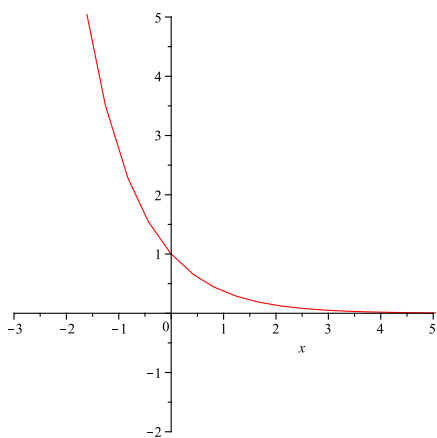


### Osservazione 3.2

*Punto di minimo e inf non sono la stessa cosa!*

*Ad esempio  $f(x) = e^{-x}$  ha estremo inferiore ma non ha minimo. Quindi possiamo affermare che il valore ottimo esiste (inteso come  $\inf\{f(x) : x \in D\}$ ), ma non esiste alcun punto di minimo.*

*Prendiamo invece la funzione  $f(x) = -x^2$ : il valore ottimo non esiste in quanto  $f(x) \rightarrow -\infty$ .*

Figura 3.1:  $f(x) = e^{-x}$ **Definizione 3.3 (Punto di minimo locale)**

Un  $\bar{x}$  si dice punto di minimo locale se

- $\bar{x} \in D$
- $\exists \varepsilon > 0 \mid f(\bar{x}) \leq f(x) \quad \forall x \in D \cap B(\bar{x}, \varepsilon)$

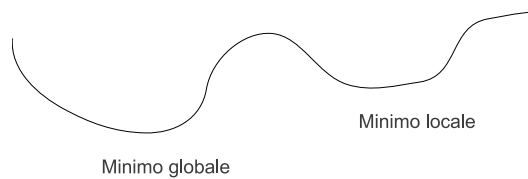


Figura 3.2: Minimo locale e globale

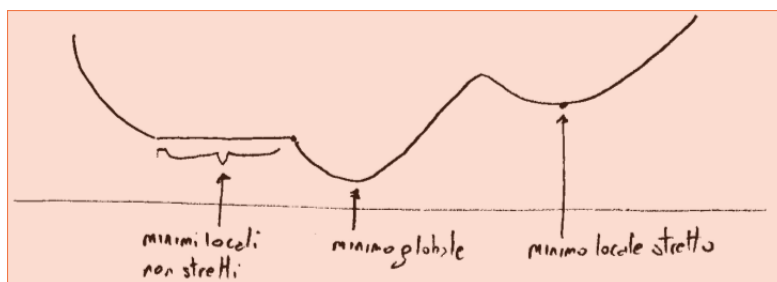


Figura 3.3: Minimo locale e globale

**Definizione 3.4 (Punto di minimo locale stretto)**

Un  $\bar{x}$  si dice punto di minimo locale stretto di  $(P)$  se

- $\bar{x} \in D$
- $\exists \varepsilon > 0 \mid f(\bar{x}) < f(x) \quad \forall x \in D \cap B(\bar{x}, \varepsilon), x \neq \bar{x}$

Cioè il minimo è unico nell'intervallo selezionato.

**Definizione 3.5 (Funzione convessa)**

Una  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  si dice convessa se

$$\forall x, y \in \mathbb{R}^n, \forall \lambda \in [0, 1] \quad \underbrace{f(\lambda x + (1 - \lambda)y)}_{\text{segmento } [x, y] \subseteq \mathbb{R}^n} \leq \underbrace{\lambda f(x) + (1 - \lambda)f(y)}_{\text{segmento } [f(x), f(y)] \subseteq \mathbb{R}} \quad (3.1)$$

**Definizione 3.6 (Funzione concava)**

Una  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  si dice concava se la funzione  $-f$  è convessa.

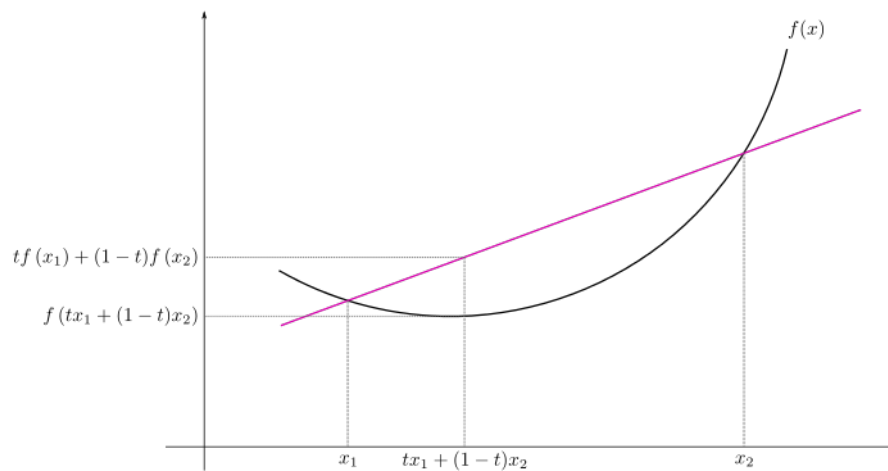


Figura 3.4: Funzione convessa

Si parla di *funzioni strettamente convesse (concave)* se nella (3.1) al posto di  $\leq$  sostituiamo  $<$  (simmetrico nel caso convesso).

 **Osservazione 3.7**

$$f \text{ è convessa} \iff f\left(\sum_{i=1}^k \lambda_i x_i\right) \leq \sum_{i=1}^k \lambda_i f(x_i) \quad \forall x_1, \dots, x_k \in \mathbb{R}^n, \forall \lambda_i \geq 0 \text{ con } \sum_{i=1}^k \lambda_i = 1$$

 **Teorema 3.8 (Locale  $\equiv$  globale)**

Sia  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  convessa,  $D \subseteq \mathbb{R}^n$  convesso. Allora ogni punto di minimo locale di  $(P)$  è anche un punto di minimo globale.

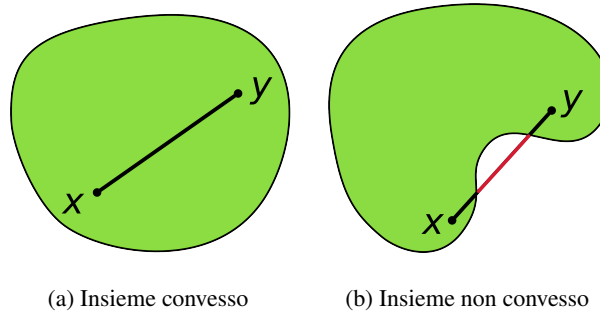


Figura 3.5: Insiemi convessi e non convessi in  $\mathbb{R}^2$

Cioè nei problemi di ottimizzazione, se la regione ammissibile è convessa, e la funzione obiettivo è convessa, non c'è distinzione fra minimo locale e minimo globale.

*Dimostrazione.*

Sia  $\bar{x} \in D$  un minimo locale. Supponiamo, per assurdo, che non sia un punto di minimo globale. Allora  $\exists \hat{x}$  tale che  $f(\hat{x}) < f(\bar{x})$ .

Consideriamo il punto  $x_\lambda = \lambda x + (1 - \lambda)\hat{x}$  con  $\lambda \in [0, 1]$ ,  $x_\lambda \in D$

Quanto vale la funzione in  $x_\lambda$ ?

$$f(x_\lambda) \underset{\text{convessità}}{\leq} \lambda f(x) + (1 - \lambda)f(\hat{x}) < \lambda f(\bar{x}) + (1 - \lambda)f(\bar{x}) = f(\bar{x})$$

$$x_\lambda \xrightarrow{\lambda \rightarrow 0} \bar{x} \quad \forall \epsilon \exists \bar{\lambda} \in [0, 1] \quad \text{t.c.} \quad x_{\bar{\lambda}} \in B(\bar{x}, \epsilon)$$

ed inoltre

$$f(x_{\bar{\lambda}}) < f(\bar{x})$$

Quindi  $\bar{x}$  non è un minimo locale (contraddizione)  $\square$

### Definizione 3.9 (Insieme convesso)

$D \subseteq \mathbb{R}^n$  si dice convesso se  $\forall x, y \in D \forall \lambda \in [0, 1]$ :

$$\underbrace{\lambda x + (1 - \lambda)y}_{\text{segmento } [x,y] \subseteq \mathbb{R}^n} \in D$$

### Osservazione 3.10

Sono insiemi convessi

- $\mathbb{R}^n$
- $\emptyset$

## 3.0.5 Risultati importanti su insiemi convessi e funzione convesse

### Teorema 3.11

Sia  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  strettamente convessa,  $D \subseteq \mathbb{R}^n$  convesso.

Se  $(P)$  ammette un punto di minimo, allora è l'unico punto di minimo.

*Dimostrazione.* Sia  $\bar{x} \in D$  minimo (globale  $\equiv$  locale) di (P) e supponiamo esista  $\hat{x} \in D$ ,  $\hat{x} \neq \bar{x}$  tale che  $f(\hat{x}) = f(\bar{x})$ . Sia  $\lambda \in [0, 1]$ :  $\lambda\hat{x} + (1 - \lambda)\bar{x} \in D$  (convessità di D) e

$$f(\lambda\hat{x} + (1 - \lambda)\bar{x}) < \lambda f(\hat{x}) + (1 - \lambda)f(\bar{x}) = \lambda f(\bar{x}) + (1 - \lambda)f(\bar{x}) = f(\bar{x})$$

Quindi  $\bar{x}$  non è un minimo (globale  $\equiv$  locale): contraddizione!  $\square$



### Nota

Ricordiamo che nel caso  $D = \mathbb{R}^n$ ,  $D$  è convesso, quindi abbiamo automaticamente le proprietà sopra citate.



### Nota

Ricordiamo inoltre che i poliedri sono un insieme convesso, anche le funzioni lineari, quindi nella programmazione lineare vale il primo teorema.



### Proprietà 3.1

$f$  convessa  $\implies f$  continua su  $\mathbb{R}^n$



### Proprietà 3.2

Sia  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ . Se  $f$  è convessa, allora l'insieme di sottolivello

$$C_\alpha = \{x \in \mathbb{R}^n \mid f(x) \leq \alpha\}$$

è convesso per ogni valore di  $\alpha$

*Dimostrazione.* Sia  $C_\alpha \neq \emptyset$  e siano  $x, y \in C_\alpha$ . Allora  $\forall \lambda \in [0, 1]$  risulta

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) \leq \lambda \alpha + (1 - \lambda)\alpha = \alpha$$

da cui  $\lambda x + (1 - \lambda)y \in C_\alpha$   $\square$



### Osservazione 3.12

Per vedere che quest'ultima proprietà vale solo in un senso, basta prendere la funzione  $x^3$ , che non è convessa. I sottolivelli sono tutti insiemi convessi.

$$C_\alpha = (-\infty, \sqrt[3]{\alpha})$$



### Proprietà 3.3

Siano  $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$  una famiglia di funzione convesse,  $k \in \mathbb{N}$  finito. Allora

1.  $\sum_{i=1}^k f_i$  è convessa
2.  $\sup_{i \in I} f_i$  è convessa

*Dimostrazione.*

1. Ovvio: basta applicare la definizione ad ogni  $f_i$  e sommare membro a membro
2. Nel caso  $I$  sia finito allora

$$(\sup f_i) = (\lambda x + (1 - \lambda)y) = f_k(\lambda x + (1 - \lambda)y) \leq f_k(x) + (1 - \lambda)f_k(y) \leq \lambda(\sup f_i)(x) + (1 - \lambda)(\sup f_i)(y)$$

per un  $k$  opportuno

□



### Teorema 3.13

Sia  $f$  differenziabile (su  $\mathbb{R}^n$ )

$$f \text{ è convessa} \iff f(y) \geq f(x) + \underbrace{\nabla f(x)^T (y - x)}_{\text{piano tangente}} \quad \forall x, y \in \mathbb{R}^n$$

Informalmente: il grafico di  $f$  sta sopra il piano tangente. Ricordiamo che  $\|x\|$  è convessa ma non è differenziabile

*Dimostrazione.*  $\implies$

$x, y \in \mathbb{R}^n, \lambda \in [0, 1]$

$$\begin{aligned} \lambda f(y) + (1 - \lambda)f(x) &\geq f(\lambda y + (1 - \lambda)x) && \implies \\ \lambda f(y) - \lambda f(x) &\geq f(\lambda y + (1 - \lambda)x) - f(x) && \implies \\ f(y) - f(x) &\geq \frac{f(\lambda x + (1 - \lambda)y) - f(x)}{\lambda} \xrightarrow{\lambda \rightarrow 0} \nabla f(x)^T (y - x) \end{aligned}$$

$\longleftarrow$

$x, y \in \mathbb{R}^n, \lambda \in [0, 1]$

$$f(x) - f(\lambda y + (1 - \lambda)x) \geq \nabla f(\lambda y + (1 - \lambda)x)^T [\lambda(x - y)] \quad (3.2)$$

$$f(y) - f(\lambda y + (1 - \lambda)x) \geq \nabla f(\lambda y + (1 - \lambda)x)^T [(\lambda - 1)(x - y)] \quad (3.3)$$

$$(1 - \lambda)(3.2) + \lambda(3.3) \implies (1 - \lambda)f(x) + \lambda f(y) \geq (1 - \lambda)f(\lambda y + (1 - \lambda)x) + \lambda f(\lambda y + (1 - \lambda)x) = f(\lambda y + (1 - \lambda)x)$$

che è la disuguaglianza di convessità □

### TODO

Ho lasciato commentato la dimostrazione presa a lezione di quest'ultimo teorema perchè più verbosa. Vedere se ha senso fare un'integrazione con quella attuale.



### Teorema 3.14

Sia  $f$  differenziabile 2 volte (su  $\mathbb{R}^n$ ). Allora

$$f \text{ è convessa} \iff \nabla^2 f(x) \text{ è semidefinita positiva} \quad \forall x \in \mathbb{R}^n, \text{ ovvero } y^T \nabla^2 f(x) y \geq 0$$

(Matrice Hessiana)

*Dimostrazione.*

$\Leftarrow x, y \in \mathbb{R}^n, \lambda \in \mathbb{R}$

$$f(x + \lambda y) - \underbrace{f(x)} - \nabla f(x)^T y \geq 0 \quad (\text{Teorema (3.13)}) \quad (3.4)$$

Sfruttando Taylor di secondo ordine

$$f(x) = \frac{1}{2} \lambda^2 y^T \nabla^2 f(x) y + r(\lambda y) \quad (3.5)$$

1. dividendo (3.5) per  $\lambda^2$
2. utilizzando la relazione  $\|\lambda y\|_2^2 = \lambda^2 \|y\|_2^2 = \lambda^2$
3. mettendo insieme (3.4) e (3.5)
4.  $\frac{r(h)}{\|h\|_2} \xrightarrow{\|h\|_2 \rightarrow 0} 0$

otteniamo

$$\frac{1}{2} y^T \nabla^2 f(x) y + \frac{r(\lambda y)}{\lambda^2} \geq 0 \xrightarrow{\lambda \rightarrow 0} \frac{1}{2} y^T \nabla^2 f(x) y \geq 0 \implies y^T \nabla^2 f(x) y \geq 0$$

che era quello che volevamo dimostrare.

$\implies$  Siano  $x, y \in \mathbb{R}^n$ .

$$f(y) - f(x) - \nabla f(x)^T (y - x) \stackrel{\text{Taylor}}{=} \frac{1}{2} (y - x)^T \nabla^2 f(x + t(y - x)) (y - x) \stackrel{\text{ipotesi}}{\geq} 0$$

Questo vale per un  $t \in (0, 1)$  opportuno. Dal teorema (3.13) otteniamo immediatamente che  $f$  è convessa, che era la nostra tesi.

□

### Osservazione 3.15 (Casi particolari dei precedenti teoremi/proposizioni)

$$f \text{ concava} \begin{cases} \text{Teorema (3.13) con } f(x) \leq f(x) + \nabla f(x)^T (y - x) \\ \text{Teorema (3.14) con } \nabla^2 f(x) \text{ semidefinita negativa (} y^T \nabla f(x) y \leq 0 \text{)} \end{cases}$$

$$f \text{ strettamente convessa} \begin{cases} \text{Teorema (3.13) con } f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad (y \neq x) \\ \text{Teorema (3.14) vale la parte necessaria con } \nabla^2 f(x) \text{ semidefinita positiva} \\ \text{[e non definita positiva]: } \nabla^2 f(x) \implies f \text{ strettamente convesso} \end{cases}$$

Ad esempio:  $n = 1, f(x) = x^4$  è strettamente convessa ma  $\nabla^2 f(0) = 0$  [ $\nabla^2 f(x) = 12x^2$ ]

### Definizione 3.16 (Epigrafico)

$$\text{epi}(f) = \{(x, t) \mid x \in \mathbb{R}^n, t \geq f(x)\} \subseteq \mathbb{R}^{n+1}$$

### Proprietà 3.4

Sia  $f : \mathbb{R}^n \rightarrow \mathbb{R}$

$$f \text{ è convessa} \iff \text{epi}(f) \text{ è convesso}$$

*Dimostrazione.*

$\implies$  Siano  $(x, t), (y, \tau) \in \text{epi}(f), \lambda \in [0, 1]$ :

$$\lambda t + (1 - \lambda)\tau \geq \lambda f(x) + (1 - \lambda)f(y) \stackrel{\text{convessità}}{\geq} f(\lambda x + (1 - \lambda)y)$$

da cui

$$(\lambda x + (1 - \lambda)y, \lambda t + (1 - \lambda)\tau) \in \text{epi}(f)$$

$\Leftarrow$  Siano  $x, y \in \mathbb{R}^n, \lambda \in [0, 1] : (x, f(x)), (y, f(y)) \in \text{epi}(f)$

$$\text{epi}(f) \text{ convesso} \implies (\lambda x + (1 - \lambda)y, \lambda f(x) + (1 - \lambda)f(y)) \in \text{epi}(f)$$

ovvero

$$\lambda f(x) + (1 - \lambda)f(y) \geq f(\lambda x + (1 - \lambda)y)$$

□



### Nota

Tali teoremi possono riscritti per le funzioni convesse

### TODO

Scrivere bene proprietà per funzioni convesse e strettamente convesse

$$f(x) = x^4 \quad \nabla f(x) = 4x^3 \quad \nabla^2 f(x) = 12x^2$$

La matrice hessiana non è definita positiva.

$$f(x) = \frac{1}{2}x^T Qx + b^T x + c$$

$$\nabla f(x) = Qx + b$$

$$\nabla^2 f(x) = Q$$

### Proposizione 3.1

Siano  $f$  convessa,  $D$  convesso. Allora l'insieme dei punti di minimo di  $(P)$  è convesso.

Infatti  $\bar{x}, \hat{x} \in D$  minimi  $f(\bar{x}) = f(\hat{x})$ .

$$f(x_\lambda) \leq \lambda f(\bar{x}) + (1 - \lambda) \underbrace{f(\hat{x})}_{=f(\bar{x})} = f(\bar{x}) \quad \rightarrow \quad f(x_\lambda) = f(\bar{x})$$

Cioè  $x_\lambda$  è minimo.

### Funzioni quadratiche e convessità

$Q \in \mathbb{R}^{n \times n}, b \in \mathbb{R}^n, Q$  simmetrica,  $c \in \mathbb{R}$

$$f(x) = \frac{1}{2}x^T Qx + b^T x + c \implies f \text{ differenziabile 2 volte: } \nabla^2 f(x) \equiv Q \quad \forall x$$

Dal teorema (3.14) seguono:



 **Proprietà 3.5**

$f$  è convessa  $\iff Q$  è semidefinita positiva, ossia gli autovalori di  $Q$  sono tutti  $\geq 0$

 **Proprietà 3.6**

$Q$  definita positiva  $\implies f$  è strettamente convessa

## 3.1 Ottimizzazione non vincolata

$D = \mathbb{R}^n$  ottimizzazione non vincolata:  $(P) \quad \min\{f(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n\}$

### 3.1.1 Condizioni di ottimalità

**Teorema 3.17 (Condizioni necessarie del primo e del secondo ordine)**

Sia  $\bar{\mathbf{x}} \in \mathbb{R}^n$  un punto di minimo locale di  $(P)$ .

1. Se  $f$  è differenziabile in  $\bar{\mathbf{x}}$ , allora  $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$
2. Se  $f$  è differenziabile 2 volte in  $\bar{\mathbf{x}}$ , allora  $\nabla^2 f(\bar{\mathbf{x}})$  (la matrice Hessiana) è semidefinita positiva.

*Dimostrazione.* Sia  $\epsilon > 0$  tale che  $f(\bar{\mathbf{x}}) = \min\{f(\mathbf{x}) : \mathbf{x} \in B(\bar{\mathbf{x}}, \epsilon)\}$ , e siano  $\mathbf{d} \in \mathbb{R}^n$  con  $\|\mathbf{d}\|_2 = 1$  una direzione e  $t \leq \epsilon$ :

1. Per ipotesi  $\bar{\mathbf{x}}$  è punto di minimo locale, quindi

$$\exists \epsilon > 0. \quad f(\bar{\mathbf{x}}) \leq f(\mathbf{x}) \quad \forall \mathbf{x} \in B(\bar{\mathbf{x}}, \epsilon)$$

sappiamo che vale la seguente relazione (Taylor):

$$0 \leq f(\bar{\mathbf{x}} + t\mathbf{d}) - f(\bar{\mathbf{x}}) = t\nabla f(\bar{\mathbf{x}})^T \mathbf{d} + r(t\mathbf{d})$$

da cui

$$0 \leq \nabla f(\bar{\mathbf{x}})^T \mathbf{d} + \frac{r(t\mathbf{d})}{t} \xrightarrow{t \rightarrow 0} \nabla f(\bar{\mathbf{x}})^T \mathbf{d}$$

ovvero  $\nabla f(\bar{\mathbf{x}})^T \mathbf{d} \geq 0$  Analogamente, utilizzando la direzione  $(-\mathbf{d})$  si ottiene  $\nabla f(\bar{\mathbf{x}})^T \mathbf{d} \leq 0$ . Quindi

$$\nabla f(\bar{\mathbf{x}})^T \mathbf{d} = 0 \quad \forall \mathbf{d} \in \mathbb{R}^n$$

(con  $\|\mathbf{d}\|_2 = 1$ ), da cui  $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$  (considerare  $\mathbf{d} = \nabla f(\bar{\mathbf{x}})$ )

2. La prima parte della dimostrazione è analoga.

Riscriviamo lo sviluppo di Taylor, questa volta del secondo ordine. (La funzione è differenziabile 2 volte)

$$0 \leq f(\bar{\mathbf{x}} + t\mathbf{d}) - f(\bar{\mathbf{x}}) = \underbrace{t\nabla f(\bar{\mathbf{x}})^T \mathbf{d}}_{(*)} + \frac{1}{2}t^2 \mathbf{d}^T \nabla^2 f(\bar{\mathbf{x}}) \mathbf{d} + r(t\mathbf{d}) = \frac{1}{2}t^2 \mathbf{d}^T \nabla^2 f(\bar{\mathbf{x}}) \mathbf{d} + r(t\mathbf{d})$$

Dividendo per  $t^2$  otteniamo:

$$0 \leq \frac{f(\bar{\mathbf{x}} + t\mathbf{d})}{t^2} = \frac{1}{2} \underbrace{\mathbf{d}^T \nabla^2 f(\bar{\mathbf{x}}) \mathbf{d}} + \frac{r(t\mathbf{d})}{t^2}$$

Facendo tendere  $t$  a 0 otteniamo

$$t \rightarrow 0 \quad \implies \quad 0 \leq \frac{1}{2} \mathbf{d}^T \nabla^2 f(\bar{\mathbf{x}}) \mathbf{d} + \cancel{\frac{r(t\mathbf{d})}{t^2}}$$

Concludiamo quindi che

$$\mathbf{d}^T \nabla^2 f(\bar{\mathbf{x}}) \mathbf{d} \geq 0 \quad \forall \mathbf{d} \in \mathbb{R}^n, \|\mathbf{d}\|_2 = 1$$

L'ultima è la definizione di matrice semidefinita positiva.



Nota

(\*) Questa quantità sparisce per il punto 1

□

**Teorema 3.18 (Caso Convesso)**

Sia  $f$  una funzione convessa e differenziabile su  $\mathbb{R}^n$ . Allora

$$x \in \mathbb{R}^n \text{ è un di minimo (locale e globale) di } (P) \iff \nabla f(\bar{x}) = 0$$

*Dimostrazione.*  $\Rightarrow$ : dimostrato dal teorema precedente

$\Leftarrow$ :

$$f \text{ convessa} \implies f(y) \geq f(\bar{x}) + \underbrace{\nabla f(\bar{x})^T (y - x)}_{\text{questo pezzo sparisce}} \quad \forall y \in \mathbb{R}^n$$

$$f(\bar{x}) = 0 \implies f(y) \geq f(\bar{x}) \quad \forall y \in \mathbb{R}^n \implies \bar{x} \text{ è un punto di minimo}$$

□

**Esempio 3.19**

$$f(x_1, x_2) = (x_2 - x_1^2)(x_2 - 4x_1^2) [= x_2^2 - 5x_1^2x_2 + 4x_1^4]$$

$$\nabla f(\mathbf{x}) = \begin{pmatrix} -10x_1x_2 + 16x_1^3 \\ 2x_2 - 5x_1^2 \end{pmatrix}$$

$$\nabla f(\mathbf{x}) = 0 \iff \begin{cases} 2x_2 = 5x_1^2 \\ 16x_1^3 = 10x_1x_2 \end{cases} \iff x_1 = x_2 = 0$$

Il gradiente si annulla se  $\mathbf{x}$  è il vettore nullo.

Vediamo la matrice Hessiana:

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} -10x_2 + 48x_1^2 & -10x_1 \\ -10x_1 & 2 \end{pmatrix}$$

Calcoliamo in  $(0, 0)$ .

$$\nabla^2 f((0, 0)) = \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix}$$

Semidefinita positiva, ma non è definita positiva (in quanto ha uno zero nella diagonale)

Calcoliamo in  $(0, 1)$ .

$$\nabla^2 f((0, 1)) = \begin{pmatrix} -10 & 0 \\ 0 & 2 \end{pmatrix}$$

Questa non è definita positiva (negativa) quindi  $f$  non è convessa.  $f(0, 0) = 0$

$$P = \{x_2 = 2x_1^2\} \quad f(x_1, 2x_1^2) = -2x_1^4 < 0 \quad \text{se } x_1 \neq 0$$

$f$  è negativa su  $P \setminus \{\bar{x}\}$ :  $x_k = (\frac{1}{k}, \frac{2}{k^2})$   $x_k \rightarrow \bar{x}$  e  $f(x_k) = \frac{-2}{k^4} < 0$

Possiamo concludere che  $\bar{\mathbf{x}} = (0, 0)$  non è un minimo locale.

**Definizione 3.20 (Punto Stazionario)**

$\bar{x} \in \mathbb{R}^n$  si dice punto stazionario per  $f$  se  $\nabla f(\bar{x}) = 0$

**Teorema 3.21 (Condizione sufficiente)**

Sia  $f$  differenziabile 2 volte in  $\bar{x} \in \mathbb{R}^n$  e valga  $\nabla f(\bar{x}) = 0$ . (Punto stazionario)

Se  $\nabla^2 f(\bar{x})$  è definita positiva, allora  $\bar{x}$  è un punto di minimo locale stretto di  $(P)$ , ed inoltre esistono  $\gamma > 0$  e  $\delta > 0$  tali che:

$$f(x) \geq f(\bar{x}) + \gamma \|x - \bar{x}\|_2^2 \quad \forall x \in B(\bar{x}, \delta)$$

*Dimostrazione.*

**Nota**

La dimostrazione data a lezione è lievemente diversa, ma solo per la notazione: lascio commentata nel sorgente Tex

Sia  $x \in \mathbb{R}^n$

$$\begin{aligned} f(x) - f(\bar{x}) &= \\ \nabla f(\bar{x})^T (x - \bar{x}) &+ \frac{1}{2} (x - \bar{x})^T \nabla^2 f(\bar{x}) (x - \bar{x}) + r(x - \bar{x}) \quad \square_{*}) \\ &= \frac{1}{2} (x - \bar{x})^T \nabla^2 f(\bar{x}) (x - \bar{x}) + r(x - \bar{x}) \end{aligned}$$

Sia  $\lambda_{min} > 0$  il più piccolo autovalore di  $\nabla^2 f(\bar{x}) (\Rightarrow y^T \nabla^2 f(\bar{x}) y \geq \lambda_{min} \|y\|_2^2 \quad \forall y \in \mathbb{R}^n)$

$$f(x) - f(\bar{x}) \geq \frac{\lambda_{min}}{2} \|x - \bar{x}\|_2^2 + r(x - \bar{x})$$

$$\frac{f(x) - f(\bar{x})}{\|x - \bar{x}\|_2^2} \geq \frac{\lambda_{min}}{2} + \frac{r(x - \bar{x})}{\|x - \bar{x}\|_2^2} \xrightarrow{x \rightarrow \bar{x}} \frac{\lambda_{min}}{2}$$

Sia  $0 < \epsilon < \frac{\lambda_{min}}{2}$ : per la definizione di limite esiste  $\delta > 0$  tale che

$$\frac{f(x) - f(\bar{x})}{\|x - \bar{x}\|_2^2} \geq \left(\frac{\lambda_{min}}{2} - \epsilon\right) = \gamma \quad \forall x \in B(\bar{x}, \delta)$$

da cui

$$f(x) \geq f(\bar{x}) + \gamma \|x - \bar{x}\|_2^2 \quad \forall x \in B(\bar{x}, \delta)$$

\*)  $\nabla f(\bar{x}) = 0$

□

**Osservazione 3.22 (Risultati per i punti di massimo)**

$$\max\{f(x) : x \in \mathbb{R}^n\} = -\min\{-f(x) : x \in \mathbb{R}^n\}$$

Quindi per le condizioni di massimo non dobbiamo dimostrare nulla.

Cambia il punto (2) (semidefinita positiva) che diventa semidefinita negativa. Per i punti di massimo locale la matrice deve essere la matrice definita negativa, otterremo un punto di massimo locale stretto.

 **Osservazione 3.23**

Ad eccezione di quelle costanti, le funzioni convesse (differenziabili) non ammettono punti di massimo locale/globale su  $\mathbb{R}^n$ :

$$\bar{x} \text{ è massimo di } f \implies \nabla f(\bar{x}) = 0$$

ma

$$f \text{ convessa} \wedge \nabla f(\bar{x}) = 0 \implies \bar{x} \text{ minimo di } f$$

Abbiamo un punto di massimo e di minimo, quindi  $f$  è costante.

**Esercizio 3.1**

$$\nabla f \neq 0 \quad \nabla f(x) \neq 0 \quad y(t) = t\nabla f(x) + x$$

Dimostrare che  $f(y(t)) \rightarrow_{t \rightarrow +\infty} +\infty$

Suggerimento : usare

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

 **Osservazione 3.24 (Funzioni Quadratiche)**

La funzioni quadratiche non convesse

$$f(x) = \frac{1}{2}x^T Qx + b^T x + c \quad Q \in \mathbb{R}^{n \times n} \text{ simmetrica, } b \in \mathbb{R}^n, c \in \mathbb{R}$$

non ammettono minimo locale in quanto  $\nabla^2 f(x) \equiv Q$  non è definita positiva

**Legami con l'analisi numerica**

**Funzioni quadratiche** Chiedere ad una funzione quadratica che

$$\nabla f(x^*) = 0 \iff Qx^* = -b$$

è la stessa cosa di risolvere il seguente sistema lineare:

$$Qx = -b$$

Questo è un primo legame con l'analisi numerica.

**Caso generale**

$$\nabla f : \mathbb{R}^n \rightarrow \mathbb{R}^n$$

In generale il gradiente è una funzione non lineare. Trovare i punti stazionari è sostanzialmente risolvere un sistema di equazioni non lineari.

$$\begin{cases} \frac{\partial f}{\partial x_1}(x) = 0 \\ \frac{\partial f}{\partial x_2}(x) = 0 \\ \dots \\ \frac{\partial f}{\partial x_n}(x) = 0 \end{cases}$$

**Condizioni del secondo ordine** Altro legame con analisi numerica è il calcolo di autovalori: infatti verificare se  $\nabla^2$  è (semi)definita positiva/negativa richiede il calcolo del più grande/piccolo autovalore di  $\nabla^2 f(\bar{x})$

**Nota**

Nel caso quadratico la matrice  $\nabla^2 f(x)$  è nota una volta nota

**TODO**

cosa??

la funzione

Per il calcolo di  $\nabla f$ ,  $\nabla^2$  possiamo usare

- differenze finite
- differenziazione automatica



## 4 Risoluzione di Sistemi Lineari

Siano  $A \in \mathbb{C}^{n \times n}$ ,  $x, b \in \mathbb{C}^n$  e si supponga che il sistema lineare

$$Ax = b$$

sia consistente (ovvero, ammetta almeno una soluzione).

I metodi per risolvere il sistema si possono dividere in

- metodi diretti:
  - Gauss
  - Cholesky (Hermitiane definite positive)
  - Householder (Maggiori garanzie di stabilità)
- Metodi iterativi



### Nota

È possibile trattare tutti i metodi diretti come fattorizzazioni della matrice dei coefficienti della matrice lineare.

### 4.1 Propagazione dell'errore

In un metodo diretto, se non ci fossero errori di rappresentazione dei dati e di arrotondamento nei calcoli, la soluzione del sistema verrebbe calcolata esattamente. Invece in un metodo iterativo, anche nell'ipotesi che non ci siano errori di rappresentazione dei dati e di arrotondamento nei calcoli, si deve comunque operare un troncamento del procedimento, commettendo un errore, detto *errore analitico*.

Una maggiorazione dell'errore da cui è affetta la soluzione può essere rappresentata da due termini distinti:

- l'*errore inerente*, dovuto agli errori di rappresentazione dei dati, che non dipendono dal particolare metodo usato, e
- l'*errore algoritmico*, dovuto agli errori di arrotondamento nei calcoli, che dipende dal metodo usato, ma non dagli errori sui dati.

Lo studio dell'errore inerente può essere fatto perturbando i dati ed esaminando gli effetti indotti sulla soluzione.



#### Teorema 4.1 (Numero di condizionamento di $A$ )

Siano  $\delta A \in \mathbb{C}^{n \times n}$  e  $\delta b \in \mathbb{C}^n$  rispettivamente la matrice e il vettore delle perturbazioni sui dati del sistema, dove  $b \neq 0$  e sia  $\|\cdot\|$  una qualunque norma indotta. Se  $A$  non è singolare e se  $\|A^{-1}\| \|\delta A\| < 1$  allora anche la matrice  $A + \delta A$  è non singolare. Indicata con  $x + \delta x$  la soluzione del sistema perturbato

$$(A + \delta A)(x + \delta x) = b + \delta b$$

risulta

$$\frac{\|\delta x\|}{\|x\|} \leq \mu(A) \frac{\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}}{1 - \underbrace{\mu(A) \frac{\|\delta A\|}{\|A\|}}_{<1}}$$

in cui  $\mu(A) = \|A\| \|A^{-1}\|$  è detto numero di condizionamento della matrice  $A$

Si osservi che il numero di condizionamento è sempre maggiore o uguale a 1. Inoltre se  $\mu(A)$  assume valori piccoli, allora piccole perturbazioni sui dati inducono piccole perturbazioni sulla soluzione e la matrice del sistema è ben condizionata.

Un metodo risulta più *stabile* di un altro se è meno sensibile agli errori indotti dai calcoli. Lo studio della stabilità di un metodo perde di significatività quando il problema è fortemente mal condizionato, poiché in questo caso l'errore inerente prevale sull'errore algoritmico.

## 4.2 Fattorizzazione

I metodi diretti che vedremo utilizzano una fattorizzazione della matrice  $A$ , nel prodotto di due matrici  $B$  e  $C$ , facilmente invertibili.

$$A = BC$$

Possiamo dunque riscrivere il sistema, e la sua soluzione si riduce alla risoluzione di due sottoproblemi:

$$BCx = b \quad \begin{cases} By = b \\ Cx = y \end{cases}$$

$B, C$  possono appartenere alle seguenti classi:

- Triangolari (Si risolve per sostituzione all'indietro).  $O(n^2)$  operazioni.
- Unitarie (Sono facili da invertire). Infatti  $Qx = b \rightarrow x = Q^H b$ .  $O(n^2)$  operazioni.

Vedremo le seguenti 3 fattorizzazioni classiche.

$$A = LU$$

- $L$  è triangolare inferiore (Lower) con diagonale unitaria ( $l_{ii} = 1$ )
- $U$  è triangolare superiore (Upper)
- associata al metodo di Gauss.

$$A = LL^H$$

- $L$  è triangolare inferiore con elementi positivi sulla diagonale ( $\mathbb{R} \ni l_{ii} > 0$ )
- associata al metodo di Cholesky.

$$A = QR$$

- $Q$  unitaria
- $R$  triangolare superiore
- associata al metodo di Householder.

Se la matrice  $A$  è reale, le matrici delle tre fattorizzazioni, quando esistono, sono reali.

Il costo computazionale della fattorizzazione è  $O(n^3)$  operazioni, mentre il costo computazionale della risoluzione dei sistemi è  $O(n^2)$  operazioni.

La fattorizzazione  $QR$  esiste per ogni matrice  $A$ , mentre non sempre è possibile ottenere le fattorizzazioni  $LU$  e  $LL^H$ . Valgono infatti i seguenti teoremi.

### Fattorizzazione LU



 **Teorema 4.2 (Esistenza della fattorizzazione  $A = LU$ )**

Sia  $A$  una matrice di ordine  $n$  e siano  $A_k$  le sue sottomatrici principali di testa di ordine  $k$ . Se  $A_k$  è non singolare per  $k = 1, \dots, n-1$  allora esiste ed è unica la fattorizzazione  $LU$  di  $A$ .

*Dimostrazione.* Per induzione su  $n$

- $n = 1$

$$\begin{array}{ccc} a_{11} & = & 1 \cdot a_{11} \\ A & & L \quad U \end{array}$$

(oppure  $a_{11}$  potrebbe essere uguale a 0).

- $n > 1$

$$A = \left| \begin{array}{c|c} A_{n-1} & \mathbf{d} \\ \mathbf{c}^H & \alpha \end{array} \right| = \left| \begin{array}{c|c} L_{n-1} & \mathbf{0} \\ \mathbf{u}^H & 1 \end{array} \right| \cdot \left| \begin{array}{c|c} U_{n-1} & \mathbf{v} \\ \mathbf{0} & \beta \end{array} \right|$$

Otteniamo le seguenti equazioni:

$$\begin{cases} A_{n-1} = L_{n-1} \cdot U_{n-1} & (1) \\ d = L_{n-1} \cdot v & (2) \\ c^H = u^H \cdot U_{n-1} & \rightarrow c = U_{n-1}^H \cdot u & (3) \\ \alpha = u^H \cdot v + \beta & (4) \end{cases}$$

Da cui traiamo le seguenti:

- (1) essendo  $L_{n-1}$  triangolare superiore  $\det(L_{n-1}) = \prod_{i=0}^{n-1} l_{ii} = 1$  allora  $\det(U_{n-1}) = \det(A_{n-1})$  che per ipotesi è diverso da zero. Quindi esistono le inverse di  $L$  e  $U$  e i seguenti valori sono determinati univocamente.
- (2)  $v = L_{n-1}^{-1} \cdot d$
- (3)  $u = (U_{n-1}^H)^{-1} \cdot c$
- (4)  $\beta = \alpha - u^H \cdot v$

□

Se cade l'ipotesi del teorema perdiamo l'unicità, possiamo non avere alcuna fattorizzazione o averne più di una.

**Esempio 4.3 (Fattorizzazione non unica)**

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 2 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 \\ l & 1 \end{pmatrix}}_L \cdot \underbrace{\begin{pmatrix} 0 & 1 \\ 0 & u \end{pmatrix}}_U$$

esistono infinite combinazioni per cui vale  $l + u = 2$ .

 **Teorema 4.4**

Per ogni matrice  $A$ , esiste una matrice di permutazione  $\Pi$  per cui si può ottenere la fattorizzazione  $LU$  di  $\Pi A$ , cioè

$$\Pi A = LU$$

**Fattorizzazione  $LL^H$**


**Teorema 4.5 (Esistenza fattorizzazione  $LL^H$ )**

Sia  $A$  una matrice hermitiana definita positiva, allora esiste ed è unica la fattorizzazione  $LL^H$  di  $A$ .

*Dimostrazione.* con  $l_{ii} > 0$ .

Per il teorema ??,  $A$  ha le sottomatrici  $A_k$  con  $\det(A_k) > 0$  e quindi per il teorema 4.2 esiste unica la fattorizzazione (per evitare confusione rinominiamo  $L$  con  $M$ ):

$$A = M \cdot U$$

in cui ricordiamo che  $M$  è triangolare inferiore  $m_{ii} = 1$  e  $U$  è triangolare superiore. Possiamo scomporre ancora:

$$A = M \cdot \underbrace{D \cdot R}_U$$

Con  $D$  matrice diagonale i cui elementi principali sono quelli di  $U$ , e  $R$  triangolare superiore con  $r_{ii} = 1$ .

$$\underbrace{\begin{pmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{pmatrix}}_U = \underbrace{\begin{pmatrix} u_{11} & 0 \\ 0 & u_{22} \end{pmatrix}}_D \underbrace{\begin{pmatrix} 1 & \frac{u_{12}}{u_{11}} * \\ 0 & 1 \end{pmatrix}}_R$$

\* è possibile solo se  $\det(A) \neq 0$  ed è effettivamente così nel nostro caso, infatti  $\begin{cases} \det(U) \neq 0 \\ \det(A) > 0 \end{cases}$

Dalla proprietà delle hermitiane  $A^H = A$  otteniamo:

$$\underbrace{R^H}_L \underbrace{D^H M^H}_U = \underbrace{M}_L \underbrace{DR}_U$$

dato che abbiamo da entrambi i lati la fattorizzazione LU, dalla sua unicità otteniamo:

$$\begin{aligned} R^H = M &\Rightarrow M^H = R \\ DR = D^H M^H = D^H R &\Rightarrow D = D^H \end{aligned}$$

Sappiamo quindi che  $d_{ii} = \overline{d_{ii}}$ , quindi  $D \in \mathbb{R}$ .

Vogliamo ora dimostrare che  $D$  è hermitiana definita positiva, partiamo dalla definizione:

$$x^H D x = \underbrace{x^H M^{-1}}_{y^H} \underbrace{MDM^H}_A \underbrace{(M^H)^{-1} x}_y = y^H A y \underset{A \text{ def pos}}{\geq} 0$$

Quindi possiamo concludere che anche  $D$  è diagonale ed hermitiana definita positiva, ed avendo solo autovalori positivi, avrà  $d_{ii} > 0$ .

Esiste allora un'unica matrice diagonale  $F$  ad elementi principali reali e positivi, tale che  $D = F^2$ , e quindi  $F = \text{diag}\{\sqrt{d_{ii}}\}$ . Da cui possiamo ricavare:

$$A = MDR = \underbrace{MF}_L \underbrace{F^H M^H}_U = LL^H$$

Notare che la fattorizzazione di Cholesky è *caratterizzante*: se  $A$  è fattorizzabile  $LL^H$  allora è hermitiana definita positiva.  $\square$

**Fattorizzazione QR**

A differenza della fattorizzazione  $LU$  e  $LL^H$ , la fattorizzazione  $QR$  di una matrice  $A$  non è unica.


**Osservazione 4.6**

Se  $Q$  e  $R$  esistono, non sono uniche.

*Dimostrazione.* Se  $A = QR = Q \underbrace{DD^{-1}}_I R$

e scegliamo  $D$  come matrice diagonale e unitaria, quindi con  $|d_{ii}| = 1$ , abbiamo

$$DD^H = \begin{vmatrix} \ddots & & & \\ & d_{ii} \cdot \bar{d}_{ii} & & \\ & & \ddots & \\ & & & \ddots \end{vmatrix} = I$$

e possiamo scrivere

$$A = QR = \underbrace{QD}_{\text{unit}} \underbrace{D^H R}_{\text{tri. sup}} = Q_1 R_1$$

quindi abbiamo ottenuto un'altra fattorizzazione.

Le matrici  $D$  vengono dette *matrici di fase* ed solo con queste che la fattorizzazione non è unica, quindi possiamo dire che la fattorizzazione  $QR$  è unica a meno di matrici di fase.  $\square$

La determinazione delle matrici della fattorizzazione di  $A$  viene generalmente effettuata nei due modi seguenti:

- applicando alla matrice  $A$  una *successione di matrici elementari* (metodo di Gauss, metodo di Householder);
- con *tecniche compatte* (metodo di Cholesky).

### 4.3 Metodo di Gauss

Nel metodo si ricava una incognita da una delle equazioni e si sostituisce in tutte le altre, ripetendo il procedimento fino a trovare la soluzione (nella variante classica impone una certa strategia nella risoluzione). Il cambiamento della matrice dovuto alla sostituzione dei coefficienti può essere espresso come la moltiplicazione per una matrice  $E$ :

$$EA^{(1)} = A^{(2)}$$

Dove  $A^{(1)}$  e  $A^{(2)}$  sono la matrici prima e dopo la sostituzione. La matrice  $E$  è detta *elementare* ed ha diagonale unitaria e la prima colonna diversa da zero.

$$E = \begin{vmatrix} 1 & & & \mathbf{0} \\ x_2 & 1 & & \\ \vdots & & \ddots & \\ x_n & \mathbf{0} & & 1 \end{vmatrix}$$

Facendo i passi sequenti avremo:

$$E^{(S)} \dots E^{(2)} \cdot E^{(1)} \cdot A^{(1)} = U$$

con  $U$  triangolare superiore. Ora portando al secondo membro:

$$A = A^{(1)} = \underbrace{(E^{(S)} \dots E^{(1)})^{-1}}_L U$$

$$A^{(n)} = E^{(n-1)} \dots E^{(2)} E^{(1)} A$$

#### Definizione 4.7 (Matrice Elementare)

Siano  $\sigma \in \mathbb{R}$  e  $u, v \in \mathbb{C}^n$ ,  $u, v, \neq 0$ . Si definisce matrice elementare:

$$E(\sigma, \mathbf{u}, \mathbf{v}) = I - \sigma \mathbf{u} \mathbf{v}^H$$

$$\left| \begin{array}{ccc|c} 1 & & & x \\ & 1 & & x \\ & & \ddots & \vdots \\ 0 & & & 1 \end{array} \right| - \sigma \left| \begin{array}{c} x \\ x \\ \vdots \\ x \end{array} \right| = |x \ x \ \dots \ x|$$

**Nota**

**Perché calcola questi autovalori come se la diade fosse singolare?**

$\sigma uv^H$  è detta *diade* ed ha rango  $\leq 1$ ,  $\text{det}(\sigma uv^H) = 0$ .

I suoi autovalori sono:

- $\lambda_1 = 0$  (molteplicità  $n - 1$ )
- $\lambda_2 = \sigma v^H u$  (molteplicità 1)

Quindi gli autovalori di  $E$  sono:  $\begin{cases} 1 - 0 = 1 & \text{molt. } n-1 \\ 1 - \sigma v^H u & \text{molt. } 1 \end{cases}$

dai quali ricaviamo il determinante di  $E$ :  $\text{det}(E) = 1(1 - \sigma v^H u) \neq 0$ .

Quindi  $E$  è invertibile se e solo se  $\sigma v^H u \neq 1$ , infatti questa condizione è l'ipotesi del seguente teorema che ci permette di calcolare l'inversa di una matrice elementare.

La classe delle matrici elementari non singolari è chiusa rispetto all'operazione di inversione. Vale infatti il seguente teorema.

**Teorema 4.8 (Invertibilità matrici elementari)**

Ogni matrice elementare  $E(\sigma, \mathbf{u}, \mathbf{v})$  per cui  $\sigma v^H \mathbf{u} \neq 1$  è invertibile e la sua inversa è ancora una matrice elementare della forma  $E(\tau, \mathbf{u}, \mathbf{v})$ ,  $\tau \in \mathbb{C}$ .

*Dimostrazione.* Se  $\sigma = 0$  la tesi è ovvia. Per  $\sigma \neq 0$  dimostriamo che esiste un  $\tau$  tale che:

$$\begin{aligned} (I - \sigma uv^H)(I - \tau uv^H) &= I \\ I - (\sigma + \tau)uv^H + \sigma\tau u \underbrace{v^H u}_{\text{scalare}} v^H &= I \\ \underbrace{(-\sigma - \tau + \sigma\tau v^H u)}_{\text{scalare}} \underbrace{uv^H}_{\text{matrice}} &= 0 \end{aligned}$$

Per non cadere nel caso in cui  $E = I$ , poniamo che  $u$  e  $v$  siano nulli. Allora possiamo concludere che

$$\begin{aligned} -\sigma - \tau + \sigma\tau v^H u &= 0 \\ \tau(-1 + \sigma v^H u) &= \sigma \\ \tau &= \frac{\sigma}{\sigma v^H u - 1} \end{aligned}$$

che esiste perché per ipotesi  $\sigma v^H u \neq 1$ .  $\square$

Vogliamo usare le matrici elementari per il metodo Gauss, e quindi ci chiediamo se esista una  $E$  tale che  $A^{(2)} = EA^{(1)}$ , questo ci è garantito dal teorema seguente.

**Teorema 4.9**

Siano  $x, y \in \mathbb{C}^n$ ,  $x, y \neq \mathbf{0}$ . Esistono matrici elementari non singolari  $E(\sigma, u, v)$  tali che

$$E(\sigma, u, v)x = y$$

*Dimostrazione.* Vogliamo mostrare che:

$$(I - \sigma uv^H)x = y$$

$$x - \sigma uv^H x = y$$

$$\sigma u = \frac{x - y}{v^H x}$$

Quindi la prima condizione è che  $v^H x \neq 0$ .

Inoltre per avere  $E$  invertibile è necessario che  $\sigma v^H u \neq 1$ , da cui

$$v^H \underline{\sigma u} = v^H \frac{x - y}{v^H x} = \frac{v^H x}{v^H x} - \frac{v^H y}{v^H x} = 1 - \frac{v^H y}{v^H x} \neq 1$$

$$\frac{v^H y}{v^H x} \neq 0 \Rightarrow v^H y \neq 0$$

Quindi per ottenere la matrice  $E$  basta rispettare il seguente sistema:

$$v : \begin{cases} v^H x \neq 0 & \text{esistenza} \\ v^H y \neq 0 & \text{invertibilità} \end{cases}$$

□

## 4.4 Matrici Elementari di Gauss

Vediamo ora quali matrici elementari sono adatte per il metodo di Gauss.

Sia  $\mathbf{x}$ , con  $x_1 \neq 0$ . Si vuole determinare una matrice

$$M = E(\sigma, \mathbf{u}, \mathbf{e}_1) = I - \sigma \underline{\mathbf{u} \mathbf{e}_1^H} \quad \mathbf{e}_1 = | 1 \ 0 \ \dots \ 0 |^T$$

tale che

$$M \mathbf{x} = \underline{\mathbf{x} \mathbf{e}_1} = | x_1 \ 0 \ \dots \ 0 |^T$$

cioè proietti la prima componente del vettore  $\mathbf{x}$ . Dobbiamo quindi trovare  $\sigma$  ed  $\mathbf{u}$  per cui valga questa condizione.

### Esistenza

Per stabilire l'esistenza di  $M$ : sappiamo che vale la condizione

$$\mathbf{e}_1^H \mathbf{x} = x_1 \neq 0$$

Ed il teorema 4.9 ci dice che questa condizione è sufficiente per l'esistenza.

Quindi

$$\sigma \mathbf{u} = \frac{\mathbf{x} - \mathbf{y}}{v^H \mathbf{x}} = \frac{\mathbf{x} - x_1 \mathbf{e}_1}{\mathbf{e}_1^H \mathbf{x}} = \frac{\mathbf{x} - x_1 \mathbf{e}_1}{x_1} = \frac{1}{x_1} \begin{vmatrix} 0 \\ x_2 \\ \vdots \\ x_n \end{vmatrix} = \begin{vmatrix} 0 \\ \frac{x_2}{x_1} \\ \vdots \\ \frac{x_n}{x_1} \end{vmatrix}$$

La matrice  $M^{(1)}$  è perciò:

$$M^{(1)} = \begin{pmatrix} 1 & & & \\ -\frac{x_2}{x_1} & 1 & & \\ \vdots & & \ddots & \\ -\frac{x_n}{x_1} & & & 1 \end{pmatrix}$$

### Matrice inversa

La matrice  $M$  risulta invertibile sfruttando 4.8, in quanto

$$\sigma \mathbf{e}_1^H \mathbf{u} = 0$$

Possiamo quindi calcolare la sua inversa, partendo dal valore di  $\tau$ :

$$\tau = \frac{\sigma}{\underbrace{\sigma \mathbf{v}^H \mathbf{u}}_{=0} - 1} = -\sigma$$

Nel caso del metodo di Gauss  $\mathbf{v}^H \mathbf{u} = 0$  dato che  $\mathbf{v}^H = |1 \ 0 \ \dots \ 0|$  e  $\mathbf{u} = |0 \ x \ \dots \ x|^{-1}$ .  
Quindi in generale una matrice inversa della elementare di Gauss è data da

$$M^{-1} = I + \sigma \mathbf{u} \mathbf{e}_1^H$$

nel nostro caso:

$$(M^{(1)})^{-1} = \begin{pmatrix} 1 & & & \\ \frac{x_2}{x_1} & 1 & & \\ \vdots & & \ddots & \\ \frac{x_n}{x_1} & & & 1 \end{pmatrix}$$

Le matrici elementari hanno la proprietà di trasformare un qualunque vettore non nullo in un vettore con al più una componente diversa da zero. Quindi possiamo sfruttarle per trasformare in forma triangolare superiore una matrice, moltiplicandola successivamente per opportune matrici elementari.

Al  $k$ -esimo passo, posto  $a_{kk}^{(k)} \neq 0$  si considera il vettore

$$m^{(k)} = | \underbrace{0 \ \dots \ 0}_{k \text{ componenti}} \ m_{k+1,k} \ \dots \ m_{nk} |^T$$

dove  $m_{ik} = a_{ik}^{(k)} / a_{kk}^{(k)}$  con  $i = k + 1, \dots, n$  da cui ricaviamo

$$E^k = E(1, m^{(k)}, e_k) = M^{(k)} = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -m_{k+1,k} & \ddots & \\ & & \vdots & \ddots & \\ & & -m_{nk} & & 1 \end{pmatrix}$$

Poniamo ora

$$U = A^{(n)} = M^{(n-1)} \dots M^{(2)} M^{(1)} A$$

Possiamo utilizzare l'inversione delle  $M$ , ottenendo quindi

$$\underbrace{(M^{(1)})^{-1} \dots (M^{(n-2)})^{-1} (M^{(n-1)})^{-1}}_L U = A$$

$$L = (M^{(1)})^{-1} \dots (M^{(n-1)})^{-1} = \begin{pmatrix} 1 & & & & \\ m_{21} & \ddots & & & \\ & \ddots & 1 & & \\ \vdots & & m_{k+1,k} & \ddots & \\ & & \vdots & \ddots & \\ m_{n1} & \dots & m_{nk} & \dots & m_{n,n-1} & 1 \end{pmatrix}$$

Quindi il procedimento consiste nel prodotto di  $n$  matrici triangolari inferiori con elementi diagonali uguali a 1, ed il risultato è ancora una matrice di questa forma, quindi compatibile con la nostra definizione di matrice  $L$  del metodo  $LU$ .

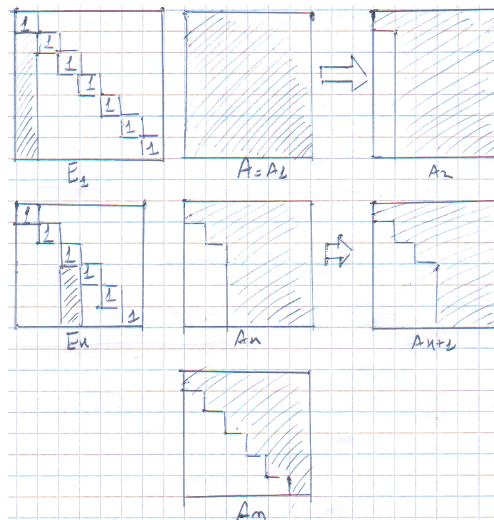


Figura 4.1: Passi della trasformazione LU per mezzo di matrici elementari di Gauss

**Esempio 4.10**

Per  $n = 4$

$$(M^{(1)})^{-1}(M^{(2)})^{-1}(M^{(3)})^{-1} = \begin{pmatrix} 1 & & & \\ m_{21} & 1 & & \\ m_{31} & & 1 & \\ m_{41} & & & 1 \end{pmatrix} \begin{pmatrix} 1 & & & \\ & 1 & & \\ & m_{32} & 1 & \\ & m_{42} & & 1 \end{pmatrix} \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & m_{43} & 1 \end{pmatrix} = \begin{pmatrix} 1 & & & \\ m_{21} & 1 & & \\ m_{31} & m_{32} & 1 & \\ m_{41} & m_{42} & m_{43} & 1 \end{pmatrix} = L$$

L'elemento  $a_{kk}$  della matrice  $A^{(k)}$ , detto *pivot* al k-esimo passo, è per ipotesi diverso da zero e dato che  $A^{(k)}$  è triangolare superiore per 1.1.2.1

$$\det(A_k) = a_{11}^{(1)} \cdot \dots \cdot a_{kk}^{(k)} \neq 0$$

quindi  $A^{(k)}$ , detta minore principale di testa, è non singolare.

**4.5 Matrici elementari di Householder**

Il procedimento di fattorizzazione della matrice A con matrici di Householder è sempre applicabile. Queste matrici ci servono per costruire la fattorizzazione QR.

$$A = \underbrace{Q}_{\text{unitaria}} \cdot \underbrace{R}_{\text{tr. sup.}}$$

Dobbiamo trovare una matrice elementare P unitaria tale che

$$P_{n-1} \dots P_1 \cdot A = R$$

$$A = \underbrace{(P_{n-1} \dots P_1)^{-1}}_Q R$$

**Definizione 4.11 (Matrice elementare di Householder)**

Una matrice elementare hermitiana

$$P = I - \beta \mathbf{v} \mathbf{v}^H$$

con  $\beta \in \mathbb{R}$  e  $\mathbf{v} \in \mathbb{C}^n, \mathbf{v} \neq 0$ , è detta matrice di Householder se è unitaria, cioè se  $P^H P = P P^H = I$ .

Per ogni vettore  $\mathbf{x} \in \mathbb{C}^n$ , con  $\mathbf{x} \neq 0$ , si può determinare una matrice di Householder  $P$  tale che

$$P\mathbf{x} = \alpha \mathbf{e}_1$$

dove  $\alpha$  è una opportuna costante e  $\mathbf{e}_1$  è il primo vettore della base canonica. Nel metodo,  $x$  assumerà i valori delle colonne di  $A$ , dalla prima all' $n$ -esima, che si vorranno trasformare.

#### Proprietà 4.1

La matrici unitarie godono della seguenti proprietà:

•

$$U^{-1} = U^H$$

• moltiplicare per una matrice unitaria non cambia la norma 2

$$\|U\mathbf{x}\|_2 = \|\mathbf{x}\|_2 \quad \text{vettori}$$

$$\|UM\|_2 = \|M\|_2 \quad \text{matrici}$$

Per quanto riguarda le norme di vettori infatti, ricordando che

$$\|\mathbf{x}\| \stackrel{\text{def}}{=} \sqrt{\mathbf{x}^H \mathbf{x}} = \sqrt{\sum_i^n |x_i|^2}$$

otteniamo

$$\|U\mathbf{x}\|_2 = \sqrt{(U\mathbf{x})^H U\mathbf{x}} = \sqrt{\mathbf{x}^H U^H U \mathbf{x}} = \sqrt{\mathbf{x}^H \mathbf{x}} = \|\mathbf{x}\|_2$$

Invece per le matrici consideriamo la norma indotta (spettrale)  $\rho$ . Ricordando che

$$\|M\|_2 \stackrel{\text{def}}{=} \sqrt{\rho(M^H M)}$$

otteniamo

$$\|UM\|_2 = \sqrt{\rho((UM)^H UM)} = \sqrt{\rho(M^H U^H U M)} = \sqrt{\rho(M^H M)} = \|M\|_2$$

Quindi la matrice unitaria è una trasformazione che conserva le lunghezze: una isometria

Quindi, nel procedimento che useremo, conserveremo le lunghezze, e questo è importante per la stabilità.

#### Condizioni per $\beta$

Dal fatto che  $P$  è unitaria e hermitiana ricaviamo:

$$I \stackrel{\text{unit.}}{=} PP^H \stackrel{\text{herm.}}{=} PP = (I - \beta \mathbf{v}\mathbf{v}^H)(I - \beta \mathbf{v}\mathbf{v}^H) = I - 2\beta \mathbf{v}\mathbf{v}^H + \beta^2 \underbrace{\mathbf{v}^H \mathbf{v}}_{\text{scalare}} \mathbf{v}^H = I$$

In sostanza dobbiamo rispettare la condizione

$$\beta(-2\mathbf{v}\mathbf{v}^H + \beta \mathbf{v}^H \mathbf{v}\mathbf{v}^H) = 0$$

Ponendo la condizione  $\beta \neq 0$

$$(-2 + \beta \mathbf{v}^H \mathbf{v}) \mathbf{v}\mathbf{v}^H = 0$$

e ponendo  $\mathbf{v} \neq 0$ , come richiesto dalla definizione di elementare di Householder, otteniamo

$$\beta = \frac{2}{\mathbf{v}^H \mathbf{v}} = \frac{2}{\|\mathbf{v}\|_2^2}$$



**Condizioni per  $\alpha$** 

Inoltre dato che  $P\mathbf{x} = \alpha\mathbf{e}_1$  e  $P$  è unitaria abbiamo la seguente condizione su  $|\alpha|$

$$\|\mathbf{x}\|_2 = \|P\mathbf{x}\|_2 = \|\alpha\mathbf{e}_1\|_2 = |\alpha|$$

Inoltre dalle proprietà delle hermitiane risulta  $\mathbf{x}^H P\mathbf{x} \in \mathbb{R}$ , da cui:

$$\mathbb{R} \ni \mathbf{x}^H P\mathbf{x} = \mathbf{x}^H \alpha\mathbf{e}_1 = \bar{x}_1 \alpha = \dots$$

Ora esprimiamo

$$\begin{aligned} \alpha &= |\alpha| (\cos(\varphi) + i \sin(\varphi)) \\ x_1 &= |x_1| (\cos(\psi) + i \sin(\psi)) \\ \bar{x}_1 &= |x_1| (\cos(-\psi) + i \sin(-\psi)) \end{aligned}$$

Ricordando dalla trigonometria che

$$\begin{aligned} \text{sen}(\alpha - \beta) &= \text{sen } \alpha \cos \beta - \cos \alpha \text{sen } \beta \\ \cos(\alpha - \beta) &= \cos \alpha \cos \beta + \text{sen } \alpha \text{sen } \beta \end{aligned}$$

otteniamo

$$\dots = |x_1| |\alpha| (\cos(\varphi - \psi) + i \sin(\varphi - \psi))$$

Affinché questo numero sia reale è necessario annullare la sua parte immaginaria.

$$\sin(\varphi - \psi) = 0 \quad \text{da cui due possibilità} \quad \begin{cases} \varphi = \psi & (1) \\ \varphi = \pi + \psi & (2) \end{cases}$$

Poniamo  $\theta = (\cos(-\psi) + i \sin(-\psi))$  ed esprimiamolo in funzione di  $\mathbf{x}$  come  $\theta = \frac{x_1}{|x_1|}$ , ed esprimiamo a sua volta  $\alpha$  in funzione di  $\theta$ . In pratica  $\theta$  è il *versore*, che rappresenta la direzione di  $x$ . Vogliamo che  $\alpha$  abbia la stessa direzione di  $x$ , oppure quella opposta. Vediamo i vincoli su  $\alpha$  delle precedenti condizioni:

$$\alpha = |\alpha|\theta \quad (1)$$

$$\alpha = |\alpha|(-\theta) \quad (2)$$

**Condizioni per  $\mathbf{v}$** 

A questo punto manca da trovare  $\mathbf{v}$  tale che  $\beta = \frac{2}{\|\mathbf{v}\|_2^2}$

$$\begin{aligned} P\mathbf{x} &= \alpha\mathbf{e}_1 \\ (I - \beta\mathbf{v}\mathbf{v}^H)\mathbf{x} &= \alpha\mathbf{e}_1 \\ \mathbf{x} - \beta\mathbf{v} \underbrace{\mathbf{v}^H \mathbf{x}}_{\text{scalare}} &= \alpha\mathbf{e}_1 \Rightarrow \mathbf{v} = \frac{\mathbf{x} - \alpha\mathbf{e}_1}{\beta\mathbf{v}^H \mathbf{x}} = c(\mathbf{x} - \alpha\mathbf{e}_1) \end{aligned}$$

Teoricamente non possiamo ricavare  $\mathbf{v}$  dalla prima forma, perché è in funzione di  $\mathbf{v}^H$ , proviamo allora a metterlo dentro un fattore moltiplicativo  $c = \frac{1}{\beta\mathbf{v}^H \mathbf{x}}$ . Calcoliamo ora

$$\beta\mathbf{v}\mathbf{v}^H = \frac{2}{\cancel{\mathbf{v}^H(\mathbf{x} - \alpha\mathbf{e}_1)^H(\mathbf{x} - \alpha\mathbf{e}_1)}} \cdot \cancel{\mathbf{v}^H(\mathbf{x} - \alpha\mathbf{e}_1)} \mathbf{v}^H(\mathbf{x} - \alpha\mathbf{e}_1) = \frac{2}{\|\mathbf{x} - \alpha\mathbf{e}_1\|_2^2} \cdot (\mathbf{x} - \alpha\mathbf{e}_1)(\mathbf{x} - \alpha\mathbf{e}_1)^H$$

vediamo come questo termine sia indipendente da  $c$  che viene semplificato, quindi risolto il dubbio precedente possiamo esprimere

$$\mathbf{v} = \mathbf{x} - \alpha\mathbf{e}_1$$

Ricapitolando:

- $\beta = \frac{2}{\mathbf{v}^H \mathbf{v}} = \frac{2}{\|\mathbf{v}\|_2^2}$
- $|\alpha| = \|\mathbf{x}\|_2$

- $\alpha = \pm|\alpha|\theta, \quad \theta = \frac{x_1}{|x_1|} = \text{sgn}(x_1)$
- $\mathbf{v} = \mathbf{x} - \alpha\mathbf{e}_1$

**Esempio 4.12 (Calcolo di P)**

Calcoliamo la matrice  $P = I - \beta\mathbf{v}\mathbf{v}^H$  relativa a

$$A = \begin{pmatrix} 1 & x & x \\ 1 & x & x \\ 1 & x & x \end{pmatrix}$$

$\mathbf{x}$  è la prima colonna di  $A$ , da cui otteniamo

$$\begin{cases} \|\mathbf{x}\| = \sqrt{3} \\ \theta = \text{sgn}(x_1) = 1 \end{cases} \implies \alpha = \pm\sqrt{3}$$

$$\mathbf{v} = \mathbf{x} - \alpha\mathbf{e}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \mp \sqrt{3} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \mp \sqrt{3} \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 + \sqrt{3} \\ 1 \\ 1 \end{pmatrix}$$

Dato che sono positivi sia 1 che  $\sqrt{3}$  scegliamo il segno + per ragioni di stabilità. Calcoliamo infine  $\beta$

$$\|\mathbf{v}\|^2 = 1 + 3 + 2\sqrt{3} + 1 + 1 = 6 + 2\sqrt{3} = 2(3 + \sqrt{3})$$

$$\beta = \frac{2}{2(3 + \sqrt{3})} = \frac{1}{3 + \sqrt{3}}$$

Da cui otteniamo

$$P = I - \beta\mathbf{v}\mathbf{v}^H = \begin{pmatrix} 1 - \frac{1}{3+\sqrt{3}}(1+\sqrt{3})^2 & -\frac{1+\sqrt{3}}{3+\sqrt{3}} & -\frac{1+\sqrt{3}}{3+\sqrt{3}} \\ -\frac{1+\sqrt{3}}{3+\sqrt{3}} & 1 - \frac{1}{3+\sqrt{3}} & -\frac{1}{3+\sqrt{3}} \\ -\frac{1+\sqrt{3}}{3+\sqrt{3}} & -\frac{1}{3+\sqrt{3}} & 1 - \frac{1}{3+\sqrt{3}} \end{pmatrix}$$

**4.5.1 Massimo Pivot****Work in progress**

In generale si parte dalla prima colonna, si calcola la matrice elementare di Householder, e si moltiplica per  $A$ , il risultato ha la prima colonna triangolarizzata. Notare che il valore che compare sulla diagonale, ha il modulo uguale alla norma della colonna che ha sostituito!

Quindi se invece di partire dalla prima colonna, prendiamo di volta in volta quella con norma massima, andiamo ad ordinare i valori principali in modo decrescente.

Questo è equivalente a permutare le colonne della matrice  $A$ , ordinandole per norma:

$$A\Pi = QR$$

dove  $R$  ha i valori principali ordinati.

Questo può essere utile quando vogliamo gli autovalori nulli in fondo alla matrice, come nell'uso del metodo QR per il problema dei minimi quadrati.

### 4.5.2 Complessità

Tornando al metodo di Cholesky, per fare  $PA$  non faremo esplicitamente il prodotto (costo  $O(n^3)$ ), ricordiamo che:

$$PA = (I - \beta \mathbf{v}\mathbf{v}^H)A = A - \beta \mathbf{v}\mathbf{v}^H A$$

$$1. \mathbf{v}^H A \quad O(n^2)$$

$$2. \beta \mathbf{v} \quad O(n^2)$$

Quindi eseguendo il calcolo in questo modo abbiamo un costo quadratico invece che cubico.

Complessità di Gauss:  $O(\frac{2}{3}n^3)$

Complessità di Householder:  $O(\frac{4}{3}n^3)$

Gauss ha una possibile instabilità quando  $L$  e  $U$  hanno coefficienti grandi rispetto a quelli di partenza mentre Householder non ha questo rischio perché esegue trasformazioni unitarie, ma costa il doppio.

## 4.6 Fattorizzazione di Cholesky

$$A = LL^H$$

Dal precedente risultato 4.5 Sappiamo che per  $A$  hermitiana definita positiva  $L$  esiste unica (con  $l_{ii} > 0$ ). Ogni componente  $l_{ij}$  è il prodotto della riga  $i$  della matrice  $L$  e della colonna  $j$  delle matrice  $L^H$ :

$$a_{ij} = \sum_{k=1}^j l_{ik} \bar{l}_{jk}$$

Vediamo ad esempio la prima colonna, analizzando a parte l'elemento principale:

$$a_{11} = l_{11} \bar{l}_{11} = |l_{11}|^2 = l_{11}^2 \quad \implies \quad l_{11} = \sqrt{a_{11}}$$

$$a_{i1} = l_{i1} l_{11} \quad \implies \quad l_{i1} = \frac{a_{i1}}{l_{11}} \quad i = 2, \dots, n$$

Se non avessimo avuto una matrice definita positiva, avremmo avuto un radicando negativo. Notare che per verificare che una matrice sia definita positiva è sufficiente applicare Cholesky, in caso negativo il metodo si blocca.

Estendiamo al caso generale

$$\begin{aligned} a_{jj} &= \sum_{k=1}^j |l_{jk}|^2 = \sum_{k=1}^{j-1} |l_{jk}|^2 + l_{jj}^2 \quad \implies \quad l_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} |l_{jk}|^2} \\ a_{ij} &= \sum_{k=1}^{j-1} l_{ik} \bar{l}_{jk} + l_{ij} l_{jj} \quad \implies \quad l_{ij} = \frac{1}{l_{jj}} (a_{ij} - \sum_{k=1}^{j-1} l_{ik} \bar{l}_{jk}) \end{aligned}$$

### 4.6.1 Complessità

Il sistema lineare risultante si risolve nel seguente modo:

$$\begin{cases} \mathbf{A}\mathbf{x} = \mathbf{b} \\ \mathbf{L}\mathbf{L}^H \mathbf{x} = \mathbf{b} \end{cases} \quad \begin{cases} \mathbf{L}\mathbf{y} = \mathbf{b} & (1) \\ \mathbf{L}^H \mathbf{x} = \mathbf{y} & (2) \end{cases} \quad O(n^2)$$

Queste due equazioni sono più facili da risolvere, dato che le matrici sono triangolari.

Il costo del calcolo di  $L$  è  $O(\frac{n^3}{3})$ , la matrice è simmetrica quindi i dati sono circa la metà e infatti anche il costo è dimezzato.

### 4.6.2 Stabilità

$$\forall j, k \quad a_{jj} = \sum |l_{jk}|^2 \geq |l_{jk}|^2$$

$$|l_{jk}| \leq \sqrt{a_{jj}} \quad \max |l_{jk}| \leq \max \sqrt{a_{jj}}$$

Questa è una indicazione del fatto che gli elementi di  $L$  sono limitati superiormente dagli elementi di  $A$ , al contrario del metodo di Gauss dove gli elementi possono crescere molto.

Quindi la stabilità del metodo di Cholesky è garantita dal fatto che  $L$  non può avere elementi più grandi di quelli di  $A$ .

## 4.7 Complessità sui sistemi lineari

I metodi visti precedentemente hanno costo  $O(n^3)$  e vale il seguente risultato:

### Teorema 4.13

Se  $kn^\theta$  operazioni aritmetiche sono sufficienti per moltiplicare due matrici quadrate di ordine  $n$ , allora  $hn^\theta$  operazioni sono sufficienti per invertire la una matrice (o per risolvere un sistema lineare).

Costo del prodotto di due matrici  $n \times n$  con l'algoritmo di Strassen:  $kn^{\log_2 7}$ ,  $\theta = \log_2 7 = 2,8\dots$

### 4.7.1 Tabella riassuntiva

Metodo	Costi	Output	Requisiti	Pregi	Difetti
Gauss ( $LU$ )	$n^3/3$	$L$ è una matrice triangolare inferiore con elementi principali uguali ad 1 ed $U$ una matrice triangolare superiore.	$A$ di ordine $n$ : sottomatrici di testa $A_k$ non singolari ( $k = 1 \dots n - 1$ )	È unica	Poco stabile
Householder ( $QR$ )	$2n^3/3$	$Q$ è una matrice unitaria ed $R$ è una matrice triangolare superiore.	Sempre applicabile	Usa matrici unitarie: stabile	Non unica a meno di matrici di fase. Lento.
Cholesky ( $LL^H$ )	$n^3/6$	$L$ triangolare inferiore con elementi diagonali positivi	Hermitiana definita positiva	Stabile e veloce	

## 4.8 Sistemi Lineari: metodi iterativi

I Metodi iterativi si distinguono in

- Metodo associativi e decomposizione additiva (dati per fatti)
- Metodo del gradiente coniugato
- Metodi iterativi per sistemi non lineari

questi metodi risultano particolarmente convenienti se la matrice  $A$  è sparsa, cioè se il numero degli elementi non nulli di  $A$  è dell'ordine della dimensione della matrice.

### 4.8.1 Convergenza

**Definizione 4.14 (Convergenza)**

Una successione  $\{\mathbf{x}^{(k)}\}$  di vettori di  $\mathbb{C}^n$  si dice convergente al vettore  $\mathbf{x}^*$  di  $\mathbb{C}^n$  se esiste una norma per cui risulta

$$\lim_{k \rightarrow \infty} \|\mathbf{x}^{(k)} - \mathbf{x}^*\| = 0$$

che si può scrivere anche

$$\lim_{k \rightarrow \infty} \mathbf{x}^{(k)} = \mathbf{x}^*$$

Questa condizione di convergenza si traduce in una condizione di convergenza delle successioni formate dalle singole componenti. Infatti

$$\forall i = 1, \dots, n. \lim_{k \rightarrow \infty} |x_i^{(k)} - x_i^*| = 0 \iff \lim_{k \rightarrow \infty} x_i^{(k)} = x_i^*$$

Il seguente teorema è di fondamentale importanza nello studio della convergenza dei metodi iterativi per la risoluzione dei sistemi lineari.

**4.8.2 Richiami: decomposizione additiva**

Sia  $A \in \mathbb{C}^{n \times n}$  una matrice non singolare e si consideri la decomposizione di  $A$ , con  $M$  non singolare:

$$A = M - N \quad \det(M) \neq 0$$

Dal sistema  $A\mathbf{x} = b$  risulta

$$(M - N)\mathbf{x} = b$$

$$M\mathbf{x} - N\mathbf{x} = b$$

Quindi

$$\mathbf{x} = \underbrace{M^{-1}N}_{P} \mathbf{x} + \underbrace{M^{-1}b}_{q}$$

Si ottiene il seguente sistema equivalente a quello iniziale:

$$\mathbf{x} = P\mathbf{x} + q$$

Dato un vettore iniziale  $\mathbf{x}^{(0)}$ , si considera la successione  $\mathbf{x}^{(1)} \mathbf{x}^{(2)} \dots$  così definita

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

Se la successione converge si indica con

$$\mathbf{x}^* = \lim_{k \rightarrow \infty} \mathbf{x}^k$$

e passando al limite risulta

$$\mathbf{x}^* = P\mathbf{x}^* + q$$

Abbiamo un metodo iterativo in cui, partendo da un vettore iniziale  $\mathbf{x}^{(0)}$ , la soluzione viene approssimata utilizzando una successione  $\{\mathbf{x}^{(k)}\}$  di vettori. La matrice  $P$  si dice *matrice di iterazione del metodo*.

Al variare del vettore iniziale  $\mathbf{x}^{(0)}$  si ottengono diverse successioni, alcune delle quali possono essere convergenti ed altre no. Un metodo iterativo è detto *convergente* se, qualunque sia il vettore iniziale  $\mathbf{x}^{(0)}$ , la successione è convergente.

**Teorema 4.15 (Convergenza)**

Sia  $P \in \mathbb{C}^{n \times n}$ , allora

$$\lim_{k \rightarrow \infty} P^k = 0 \text{ se e solo se } \rho(P) < 1$$

(Il metodo iterativo è convergente se e solo se  $\rho(P) < 1$ ).

*Dimostrazione.* (Riduzione di  $P$  in forma canonica di Jordan)

Sappiamo esistere una matrice non singolare  $S \in \mathbb{C}^{n \times n}$ , tale che  $P = SJS^{-1}$ , dove  $J$  è la forma normale di Jordan di  $P$ ; risulta allora

$$J = S^{-1}PS = \begin{pmatrix} J_1 & & \\ & J_2 & \\ & & \ddots \end{pmatrix} \quad \text{dove} \quad J_i = \begin{pmatrix} C_i^{(1)} & & \\ & C_i^{(2)} & \\ & & \ddots \end{pmatrix}$$

Per passare al limite, vediamo cosa accade moltiplicando  $k$  volte:

$$J^k = \dots S^{-1}P \underbrace{S S^{-1}}_I P S = S^{-1}P^k S$$

notare che rimangono solo il primo  $S^{-1}$  e l'ultimo  $S$ .

$$P = SJS^{-1} \quad P^k = S J^k S^{-1}$$

dove  $P^k \rightarrow 0$  se  $J^k \rightarrow 0$ .

Abbiamo visto che  $J_i$  è diagonale a blocchi, dove i blocchi sono della forma:

$$C_i^{(j)} = \lambda_i I + U = \begin{pmatrix} \lambda_i & 1 & & & 0 \\ & \lambda_i & 1 & & \\ & & \lambda_i & 1 & \\ & & & \lambda_i & 1 \\ 0 & & & & \lambda_i \end{pmatrix}$$

dove le matrici  $U$  sono della forma

$$U = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & & 1 \\ & & & & 0 \end{pmatrix}$$

e sono dette matrici di *shifting* perché nei passi successivi  $U^k$  fanno shiftare la diagonale di uni in alto a destra, fino a farla scomparire.

Infatti, assumendo  $U^0 = I$ , possiamo fermare la sommatoria ad un certo  $\nu$  tanto poi resterà uguale dato che per  $r > \nu$   $U^r = \mathbf{0}$ .



#### Nota

Ricordiamo dall'algebra la seguente proprietà (Triangolo di Tartaglia)

$$(a + b)^n = \sum_{k=0}^n \binom{n}{k} a^{n-k} b^k$$

Al  $k$ -esimo passo diventano della forma:

$$[C_i^{(j)}]^k = (\lambda I + U)^K = \sum_{r=0}^k \binom{k}{r} (\lambda^{k-r} I) U^r = \sum_{r=0}^{\nu-1} \binom{k}{r} \lambda^{k-r} U^r \stackrel{*}{=} \begin{pmatrix} \lambda^k & k\lambda^{k-1} & \binom{k}{2}\lambda^{k-2} & \dots \\ & \lambda^k & \ddots & \ddots \\ & & \ddots & \ddots \\ 0 & & & \lambda^k \end{pmatrix}$$

**Nota**

\*) : In questo passaggio stiamo intendendo che la matrice finale è ottenuta dalla somma di matrici. Inoltre i vari elementi delle diagonali sono shiftati grazie alle matrici di *shifting*

Ne segue che condizione necessaria e sufficiente affinché  $\lambda_i^k$  e  $\binom{k}{r} \lambda_i^{k-r}$  tendano a zero per  $k \rightarrow \infty$  è che sia  $|\lambda_i| < 1$ , cioè  $\rho(P) < 1$ .

Convergenza  $\iff$  P convergente  $\iff \rho(P) < 1$ .  $\square$

## 4.9 Particolari decomposizioni additive

Fra i metodi iterativi individuati da una particolare scelta della decomposizione sono particolarmente importanti il metodo di Jacobi e il metodo di Gauss-Seidel, per i quali è possibile dare delle condizioni sufficienti di convergenza verificate da molte delle matrici che si ottengono risolvendo problemi differenziali. Si consideri la decomposizione della matrice A

### 4.9.1 Metodi iterativi di Jordan e Gauss-Seidel

$$A = D - B - C$$

$$D = \begin{pmatrix} a_{ii} & & \\ & a_{ii} & \\ & & a_{ii} \end{pmatrix}, \quad B = \begin{pmatrix} -a_{ij} & & \\ & -a_{ij} & \\ & & -a_{ij} \end{pmatrix}, \quad C = \begin{pmatrix} -a_{ij} & -a_{ij} \\ & -a_{ij} \end{pmatrix}$$

Scegliendo  $M = D$ ,  $N = B + C$ , si ottiene il metodo di *Jacobi*.

Scegliendo  $M = D - B$ ,  $N = C$ , si ottiene il metodo di *Gauss-Seidel*.

Per queste decomposizioni risulta  $\det M \neq 0$  se e solo se tutti gli elementi principali di A sono non nulli.

$$A = M - N$$

$$A = D - (B + C) \quad P = M^{-1}N = D^{-1}(B + C)$$

### 4.9.2 Metodo di Jacobi

Indicando con J la matrice di iterazione del metodo di Jacobi, abbiamo

$$J = D^{-1}(B + C)$$

da cui otteniamo la successione

$$\mathbf{x}^{(k)} = \underbrace{D^{-1}(B + C)}_{J=P} \mathbf{x}^{(k-1)} + \underbrace{D^{-1}b}_q$$

$$J = \begin{pmatrix} 0 & -\frac{a_{ij}}{a_{ii}} & -\frac{a_{ij}}{a_{ii}} \\ -\frac{a_{ij}}{a_{ii}} & 0 & -\frac{a_{ij}}{a_{ii}} \\ -\frac{a_{ij}}{a_{ii}} & -\frac{a_{ij}}{a_{ii}} & 0 \end{pmatrix}$$

### 4.9.3 Metodo di Gauss-Seidel

Indicando con G la matrice di iterazione del metodo di Gauss-Seidel, abbiamo

$$A = \underbrace{D - B}_M - \underbrace{C}_N \quad G = P = M^{-1}N = (D - B)^{-1}C$$

notare che  $D$  e  $D - B$  sono invertibili dato che  $a_{ii} \neq 0$ .

Otteniamo la successione

$$\mathbf{x}^{(k)} = \underbrace{(D - B)^{-1}C}_{G=P} \mathbf{x}^{(k-1)} + \underbrace{(D - B)^{-1}b}_q$$

#### 4.9.4 Condizioni sufficienti per convergenza Jacobi e Gauss-Seidel

Predominanza diagonale

- Predominanza diagonale forte di A:

$$|a_{ii}| > \sum_{j=1; j \neq i}^n |a_{ij}| \quad \forall i$$

- Predominanza diagonale debole:

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

$$\exists r. |a_{rr}| > \sum_{j=1; j \neq r}^n |a_{rj}|$$

cioè la indebolisco per tutti gli i tranne uno (r).



#### Teorema 4.16 (Convergenza Jacobi e Gauss-Seidel)

Sia  $A = M - N$  la decomposizione della matrice A corrispondente al metodo di Jacobi o al metodo di Gauss-Seidel. Se vale una delle seguenti ipotesi

- A è a predominanza diagonale forte
- A è a predominanza diagonale ed è irriducibile
- A è a predominanza diagonale forte per colonne
- A è a predominanza diagonale ed è irriducibile per colonne

allora  $\rho(M^{-1}N) < 1$  e quindi i metodi sono convergenti.

Passiamo adesso a vedere un teorema che riguarda le matrici hermitiane e la convergenza del metodo di Gauss-Seidel applicato a tali matrici. Per dimostrare tale teorema, avremo bisogno di un lemma.

#### Lemma 4.17

Data A matrice hermitiana tale che  $\forall i. a_{ii} > 0$  e G sua matrice di iterazione di Gauss-Seidel, la matrice  $A - G^H A G$  è definita positiva.

*Dimostrazione.* A è una matrice hermitiana, dunque nella sua decomposizione vista nella definizione dei metodi di Jacobi e Gauss-Seidel possiamo prendere  $C = B^H$ . Abbiamo dunque  $A = D - B - B^H \Leftrightarrow B^H = D - B - A$ . Vale dunque

$$G = (D - B)^{-1} B^H = (D - B)^{-1} ((D - B) - A) = I - (D - B)^{-1} A$$

Poniamo  $F = (D - B)^{-1} A$ , e scriviamo dunque  $G = I - F$ .

Prendiamo la matrice  $A - G^H A G$ : possiamo scrivere

$$\begin{aligned} A - G^H A G &= \\ A - (I - F^H) A (I - F) &= \\ A - (A - F^H A + A F + F^H A F) &= \\ F^H A + A F + F^H A F &= \end{aligned}$$



$$\begin{aligned}
F^H(AF^{-1} + F^{-H}A - A)F &= \quad \text{posto } F^{-1} = A^{-1}(D - B) \quad F^{-H} = (D - B^H)A^{-1} \\
F^H(AA^{-1}(D - B) + (D - B^H)A^{-1}A - (D - B - B^H))F &= \\
F^H(D - B + D - B^H - D - B - B^H)F &= \\
F^HDF &
\end{aligned}$$

Questa matrice è definita positiva: infatti

$$\begin{aligned}
x^H F^H D F x &= \quad \text{posto } y = Fx \\
y^H D y &= \sum_{i=1}^n a_{ii} |y_{ii}|^2 > 0
\end{aligned}$$

in quanto  $a_{ii} > 0$  per ipotesi.  $\square$

Passiamo adesso al teorema.



#### Teorema 4.18

Sia  $A$  una matrice hermitiana tale che  $\forall i. a_{ii} > 0$ .

Allora

$$\text{Gauss- Seidel converge} \iff A \text{ è definita positiva}$$



#### Osservazione 4.19

Una piccola nota a margine del teorema.

*Dimostrazione.* Dobbiamo dimostrare i due sensi della doppia implicazione.

$\Leftarrow$  Dobbiamo dimostrare che se la matrice  $A$  è definita positiva allora Gauss-Seidel converge. Sappiamo che la matrice  $A - G^H A G$  è definita positiva, come visto nel lemma 4.17. Allora possiamo derivarne

$$\begin{aligned}
0 < x^H (A - G^H A G) x &\iff \\
x^H A x - x^H G^H A G x &\iff \quad \text{posto } Gx = \lambda_G x \\
x^H A x - x^H \overline{\lambda_G} A \lambda_G x &= \\
x^H A x - |\lambda_G|^2 x^H A x &= \\
(1 - |\lambda_G|^2) x^H A x &> 0
\end{aligned}$$

$A$  è definita positiva per ipotesi, dunque perché si verifichi la relazione deve valere

$$(1 - |\lambda_G|^2) > 0 \Rightarrow |\lambda_G| < 1 \Rightarrow \rho(G) < 1$$

come volevasi dimostrare.

$\Rightarrow$  Dobbiamo adesso dimostrare che, sotto le ipotesi del teorema, se Gauss-Seidel converge, allora la matrice  $A$  è definita positiva.

Abbiamo visto nel lemma 4.17 che la matrice  $A - G^H A G$  è definita positiva. Questo significa che vale la disuguaglianza  $e^{(k)H} (A - G^H A G) e^{(k)} > 0$ . Chiameremo questa quantità  $\alpha(k)$ .

Consideriamo la successione degli  $\alpha^{(k)}$ : otteniamo che

$$\begin{aligned}
e^{(k)H} (A - G^H A G) e^{(k)} &> 0 \iff \\
e^{(k)H} A e^{(k)} - e^{(k)H} G^H A G e^{(k)} &> 0 \iff \\
e^{(k)H} A e^{(k)} &> e^{(k+1)H} A e^{(k+1)}
\end{aligned}$$

in quanto  $e^{(k+1)} = Ge^{(k)}$ . La successione degli  $\alpha^k$  è dunque strettamente decrescente, in quanto questa relazione vale per un qualunque valore di  $k$ .

Supponiamo adesso per assurdo che  $A$  non sia definita positiva. Questo significherebbe che  $\exists z \neq 0. z^H A z \leq 0$ . Visto che Gauss-Seidel converge, posso prendere un qualunque vettore come vettore iniziale del metodo: se prendessi proprio  $x^{(0)} = z$  otterrei la relazione

$$0 \geq z^H A z = e^{(0)H} A e^{(0)} > e^{(1)H} A e^{(1)} > \dots$$

Tale successione non può ovviamente convergere a 0, e dunque Gauss-Seidel non convergerebbe. Questo contraddice evidentemente l'ipotesi, dunque la matrice  $A$  è sicuramente definita positiva.

□

Passiamo adesso ad un teorema importante riguardante il confronto fra i raggi spettrali delle matrici di iterazione di Jacobi e Gauss-Seidel.



#### Teorema 4.20 (Stein-Rosenberg)

Sia  $A$  una matrice tale che

- $a_{ii} \neq 0$
- la matrice di iterazione  $J$  di Jacobi non contenga elementi negativi

Allora si verifica uno dei seguenti casi:

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

Non dimostreremo questo teorema, ma ci limiteremo ad analizzarne le conseguenze.

In generale  $\rho(P)$  può essere considerato come una misura della velocità di convergenza. Infatti, chiamato  $e^{(k)}$  l'errore alla  $k$ -esima iterazione,

$$e^{(k)} = P^k e^{(0)} \Rightarrow \|e^{(k)}\| = \|P^k e^{(0)}\| \leq \|P^k\| \|e^{(0)}\|$$

Se  $\|P\| < 1$  questa disuguaglianza è ottima: la norma di  $e^k$  diminuirà all'aumentare di  $k$ . Ma come si lega la norma al raggio spettrale? Ci viene in aiuto il seguente teorema, che non dimostreremo.



#### Teorema 4.21 (Legame fra raggio spettrale e norme indotte)

Per ogni matrice  $P$  e per ogni norma indotta  $\|\cdot\|$ , vale

$$\rho(P) = \inf \{\|P\|\}$$

Per questo teorema  $\forall \bar{\epsilon}. \exists \|\cdot\|_*. \rho(P) \leq \|P\|_* \leq \rho(P) + \bar{\epsilon}$

In pratica,

- se ho un raggio spettrale strettamente minore di 1 posso trovare una norma che assicuri la convergenza
- più basso è il raggio spettrale (e dunque la norma) meglio è

Il teorema dice dunque che, in caso di matrici di Jacobi ad elementi non negativi, se il metodo di Jacobi converge allora converge anche il metodo di Gauss Seidel, ed inoltre Gauss Seidel converge più velocemente.

## 4.10 Matrici tridiagonali

Vediamo adesso un teorema che ci dice come si comportano i due metodi su una classe particolare di matrici: le matrici *tridiagonali*

### Definizione 4.22 (Matrice Tridiagonale)

Una matrice  $A$  è detta *tridiagonale* se  $\forall i, j, |i - j| > 1. a_{ij} = 0$

In pratica, è una matrice che ha tutti zeri tranne

- sulla diagonale principale
- sulle due diagonali adiacenti

Tipicamente la notazione utilizzata per indicare i singoli elementi delle matrici tridiagonali é

$$\begin{bmatrix} a_1 & c_1 & 0 & 0 & 0 & 0 \\ b_1 & a_2 & c_2 & 0 & 0 & 0 \\ 0 & b_2 & a_3 & c_3 & 0 & 0 \\ 0 & 0 & b_3 & a_4 & c_4 & 0 \\ 0 & 0 & 0 & b_4 & a_5 & c_5 \\ 0 & 0 & 0 & 0 & b_5 & a_6 \end{bmatrix}$$

Su tali matrici vale il seguente teorema, che in sostanza afferma che il tasso asintotico di convergenza del metodo di Gauss-Seidel è doppio di quello del metodo di Jacobi e, asintoticamente, sono necessarie metà iterazioni del metodo di Gauss-Seidel per ottenere la stessa precisione che con il metodo di Jacobi.



### Teorema 4.23

Sia  $T$  tridiagonale, tale che  $\forall i. a_i \neq 0$ . Allora

- Se  $\lambda$  è autovalore di  $J$ ,  $\lambda^2$  è autovalore di  $G$
- Se  $\mu$  è autovalore di  $G$ ,  $\mu \neq 0$ , le radici quadrate di  $\mu$  sono autovalori di  $J$

Infatti cadiamo nel caso del teorema di Stein Rosenberg

$$\rho(G) = \rho^2(J)$$

*Dimostrazione.* Piuttosto che dimostrare formalmente per matrici di dimensione qualunque, vediamo la dimostrazione per una matrice  $4 \times 4$ . La sua matrice di Jordan avrà forma

$$J = \begin{bmatrix} 0 & -\frac{c_1}{a_1} & 0 & 0 \\ -\frac{b_1}{a_2} & 0 & -\frac{c_2}{a_2} & 0 \\ 0 & -\frac{b_2}{a_3} & 0 & -\frac{c_3}{a_3} \\ 0 & 0 & -\frac{b_3}{a_4} & 0 \end{bmatrix}$$

Prendiamo la matrice  $S$  e la sua inversa  $S^{-1}$  così fatte

$$S = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \alpha & 0 & 0 \\ 0 & 0 & \alpha^2 & 0 \\ 0 & 0 & 0 & \alpha^3 \end{bmatrix} \quad S^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{\alpha} & 0 & 0 \\ 0 & 0 & \frac{1}{\alpha^2} & 0 \\ 0 & 0 & 0 & \frac{1}{\alpha^3} \end{bmatrix} \quad \alpha \neq 0$$

e calcoliamo  $SJS^{-1}$ , che è una trasformazione per similitudine (che preserva dunque gli autovalori).

$$SJS^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \alpha & 0 & 0 \\ 0 & 0 & \alpha^2 & 0 \\ 0 & 0 & 0 & \alpha^3 \end{bmatrix} \begin{bmatrix} 0 & -\frac{c_1}{a_1} & 0 & 0 \\ -\frac{b_1}{a_2} & 0 & -\frac{c_2}{a_2} & 0 \\ 0 & -\frac{b_2}{a_3} & 0 & -\frac{c_3}{a_3} \\ 0 & 0 & -\frac{b_3}{a_4} & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{\alpha} & 0 & 0 \\ 0 & 0 & \frac{1}{\alpha^2} & 0 \\ 0 & 0 & 0 & \frac{1}{\alpha^3} \end{bmatrix} =$$

$$\begin{bmatrix} 0 & -\frac{c_1}{a_1} & 0 & 0 \\ -\frac{ab_1}{a_2} & 0 & -\frac{ac_2}{a_2} & 0 \\ 0 & -\frac{a^2b_2}{a_3} & 0 & -\frac{a^2c_3}{a_3} \\ 0 & 0 & -\frac{a^3b_3}{a_4} & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{\alpha} & 0 & 0 \\ 0 & 0 & \frac{1}{\alpha^2} & 0 \\ 0 & 0 & 0 & \frac{1}{\alpha^3} \end{bmatrix} =$$

$$\begin{bmatrix} 0 & -\frac{c_1}{\alpha a_1} & 0 & 0 \\ -\frac{ab_1}{a_2} & 0 & -\frac{c_2}{\alpha a_2} & 0 \\ 0 & -\frac{ab_2}{a_3} & 0 & -\frac{c_3}{\alpha a_3} \\ 0 & 0 & -\frac{ab_3}{a_4} & 0 \end{bmatrix}$$

Possiamo dunque esprimere  $SJS^{-1}$  come  $\alpha D^{-1}B + \frac{1}{\alpha}D^{-1}C$ , dove B, C e D sono le matrici relative alla decomposizione di Jacobi e Gauss-Seidel viste qualche pagina fa.

Se  $\lambda$  è autovalore di  $J$ , allora lo è anche di  $SJS^{-1}$ : posso dunque scrivere

$$\begin{aligned} 0 &= \det(SJS^{-1} - \lambda I) = \\ &= \det(\alpha D^{-1}B + \frac{1}{\alpha}D^{-1}C - \lambda I) = \\ &= \det(\frac{1}{\alpha}(\alpha^2 D^{-1}B + D^{-1}C - \lambda \alpha I)) = \\ &= \frac{1}{\alpha^n} \det((\alpha^2 D^{-1}B + D^{-1}C - \lambda \alpha I)) = 0 \end{aligned}$$

Prendendo  $\alpha = \lambda$ , ed essendo sicuramente  $\frac{1}{\alpha^n} \neq 0$ , abbiamo che l'ultima riga è verificata quando

$$\det((\lambda^2 D^{-1}B + D^{-1}C - \lambda^2 I)) = 0 \quad (4.1)$$

Prendiamo ora la G di Gauss-Seidel e calcoliamone gli autovalori.

$$\begin{aligned} \det(G - \mu I) &= \\ &= \det((D - B)^{-1}C - \mu I) = \\ &= \det[(D - B)^{-1}[C - \mu(D - B)]] = \\ &= \det\left[\left[D(I - D^{-1}B)\right]^{-1}[C - \mu(D - B)]\right] = \\ &= \det\left[(I - D^{-1}B)^{-1}D^{-1}(C - \mu D + \mu B)\right] = \\ &= \det\left[(I - D^{-1}B)^{-1}(D^{-1}C - \mu I + \mu D^{-1}B)\right] = \\ &= \det(I - D^{-1}B)^{-1} \det(D^{-1}C - \mu I + \mu D^{-1}B) = 0 \end{aligned}$$

Osserviamo l'ultima riga: si tratta del prodotto di due determinanti che deve avere come risultato 0. Il primo fattore non può essere 0, essendo il determinante di un'inversa, in quanto le matrici inverse sono non singolari. Deve quindi valere

$$\det(D^{-1}C - \mu I + \mu D^{-1}B) = 0 \quad (4.2)$$

Confrontando le equazioni 4.1 e 4.2, notiamo che gli autovalori  $\mu$  di G sono i quadrati degli autovalori  $\lambda$  di J, come volevasi dimostrare.  $\square$

## 5 Metodi di risoluzione per sistemi non lineari

Andremo adesso a vedere come si possono risolvere i sistemi non lineari. Per prima cosa caratterizziamoli. Possiamo definire un sistema non lineare in due maniere: dato  $\Omega \subseteq \mathbb{R}^n$ , con  $\Omega$  sottoinsieme aperto di  $\mathbb{R}^n$ :

- $F(\mathbf{x}) = \mathbf{0}$  dove  $F : \Omega \rightarrow \mathbb{R}^n$ , dove

$$F(\mathbf{x}) = \begin{cases} f_1(\mathbf{x}) = 0 \\ f_2(\mathbf{x}) = 0 \\ \vdots \\ f_n(\mathbf{x}) = 0 \end{cases}$$

Le  $f_i$  sono quindi funzioni del tipo  $\Omega \rightarrow \mathbb{R}$ .

- $\mathbf{x} = G(\mathbf{x})$  dove  $G : \Omega \rightarrow \mathbb{R}^n$ , ovvero

$$\begin{cases} x_1 = g_1(\mathbf{x}) \\ x_2 = g_2(\mathbf{x}) \\ \vdots \\ x_n = g_n(\mathbf{x}) \end{cases}$$

Anche le  $g_i$  sono quindi funzioni del tipo  $\Omega \rightarrow \mathbb{R}$ .

Inoltre, se  $F(\mathbf{x}), G(\mathbf{x}) \in C^1(\Omega)$ , le due funzioni generano le matrici Jacobiane, che chiameremo rispettivamente  $J(\mathbf{x})$  ed  $H(\mathbf{x})$ .

Per arricchire le possibilità di personalizzare il metodo risolutivo, possiamo porre

$$\mathbf{x} = G(\mathbf{x}) = \mathbf{x} - A(\mathbf{x})F(\mathbf{x})$$

per qualche  $A(x)$ .

Tale relazione si verifica in quanto  $F(\mathbf{x}) = \mathbf{0}$  nella soluzione.

### 5.1 Generalità sui metodi iterativi per sistemi non lineari

Possiamo dunque studiare le proprietà generali della famiglia dei metodi iterativi che seguono lo schema

$$x^{(i+1)} = G(x^{(i)}) = x^{(i)} - A(x^{(i)})F(x^{(i)})$$

e che definiscono una successione di  $\mathbf{x}^{(i)} \in \mathbb{R}^n$  che converge ad  $\alpha$  se  $\lim_{i \rightarrow \infty} \|\mathbf{x}^{(i)} - \alpha\| = 0$

Vediamo adesso di capire se e quando questo schema iterativo può convergere alla soluzione. Enunciamo un teorema che ci da una condizione sufficiente per lo schema iterativo  $\mathbf{x}^{(i+1)} = G(\mathbf{x}^{(i)})$ .



#### **Teorema 5.1 (Teorema del punto fisso (sufficiente))**

Sia  $\mathbf{x} = G(\mathbf{x})$  un sistema non lineare, e  $\alpha$  una soluzione di tale sistema (tale dunque che  $\alpha = G(\alpha)$ ). Sia  $S$  un intorno di  $\alpha$ , tale cioè che per un certo valore di  $\rho > 0$  vale

$$S = \{x \in \mathbb{R}^n, \|\mathbf{x} - \alpha\|_\infty \leq \rho\}$$

Sia inoltre

$$\forall \mathbf{x} \in S, \|H(\mathbf{x})\|_\infty < 1$$

Allora si ha convergenza per  $\mathbf{x}^{(0)} \in S$ .

*Dimostrazione.* Per induzione si dimostra che

$$\|\mathbf{x}^{(i)} - \alpha\|_\infty \leq \lambda^i \rho \quad \text{dove } \lambda = \max_S \|H(\mathbf{x})\|_\infty$$

$P(0)$  banalmente  $\|\mathbf{x}^{(0)} - \alpha\|_\infty \leq 1 \cdot \rho$  è vera per ipotesi

$P(i) \Rightarrow P(i+1)$  So che  $\mathbf{x}^{(i)} - \alpha = G(\mathbf{x}^{(i-1)}) - G(\alpha)$ . Si tratta di un vettore: possiamo descrivere la  $r$ -esima componente, utilizzando il teorema del valor medio (16.2), come

$$\mathbf{x}_r^{(i)} - \alpha_r = g_r(\mathbf{x}^{(i-1)}) - g_r(\alpha) = \nabla g_r(\xi_r)(\mathbf{x}^{(i-1)} - \alpha) = \sum_{s=1}^n \frac{\partial g_r}{\partial x_s}(\xi_r)(x_s^{(i-1)} - \alpha_s)$$

e  $\nabla g_r(\xi_r)$  è la riga  $r$ -esima della matrice jacobiana di  $G$  calcolata in  $\xi_r$ . Passiamo al modulo:

$$|x_r^{(i)} - \alpha_r| = \left| \sum_{s=1}^n \frac{\partial g_r}{\partial x_s}(\xi_r)(x_s^{(i-1)} - \alpha_s) \right|$$

Per la disuguaglianza triangolare tra moduli (16.1 a pag. 226) si ha:

$$\left| \sum_{s=1}^n \frac{\partial g_r}{\partial x_s}(\xi_r)(x_s^{(i-1)} - \alpha_s) \right| \leq \sum_{s=1}^n \left| \frac{\partial g_r}{\partial x_s}(\xi_r)(x_s^{(i-1)} - \alpha_s) \right|$$

Ricordando che  $|ab| = |a| \cdot |b|$ , si ha infine

$$|x_r^{(i)} - \alpha_r| \leq \sum_{s=1}^n \left| \frac{\partial g_r}{\partial x_s}(\xi_r) \right| |x_s^{(i-1)} - \alpha_s|$$

Considerando che la norma infinito di un vettore è  $\|y\|_\infty = \max_{i=1 \dots n} |y_i|$ , si ha

$$\sum_{s=1}^n \left( \left| \frac{\partial g_r}{\partial x_s}(\xi_r) \right| |x_s^{(i-1)} - \alpha_s| \right) \leq \sum_{s=1}^n \left( \left| \frac{\partial g_r}{\partial x_s}(\xi_r) \right| \|x^{(i-1)} - \alpha\|_\infty \right) = \sum_{s=1}^n \left( \left| \frac{\partial g_r}{\partial x_s}(\xi_r) \right| \right) \|x^{(i-1)} - \alpha\|_\infty$$

Inoltre, considerando che la norma infinito di una matrice è  $\|A\|_\infty = \max_{i=1 \dots n} \sum_{j=1}^n |A_{ij}|$ , e quindi

$$\|H(x)\|_\infty = \max_{i=1 \dots n} \sum_{s=1}^n \left| \frac{\partial g_i}{\partial x_s}(x) \right|$$

si ha

$$\sum_{s=1}^n \left( \left| \frac{\partial g_r}{\partial x_s}(\xi_r) \right| \right) \|x^{(i-1)} - \alpha\|_\infty \leq \|H(\xi_r)\|_\infty \|x^{(i-1)} - \alpha\|_\infty \leq \lambda \|x^{(i-1)} - \alpha\|_\infty$$

In definitiva sappiamo che

$$|x_r^{(i)} - \alpha_r| \leq \lambda \|x^{(i-1)} - \alpha\|_\infty \quad r = 1 \dots n$$

questo vale in particolare per  $r$  che massimizza  $|x_r^{(i)} - \alpha_r|$

$$\|x^{(i)} - \alpha\|_\infty \leq \lambda \|x^{(i-1)} - \alpha\|_\infty$$

Possiamo quindi scrivere

$$\|x^{(i)} - \alpha\|_\infty \leq \lambda \|x^{(i-1)} - \alpha\|_\infty \leq \lambda^2 \|x^{(i-2)} - \alpha\|_\infty \leq \dots \leq \lambda^i \|x^{(0)} - \alpha\|_\infty \leq \lambda^i \rho$$

in quanto  $x^{(0)} \in S \Rightarrow \|x^{(0)} - \alpha\|_\infty \leq \rho$ .

Detto questo, è facile vedere che, essendo  $\lambda < 1$ ,  $\lambda^i \rho \xrightarrow{i \rightarrow \infty} 0$  e che quindi  $\|x^{(i)} - \alpha\|_\infty \xrightarrow{i \rightarrow \infty} 0$   $\square$

Questo teorema fornisce una condizione sufficiente per la convergenza. Se quindi si verificano le ipotesi, possiamo affermare con certezza che il metodo converge. Ma il metodo può convergere anche se le condizioni non si verificano!

Vediamo un esempio.

**Esempio 5.2**

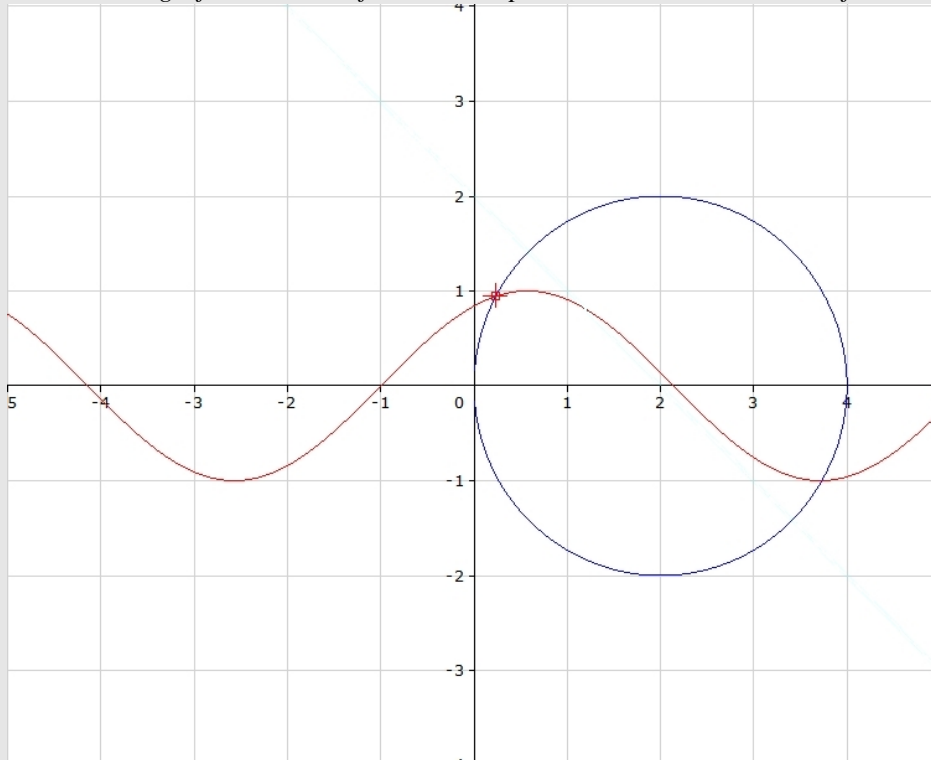
Consideriamo il sistema non lineare

$$\begin{cases} x_1 = g_1(x) = \frac{1}{4}(x_1^2 + x_2^2) \\ x_2 = g_2(x) = \sin(x_1 + 1) \end{cases} \quad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

Vogliamo

- Verificare l'esistenza di punti fissi
- Applicare il teorema del punto fisso

Tracciamo il grafico delle due funzioni. La prima delle due è una circonferenza, la seconda un seno.



Come possiamo vedere, ci sono due punti di incontro, che chiameremo  $\alpha$  (quello più vicino all'asse delle ordinate) e  $\beta$ . Ora devo trovare un intorno  $S$  di uno dei due punti tale che  $\forall x \in S. \|H(x)\|_\infty < 1$ . Calcoliamo  $H(x)$ , che ricordiamo essere la matrice la quale, sulle righe, ha i gradienti delle funzioni che compongono  $G$ .

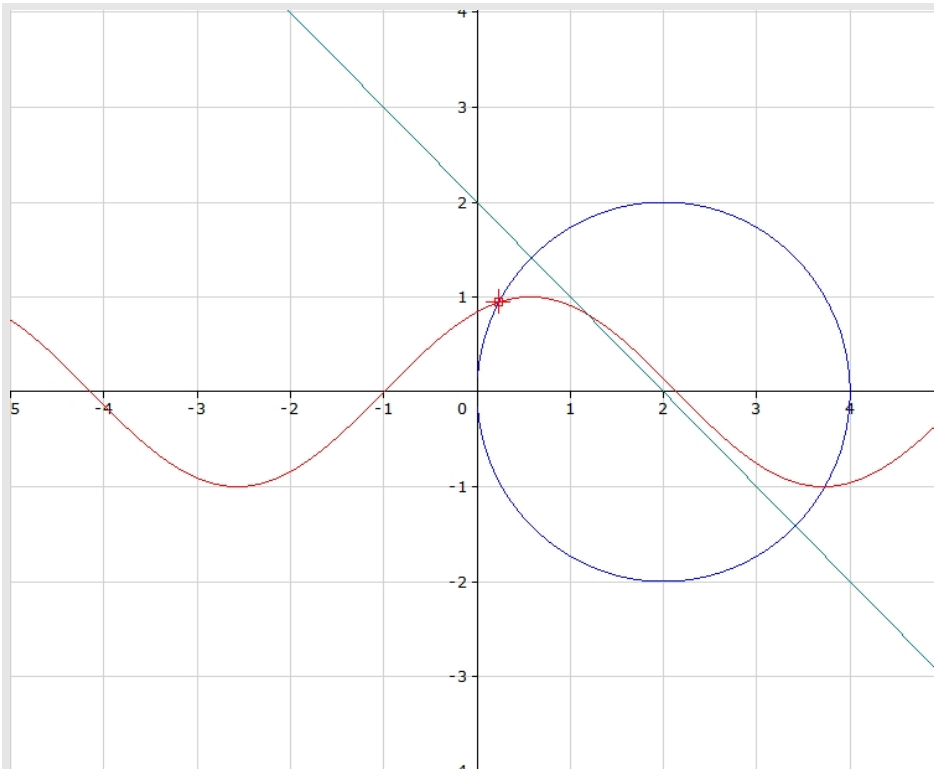
$$H(x) = \begin{bmatrix} \frac{x_1}{2} & \frac{x_2}{2} \\ \cos(x_1 + 1) & 0 \end{bmatrix}$$

Devo ora calcolare la norma infinito della matrice, lasciando  $x$  parametrico. La norma infinito è la massima somma degli elementi delle righe di

$$\|H(x)\|_\infty = \max\left(\frac{1}{2}(|x_1| + |x_2|), |\cos(x_1 + 1)|\right)$$

Voglio che  $\|H(x)\|_\infty$  sia minore di 1: ovviamente  $|\cos(x_1 + 1)| < 1$ , quindi rimane da vedere che  $\frac{1}{2}(|x_1| + |x_2|) < 1$ . Questo accade quando  $x_1 + x_2 < 2$ . Possiamo dunque tracciare l'equazione della retta  $y = -x + 2$  sotto la quale vale sempre  $\|H(x)\|_\infty < 1$ . Possiamo vedere il grafico di questa retta nella pagina successiva.

Se prendiamo  $\alpha$ , appare evidente che possiamo costruirne un intorno  $S$  non vuoto tale che  $\forall x \in S. H(x) < 1$ : basta che tale intorno non oltrepassi la retta  $x_1 + x_2 < 2$ . Non possiamo dire la stessa cosa per  $\beta$ : infatti  $H(\beta) > 1$  ed è quindi impossibile costruirne un intorno in cui la proprietà non vada. Ma può darsi che il metodo funzioni comunque, in quanto la proprietà enunciata non è necessaria.



Enunciamo un secondo teorema del punto fisso, che è necessario e sufficiente.



### Teorema 5.3 (Teorema del punto fisso (necessario e sufficiente))

Siano

- $g(\mathbf{x})$  di classe  $C^1(\Omega)$
- $\Omega$  insieme aperto
- $\alpha$  punto di  $\Omega$  tale che  $\alpha = g(\alpha)$  ( $\alpha$  è punto fisso di  $g$ )

Allora esiste una norma  $\|\cdot\|_*$  vettoriale ed un intorno  $S = \{x : \|x - \alpha\|_* < \pi\}$  tale che per  $\mathbf{x}^{(0)} \in S$  si ha convergenza se e solo se  $\rho(H(\alpha)) \leq 1$ .

Non dimostreremo questo teorema. Piuttosto viene da chiedersi il perché delle differenze fra i due teoremi: uno pone condizioni sulla norma, l'altro sul raggio spettrale.

Effettivamente esiste un legame fra queste due grandezze: si veda il Teorema 4.21 a pagina 82, il quale asserisce che per ogni matrice  $B$  e per ogni norma indotta  $\|\cdot\|$  vale

$$\rho(B) = \inf_{\|\cdot\| \text{ indotta}} \{\|B\|\}$$

In particolare, se  $\rho(B) < 1$ , possiamo trovare, prendendo  $\epsilon < 1 - \rho(B)$ , una norma  $\|\cdot\|_*$  tale che

$$\rho(B) \leq \|B\|_* \leq \rho(B) + \epsilon$$

verificando dunque la condizione del teorema del punto fisso nella sua versione sufficiente, in quanto per la continuità di  $g$  esisterà un intorno di  $\alpha$  nel quale la matrice hessiana avrà norma minore di 1.



## 5.2 Metodo di Newton-Raphson (delle tangenti)

Vediamo adesso un metodo per la risoluzione di sistemi non lineari che segue lo schema presentato nel paragrafo precedente: il metodo di Newton-Raphson. Si tratta di un'estensione del metodo di Newton. Prendiamo il nostro problema, espresso al solito nelle due forme alternative  $F(\mathbf{x}) = \mathbf{0}$  e  $\mathbf{x} = G(\mathbf{x})$ . Abbiamo visto che possiamo sempre ottenere da tali espressioni la relazione

$$\mathbf{x} = G(\mathbf{x}) = \mathbf{x} - A(\mathbf{x})F(\mathbf{x})$$

per una qualche funzione  $A(\mathbf{x})$ . Il metodo di Newton-Raphson prende  $A(\mathbf{x}) = J^{-1}(\mathbf{x})$ , dove  $J(\mathbf{x})$  è la matrice Jacobiana di  $F$ . Manipolando otteniamo che

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - J^{-1}(\mathbf{x}^{(i)})F(\mathbf{x}^{(i)}) \quad \Rightarrow \quad J(\mathbf{x}^{(i)})(\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}) = -F(\mathbf{x}^{(i)})$$

che è un sistema lineare che ha  $J(\mathbf{x}^{(i)})$  come matrice dei coefficienti,  $\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}$  come incognite e  $-F(\mathbf{x}^{(i)})$  come vettore dei termini noti.

### Esempio 5.4 (Matrice associata al metodo di Newton-Raphson)

Nel caso delle matrici  $2 \times 2$  ottengo il sistema

$$\begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x^{(i)}) & \frac{\partial f_1}{\partial x_2}(x^{(i)}) \\ \frac{\partial f_2}{\partial x_1}(x^{(i)}) & \frac{\partial f_2}{\partial x_2}(x^{(i)}) \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} -f_1(x^{(i)}) \\ -f_2(x^{(i)}) \end{bmatrix}$$

dove  $\theta = x^{(i+1)} - x^{(i)}$

Al solito, dobbiamo dimostrare che il metodo converge.



### Teorema 5.5 (Convergenza di Newton-Raphson)

Siano

- $F(\mathbf{x}) \in C^2(\Omega)$  con matrice jacobiana  $J(\mathbf{x})$
- $\alpha \in \Omega$  tale che  $F(\alpha) = \mathbf{0}$

Allora, se  $J(\mathbf{x})$  è non singolare in  $\Omega$ ,  $\exists S \subseteq \Omega$  intorno di  $\alpha$  tale che, preso  $\mathbf{x}^{(0)} \in S$ :

(a) la successione  $\{\mathbf{x}^{(i)}\}_i$  generata dal metodo di Newton-Raphson a partire dal punto  $\mathbf{x}^{(0)}$  converge a  $\alpha$

(b)  $\forall \epsilon > 0, \exists \beta > 0, \forall i > 0, \|\mathbf{x}^{(i+1)} - \alpha\| \leq \beta \|\mathbf{x}^{(i)} - \alpha\|^2$

*Dimostrazione.* Dimostriamo i due punti del teorema.

(a) Per comodità di notazione nella dimostrazione scriveremo  $K(\mathbf{x}) = J^{-1}(\mathbf{x})$ . Abbiamo dunque

$$G(\mathbf{x}) = \mathbf{x} - K(\mathbf{x})F(\mathbf{x})$$

Si tratta di un'uguaglianza fra vettori: prendiamone la  $r$ -esima componente.

$$g_r(\mathbf{x}) = x_r - \sum_{s=1}^n K_{rs}(\mathbf{x})f_s(\mathbf{x})$$

Deriviamo rispetto ad  $x_t$ , per  $t \in [1, n]$ , ottenendo i valori della matrice jacobiana  $H$  di  $g$ .

$$h_{rt}(\mathbf{x}) = \frac{\partial g_r}{\partial x_t}(\mathbf{x}) = \delta_{rt} - \sum_{s=1}^n \frac{\partial K_{rs}}{\partial x_t}(\mathbf{x}) f_s(\mathbf{x}) - \sum_{s=1}^n K_{rs}(\mathbf{x}) \frac{\partial f_s}{\partial x_t}(\mathbf{x})$$

dove  $\delta_{rt}$  vale 1 se  $r = t$  e 0 altrimenti. Analizziamo il terzo addendo: si tratta del prodotto fra la  $r$ -esima riga di  $K$  (ovvero la  $r$ -esima riga di  $J^{-1}$ ) e la  $t$ -esima colonna di  $J$ . Visto che  $JJ^{-1} = I$  per definizione di matrice inversa, so che tale prodotto deve fare 1 se  $r = t$  e 0 altrimenti. Ho quindi che  $\delta_{rt} = \sum_{s=1}^n K_{rs}(\mathbf{x}) \frac{\partial f_s}{\partial x_t}(\mathbf{x})$ , e posso annullarli uno con l'altro. Rimane dunque

$$h_{rt}(\mathbf{x}) = - \sum_{s=1}^n \frac{\partial K_{rs}}{\partial x_t}(\mathbf{x}) f_s(\mathbf{x})$$

che implica che  $h_{rt}(\alpha) = 0$ , in quanto  $\forall s. f_s(\alpha) = 0$ . Quindi  $\rho(H(\alpha)) = 0$ , e posso applicare il teorema 5.3 del punto fisso che mi assicura la convergenza in un intorno  $S$  di  $\alpha$ .

(b) Vale

$$\begin{aligned} J(x^{(i)})(x^{(i+1)} - x^{(i)}) &= -F(x^{(i)}) \\ J(x^{(i)})[(x^{(i+1)} - \alpha) - (x^{(i)} - \alpha)] &= -F(x^{(i)}) \\ J(x^{(i)})(x^{(i+1)} - \alpha) &= J(x^{(i)})(x^{(i)} - \alpha) - F(x^{(i)}) \end{aligned} \quad (5.1)$$

Indichiamo con  $S_r$  le matrici hessiane delle funzioni  $f_r$ .

$$(S_r(x))_{st} = \frac{\partial^2 f_r}{\partial x_s \partial x_t}(x)$$

Applicando la sostituzione di Taylor posso porre

$$\begin{aligned} F(\alpha) &= F(x^{(i)}) + J(x^{(i)})(\alpha - x^{(i)}) + v \\ -F(x^{(i)}) &= -F(\alpha) + J(x^{(i)})(\alpha - x^{(i)}) + v \end{aligned} \quad (5.2)$$

dove  $v = \frac{1}{2}(x^{(i)} - \alpha)^T S_r(\xi_r)(x^{(i)} - \alpha)$ , e passando a moduli e norme,

$$|v| \leq \frac{n}{2} \|S_r(\xi_r)\|_{\infty} \|x^{(i)} - \alpha\|_{\infty}^2 \quad (5.3)$$

Abbiamo tutti i pezzi per trarre la conclusione: prendiamo l'equazione 5.1 ed applichiamo la sostituzione 5.2:

$$\begin{aligned} J(x^{(i)})(x^{(i+1)} - \alpha) &= J(x^{(i)})(x^{(i)} - \alpha) - F(x^{(i)}) \\ J(x^{(i)})(x^{(i+1)} - \alpha) &= J(x^{(i)})(x^{(i)} - \alpha) - F(\alpha) + J(x^{(i)})(\alpha - x^{(i)}) + v \end{aligned}$$

Essendo  $F(\alpha) = 0$  e  $(x^{(i)} - \alpha) = -(\alpha - x^{(i)})$ , ottengo  $(x^{(i+1)} - \alpha) = J^{-1}(x^{(i)})v$ . Passando alle norme ed applicando 5.3 ottengo

$$\|x^{(i+1)} - \alpha\|_{\infty} \leq \gamma \|x^{(i)} - \alpha\|_{\infty}^2$$

per  $\gamma = \frac{n}{2} \max_{x \in S} \|K(x)\|_{\infty} \max_{x \in S, r \in [1, n]} \|S_r(x)\|_{\infty}$

Notare che  $\gamma$  è costante: per l'equivalenza topologica fra le norme, esisterà un  $\beta$  per cui questa relazione vale per qualunque norma  $\|\cdot\|$ , come volevasi dimostrare.

□

### 5.2.1 Newton-Raphson su funzioni convesse

Per via delle proprietà delle funzioni convesse, possiamo utilizzare Newton-Raphson con risultati di convergenza sensibilmente migliori rispetto al normale. Otterremo questo risultato grazie ai due teoremi seguenti.

Il primo teorema lega le proprietà della matrice Jacobiana alla convessità.


**Teorema 5.6 (Legame fra matrice Jacobiana e convessità)**

Siano

- $D$  insieme convesso
- $f \in C^1(D)$

Allora  $f$  è convessa su  $D$  se e solo se  $f(v) - f(u) \geq J(u)(v - u)$

*Dimostrazione.* Dimostriamo i due sensi della doppia implicazione

$\Leftarrow$  Assumiamo  $f(v) - f(u) \geq J(u)(v - u)$ . Siano

- $x, y \in D$
- $\lambda \in [0, 1]$
- $z(\lambda) = \lambda \cdot x + (1 - \lambda) \cdot y$

Ovviamente, essendo  $D$  convesso, avremo che  $z(\lambda) \in D$  per definizione di insieme convesso. Possiamo dire che

$$\begin{aligned} f(x) - f(z(\lambda)) &\geq J(z(\lambda))(x - z(\lambda)) \\ f(y) - f(z(\lambda)) &\geq J(z(\lambda))(y - z(\lambda)) \end{aligned}$$

Moltiplichiamo la prima equazione per  $\lambda$  e la seconda per  $(1 - \lambda)$ , e sommiamo membro a membro. Otterremo

$$\lambda f(x) + (1 - \lambda)f(y) - f(z(\lambda)) \geq J(z(\lambda))(\lambda x + (1 - \lambda)y - z(\lambda))$$

Ma

$$\lambda x + (1 - \lambda)y - z(\lambda) = \lambda x + (1 - \lambda)y - \lambda x - (1 - \lambda)y = 0$$

La parte destra è dunque uguale a zero: rimane la relazione

$$\lambda f(x) + (1 - \lambda)f(y) \geq f(z(\lambda)) = f(\lambda x + (1 - \lambda) \cdot y)$$

che corrisponde alla definizione di funzione convessa vista in 3.1 a pagina 51.

$\Rightarrow$  Supponiamo adesso che  $f$  sia convessa in  $D$ . Siano

- $u, v \in D$
- $\lambda \in [0, 1]$
- $w(\lambda) = \lambda v + (1 - \lambda)u$

Essendo  $f \in C^1(D)$ , per una qualunque norma  $\|\cdot\|$  esiste il *differenziale totale*

$$\lim_{x' \rightarrow x} \frac{\|f(x') - f(x) - J(x)(x' - x)\|}{\|x' - x\|} = 0$$

Preso  $x' = w(\lambda)$ ,  $x = u$  ho, considerando che  $\|w(\lambda) - u\| = \|\lambda v + (1 - \lambda)u - u\| = \|\lambda v - \lambda u\| = |\lambda|\|v - u\|$  e che  $w(\lambda) \rightarrow u \Leftrightarrow \lambda \rightarrow 0$ ,

$$\begin{aligned} \lim_{\lambda \rightarrow 0} \frac{\|f(w(\lambda)) - f(u) - J(u)(w(\lambda) - u)\|}{\|w(\lambda) - u\|} &= 0 \Leftrightarrow \\ \Leftrightarrow \lim_{\lambda \rightarrow 0} \frac{1}{\lambda} \|f(w(\lambda)) - f(u) - \lambda J(u)(v - u)\| &= 0 \Leftrightarrow \\ \Leftrightarrow \lim_{\lambda \rightarrow 0} \frac{1}{\lambda} [f(w(\lambda)) - f(u)] &= J(u)(v - u) \end{aligned}$$

Ma abbiamo che, per la definizione di funzione convessa,

$$\lambda f(v) + (1 - \lambda)f(u) \geq f(w(\lambda)) \Leftrightarrow f(v) - f(u) \geq \frac{1}{\lambda} [f(w(\lambda)) - f(u)]$$

e che quindi

$$\lambda f(v) + (1 - \lambda)f(u) \geq J(u)(v - u)$$

come volevasi dimostrare

□

Detto questo, possiamo dimostrare che



### Teorema 5.7 (Convergenza di Newton-Raphson per funzioni convesse)

Siano

- $a_i, b_i \in \mathbb{R}^n, i \in [1, n]$
- $D$  intervallo di  $\mathbb{R}^n$
- $f \in C^1(D)$  convessa su  $D$
- $\alpha \in D$  tale che  $f(\alpha) = 0$

Allora, se si verificano le seguenti condizioni

- $\forall x \in D. J(x)$  non è singolare
- $[J(x)]^{-1} \geq 0$

e considerata la successione  $\{x^{(i)}\}_i$  generata dal metodo di Newton-Raphson, vale la seguente relazione:

$$\forall x^{(0)} \in D \text{ tale che } f(x^{(0)}) \geq 0 \quad \lim_{i \rightarrow \infty} x^{(i)} \rightarrow \alpha$$

Inoltre  $\alpha$  è l'unica soluzione del problema.

*Dimostrazione.* Per prima cosa dimostriamo per induzione su  $i$  che

- $\forall i. \alpha \leq x^{(i+1)} \leq x^{(i)}$
- $f(x^{(i+1)}) \geq 0$

Il secondo punto servirà solamente per applicare l'induzione. Il primo punto è quello importante: ci dice che la successione di  $x^{(i)}$  si avvicina verso  $\alpha$ .

**Passo base** Dal teorema 5.6 sappiamo che, se  $f$  è convessa, vale  $f(\alpha) - f(x^{(0)}) \geq J(x^{(0)})(\alpha - x^{(0)})$ , e che dunque

$$- [J(x^{(0)})]^{-1} f(x^{(0)}) \geq \alpha - x^{(0)} \quad (5.4)$$

Essendo  $[J(x^{(0)})]^{-1} \geq 0$  e  $f(x^{(0)}) \geq 0$  per ipotesi, ottengo

- $\alpha \leq x^{(1)}$  in quanto  $\alpha - x^{(0)}$  è un valore negativo
- $\alpha \leq x^{(1)} = x^{(0)} - [J(x^{(0)})]^{-1} f(x^{(0)}) \leq x^{(0)}$

Inoltre, sempre per il fatto che  $f$  è convessa, posso applicare il teorema 5.6 ottenendo

$$f(x^{(1)}) - f(x^{(0)}) \geq J(x^{(0)})(x^{(1)} - x^{(0)})$$

Questa relazione, essendo  $J(x^{(0)})(x^{(1)} - x^{(0)}) = -f(x^{(0)})$  per definizione del metodo, implica  $f(x^{(1)}) \geq 0$ . Inoltre, essendo  $x^{(1)}$  compreso fra  $x^{(0)}$  e  $\alpha$ , abbiamo che  $x^{(1)} \in D$  e che quindi  $[J(x^{(1)})]^{-1} \geq 0$

**Passo induttivo** Essendo  $[J(x^{(1)})]^{-1} \geq 0$  e  $f(x^{(1)}) \geq 0$  posso applicare esattamente la stessa dimostrazione vista per il passo base.

La successione è quindi decrescente e inferiormente limitata: possiamo esser sicuri che converge ad un valore  $\beta$  per il quale

$$\begin{aligned} f(\beta) &= \lim_{i \rightarrow \infty} f(x^{(i)}) = \\ &= \lim_{i \rightarrow \infty} J(x^{(i)}) [x^{(i)} - x^{(i-1)}] = \\ &= \lim_{i \rightarrow \infty} J(x^{(i)}) [\beta - \beta] = 0 \end{aligned}$$

Quindi  $\beta$  è un'altra soluzione dell'equazione: ma dal teorema 5.6 posso concludere che

$$0 = f(\alpha) - f(\beta) \geq J(\alpha)(\beta - \alpha)$$

$$0 = f(\beta) - f(\alpha) \geq J(\beta)(\alpha - \beta)$$

ed essendo  $[J(\alpha)]^{-1} \geq 0$ ,  $[J(\beta)]^{-1} \geq 0$  ho

$$\begin{aligned} 0 &\geq \beta - \alpha \\ 0 &\geq \alpha - \beta \end{aligned} \Rightarrow \alpha = \beta$$

Quindi, ricapitolando

- la successione converge ad un valore
- tale valore è  $\alpha$ , che è soluzione *unica*

□



## 6 Ottimizzazione non vincolata

Il problema che affronteremo in questo capitolo è quello già enunciato nel Capitolo 3, ovvero quello dell'ottimizzazione *non vincolata*, nel quale si cerca il minimo di una funzione in tutto lo spazio  $\mathbb{R}^n$ :

$$\min\{f(x) : x \in \mathbb{R}^n\} \quad f : \mathbb{R}^n \rightarrow \mathbb{R}$$

Abbiamo già visto che in questo tipo di problemi il minimo si trova necessariamente in un punto  $\bar{x}$  stazionario, ovvero

$$\bar{x} \text{ punto di minimo} \implies \nabla f(\bar{x}) = 0$$

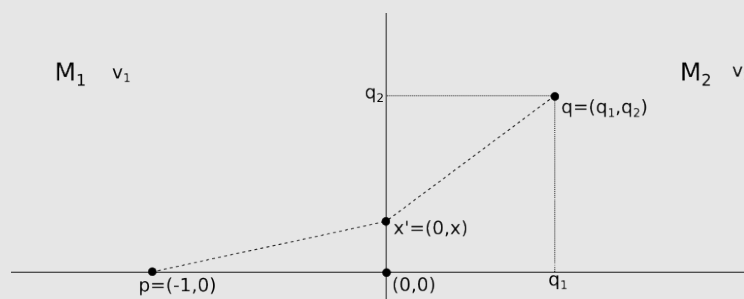
Nel caso che la funzione sia convessa, questa è condizione anche sufficiente. Nel caso che la funzione sia quadratica, il soddisfacimento della condizione si riconduce alla risoluzione di un sistema lineare. Altrimenti, in generale il problema che si forma è non-lineare, e lo possiamo risolvere ad esempio con il metodo Newton-Raphson visto nella Sezione 5.2. In questo capitolo vedremo alcuni metodi iterativi (di cui uno basato su Newton-Raphson) per risolvere il problema nel caso che  $f$  sia non lineare.

Ma guardiamo innanzi tutto degli esempi di questi problemi.

### Esempio 6.1

Vediamo un semplice problema di ottimizzazione non vincolata. Abbiamo un piano diviso da una retta (in 2 semipiani). In questi piani un oggetto si muove attraverso due mezzi  $M_1$  e  $M_2$  (es. acqua, aria) che hanno velocità di percorrenza rispettivamente  $v_1, v_2$ . L'oggetto deve muoversi da un punto  $p \in M_1$  a  $q \in M_2$  minimizzando il tempo di percorrenza.

Per semplificare il modello, si prende l'asse  $x$  passante per  $p$  e scegliamo un'unità di misura tale che sia  $p = (-1, 0)$ .



All'interno dello stesso mezzo, il percorso più breve tra due punti è ovviamente la retta. Il percorso sarà dunque determinato dal parametro  $x$ , ordinata del punto in cui il percorso passa da  $M_1$  a  $M_2$ , e sarà composto da due segmenti che uniscono i punti  $p = (-1, 0), x' = (0, x), q = (q_1, q_2)$ .

I due segmenti sono lunghi rispettivamente  $\sqrt{1+x^2}$  e  $\sqrt{q_1^2 + (q_2 - x)^2}$ .

La funzione che calcola il tempo di percorrenza, da minimizzare, è la seguente:

$$f(x) = \frac{\sqrt{1+x^2}}{v_1} + \frac{\sqrt{q_1^2 + (q_2 - x)^2}}{v_2}$$

Dobbiamo risolvere il problema di ottimizzazione senza vincoli:

$$\min\{f(x) : x \in \mathbb{R}\}$$

Cerchiamo i punti in cui il gradiente della funzione si annulla. Dato che la funzione è a una variabile, il gradiente è la derivata:

$$\nabla f(x) = f'(x) = 0$$

$$f'(x) = \frac{1}{v_1} \frac{2x}{2 \cdot \sqrt{1+x^2}} + \frac{1}{v_2} \frac{2x-2q_2}{2 \sqrt{q_1^2 + (q_2-x)^2}} = \frac{x}{v_1 \sqrt{1+x^2}} - \frac{q_2-x}{v_2 \sqrt{q_1^2 + (q_2-x)^2}}$$

Notiamo che i denominatori non si annullano mai e risolviamo  $f'(x) = 0$ .

$$v_2 x \sqrt{q_1^2 + (q_2-x)^2} - v_1 (q_2-x) \sqrt{1+x^2} = 0$$

Abbiamo ottenuto una funzione non lineare che da il punto di stazionarietà. Inoltre la funzione è convessa, quindi in questo caso la condizione  $f'(x) = 0$  è necessaria e sufficiente perché la funzione sia minima.

**Esercizio:** dimostra che  $f'(x)$  è convessa.

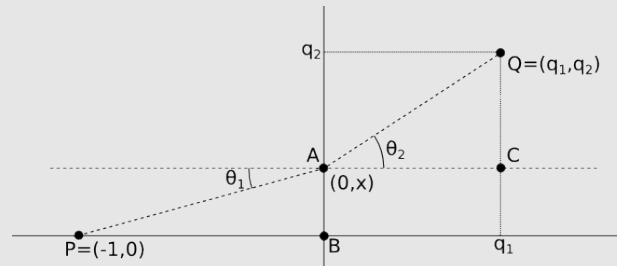
Riscriviamo l'equazione come:

$$\frac{x}{\sqrt{1+x^2}} \cdot \frac{\sqrt{q_1^2 + (q_2-x)^2}}{q_2-x} = \frac{v_1}{v_2} \quad (6.1)$$

Facciamo intanto un'osservazione. Dato che le velocità sono positive, è

$$\frac{v_1}{v_2} > 0 \implies \frac{x}{q_2-x} \geq 0 \implies 0 \leq x \leq q_2$$

Abbiamo ristretto lo spazio in cui cercare la soluzione ottimale.



Diamo dei nomi ai punti come nella figura sopra e riscriviamo la 6.1 rinominando i membri delle frazioni come:

$$\frac{\overline{AB}}{\overline{PA}} \cdot \frac{\overline{AQ}}{\overline{QC}} = \frac{v_1}{v_2}$$

Guardando la figura, si noti come

$$\frac{\overline{AB}}{\overline{PA}} = \sin \theta_1, \quad \frac{\overline{AQ}}{\overline{QC}} = \frac{1}{\sin \theta_2}$$

Abbiamo ottenuto che la durata di percorrenza  $f(x)$  si minimizza quando

$$\frac{\sin \theta_1}{\sin \theta_2} = \frac{v_1}{v_2}$$

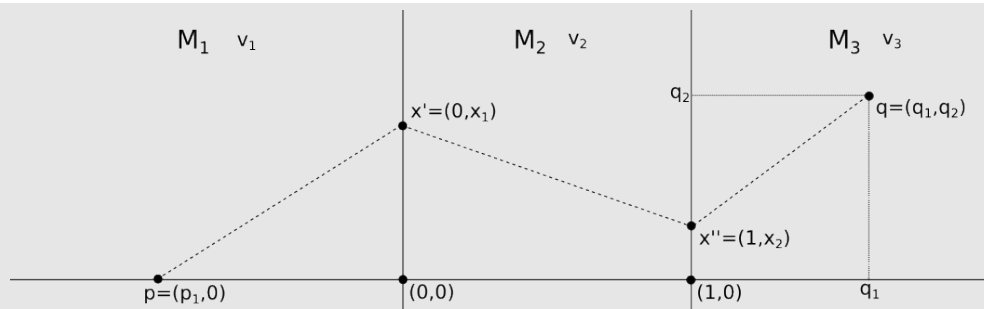
(Tra l'altro, questa è la legge della rifrazione di Snell–Cartesio.)

È un'equazione non lineare ad una variabile che si può risolvere con i classici metodi del calcolo numerico, ad esempio applicando il Metodo delle Tangenti nell'intervallo  $[0, q_2]$ .

## Esempio 6.2

Nell'esempio 6.1 la funzione che abbiamo ottenuto per la ricerca del punto stazionario era ad una variabile. Complichiamo il problema in modo da ottenere una funzione a due variabili.





Il problema è simile ma invece di due, i mezzi sono tre:  $M_1, M_2, M_3$ , divisi da due rette parallele. Per semplificare i conti, posizioniamo gli assi in modo che le rette di separazione dei piani siano verticali e il punto  $p$  di partenza sia sull'asse delle ascisse, e scegliamo un'unità di misura tale che il semipiano  $M_2$  sia largo 1.

Stavolta per individuare un tragitto sono necessari i due parametri  $x_1$  e  $x_2$ , ovvero il valore dell'ordinata dei punti  $x'$  e  $x''$  in cui il tragitto passa da un mezzo all'altro.

La funzione che calcola la velocità di percorrenza è

$$f(x_1, x_2) = \frac{\sqrt{p_1^2 + x_1^2}}{v_1} + \frac{\sqrt{1 + (x_2 - x_1)^2}}{v_2} + \frac{\sqrt{(q_1 - 1)^2 + (q_2 - x_2)^2}}{v_3}$$

Abbiamo un problema di minimizzazione, anche questo non vincolato, in 2 variabili:

$$\min\{f(x) : x \in \mathbb{R}^2\}$$

$f(x)$  è convessa, quindi  $\nabla f(x) = 0$  è condizione necessaria e sufficiente per la minimalità di  $f(x)$ .

**Esercizio:** dimostrare che  $f(x)$  è convessa.

Cerchiamo dove

$$\nabla f(x) = 0$$

Stavolta  $\nabla f(x)$  è un vettore a due dimensioni. Vediamone le componenti, che si devono annullare:

$$\frac{\partial f}{\partial x_1}(x_1, x_2) = \frac{x_1}{v_1 \sqrt{p_1^2 + x_1^2}} - \frac{x_2 - x_1}{v_2 \sqrt{1 + (x_2 - x_1)^2}} = 0$$

$$\frac{\partial f}{\partial x_2}(x_1, x_2) = \frac{x_2 - x_1}{v_2 \sqrt{1 + (x_2 - x_1)^2}} - \frac{q_2 - x_2}{v_3 \sqrt{(q_1 - 1)^2 + (q_2 - x_2)^2}} = 0$$

Questo è un sistema non lineare di due equazioni in due incognite. Condizione necessaria e sufficiente perché la soluzione  $x = (x_1, x_2)$  sia minima è che il sistema sia soddisfatto.

Questo particolare sistema si può risolvere con i metodi classici del calcolo numerico, ad esempio il metodo delle tangenti (detto anche Newton–Raphson), cercando sia  $x_1$  che  $x_2$  nell'intervallo  $[0, q_2]$ .

## 6.1 Metodi risolutivi per i problemi di ottimizzazione non vincolata

Abbiamo visto negli esempi dei problemi che si riducono alla risoluzione di un sistema non vincolato di equazioni non lineari, e abbiamo accennato al fatto che questi sistemi, in alcuni casi, possono essere risolti con i metodi del calcolo numerico (es. Metodo delle tangenti, delle secanti o di bisezione). Va notato però che questi metodi hanno bisogno come dato di input dell'intervallo entro cui cercare la soluzione, e della garanzia che questo intervallo contenga una sola soluzione. Sebbene questo sia possibile nei due esempi presi in considerazione, non è possibile in generale.

Guardiamo quindi a dei metodi iterativi in grado di risolvere un generico problema di ottimizzazione non

vincolata. Il problema si pone come

$$(P) \quad \min\{f(x) : x \in \mathbb{R}^n\} \quad f : \mathbb{R}^n \rightarrow \mathbb{R} \text{ differenziabile}$$

I metodi che vedremo sono tutti di natura iterativa, genereranno dunque una successione di valori  $x^0, x^1, \dots, x^k$ . Quello che cerchiamo, come sempre, sono i punti stazionari  $x^*$ , ovvero tali che

$$\nabla f(x^*) = 0$$

Se la funzione è convessa, come sappiamo  $\nabla f(x^*) = 0$  è condizione sufficiente, oltre che necessaria, perché  $x$  sia punto di minimo, dunque nelle funzioni convesse i punti stazionari sono anche punti di minimo. Altrimenti, se  $f$  non è convessa,  $\nabla f(x^*) = 0$  è comunque condizione necessaria per essere punto di minimo, quindi siamo comunque interessati a cercare i punti in cui il gradiente si annulla perché candidati ad essere minimi.

L'unica ipotesi di cui abbiamo bisogno per questo tipo di metodi è che  $f$  sia differenziabile (ma in alcuni metodi porremo il vincolo che  $f$  sia differenziabile con continuità, cioè che il suo gradiente  $\nabla f$  sia una funzione continua, oppure che sia differenziabile due volte).

In questi metodi, trovare  $\bar{k}$  t.c.  $\nabla f(x^{\bar{k}}) = 0$  può essere un processo infinito. Non potendo garantire la finitezza, ciò che chiediamo ai nostri metodi è la convergenza, che può essere definita in diversi modi:

1. Il limite esiste ed è un punto stazionario:

$$\exists \lim_{k \rightarrow \infty} x^k = x^* \quad \text{e} \quad \nabla f(x^*) = 0$$

2. Definizione più debole: se la successione converge, allora converge al punto stazionario (ma non è detto che converga!). Quindi tutti i punti di accumulazione di  $\{x^k\}$  sono stazionari (ma non è detto che ne esistano):

$$\lim_{k \rightarrow \infty} x^k = x^* \quad \implies \quad \nabla f(x^*) = 0$$

3. Definizione ancora più debole: almeno un punto di accumulazione della successione  $\{x^k\}$  è stazionario.

Tra i metodi iterativi, daremo particolare attenzione ai Metodi di discesa:

### Definizione 6.3 (Metodo di discesa)

*Quei metodi iterativi tali che ad ogni iterazione, il valore della funzione obiettivo decresce:*

$$f(x^0) > f(x^1) > f(x^2) > \dots > f(x^k) > \dots$$

In particolare, si dice Metodo di discesa *non monotona*<sup>1</sup> se:

### Definizione 6.4 (Metodo di discesa non monotona)

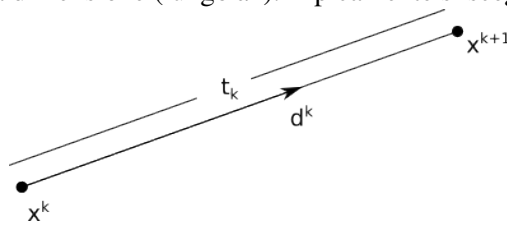
*Dopo  $n$  iterazioni ( $n$  fissato), il valore della funzione obiettivo decresce:*

$$\exists n \in \mathbb{N} \quad \text{t.c.} \quad \forall k \quad f(x^k) > f(x^{k+n})$$

<sup>1</sup>Utilizzeremo raramente questa definizione.

### 6.1.1 Ricerca monodimensionale

Nei metodi di ricerca monodimensionale (line search), se nella  $k$ -esima iterazione abbiamo ottenuto il punto  $x^k$ , otteniamo il punto  $x^{k+1}$  muovendoci dal punto  $x^k$  lungo una direzione  $d^k$  di un passo lungo  $t_k$  (direzione e ampiezza del passo sono scelti ad ogni iterazione). Si chiama *monodimensionale* proprio perché l'ampiezza del passo  $t_k$  è scelta su un'unica dimensione (lungo  $d^k$ ). Tipicamente si sceglie  $d^k$  t.c.  $\|d^k\| = 1$ .



In formule:

$$x^{k+1} = x^k + t_k d^k \quad t_k \in \mathbb{R} \quad \begin{array}{l} t_k \in \mathbb{R} \text{ passo di spostamento} \\ d^k \in \mathbb{R}^n \text{ direzione} \end{array}$$



#### Nota

Anche l'algoritmo del simplesso per la programmazione lineare vincolata è un metodo di line search, visto che da un vertice ci si muove lungo una direzione (cioè lungo il bordo dei vincoli) per giungere ad un altro vertice.

I metodi di ricerca monodimensionale si differenziano tra loro per il modo in cui, ad ogni passo, viene scelta la direzione  $d$  e l'ampiezza  $t$  del passo successivo.

Vediamo innanzi tutto come ottenere un metodo di discesa. Supponiamo, ad un generico passo, di aver ottenuto  $\bar{x} \in \mathbb{R}^n$  (non stazionario, cioè tale che  $\nabla f(\bar{x}) \neq 0$ , altrimenti abbiamo già la soluzione).

Dato che  $f$  è differenziabile, abbiamo lo sviluppo di Taylor:

$$f(\bar{x} + t \cdot d) - f(\bar{x}) = t \cdot \nabla f(\bar{x})^T \cdot d + r(t \cdot d)$$

Dove  $r(td)$  è un resto tale che  $\frac{r(td)}{t}$  va a 0. Inoltre sappiamo che

$$\lim_{t \rightarrow 0} \frac{f(\bar{x} + td) - f(\bar{x})}{t} = \nabla f(\bar{x})^T \cdot d$$

Da questa equazione si ricava che, se scegliamo la direzione  $d$  in modo che sia  $\nabla f(\bar{x})^T \cdot d < 0$ , in un intorno di  $t = 0$  sufficientemente piccolo l'argomento del limite deve essere negativo:

$$\exists \varepsilon \text{ t.c. } \forall t \in [0, \varepsilon] \quad \frac{f(\bar{x} + td) - f(\bar{x})}{t} < 0$$

ma dato che  $t > 0$ , deve necessariamente essere  $f(\bar{x} + td) - f(\bar{x}) < 0$  e dunque  $f(\bar{x} + td) < f(\bar{x})$ , che è proprio ciò che cercavamo.

Ricapitolando:

#### Proprietà 6.1

se si sceglie

$d$  t.c.  $\nabla f(\bar{x})^T \cdot d < 0$  e  $t \in (0, \varepsilon)$  con  $\varepsilon$  sufficientemente piccolo  $\implies$  il passo  $x^k + t_k d^k$  è di discesa

### 6.1.2 La massima decrescita

Naturalmente, il nostro interesse è che il metodo non solo sia di discesa, ma che discenda più velocemente possibile. Cerchiamo quindi di prendere  $d$  in modo che il prodotto non solo sia negativo, ma il più negativo possibile. Per andare incontro a questa esigenza, si forma un altro problema:

$$(Pa) \quad \min\{\nabla f(\bar{x})^T d : \|d\|_2 = 1\} \quad (6.2)$$



#### Nota

La norma di  $d$  deve essere limitata, ad es. con  $\|d\|_2 = 1$ , altrimenti supponiamo di trovare una direzione

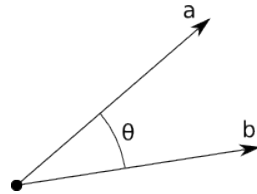
$$\bar{d} = dt \quad \text{t.c.} \quad \nabla f(\bar{x})^T \bar{d} < 0 \quad \implies \quad \nabla f(\bar{x})^T \bar{d} = \nabla f(\bar{x})^T (td) = t(\nabla f(\bar{x})^T d)$$

Ma se mandiamo  $t$  a  $+\infty$ :

$$t(\nabla f(\bar{x})^T d) \longrightarrow_{t \rightarrow +\infty} -\infty$$

E il problema risulta inferiormente illimitato.

Il problema di ottimizzazione 6.2 lo sappiamo risolvere per via teorica. Si ricorda che il prodotto scalare tra due vettori  $a$  e  $b$  si calcola come



$$a^T b = \|a\|_2 \|b\|_2 \cos \theta$$

Quindi tenendo conto che  $\|d\|_2 = 1$ , il prodotto scalare  $\nabla f(\bar{x})^T d$  si calcola come

$$\nabla f(\bar{x})^T d = \|\nabla f(\bar{x})\|_2 \cos \theta$$

In cui  $\bar{x}$  e  $f$  sono fissati, quindi  $\nabla f(\bar{x})$  è un numero fissato nel nostro problema.

Quindi il problema si può riscrivere come:

$$\min\{\nabla f(\bar{x})^T d : \|d\|_2 = 1\} = \|\nabla f(\bar{x})\|_2 \cdot \min\{\cos \theta : \|d\|_2 = 1\}$$

Notare che  $d$  sembra scomparire nella formula, in realtà  $\theta$  dipende da  $d$  (e da  $\nabla f(\bar{x})$  che è fisso per il problema). Il minimo di questo problema si ottiene quando

$$\cos \theta = -1 \quad \iff \quad \theta = \pi$$

Ma perché questo angolo sia  $\pi$ , è necessario che si prenda  $d$  nella direzione opposta rispetto al gradiente. Inoltre vogliamo  $d$  di norma unitaria, e per questo basta dividerlo per la norma del gradiente.

$$\iff \quad d = \frac{-\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|_2}$$

In altre parole, sto chiedendo che  $d$  sia il versore della direzione opposta al gradiente.

Ancora una volta, il gradiente è la direzione di massima crescita, quindi la massima decrescita è la direzione opposta al gradiente.

## 6.2 Metodi del gradiente

Da quanto visto sopra, si deduce facilmente un primo metodo di ricerca monodimensionale in cui la direzione  $d$  scelta ad ogni passo è l'opposto del gradiente:

$$x^{k+1} = x^k + t_k \cdot d^k \quad \text{con } d^k = -\nabla f(x^k)$$

Ovvero

$$x^{k+1} = x^k - t_k \cdot \nabla f(x^k)$$

(Avremmo potuto prendere  $d^k = -\frac{\nabla f(x^k)}{\|\nabla f(x^k)\|_2}$ , ma lasciamo perdere la divisione per la norma del gradiente perché la consideriamo “inglobata” nel passo di spostamento.)

Per quanto riguarda la scelta del passo di spostamento  $t_k$ , la scelta ideale è ovviamente di prenderlo esatto:

### Definizione 6.5 (Metodo della Ricerca Esatta)

Metodo nel quale la lunghezza del passo si calcola con un nuovo problema di ottimizzazione:

$$t_k \in \arg \min \{f(x^k - t_k \nabla f(x^k)) : t_k \geq 0\}$$

Ovvero il passo è cercato lungo la semiretta in modo che minimizzi la funzione obiettivo.

### Definizione 6.6 (Metodo del Gradiente Esatto)

Metodo del gradiente (ovvero tale che  $d^k = -\nabla f(x^k)$ ) in cui si sceglie il passo col metodo della ricerca esatta.

Sotto forma di algoritmo si può enunciare come segue:

1. Poniamo  $k = 0$ ; Si sceglie un punto di partenza  $x_0$ ;
2. Se  $\nabla f(x^k) = 0$ , allora STOP: abbiamo trovato un punto stazionario.
3. Si calcola la direzione come l'opposto del gradiente:  $d^k = -\nabla f(x^k)$ ;
4. Si calcola il passo con la ricerca esatta:

$$t_k \in \operatorname{argmin}\{f(x^k + t_k \cdot d^k) : t_k \geq 0\}$$

ovvero

$$t_k \in \operatorname{argmin}\{f(x^k - t_k \cdot \nabla f(x^k)) : t_k \geq 0\}$$

5. Ci si sposta al punto successivo:  $x^{k+1} = x^k + t_k \cdot d^k = x^k - t_k \cdot \nabla f(x^k)$
6.  $k := k + 1$ , ritornare al passo 2.

Come vedremo in seguito (6.2.2), il passo 4 impone la soluzione di un problema di ottimizzazione esatta che, in generale, è tutt'altro che banale, per questo vedremo altri metodi del gradiente che non utilizzano la ricerca esatta ma scelgono il passo in maniera più approssimata e meno costosa. Tuttavia per alcune funzioni obiettivo particolari, ad es. le funzioni quadratiche, il calcolo della ricerca esatta è possibile e facilmente risolvibile, perché la funzione da minimizzare  $f(x^k - t_k \nabla f(x^k))$  è una parabola con il vertice verso l'alto, della quale è facile calcolare esattamente il minimo.

### 6.2.1 Proprietà del Metodo del Gradiente Esatto

Abbiamo intenzione di dimostrare tre proprietà del Metodo del Gradiente Esatto:

1. Il metodo del gradiente esatto è un metodo di discesa;
2. Due direzioni successive del metodo del gradiente esatto sono tra loro ortogonali;
3. Il metodo converge.

Per queste dimostrazioni, si introduce la seguente funzione:

$$\varphi(t) = f(x^k - t\nabla f(x^k))$$

Questa è la funzione  $f$  calcolata lungo la semiretta nella quale viene cercato il passo di spostamento, vista come funzione del passo  $t$ . È la funzione di ricerca: durante la ricerca esatta si cerca il minimo di  $\varphi$ . Si noti che è la composizione di due funzioni:

$$t \xrightarrow{\theta} x^k - t\nabla f(x^k) \xrightarrow{f} f(x^k - t\nabla f(x^k))$$

La funzione  $\theta : \mathbb{R} \rightarrow \mathbb{R}^n$  associa il passo  $t$  al punto trovato, mentre  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  associa il punto trovato al valore della funzione obiettivo.  $\varphi$  si scrive come  $\varphi = f \circ \theta : \mathbb{R} \rightarrow \mathbb{R}$ .

$\theta$  si può vedere come un insieme  $(\theta_1, \dots, \theta_i, \dots, \theta_n)$  di  $n$  funzioni, una per ogni componente, tali che  $\theta_i : \mathbb{R} \rightarrow \mathbb{R}$  assume il valore dell' $i$ -esima componente del vettore  $x^k - t\nabla f(x^k) = \theta(t)$ . Quindi

$$\theta = (\theta_1, (\theta_2, \dots, \theta_n)) \quad \theta(t) = x^k - t\nabla f(x^k) = \begin{pmatrix} \theta_1(t) \\ \theta_2(t) \\ \vdots \\ \theta_n(t) \end{pmatrix}$$

La funzione  $\theta$  è derivabile, quindi ogni  $\theta_i$  è derivabile.

$$\theta' = (\theta'_1, \theta'_2, \dots, \theta'_n)$$

Si noti quindi che

$$\begin{array}{l} f \text{ è differenziabile} \\ \theta_i \text{ sono derivabili} \end{array} \implies \varphi = f \circ \theta \text{ è derivabile}$$

e la sua derivata è <sup>2</sup>:

$$\varphi'(t) = \nabla f(x^k - t\nabla f(x^k))^T \cdot \theta'(t)$$

Ma dato che  $\theta'(t) = -\nabla f(x^k)$ ,

$$\varphi'(t) = -\nabla f(x^k - t\nabla f(x^k))^T \cdot \nabla f(x^k) \quad (6.3)$$

Questo strumento ci permetterà di fare facilmente tutte e tre le dimostrazioni.

Sappiamo inoltre che:

1. Nel punto di minimo, che è  $t_k$ , la derivata si annulla:

$$\varphi'(t_k) = 0$$

<sup>2</sup> Usando il seguente teorema



#### Teorema 6.7

Sia  $\Theta : \mathbb{R} \rightarrow \mathbb{R}^n$  derivabile in  $\bar{t} \in \mathbb{R}$ ,  $g : \mathbb{R}^n$  differenziabile in  $\Theta(\bar{t}) \in \mathbb{R}^n$ .

Allora  $g \circ \Theta$  è differenziabile in  $\bar{t}$  e  $(g \circ \Theta)'(\bar{t}) = \nabla g(\Theta(\bar{t}))^T \Theta'(\bar{t})$ . (dove  $\Theta' = (\Theta'_1, \dots, \Theta'_n)$ ,  $\Theta = (\Theta_1, \dots, \Theta_n)$ )

2. Il valore che questa derivata assume nel punto 0 (si calcola esplicitamente):

$$\varphi'(0) = -\|\nabla f(x^k)\|_2^2 < 0$$

È < 0 perché è una norma diversa da 0 (se il punto non è stazionario) con un meno davanti.

Possiamo sfruttare il risultato in 6.3 e i due sopra per dimostrare le tre proprietà enunciate sopra.

### Teorema 6.8

Il metodo del gradiente esatto è di discesa, ovvero  $\forall k \quad f(x^{k+1}) < f(x^k)$

Dimostrazione.

$$f(x^k) = \varphi(0) \geq \underbrace{\varphi(t_k)}_{\text{punto di minimo}} = f(x^{k+1})$$

Abbiamo dimostrato  $f(x^{k+1}) \leq f(x^k)$ , ma è possibile che siano uguali?

Guardiamo cosa succede nell'intorno di 0. Se  $\varphi'(0) < 0$ , la funzione  $\varphi$  è decrescente almeno nell'intorno, ne segue che 0 non può essere un punto di minimo, quindi

$$f(x^k) = \varphi(0) \neq \varphi(t_k) = f(x^{k+1})$$

Da cui

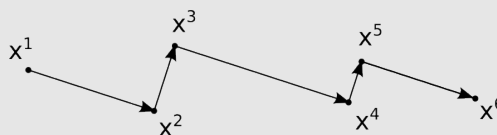
$$f(x^k) = \varphi(0) > \varphi(t_k) = f(x^{k+1})$$

□

### Teorema 6.9

Due direzioni successive del metodo del gradiente esatto sono tra loro ortogonali, ovvero

$$\nabla f(x^{k+1})^T \cdot \nabla f(x^k) = 0$$



Dimostrazione. Sappiamo che

$$\varphi'(t_k) = 0 \quad x^{k+1} = x^k - t_k \nabla f(x^k)$$

Ma sappiamo anche da 6.3 che la derivata vale

$$\varphi'(t_k) = -\nabla f(x^{k+1})^T \nabla f(x^k) = 0$$

□

### Teorema 6.10

Il metodo converge (se esiste il limite, allora il limite è un punto stazionario):

$$\lim_{k \rightarrow \infty} f(x^k) = x^* \in \mathbb{R}^n \quad \implies \quad \nabla f(x^*) = 0$$

Dimostrazione. Dal 6.9 sappiamo che

$$\nabla f(x^{k+1})^T \nabla f(x^k) = 0$$

Inoltre

$$\text{per } k \rightarrow \infty, \quad x^k \rightarrow x^* \text{ e } x^{k+1} \rightarrow x^*$$

Per ipotesi, la funzione è differenziabile con continuità (cioè la funzione gradiente è continua), quindi per la continuità delle funzioni

$$\nabla f(x^{k+1})^T \nabla f(x^k) \xrightarrow{k \rightarrow +\infty} \nabla f(x^*)^T \nabla f(x^*) = \|\nabla f(x^*)\|_2^2$$

Quindi

$$\|\nabla f(x^*)\|_2^2 = 0 \implies \nabla f(x^*) = 0$$

$\nabla f(x^*)$  è stazionario.

## 6.2.2 Critica al metodo del gradiente esatto

### Critica al Passo 4 (e caso particolare delle funzioni quadratiche)

Come accennato, il passo 4 del Metodo del Gradiente Esatto (vedi la sua definizione 6.6), nel quale viene calcolata l'ampiezza del passo di spostamento, è un problema più semplice rispetto a cercare il minimo di una funzione, ma in generale è comunque costoso. Il calcolo del minimo di  $\varphi(t)$  è un problema ad una sola variabile (invece di  $n$ ), ovvero è un problema di ottimizzazione monodimensionale, ma richiede comunque una procedura algoritmica.

Vi sono tuttavia dei casi in cui il calcolo esatto del passo di spostamento non richiede un algoritmo: nel caso in cui  $f$  sia una funzione quadratica, come vedremo non troviamo più davanti ad un problema di ottimizzazione ma semplicemente al calcolo di una formula.

Si ricorda la definizione di funzione quadratica:

$$f(x) = \frac{1}{2}x^T Qx + b^T x + c \quad Q = Q^T \in \mathbb{R}^{n \times n}, b \in \mathbb{R}^n, c \in \mathbb{R}$$

Dove  $Q$  è una matrice quadrata e che, aggiustando i coefficienti, possiamo rendere anche simmetrica. Avevamo visto che il gradiente di questa funzione è

$$\nabla f(x) = Qx + b$$

Per cui trovare i punti stazionari equivale a trovare soluzioni del sistema lineare

$$\nabla f(x) = Qx + b = 0$$

#### Proprietà 6.2

Il problema ha senso se  $Q$  è semi-definita positiva. Vediamo cosa succederebbe in caso contrario: se non fosse semi-definita positiva, allora vi sarebbe un  $\bar{x} \in \mathbb{R}^n$  tale che  $\bar{x}^T Q \bar{x} < 0$ . Inoltre  $f$  in  $t\bar{x}$  varrebbe

$$f(t\bar{x}) = \frac{1}{2}t^2(\bar{x}^T Q \bar{x}) + tb^T \bar{x} + c$$

Questa, come funzione di  $t$ , è una parabola con il coefficiente del termine quadrato ( $\bar{x}^T Q \bar{x}$ ) negativo. Dunque il suo limite per  $t \rightarrow \infty$  è

$$\lim_{t \rightarrow +\infty} f(t\bar{x}) = -\infty$$

Abbiamo trovato che se esiste un punto  $\bar{x}$  che viola il fatto che  $Q$  sia semi-definita positiva, si ottiene che in quella direzione la funzione obiettivo va a  $-\infty$ , il problema risulta dunque inferiormente illimitato. Non ha dunque senso la minimizzazione vincolata di una funzione quadrata che non sia semi-definita positiva.



Si ricorda che  $\varphi$  è definita come <sup>3</sup>

$$\varphi(t) = f(x - t\nabla f(x))$$

Dato che  $Q$  è semi-definita positiva, possiamo calcolare la derivata di  $\varphi(t)$

$$\begin{aligned}\varphi'(t) &= -\nabla f(x - t\nabla f(x))^T \nabla f(x) \\ &= -[Q(x - t\nabla f(x)) + b]^T \cdot \nabla f(x) \\ &= -[Qx - tQ\nabla f(x) + b]^T \cdot \nabla f(x) \\ &= -[\nabla f(x) - tQ\nabla f(x)]^T \cdot \nabla f(x) \\ &= -\nabla f(x)^T \cdot \nabla f(x) + t \underbrace{\nabla f(x)^T Q \nabla f(x)}_{\geq 0}\end{aligned}$$

Cerchiamo di vedere dove si annulla la derivata  $\varphi'$ . Abbiamo che  $\nabla f(x)^T Q \nabla f(x) \geq 0$  perché  $Q$  è semi-definita positiva. Distinguiamo due casi:

1. Se  $\nabla f(x)^T Q \nabla f(x) > 0$  (questo è sempre vero se  $Q$  è definita positiva) allora il gradiente si annulla nel punto

$$t = \frac{\nabla f(x)^T \nabla f(x)}{\nabla f(x)^T Q \nabla f(x)} \iff \varphi'(x) = 0$$

2. Altrimenti, se  $\nabla f(x)^T Q \nabla f(x) = 0$  (possibile solo se  $Q$  è semi-definita positiva) allora

$$\varphi'(t) = -\nabla f(x)^T \nabla f(x) = \underbrace{-\|\nabla f(x)\|_2^2}_{< 0}$$

che è costante e  $\neq 0$ . Ma se  $\varphi$  ha derivata costante e negativa, significa che è una funzione lineare (una retta) nella forma

$$\varphi(t) = f(x - t\nabla f(x)) = -\|\nabla f(x)\|_2^2 t + \text{costante}$$

E il suo limite

$$\lim_{t \rightarrow +\infty} \varphi(t) = -\infty$$

Quindi, se con una certa matrice  $Q$  si incontra la condizione  $\nabla f(x)^T Q \nabla f(x) = 0$  (può succedere solo se  $Q$  è semidefinita positiva, non se è definita positiva) allora il problema è inferiormente illimitato.

Ricapitolando, il caso che la funzione obiettivo  $f$  sia quadratica non pone problemi con la ricerca esatta:

- Se  $Q$  non è semidefinita positiva, il problema è inferiormente illimitato.
- Altrimenti abbiamo la formula esplicita per calcolare il passo di spostamento:  $t = \frac{\nabla f(x)^T \nabla f(x)}{\nabla f(x)^T Q \nabla f(x)}$
- Se si pone la condizione  $\nabla f(x)^T Q \nabla f(x) = 0$  (impossibile se  $Q$  è definita positiva) allora il problema è inferiormente illimitato.
- Se  $f$  non è quadratica, per trovare il minimo di  $\varphi$ , ovvero la lunghezza del passo di spostamento, ha bisogno di metodi iterativi (es. tangenti, secanti, newton) e dunque inesatti.

### Funzione ripida

Può succedere che nei pressi di un punto stazionario, le curve di livello siano estremamente ripide. È questo il caso della Funzione di Rosenbrock, definita come:

$$f(x, y) = (1 - x)^2 + 100(y - x^2)^2 \quad (6.4)$$

Si vede ad occhio che in  $(1, 1)$  la funzione vale  $f(1, 1) = 0$  ed essendo  $f$  la somma di due quadrati, non può esservi un punto minore. Tuttavia questa funzione ha la capacità di rompere tutti i metodi iterativi per la ricerca del minimo locale, perché la discesa verso questo minimo è molto ripida. Il metodo del gradiente esatto, una volta entrato nel “canale” (la parte blu nella Figura 6.1), essendo un metodo di decrescita ed avendo le direzioni tra passi successivi ortogonali tra loro è costretto a fare passi molto brevi per non risalire sull'altra sponda del canale, e quindi è estremamente lento.

<sup>3</sup>Rispetto alla definizione precedente abbiamo rinominato  $x^k$  in  $x$  per semplicità.

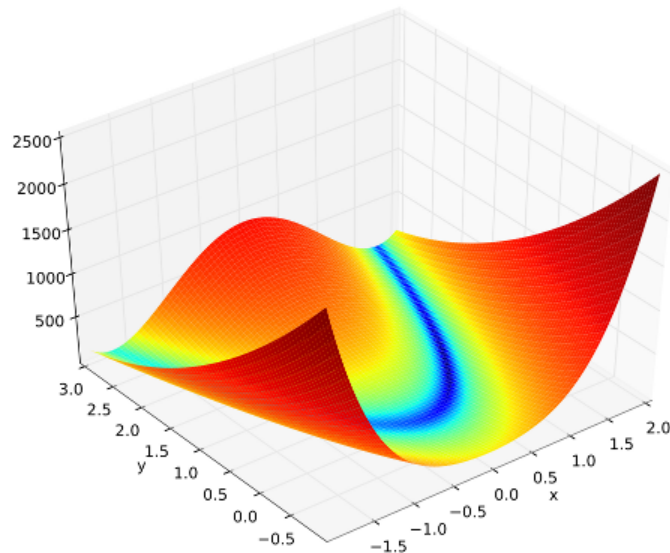


Figura 6.1: Funzione banana di Rosenbrock in tre dimensioni. Immagine presa da Wikipedia, [http://en.wikipedia.org/wiki/Rosenbrock\\_function](http://en.wikipedia.org/wiki/Rosenbrock_function).

### Critica al passo 3 (Direzioni ortogonali)

Abbiamo visto nel Teorema 6.9 che la direzione del passo all'iterazione  $k$  è ortogonale a quella del passo  $k + 1$ . Gli spostamenti sono quindi rigidi. Se siamo lontani dal punto stazionario il fatto di avere direzioni ortogonali non crea problemi, ma quando ci si avvicina può portare a dei passi molto piccoli e quindi un rallentamento della convergenza.

Si ricorda che la direzione di spostamento, scelta al passo 3 dell'algoritmo in 6.6, è ortogonale rispetto a quella del passo precedente perché è scelta come il gradiente. Vediamo dei modi alternativi per scegliere la direzione, sempre in modo che il metodo sia di decrescita.

### Direzioni alternative di decrescita

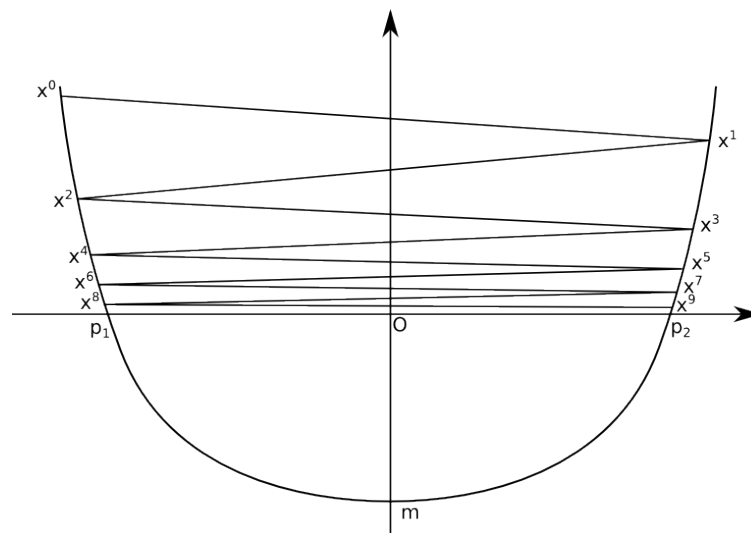
Possiamo scegliere  $d^k = -D_k \nabla f(x^k)$  con  $D_k \in \mathbb{R}^{n \times n}$  definita positiva, ovvero data la direzione del gradiente, la alteriamo moltiplicandola per una matrice scelta ad ogni passo.

Si noti che questa è la generalizzazione della scelta del gradiente usata finora, infatti per ottenere  $d^k = -\nabla f(x^k)$ , basta scegliere  $D_k = I$ .

Se  $D_k$  è definita positiva, secondo la proprietà 6.1, almeno vicino al punto  $x^k$ , la direzione  $d^k$  è di discesa perché

$$\nabla f(x^k)^T d^k = -\nabla f(x^k)^T D_k \nabla f(x^k) < 0$$

Tuttavia, sebbene finora ci siamo accontentati di decrescere, questa proprietà in generale non basta. Ad esempio, un metodo di decrescita applicato su una parabola potrebbe generare la sequenza di punti in figura:



Come si vede, il valore della funzione obiettivo decresce ad ogni iterazione, ma le due sottosuccessioni convergono ai due punti  $p_1$  e  $p_2$ , che non hanno niente a che fare con il minimo  $m$ , molto distante. Quindi la ricerca esatta può essere una richiesta eccessiva, ma comunque dobbiamo porre dei paletti più rigidi della semplice decrescita. Esistono diverse proprietà, di diversa rigidità, che i metodi possono offrire:

### Possibili scelte per il passo

- Ricerca esatta:

$$t_k \in \operatorname{argmin}\{f(x^k + td^k) : t \geq 0\}$$

Ovvero lungo la direzione  $d^k$  (che in generale non è  $\nabla f(x^k)$ ) si prende il minimo. Abbiamo visto che in generale individuare il minimo in maniera esatta è difficile o impossibile.

- Minimizzazione limitata:

$$t_k \in \operatorname{argmin}\{f(x^k + td^k) : t \in [0, T]\} \quad T \geq 0$$

$T$  è un valore fissato scelto dall'utente. Serve per limitare la ricerca a passi non troppo grandi, che potrebbero mandare troppo fuori strada.

- Costante:

$$t_k = \bar{t}, \quad \bar{t} > 0$$

Il passo è sempre ampio  $\bar{t}$  definito dall'utente. Il rischio è che se  $\bar{t}$  è troppo grande, quando si è vicini al punto stazionario si cominciano a girarci intorno, se invece è troppo piccolo l'avvicinamento nelle iterazioni iniziali sia lento.

- Passi decrescenti:

$$t_k \downarrow 0 \quad \text{con} \quad \sum_{k=0}^{\infty} t_k = +\infty$$

Ovvero i passi sono decrescenti ma la serie dei passi deve divergere, ad esempio si può scegliere  $t_k = \frac{1}{k}$  ma non  $t_k = \frac{1}{k^2}$  che converge a 0.

- Ricerca inesatta: è come la ricerca esatta, ma invece del minimo cerchiamo una sua approssimazione, ponendo delle condizioni sulla decrescita, in modo che il processo sia meno oneroso.

### 6.2.3 Metodi del gradiente con ricerca inesatta

Vediamo quali sono i paletti che dobbiamo porre per la ricerca inesatta, e come potrebbe funzionare un metodo basato sulla ricerca inesatta. La situazione che abbiamo davanti è la stessa del Metodo del gradiente esatto (6.6): da un punto  $x^k$  abbiamo individuato (passo 3) una direzione di discesa  $d^k$ . Vogliamo però sostituire il passo 4, ovvero la scelta del passo  $t_k$ .

Abbiamo detto che non si cerca il minimo, però vogliamo che la decrescita sia consistente. Cerchiamo quindi un  $t_k$  in modo che sia rispettata la

#### Definizione 6.11 (Condizione di Armijo)

$$(AJO) \quad f(x^k + td^k) \leq f(x^k) + c_1 t \nabla f(x^k)^T d^k$$

con  $c_1 \in (0, 1)$  un parametro fissato.

La funzione nel nuovo punto non solo è minore della funzione nel vecchio punto, ma è anche più piccola di almeno  $c_1 t \nabla f(x^k)^T d^k$ , dove  $c_1$  è il grado di libertà lasciato all'algoritmo.

#### Proprietà 6.3

Supponiamo la funzione da minimizzare sia inferiormente limitata (altrimenti il metodo non serve) e rispetti la condizione di Armijo, allora questo impedisce che  $t_k \rightarrow +\infty$ .

*Dimostrazione.* Se andasse  $t_k \rightarrow +\infty$ , allora

$$\lim_{k \rightarrow \infty} (f(x^k) + c_1 t_k \underbrace{\nabla f(x^k)^T d^k}_{< 0}) = -\infty$$

ma dato che

$$f(x^k + td^k) \leq f(x^k) + c_1 t \nabla f(x^k)^T d^k \implies f(x^k + td^k) \rightarrow -\infty$$

che è impossibile perché abbiamo assunto che  $f$  sia inferiormente limitata.  $\square$

Quindi la condizione di Armijo impedisce che il passo sia troppo grande. Ma non impedisce che il passo sia troppo piccolo! Vediamo per quale motivo.

#### Proprietà 6.4

Esiste sempre un intervallo, per quanto piccolo, in cui la condizione di Armijo è rispettata.

*Dimostrazione.* Abbiamo visto che

$$\varphi'(0) = \nabla f(x^k)^T d^k < 0$$

Inoltre,

$$\varphi'(0) = \lim_{t \rightarrow 0} \frac{f(x^k + td^k) - f(x^k)}{t}$$

e dalla definizione di limite, si può dire che

$$\exists \bar{t} \text{ t.c. (AJO) vale } \forall t \in [0, \bar{t}]$$

Essendo  $\nabla f(x^k)^T d^k < 0$  e  $c_1 \in (0, 1)$ , risulta  $\nabla f(x^k)^T d^k < c_1 \nabla f(x^k)^T d^k$

Quindi per un  $t$  sufficientemente piccolo, vale

$$\frac{f(x^k + td^k) - f(x^k)}{t} < c_1 \nabla f(x^k)^T d^k$$

Ma riscrivendola:

$$f(x^k + td^k) < f(x^k) + t \cdot c_1 \nabla f(x^k)^T d^k$$

abbiamo ottenuto proprio la condizione di Armijo.  $\square$

Cosa significa questo? Che sebbene la condizione di Armijo impedisca passi troppo grandi, è soddisfatta da un qualunque valore del passo che sia sufficientemente piccolo.

Quindi la condizione non basta. Dobbiamo imporre qualche altra condizione per cui il passo non sia troppo piccolo. Ad esempio, si può richiedere che le derivate calcolate in  $t$  e in 0 non siano troppo simili. Imponiamo una nuova condizione:

**Definizione 6.12 (Condizione sulla Curvatura)**

$$\varphi'(t) \geq c_2 \varphi'(0)$$

che si può scrivere anche:

$$(CUR) \quad \nabla f(x^k + td^k)^T d^k \geq c_2 \nabla f(x^k)^T d^k$$

con  $c_2$  un parametro fissato,  $c_2 \in (c_1, 1)$ .

Le condizioni (AJO) e (CUR) unite formano le

**Definizione 6.13 (Condizioni di Wolfe)**

(AJO) e (CUR) rispettate entrambe:

$$(AJO) \quad f(x^k + td^k) \leq f(x^k) + c_1 t \nabla f(x^k)^T d^k \quad c_1 \in (0, 1)$$

$$(CUR) \quad \nabla f(x^k + td^k)^T d^k \geq c_2 \nabla f(x^k)^T d^k \quad c_2 \in (c_1, 1)$$

D'ora in poi quindi cercheremo un passo  $t_k$  che rispetti le condizioni di Wolfe.

 **Proprietà 6.5**

Sia  $t_{k_{\min}}$  il passo della ricerca esatta (che quindi minimizza  $\varphi$ ), facciamo vedere che la condizione di curvatura è rispettata.

*Dimostrazione.* Vogliamo dimostrare che (CUR) è rispettata, ovvero

$$\varphi'(t_{k_{\min}}) \geq c_2 \varphi'(0)$$

Essendo  $t_{k_{\min}}$  il minimo di  $\varphi$ , la derivata vale

$$\varphi'(t_{k_{\min}}) = 0$$

inoltre la derivata in 0 vale

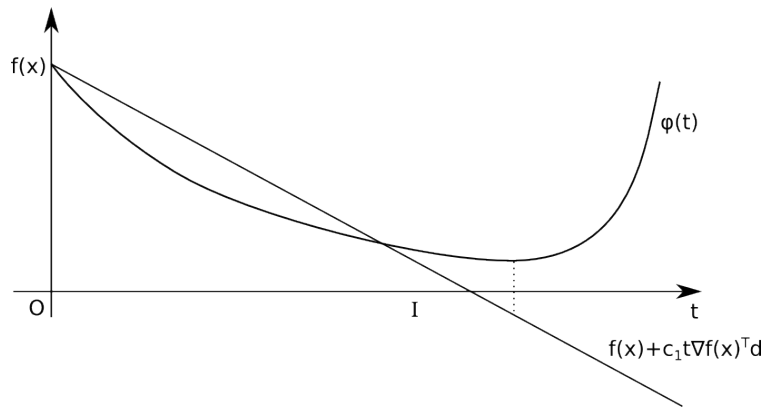
$$\varphi'(0) = \nabla f(x^k)^T d^k < 0$$

quindi (CUR) è soddisfatta.  $\square$

Ma il passo della ricerca esatta rispetta anche la condizione di Armijo (e quindi le condizioni di Wolfe)? In generale, no. La condizione di Armijo (6.11) dice che

$$\underbrace{f(x + td)}_{\varphi(t)} \leq \underbrace{f(x) + c_1 t \nabla f(x)^T d}_{\text{lineare}} \quad \underbrace{\nabla f(x)^T d}_{c_1 \in (0,1)} < 0$$

Vediamo i grafici sovrapposti di una possibile  $\varphi(t)$  e della funzione lineare  $f(x) + c_1 t \nabla f(x)^T d$  al variare di  $t$ . Si noti che nella seconda funzione, il termine  $f(x)$  è una costante, così come  $c_1 \nabla f(x)^T d$ , che è negativo. Nel punto  $t = 0$  entrambe le funzioni valgono  $f(x)$  e la derivata di  $\varphi(t)$  ha un valore minore, quindi per un tratto sarà inferiore dell'altra. I grafici avranno dunque forma:



Perché la condizione di Armijo valga per la ricerca esatta, la curva di  $\varphi(t)$ , nel suo punto di minimo deve stare sotto la curva  $f(x) + c_1 t \nabla f(x)^T d$ . Ma è possibile costruire degli esempi, come quello del grafico, in cui questo non avviene.

Tuttavia, nella pratica  $c_1$  viene scelto talmente piccolo (e dunque la funzione lineare diventa talmente “orizzontale”) che il minimo di  $\varphi(t)$  è sempre sotto.

Adesso poniamoci la domanda: esistono dei metodi che rispettino le condizioni di Wolfe? Si può dimostrare che

**Proposizione 6.1**

Supponiamo  $f$  sia inferiormente limitata. Allora esiste un intervallo  $(a, b)$  per cui le condizioni di Wolfe sono verificate per ogni  $t \in (a, b)$ .

(Gli intervalli potrebbero essere anche più di uno.)

Dimostrazione. Sia

$$\tau = \sup \{ \tau \mid \forall t \in [0, \tau] \text{ è rispettata (AJO), ovvero } f(x^k + \tau d^k) \leq f(x^k) + c_1 \tau \nabla f(x^k)^T d^k \}$$

Vediamo alcune proprietà di questo insieme:

- Può l'insieme essere vuoto? Abbiamo visto in 6.4 che la condizione di Armijo vale almeno in un piccolo intervallo, quindi l'insieme non è vuoto (esiste almeno un  $\tau$ ).
- Può il sup di questo insieme essere  $+\infty$ ? Se così fosse, allora (AJO) varrebbe per ogni  $t$ , ma (AJO) vieta che  $t$  vada a  $+\infty$  se la funzione è inferiormente limitata (si veda 6.3). Quindi il sup deve essere un numero finito.
- Potremmo avere la disuguaglianza stretta?

$$\underbrace{f(x^k + \tau_1 d^k)}_{\text{continua in } \tau_1} < \underbrace{f(x^k) + c_1 \tau_1 \nabla f(x^k)^T d^k}_{\text{lineare in } \tau_1}$$

Se due funzioni continue in un punto non hanno lo stesso valore (caso della disuguaglianza stretta), per il teorema della permanenza del segno le due funzioni continuano ad essere una maggiore o uguale dell'altra in tutto l'intorno del punto considerato. Ma allora questo punto non sarebbe più il sup, perché la condizione continuerebbe a valere in punti più grandi di  $\tau_1$ . Quindi abbiamo necessariamente che

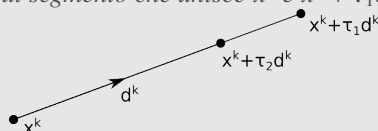
$$\begin{aligned} f(x^k + \tau_1 d^k) &= f(x^k) + c_1 \tau_1 \nabla f(x^k)^T d^k \\ f(x^k + \tau_1 d^k) - f(x^k) &= c_1 \tau_1 \nabla f(x^k)^T d^k \end{aligned} \tag{6.5}$$

Per il teorema del valor medio, abbiamo

$$f(x^k + \tau_1 d^k) - f(x^k) = \nabla f(x^k + \tau_2 d^k)^T \cdot (\tau_1 d^k) = \tau_1 \nabla f(x^k + \tau_2 d^k)^T d^k$$

per un opportuno  $\tau_2 \in (0, \tau_1)$ .

Ovvero,  $x^k + \tau_2 d^k$  è un generico punto sul segmento che unisce  $x^k$  e  $x^k + \tau_1 d^k$ :



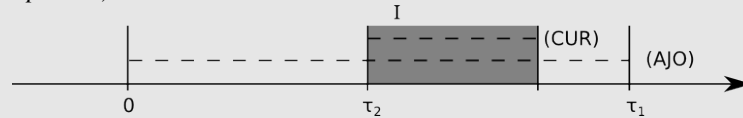
Dato che  $\tau_2 < \tau_1$ , in  $x^k + \tau_2 d^k$  vale al condizione di Armijo. Dividiamo la 6.5 per  $\tau_1$ :

$$\nabla f(x^k + \tau_2 d^k)^T d^k = c_1 \nabla f(x^k)^T d^k$$

ma dato che  $c_2 > c_1$  e  $\nabla f(x^k)^T d^k < 0$ , è

$$f(x^k + \tau_2 d^k)^T d^k = c_1 \nabla f(x^k)^T d^k > c_2 \nabla f(x^k)^T d^k$$

Abbiamo mostrato che nel punto individuato con il passo  $\tau_2$  la condizione (CUR) vale (con il maggiore stretto) e che in  $[0, \tau_1]$  vale (AJO). Dunque, dato che  $\nabla f$  è una funzione continua, esiste un intervallo in cui valgono entrambe (la condizione di Wolfe è rispettata):



Siamo in grado di sviluppare un algoritmo leggermente diverso dal Metodo del Gradiente Esatto (6.6). Funziona in modo simile ma con due differenze: la direzione non è più necessariamente l'opposto del gradiente ma è una generica direzione (passo 3) e la ricerca del passo di spostamento non è più esatta ma tale che le condizioni di Wolfe (6.13) siano rispettate (passo 4).

#### Definizione 6.14 (Metodi del Gradiente con ricerca inesatta)

Famiglia di metodi del gradiente in cui si sceglie il passo con un metodo di ricerca inesatta. Sotto forma di algoritmo si può enunciare come segue:

1. Poniamo  $k = 0$ ; Si sceglie un punto di partenza  $x_0$ ;
2. Se  $\nabla f(x^k) = 0$ , allora STOP: abbiamo trovato un punto stazionario.
3. Si sceglie una direzione  $d^k$  che sia di discesa (cioè  $\nabla f(x_k)^T d^k < 0$ )
4. Calcolare  $t_k > 0$  che soddisfi (AJO) e (CUR)
5. Ci si sposta al punto successivo:  $x^{k+1} = x^k + t_k \cdot d^k$
6.  $k := k + 1$ , ritornare al passo 2.

I metodi che rientrano in questa famiglia si differenziano per come implementano la scelta della direzione (passo 3) e per come trovano il passo che soddisfi le condizioni di Wolfe (passo 4).



#### Nota

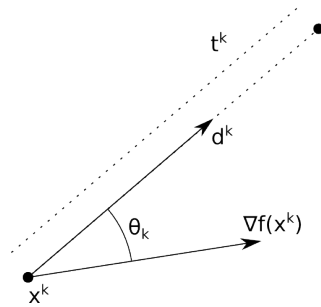
Il metodo del gradiente esatto rientra solo nella pratica, e non formalmente, dentro questa famiglia di metodi inesatti. Abbiamo visto infatti che sebbene la direzione  $d^k = -\nabla f(x^k)$  sia di discesa (è in effetti quella di massima discesa, vedi 6.1.2 a pag. 100), la ricerca esatta del passo di spostamento non soddisfa in generale (ovvero per qualunque valore di  $c_1$  e per qualunque funzione da minimizzare) le condizioni di Wolfe (vedi Proprietà 6.5), ma le soddisfa soltanto nella pratica scegliendo  $c_1$  piccolo.



#### Nota

In realtà la dizione “metodi del gradiente” per i metodi del gradiente con ricerca inesatta non è del tutto corretta, perché in questi metodi la direzione di spostamento non è necessariamente scelta come l'opposto del gradiente, come avviene nel Metodo del gradiente esatto (vedi 6.6). Tuttavia, questi metodi usano in qualche modo il gradiente per scegliere la direzione quindi la dizione può essere usata.

Vogliamo adesso dimostrare la convergenza degli algoritmi che appartengono a questa famiglia. Consideriamo le successioni  $x^k$ ,  $d^k$  e  $t^k$  generate da questa famiglia di metodi, e  $\theta_k$ , che è la successione dei valori dell'angolo compreso tra il gradiente  $\nabla f(x^k)$  e la direzione  $d^k$



Le tre grandezze sono legate dalla seguente equazione, che avevamo già incontrato in 6.1.2:

$$\nabla f(x^k)^T d^k = \|\nabla f(x^k)\|_2 \|d^k\|_2 \cos \theta_k$$

Vediamo innanzi tutto un lemma (che non dimostriamo)

### Lemma 6.15

Supponiamo che

1.  $f$  sia inferiormente limitata
2.  $\nabla f$  non solo sia una funzione continua, ma che la continuità sia Lipschitziana (2.22), ovvero

$$\exists L > 0 \quad \text{t.c.} \quad \|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \forall x, y \in \mathbb{R}^n$$

Allora, detto  $\theta_k$  l'angolo compreso tra  $d_k$  e  $\nabla f(x^k)$ ,

$$\sum_{k=0}^{+\infty} \|\nabla f(x^k)\|_2^2 \cos^2 \theta_k < +\infty$$

cioè la serie per  $k = 0, \dots, +\infty$  è convergente.



#### Nota

Cosa significa dunque che una funzione è Lipschitziana? Una funzione  $f$  è continua quando per  $y \rightarrow x \implies \nabla f(y) \rightarrow \nabla f(x)$ . Una funzione Lipschitziana è continua, ma in più dice con quale velocità  $L$  converge  $\nabla f(y) \rightarrow \nabla f(x)$ .



#### Nota

Nel Lemma 6.15 la prima condizione nella pratica è soddisfatta sempre. Per quanto riguarda la seconda, è in realtà necessario che sia soddisfatta solo nelle vicinanze di un punto stazionario. Ad esempio, tutte le funzioni polinomiali la soddisfano vicino ai punti stazionari.

Del Lemma 6.15 a noi non interessa tanto la convergenza della serie, quanto piuttosto quello che se ne può dedurre:

$$\sum_{k=0}^{+\infty} \|\nabla f(x^k)\|_2^2 \cos^2 \theta_k < +\infty \implies \|\nabla f(x^k)\|_2^2 \cos^2 \theta_k \xrightarrow{k \rightarrow +\infty} 0$$

Questo risultato porta ad una dimostrazione immediata della convergenza:



**Teorema 6.16 (Convergenza)**

Se sono soddisfatte le ipotesi del Lemma 6.15:

1.  $f$  è inferiormente limitata
2.  $\nabla f$  è Lipschitziana

e inoltre

3.  $\exists \delta > 0$  t.c.  $\cos \theta_k \leq -\delta \quad \forall k$

(ovvero la successione dei  $\cos \theta_k$  non sta andando a 0, quindi  $\theta_k$  non può avvicinarsi più di tanto a  $\frac{\pi}{2}$ . In altri termini,  $\exists \lambda \in (0, \frac{\pi}{2})$  t.c.  $\theta_k \leq \frac{\pi}{2} - \lambda$ . Notare che  $\theta_k$  lontano da  $\frac{\pi}{2}$  significa che il gradiente  $\nabla f(x_k)$  e la direzione  $d_k$  non possono essere troppo ortogonali.)

Allora ogni punto di accumulazione della successione  $\{x_k\}$  è un punto stazionario.

*Dimostrazione.* Abbiamo visto che per il Lemma 6.15,

$$\|\nabla f(x^k)\|_2^2 \cos^2 \theta_k \rightarrow_{k \rightarrow +\infty} 0$$

Ma con l'ipotesi 3 abbiamo anche impedito che  $\cos^2 \theta_k$  vada a zero (il punto più vicino a zero che il limite può raggiungere è  $-\delta$ ), dunque non può che essere:

$$\|\nabla f(x^k)\|_2^2 \rightarrow_{k \rightarrow +\infty} 0 \iff \|\nabla f(x^k)\|_2 \rightarrow_{k \rightarrow +\infty} 0$$

ma se la norma del gradiente tende a zero per qualunque sottosuccessione, questo vale in particolare per la sottosuccessione che tende al punto di accumulazione. Passando al limite, per la sua unicità si ha:

$$\|\nabla f(x^*)\|_2 = \lim_{k \rightarrow +\infty} \|\nabla f(x^k)\|_2 = 0$$

abbiamo dimostrato che

$$\|\nabla f(x^*)\|_2 = 0$$

□

### 6.2.4 Metodo di Newton-Raphson

Negli esperimenti, si è visto che per alcune funzioni il Metodo del gradiente esatto è molto lento. Ad esempio, per la funzione di Rosenbrock (6.4 a pag. 105), per ottenere un risultato con un'approssimazione di  $10^{-6}$  sono necessarie più di 1000 iterazioni. Un metodo del gradiente con ricerca inesatta non si comporterebbe meglio per via della conformazione della funzione.

Ma, come abbiamo visto, non è obbligatorio muoversi in direzione del gradiente. Si possono usare altri strumenti, ad esempio il metodo di Newton-Raphson (trattato nella Sezione 5.2), che serve per risolvere un sistema di equazioni non lineari  $F$ :

$$F : \mathbb{R}^n \rightarrow \mathbb{R}^n$$

per trovare un  $x^*$  tale che

$$F(x^*) = 0$$

Il metodo iterativo già visto nella Sezione 5.2 è

$$x^{k+1} = x^k - [JF(x^k)]^{-1} F(x^k) \quad (6.6)$$

dove  $JF(x^k)$  è la matrice Jacobiana di  $F$  calcolata in  $x^k$ . Ovvero, dopo aver ottenuto un punto  $x^k$ , quello successivo si sceglie muovendosi lungo la direzione  $-JF(x^k)^{-1} \cdot F(x^k)$  di passo unitario.

Si nota una forte somiglianza con i metodi di discesa visti finora: anche in Newton-Raphson, dato un punto, ci si muove lungo una certa direzione calcolata in funzione del punto stesso.

Possiamo usare questo metodo, che trova gli zeri di una funzione, per risolvere il nostro problema, ovvero:

$$(P) \quad \min\{f(x) : x \in \mathbb{R}^n\} \quad f : \mathbb{R}^n \rightarrow \mathbb{R} \quad \nabla f : \mathbb{R}^n \rightarrow \mathbb{R}^n$$

Infatti per risolvere (P) è necessario trovare i punti stazionari, ovvero  $x$  tale che  $\nabla f(x) = 0$ . Possiamo dunque usare il metodo di Newton–Raphson ponendo:

$$F = \nabla f$$

per cercare  $x$  t.c.  $F(x) = \nabla f(x) = 0$

Ma cosa è la matrice jacobiana della funzione gradiente? Ricordando che il gradiente è

$$\nabla f(\bar{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

La matrice jacobiana è la matrice delle derivate parziali delle  $n$  funzioni, quindi bisogna derivare  $\frac{\partial f}{\partial x_i}$  rispetto a  $x_i$ . Quindi stiamo semplicemente derivando due volte (ottenendo la matrice hessiana).

$$J(\nabla f)(x) = \nabla^2 f(x)$$

e il metodo di Newton-Raphson in 6.6 diventa

$$x^{k+1} = x^k - [\nabla^2 f(x^k)]^{-1} \nabla f(x^k)$$

Questo metodo rientra nella classe dei metodi del gradiente inesatto (vedi 6.14), con passo e direzione

$$t^k = 1 \quad d^k = -[\nabla^2 f(x^k)]^{-1} \nabla f(x^k) \quad (6.7)$$



#### Nota

Nel paragrafo sulle Direzioni alternative di decrescita (pag. 106) avevamo parlato della possibilità di scegliere la direzione  $d_k$  perturbando la direzione del gradiente con una generica matrice  $D_k$ . In questo caso, abbiamo scelto  $D_k = [\nabla^2 f(x^k)]^{-1}$ .

#### Convergenza: Vogliamo la direzione di discesa, ma a che costo!

In questo paragrafo, vedremo quali condizioni devono essere soddisfatte affinché il metodo sia di discesa, e vedremo che queste condizioni sono molto strette, così nel prossimo paragrafo vedremo che questo vincolo può essere rilassato, e si può avere un metodo che non sempre converge.

Come possiamo garantire che al passo  $k$  la direzione  $d^k$  sia di discesa? Nel paragrafo sulle direzioni di decrescita (pag. 106) avevamo visto che se  $D_k$  è definita positiva, allora la direzione  $d^k$  è di discesa. Quindi nel nostro caso, se la matrice hessiana è definita positiva, allora la sua matrice inversa è definita positiva e le direzioni  $d^k$  sono sempre di discesa.

Dato che in tutti i punti  $x_k$  della successione la matrice hessiana deve essere definita positiva, l'unico modo per avere la garanzia che il metodo sia di discesa, è che la matrice hessiana sia definita positiva in ogni punto. Ma se la matrice hessiana è definita positiva in ogni punto allora la funzione è strettamente convessa.

Dunque chiedere che il metodo sia di discesa equivale a limitare l'algoritmo a risolvere solo funzioni strettamente convesse, e neanche tutte: vi sono infatti delle funzioni strettamente convesse la cui hessiana non è sempre definita positiva. Evidentemente, questo risultato è troppo limitante.

**Convergenza: È necessario che la direzione sia di discesa?**

Proviamo a rilassare le nostre richieste: a noi non serve che il metodo sia sempre di discesa, l'importante è che converga ad un punto stazionario.

Abbiamo visto nel Teorema di convergenza del metodo Newton-Raphson 5.5 (pag. 89) che se  $F$  (ovvero ognuna delle sue componenti) è differenziabile 2 volte con continuità e data una soluzione  $x^*$  tale che  $f(x^*) = 0$ , allora se  $JF(x^*)$  è invertibile e il metodo parte da un punto  $x^0$  sufficientemente vicino alla soluzione, il metodo converge.

Abbiamo visto nel paragrafo precedente che al passo  $k$ , se la matrice hessiana  $\nabla^2 f(x^k)$  è definita positiva, allora la direzione è di discesa.

Guardiamo cosa succede se nelle ipotesi del Teorema 5.5 aggiungiamo il fatto che la matrice hessiana  $\nabla^2 f(x)$  sia definita positiva nel punto stazionario.

Si noti che se una matrice è definita positiva allora è anche invertibile, quindi abbiamo chiesto anche che  $\nabla^2 f(x) = J(\nabla f)(x^*) = JF(x^*)$  sia invertibile, le ipotesi del teorema restano quindi soddisfatte, e dunque il metodo converge a  $x^*$ .

Oltre a convergere, l'ipotesi che abbiamo aggiunto ci dà la garanzia che il metodo, una volta entrato nell'intorno di  $x^*$ , sia di discesa. Si ricorda che, perché questo sia verificato,  $D^k$  deve essere definita positiva ad ogni iterazione. E questo è verificato perché se la matrice hessiana è definita positiva in  $x^*$ , allora è definita positiva anche in un suo intorno<sup>4</sup>.

Queste considerazioni ci permettono di riscrivere il teorema in maniera diversa, nel contesto dell'ottimizzazione.

**Teorema 6.17**

Se

1.  $f$  è differenziabile 3 volte con continuità;
2.  $x^* \in \mathbb{R}^n$  è punto stazionario di  $f$  tale che  $\nabla^2 f(x^*)$  è definita positiva.

Allora, per il Teorema 5.5,

$$\exists \delta > 0 \quad t.c. \quad \forall x^0 \in B(x^*, \delta), x^k \rightarrow x^*$$

(il metodo converge) e inoltre

$\exists M > 0$  per cui:

$$\|x^{k+1} - x^*\|_2 \leq M \|x^k - x^*\|_2^2$$

(la convergenza è superlineare)

Applicando il metodo di Newton-Raphson al nostro problema di minimizzazione abbiamo ottenuto il teorema appena enunciato. Confrontiamolo con l'originale Teorema di convergenza del metodo Newton-Raphson 5.5. Le conclusioni sono le stesse, così come la seconda ipotesi. Per quanto riguarda la prima ipotesi, il metodo originale chiede che  $F$  sia differenziabile 2 volte con continuità, ma nel nostro caso  $F = \nabla f$  e quindi la prima ipotesi equivale a chiedere che  $f$  sia differenziabile 3 volte con continuità.

**Proprietà del metodo**

Facciamo alcune osservazioni su questo metodo, cominciando dal punto stazionario  $x^*$ .

**Osservazione 6.18**

Abbiamo chiesto che  $f(x^*) = 0$  e che  $\nabla^2 f(x^*)$  sia definita positiva. Dal Teorema 3.21 sappiamo che queste sono le condizioni sufficienti affinché  $x^*$  sia un punto di minimo locale.

<sup>4</sup>Questo perché abbiamo scelto  $\nabla^2 f(x)$  derivabile 2 volte con continuità.

Sul punto di partenza  $x_0$ :

 **Osservazione 6.19**

*Nei metodi visti in precedenza, non ci eravamo preoccupati della scelta del punto di partenza, che poteva essere uno qualunque, garantendo la convergenza. In Newton-Raphson invece non può essere scelto a caso ma deve essere tale che  $x^0 \in B(x^*, \delta)$ , ovvero che non sia troppo lontano dalla soluzione. È un metodo di natura locale, e se non si ha idea di dove sia il punto di minimo può darsi che le ipotesi non siano verificate e dunque il metodo non converga o converga a un punto stazionario che però è un punto di massimo oppure un punto né di minimo né di massimo.*

 **Osservazione 6.20**

*Il passo  $t_k$  è identicamente 1 (e la dimostrazione della convergenza è basata su questo). È possibile modificare il metodo di Newton perché esegua una ricerca esatta o inesatta.*

 **Osservazione 6.21**

*Il calcolo è oneroso, infatti ad ogni iterazione il calcolo della direzione ha bisogno del calcolo della matrice hessiana e della sua inversa.*

*Proprio per questo ci sono dei metodi, chiamati para-Newton, che invece di calcolare esattamente  $\nabla^2 f(x)$  ne calcolano un'approssimazione. Ad ogni iterazione, la matrice non viene ricalcolata ma aggiornata tramite una formula esplicita che stima la variazione di questa matrice rispetto al suo valore nel passo precedente in funzione del valore del gradiente. (In una dimensione, questo equivale a stimare la variazione della derivata seconda in base al valore della derivata nel punto.)*

### Un'altra derivazione del metodo di Newton

Nel paragrafo precedente, abbiamo preso il metodo di Newton da un altro contesto (originariamente serviva per trovare gli zeri di una funzione) e lo abbiamo applicato al campo dell'ottimizzazione. Per capire meglio il suo funzionamento, proviamo ad ottenerlo dal contesto dell'ottimizzazione, partendo stavolta dagli sviluppi di Taylor.

Si ricorda che col metodo del gradiente prendevamo una funzione nel punto  $x + d$  che si approssimava al primo ordine con

$$f(x + d) \approx f(x) + \nabla f(x)^T d$$

Nell'approssimazione vengono scartati i termini del secondo ordine. Cercando di minimizzare questa quantità, data la necessità di limitare  $\|d\|$  ad un valore finito, si otteneva che doveva essere  $d = -\nabla f(x)$ . Dunque sviluppando al primo ordine si ottiene il metodo del gradiente.

Ma che succede se sviluppiamo al secondo ordine? Innanzi tutto la funzione  $f$  deve essere derivabile 2 volte. Si ottiene:

$$f(x + d) \approx f(x) + \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(x) d = m_2(d)$$

(in questo caso scartiamo i termini di ordine 3)

Come prima, vogliamo minimizzare questa funzione  $m_2$ , approssimazione al secondo ordine di  $f$ . Ma  $m_2$  è una funzione quadratica:

$$m_2(d) = \frac{1}{2} d^T \underbrace{\nabla^2 f(x)}_Q d + \underbrace{\nabla f(x)^T}_b d + \underbrace{f(x)}_c$$

Sappiamo che se  $\nabla^2 f(x)$  non è semidefinita positiva, la funzione non ha limite inferiore. Quindi supponiamo che sia semidefinita positiva. Se è semidefinita positiva, allora  $m_2(d)$  è convessa. E se  $m_2$  è convessa, significa che tutti i punti stazionari sono anche punti di minimo.

Calcoliamo il gradiente della funzione quadratica<sup>5</sup> e cerchiamo dove si annulla:

$$\nabla m_2(d) \stackrel{\text{def}}{=} \nabla^2 f(x)d + \nabla f(x) = 0$$

Si forma quindi il sistema lineare:

$$\nabla^2 f(x)d = -\nabla f(x)$$

Se la matrice hessiana è invertibile (ad esempio, se  $\nabla^2 f(x)$  è definita positiva<sup>6</sup>), allora il sistema si risolve con:

$$d = -[\nabla^2 f(x)]^{-1} \nabla f(x)$$

Che è proprio la direzione usata nel metodo di Newton! Cosa significa questo? che la direzione del metodo di Newton è quella che minimizza lo sviluppo al secondo ordine della funzione  $f$  (sviluppata intorno al punto  $x$ ).

### Osservazione 6.22

*Quando è che l'approssimazione del secondo ordine usata dal metodo di Newton è buona? Quando  $\|d\|$  è molto piccola. Guardiamo infatti lo sviluppo di Taylor del primo ordine non approssimato ma con il resto esatto:*

$$f(x+d) = f(x) + \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(x+td)d \quad t \in (0, 1)$$

*(Rispetto allo sviluppo del secondo ordine approssimato, la matrice hessiana non è calcolata nel punto  $x$  ma nel punto intermedio tra  $x$  e  $x+t$ , ovvero  $x+td$  con  $t \in (0, 1)$ .)*

*La sola differenza tra lo sviluppo esatto e quello inesatto è tra  $\nabla^2 f(x+td)d$  e  $\nabla^2 f(x)d$  che ovviamente convergono per  $d \rightarrow 0$  e in particolare si avvicinano se le derivate seconde per le componenti sono piatte.*

### Osservazione 6.23

*Il Teorema 6.17 fornisce anche la velocità di convergenza: la proprietà secondo cui  $\exists M > 0$  per cui:*

$$\|x^{k+1} - x^*\|_2 \leq M \|x^k - x^*\|_2^2$$

*si chiama convergenza superlineare del secondo ordine (si può vedere che il Metodo del gradiente ha convergenza lineare).*

## Applicazioni del metodo di Newton

Abbiamo appena visto che il metodo di Newton minimizza una versione quadratica della funzione obiettivo. Se applicassimo il metodo di Newton ad una funzione quadrata, l'approssimazione sarebbe senza errore. Quindi la convergenza si ottiene in un solo passo.

### Esercizio 6.1

*Si applichi il metodo di Newton alla funzione  $f(x) = x_1^2 + 5 \cdot x_2^2$  con un punto di partenza qualunque.*

<sup>5</sup>Dall'Osservazione 2.50 si ha che con  $f$  quadratica, si ha  $\nabla f(x) = Qx + b$  e  $\nabla^2 f(x) = Q$

<sup>6</sup>Finora abbiamo supposto solo che fosse semi-definita positiva.

 **Osservazione 6.24**

*Cosa succede se si applica il metodo alla funzione banana di Rosenbrock 6.4? Come abbiamo visto, il punto iniziale deve essere sufficientemente vicino alla soluzione, che in questo caso come sappiamo è  $x^* = (1, 1)$ .*

- *Se si parte da  $x_0 = (1.1, 1.1)$  il metodo diverge, perché la partenza è troppo lontana dalla destinazione. Infatti la matrice hessiana calcolata in  $x^*$  è positiva, però, calcolata nel punto di partenza  $x_0$  abbiamo un autovalore negativo, quindi la prima iterazione non è di discesa e neanche quelle successive.*
- *Con  $x_0 = (1.01, 1.01)$  continua a divergere.*
- *Con  $x_0 = (1.001, 1.001)$  finalmente il metodo converge perché la matrice hessiana in  $x_0$  è definita positiva.*

# 7 Metodo del gradiente coniugato

## 7.1 Motivazioni e strumenti

Il metodo del gradiente coniugato è un metodo iterativo per risolvere  $Ax = b$ , che ha una proprietà: converge in al più  $n$  passi dove  $n$  è l'ordine della matrice  $A$ .

Per avere la convergenza veloce si sfruttano delle ipotesi molto forti sulla matrice del sistema, ovvero  $A$  deve essere reale, simmetrica e definita positiva. Nonostante queste assunzioni, è comunque possibile ricondurci alle ipotesi anche a partire da una matrice qualsiasi, anche rettangolare,  $C \in \mathbb{R}^{m \times n}$  inserita nel sistema  $Cx = d$ , moltiplicando entrambi i membri per la trasposta di  $C$ :

$$Cx = d \Rightarrow C^T Cx = C^T d \Rightarrow C^T C \in \mathbb{R}^{n \times n} = A \wedge C^T d \in \mathbb{R}^n = b \Rightarrow Ax = b$$

Questo metodo è considerabile anche come metodo diretto di risoluzione ma con complessità  $O(n^3)$ , rendendolo paragonabile ad altri metodi, ma nella forma iterativa può portare ad una buona approssimazione della soluzione in un numero costante di iterazioni (quindi con complessità totale  $O(n^2)$ ), e siamo sicuri di raggiungere questo livello di efficienza nel caso la matrice  $A$  sia sparsa.

Iniziamo la definizione di questo metodo considerando il funzionale quadratico  $\phi(x)$  definito come segue

$$\phi(x) = \frac{1}{2}x^T Ax - b^T x$$

dove  $A$  soddisfa le ipotesi descritte poc' anzi, in particolare

- la simmetria è necessaria perché nelle funzioni quadratiche il gradiente  $\nabla\phi(x^*) = Ax^* - b$  (vedi 2.50), e questo ci serve per collegare la soluzione del sistema lineare alla ricerca del minimo.
- il fatto che  $A$  sia definita positiva ci assicura che la funzione  $\phi(x)$  sia strettamente convessa (vedi 3.6) e quindi che esista un solo minimo.



### Teorema 7.1

$\phi(x)$  ha un solo punto  $x^*$  stazionario che è un punto di minimo per il funzionale e tale punto è anche soluzione di  $Ax = b$ .

$$x^* = \operatorname{argmin}\{ \phi(x) \mid x \in \mathbb{R}^n \} \iff Ax^* = b$$

*Dimostrazione.* Dimostriamo innanzitutto che i punti stazionari del funzionale sono soluzioni del sistema lineare. Come primo passo esplicitiamo le singole componenti che concorrono nel calcolo del funzionale

$$\phi(x) = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n x_i a_{ij} x_j - \sum_{i=1}^n b_i x_i$$

Ora calcoliamo la formula che descrive la generica componente  $k$ -esima del gradiente del funzionale.

$$\frac{\partial\phi(x)}{\partial x_k} = \frac{1}{2} \sum_{j=1}^n \underbrace{a_{kj} x_j}_{\text{per } i=k} + \frac{1}{2} \sum_{i=1}^n \underbrace{x_i a_{ik}}_{\text{per } j=k} - \underbrace{b_k}_{i=k}$$

Ma dato che  $A$  è simmetrica

$$\sum_{i=1}^n x_i a_{ik} = \sum_{i=1}^n x_i a_{ki} = \sum_{i=1}^n a_{ki} x_i$$

e sostituendo  $i$  con  $j$  possiamo riunire le due sommatorie, ottenendo

$$\sum_{j=1}^n a_{kj}x_j - b_k = (Ax)_k - b_k$$

Quindi la componente  $k$ -esima del gradiente coincide con la componente  $k$ -esima del sistema lineare originale. Se consideriamo  $x^*$  punto di minimo per il funzionale abbiamo

$$\nabla\phi(x^*) = 0 \iff Ax^* - b = 0 \iff Ax^* = b$$

Quindi  $x^*$  è soluzione del sistema.

Ora verifichiamo che  $x^*$  è anche punto di minimo del funzionale usando la formula di Taylor del secondo ordine per  $x \neq x^*$

$$\phi(x) = \phi(x^*) + \underbrace{\nabla\phi(x^*)^T}_{=0}(x - x^*) + \underbrace{\frac{1}{2}(x - x^*)^T A(x - x^*)}_{>0, A \text{ definita positiva}} > \phi(x^*)$$

Dato che la relazione è di strettamente maggiore,  $\phi(x^*)$  è anche l'unico punto di minimo.

□

## 7.2 Definizione dei parametri

Il metodo del gradiente che andremo a costruire sarà del tipo

$$(1) \quad x_{k+1} = x_k + \alpha_k p_k$$

dove  $p_k$  dovrà essere una direzione di discesa, mentre  $\alpha_k$  sarà il passo dell'algorithm.

Affinchè sia un metodo valido, la direzione  $p_k$  deve essere di discesa, quindi deve rispettare la proprietà (necessaria e sufficiente)

$$\nabla\phi(x_k)^T p_k < 0$$

Il modo più semplice per ottenere questa proprietà è ricalcare il metodo del gradiente esatto, utilizzando la direzione di massima discesa:  $p_k = -\nabla\phi(x_k) < 0$  per  $x_k \neq x_*$

Chiaramente ora diventa necessario definire in che modo devono essere calcolate le componenti del metodo ad ogni iterazione, cercando di effettuare ottimizzazioni nel loro calcolo (cercando di eliminare il più possibile le operazioni costose)

Abbiamo visto che

$$\nabla\phi(x) = Ax - b$$

Il che vale per ogni  $x$ , quindi possiamo scrivere

$$\nabla\phi(x_k) = Ax_k - b$$

Chiamiamo questa quantità negata  $r_k$  residuo. Intuitivamente, il residuo è la distanza dalla soluzione al passo  $k$ .

### 7.2.1 Passo di discesa

Ora cerchiamo di definire il passo  $\alpha_k$  che l'algorithm effettua lungo la direzione  $p_k$ . Cerchiamo di ottenere ad ogni iterazione il passo di discesa più lungo possibile, quindi  $\alpha_k$  dovrà essere tale da minimizzare il valore del gradiente al passo successivo.

Chiamiamo  $\varphi(\alpha)$  questa quantità da minimizzare la cui definizione è  $\phi(x_k + \alpha p_k)$  dove l'unica variabile è  $\alpha$ . Ora, per minimizzare  $\varphi(\alpha)$  ne cerchiamo un punto stazionario, quindi:

$$0 = \varphi'(\alpha) = \nabla\phi(x_k + \alpha p_k)^T \cdot p_k = (A(x_k + \alpha p_k) - b)^T p_k = \underbrace{(Ax_k - b)}_{-r_k} + \alpha A p_k)^T p_k = -r_k^T p_k + \alpha p_k^T A p_k = 0$$



Dall'ultimo passaggio possiamo ricavare il valore di  $\alpha$  da usare al passo  $k$

$$(2) \quad \alpha_k = \frac{r_k^T p_k}{p_k^T A p_k} \quad \text{con } p_k^T A p_k > 0 \text{ con } A \text{ definita positiva}$$

Dato che anche per il numeratore di (2) vale  $r_k^T p_k = -\nabla\phi(x_k)^T p_k > 0$  (dato che  $p_k$  è di discesa), allora il valore di  $\alpha$  sarà maggiore di 0

### 7.2.2 Aggiornamento del residuo

Una volta ottenuta la formula (2) di  $\alpha$ , cerchiamo di dare una nuova definizione al residuo in modo da ottenerne una formulazione che dipenda solo da parametri del passo precedente. Così facendo potremo utilizzare  $r_k$  come primo elemento da calcolare all'interno della singola iterazione dell'algoritmo.

$$r_{k+1} = b - Ax_{k+1} = b - A(x_k + \alpha_k p_k) = r_k - \alpha_k A p_k$$

Da cui ricaviamo

$$(3) \quad r_{k+1} = r_k - \alpha_k A p_k$$

Inoltre, utilizzando le considerazioni fatte nel calcolo di (2) possiamo notare una particolare relazione che lega  $r_{k+1}$  e  $p_k$ : avendo definito  $\alpha_k$  in modo tale da rendere  $\phi(\alpha)$  minima, otteniamo che

$$\phi'(\alpha_k) = -\nabla\phi(x_{k+1})^T p_k = 0$$

e quindi, sostituendo al gradiente l'espressione del residuo:

$$(4) \quad r_{k+1}^T p_k = 0$$

Ovvero il residuo è ortogonale alla direzione scelta all'iterazione precedente dell'algoritmo.

### 7.2.3 Scelta della direzione di discesa

#### Gradiente Esatto

$p_k$  potrebbe essere scelto in modo da ottenere la direzione di massima discesa (steepest descent) come nel metodo del gradiente esatto, prendendo quindi al passo  $k$  la direzione  $p_k = -\nabla\phi(x_k) = r_k$

Ora valuteremo questo risultato facendo delle considerazioni sulla convergenza del metodo del gradiente esatto.

Definiamo l'errore al passo  $k$  e la sua norma  $A$  di vettore:

$$e_k = x^* - x_k \quad \|e_k\|_A \stackrel{def}{=} \sqrt{e_k^T A e_k}$$

Consideriamo la relazione che lega la norma del vettore di errore al passo  $k$  con quello al passo  $k + 1$  (non dimostrata)

$$\|e_{k+1}\|_A \leq \left( \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right) \|e_k\|_A$$

dove  $\lambda_{\max}$  è il massimo autovalore di  $A$ , mentre  $\lambda_{\min}$  è l'autovalore minimo di  $A$ . Componendo i vari errori nei singoli passi dell'algoritmo, arriviamo ad ottenere

$$\|e_{k+1}\|_A \leq \left( \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right)^{k+1} \|e_0\|_A$$

Ora sostituiamo nella formula il condizionamento della matrice  $A$ , simmetrica, sfruttando la sua formula secondo la norma 2 ( $\mu_2(A) = \frac{\lambda_{\max}}{\lambda_{\min}}$ )

$$\|e_{k+1}\|_A \leq \left( \frac{\mu_2(A) - 1}{\mu_2(A) + 1} \right)^{k+1} \|e_0\|_A$$

### Direzioni alternative

Quindi, se  $\mu_2(A) - 1$  è piccolo, la convergenza è veloce, il che accade quando  $A$  è ben condizionata. Invece di utilizzare il gradiente esatto, useremo un'altra definizione che dimostreremo avere delle proprietà particolari.

$$p_k = \begin{cases} = r_0 & k = 0 \\ = r_k + \beta_k p_{k-1} & k \geq 1 \end{cases} \quad (5)$$

Questa direzione è tale che la direzione all'iterazione  $k$  dipende da quella del passo precedente e da un coefficiente  $\beta_k$  tale che  $p_k^T A p_{k-1} = 0$ , ovvero le direzioni  $p_k$  e  $p_{k-1}$  sono  $A$ -coniugate.

Applicando la formula dell'aggiornamento della direzione alla formula per le direzioni  $A$ -coniugate

$$p_k^T A p_{k-1} = (r_k + \beta_k p_{k-1})^T A p_{k-1} = r_k^T A p_{k-1} + \beta_k p_{k-1}^T A p_{k-1} = 0$$

L'ultimo passaggio ci permette di ricavare  $\beta_k$  in relazione con la direzione al passo  $k - 1$ :

$$\beta_k = -\frac{r_k^T A p_{k-1}}{p_{k-1}^T A p_{k-1}}$$

Ora è necessario dimostrare che la direzione che abbiamo scelto è in realtà una direzione di discesa:



#### Teorema 7.2

La direzione  $p_k$ , come calcolata in (5), è di discesa ( $p_k^T \nabla \phi(x_k) < 0$ )

*Dimostrazione.* Dato che il residuo al passo  $k$  è definito come il negato del gradiente del funzionale quadratico, possiamo scrivere

$$p_k^T \nabla \phi(x_k) = -p_k^T r_k =$$

Sostituendo la definizione di  $p_k$

$$-(r_k + \beta_k p_{k-1})^T r_k = -r_k^T r_k - \beta_k p_{k-1}^T r_k$$

Dato che il residuo è ortogonale alla direzione scelta al passo precedente (e quindi il loro prodotto scalare è 0) abbiamo

$$-r_k^T r_k \leq 0$$

Ma dato che il prodotto scalare può essere 0 solo nel caso in cui il residuo è 0, cioè se al passo  $k$  siamo già arrivati alla soluzione, abbiamo che  $p_k^T \nabla \phi(x_k) < 0$   $\square$

Dalla dimostrazione precedente abbiamo anche la proprietà  $p_k^T r_k = r_k^T r_k$  che ci permette di cambiare la formula (2) di  $\alpha_k$  in modo da semplificarne il calcolo

$$(2 \text{ bis}) \quad \alpha_k = \frac{r_k^T r_k}{p_k^T A p_k}$$

Ora verifichiamo alcune proprietà che riguardano i parametri coinvolti nel metodo

- $r_k^T r_{k-1} = 0$

Sostituendo a  $r_{k-1}$  l'espressione (5) abbiamo

$$r_k^T r_{k-1} = r_k^T (p_{k-1} - \beta_{k-1} p_{k-2}) = r_k^T p_{k-1} - \beta_{k-1} r_k^T p_{k-2}$$

eliminiamo il primo termine uguale a 0 e sostituiamo a  $r_k$  la sua espressione nella formula di aggiornamento del residuo

$$\begin{aligned} & -\beta_{k-1} (r_{k-1} - \alpha_{k-1} A p_{k-1})^T p_{k-2} = \\ & = -\beta_{k-1} \left( \underbrace{r_{k-1}^T p_{k-2}}_{=0 \text{ ortogonali}} - \alpha_{k-1} \underbrace{p_{k-1}^T A p_{k-2}}_{=0 \text{ A-coniugati}} \right) = 0 \end{aligned}$$

$$\bullet p_k^T r_{k-1} = \beta_k r_{k-1}^T r_{k-1} = r_k^T r_k$$

La prima parte si dimostra sostituendo a  $p_k$  la sua definizione per (5)

$$p_k^T r_{k-1} = (r_k + \beta_k p_{k-1})^T r_{k-1} = \underbrace{r_k^T r_{k-1}}_{=0 \text{ ortogonali}} + \beta_k \underbrace{p_{k-1}^T r_{k-1}}_{=r_{k-1}^T r_{k-1}} = \beta_k r_{k-1}^T r_{k-1}$$

La seconda uguaglianza sfrutta la formula di aggiornamento del residuo

$$p_k^T r_{k-1} = p_k^T (r_k + \alpha_{k-1} A p_{k-1}) = p_k^T r_k + \alpha_{k-1} \underbrace{p_k^T A p_{k-1}}_{=0 \text{ A-coniugati}} = r_k^T r_k$$

Dal primo punto ricaviamo che i residui risultanti da due passi consecutivi dell'algoritmo sono sempre ortogonali. Dal secondo punto ricaviamo una nuova formula per il calcolo di  $\beta_k$  che sfrutta esclusivamente i prodotti scalari tra residui:

$$(6) \quad \beta_k = \frac{r_k^T r_k}{r_{k-1}^T r_{k-1}}$$

### 7.3 Definizione algoritmica del metodo

Raccogliendo tutte le definizioni dei vari parametri possiamo definire l'algoritmo per il gradiente coniugato.

1. Scegliere  $x_0 \in \mathbb{R}^n$ ;  $k = 0$
2. Se  $r_k := b - Ax_k = 0$ , STOP
3.  $\beta_k = r_k^T r_k / r_{k-1}^T r_{k-1}$  ( $k \geq 1$ ),  $\beta_0$
4.  $p_k = r_k + \beta_k p_{k-1}$  ( $k \geq 1$ )  $p_0 = r_0$
5.  $\alpha_k = r_k^T r_k / p_k^T A p_k$  (ricerca esatta)
6.  $x_{k+1} = x_k + p_k$
7.  $k = k + 1$  e ritornare a 2)

### 7.4 Convergenza del metodo

Ora inizieremo un percorso che ci porterà a dimostrare che il metodo del gradiente coniugato converge alla soluzione con un numero di passi pari, al massimo, all'ordine della matrice  $A$  del sistema.

Il fondamento per la verifica del numero dei passi è il fatto che tutti i vettori  $r_k$  sono a due a due ortogonali e che le direzioni  $p_k$  sono tutte a due a due  $A$ -coniugate.

Una volta che ciò verrà dimostrato, si potrà affermare che i vettori  $r_k$  con  $k = 0..n-1$  formano necessariamente una base e soprattutto si potrà dire che nei passi successivi ad  $n$  i residui saranno sicuramente tutti pari a 0, il che comporta il raggiungimento della soluzione ottimale.

#### 7.4.1 Direzioni A-coniugate e residui ortogonali



##### Teorema 7.3

siano  $r_0 \neq 0$  e  $h \geq 1$  tale che  $r_k \neq 0$  per ogni  $k \leq h$  allora

$$\begin{cases} r_k^T r_j = 0 \\ p_k^T A p_j = 0 \end{cases}$$

per  $k \neq j \wedge k, j = 0 .. h$

*Dimostrazione.*

Si procede per induzione su  $h$

- Passo base ( $h = 1$ )

La proprietà vale per  $j = 0$  dato che, come dimostrato precedentemente, due residui successivi sono ortogonali tra loro e due direzioni successive sono A-coniugate.

- Passo induttivo ( $h \Rightarrow h + 1$ )

Come ipotesi induttiva abbiamo

$$\begin{cases} r_{h+1}^T r_j = 0 \\ p_{h+1}^T A p_j = 0 \end{cases} \quad \text{per } j = 0..h-1$$

mentre per  $j = h$  vale dalle stesse proprietà presenti al passo base.

Dalla formula di aggiornamento del residuo si ha per  $j = 0..h-1$

$$r_{h+1}^T r_j = \underbrace{r_h^T r_j}_{=0 \text{ per ip. indutt.}} - \alpha_h p_h^T A r_j = -\alpha_h p_h^T A r_j$$

Ma per la definizione della direzione  $p_j$  abbiamo

$$p_h^T A r_j = \underbrace{p_h^T A p_j}_{=0 \text{ per ip. indutt.}} - \beta_j \underbrace{p_h^T A p_{j-1}}_{=0 \text{ per ip. indutt.}} = 0$$

e quindi, tornando alla formulazione precedente abbiamo

$$-\alpha_h p_h^T A r_j = r_{h+1}^T r_j = 0$$

Ora, per quanto riguarda le direzioni, sappiamo che per  $j = 0..h-1$  abbiamo

$$A p_j = \frac{1}{\alpha_j} (r_j r_{j-1})$$

e quindi sostituendo nella nostra tesi da valutare

$$p_{h+1}^T A p_j = r_{h+1}^T A p_j = -\frac{1}{\alpha_j} (r_{h+1}^T r_j - r_{h+1}^T r_{j+1}) = -\frac{1}{\alpha_j} r_{h+1}^T r_{j+1}$$

Quest'ultimo passaggio vale 0 per  $j = 0..h-2$  per quanto ricavato dall'ipotesi induttiva, mentre per  $j = h-1$  vale per l'ortogonalità dei residui successivi

□

Ora sappiamo che  $r_0..r_n$  formano una base e quindi esiste  $s \leq n$  tale che  $r_s = 0 \wedge x_s = x^*$ , inoltre, applicando a  $r_k$  le formule di aggiornamento del residuo e di aggiornamento delle direzioni fino ad arrivare ad  $r_0$ , si nota che il vettore  $r_k$  appartiene allo spazio generato dai vettori  $\{r_0, A r_0, \dots, A^k r_0\}$ .  $r_k$  può essere espresso come polinomio di grado  $k$  di  $A$  moltiplicato per  $r_0$   $r_k = q_k(A) * r_0$

## 7.4.2 Misura dell'errore

Una stima dell'errore commesso al passo  $k$  del gradiente coniugato (senza dimostrazione) è:

$$\|e_k\|_A \leq 2 \left( \frac{\sqrt{\mu_2(A)} - 1}{\sqrt{\mu_2(A)} + 1} \right)^k \|e_0\|_A$$

Che è tanto migliore quanto la radice del numero di condizionamento di  $A$  è vicino a 1. Ciò ci fa concludere che il metodo del gradiente coniugato si comporta molto meglio del metodo del gradiente esatto per matrici mal condizionate.

In tal caso, l'unico problema potrebbe essere l'impossibilità di raggiungere una buona approssimazione entro gli  $n$  passi massimi dell'algorithm in caso di errori di arrotondamento troppo grandi.

Nel caso la matrice  $A$  sia malcondizionata, è possibile applicare tecniche di preconditionamento prima di effettuare l'algorithm, in modo da renderne più veloce la convergenza.

## 7.5 Metodo del gradiente coniugato preconditionato

Migliorare il condizionamento della matrice  $A$  nel problema  $Ax = b$  richiede trasformarla in una matrice più vicina alla matrice identità. Una possibile soluzione consiste nell'usare una matrice invertibile  $C$  per trasformare il sistema come segue:

$$C^{-1} Ax = C^{-1} b \quad \underbrace{C^{-1}A}_{B} \quad \underbrace{C^{-T}C}_I x = C^{-1}b \quad \underbrace{C^{-1}AC^{-T}}_B \quad \underbrace{C^{-T}x}_y = \underbrace{C^{-1}b}_c$$

Come detto in precedenza, per avere un miglioramento, deve verificarsi

$$B \approx I \implies C^{-1}AC^{-T} \approx I \implies CC^T \approx A$$

In tal caso si può usare come matrice  $C$  un fattore della fattorizzazione incompleta di Cholesky, che ha la proprietà di avere elementi nulli in corrispondenza degli elementi nulli di  $A$ , rimanendo così sparsa almeno quanto lo è  $A$ .

Avendo una fattorizzazione, il mal condizionamento rimane presente, spostato dalla matrice del sistema al calcolo di  $B$ . Rimane vero comunque che, avendo una matrice meglio condizionata, si arrivi ad una soluzione del sistema  $By = c$  in un numero di passi minori del problema iniziale malcondizionato.



# 8 Metodo del Gradiente coniugato non lineare

## 8.1 Considerazioni preliminari

Generalizzare il metodo del gradiente coniugato ai casi non lineari implica perdere tutte le ipotesi dovute all'uso del funzionale quadratico  $\phi(x) = \frac{1}{2}x^T Ax - b^T$ , adattando quindi il calcolo dei vari parametri del metodo al caso generico.

Un primo problema nasce dalla definizione del residuo  $r_k$  che equivale nel caso lineare a  $-\nabla\phi(x)$  e su cui si basano sia direzione che passo dell'algoritmo.

Se utilizzassimo, analogamente al gradiente coniugato, il gradiente della funzione obiettivo per definire il residuo al passo  $k$

$$r_k \rightsquigarrow -\nabla f(x^k)$$

avremmo la seguente situazione:

1. Scegliere  $x_0 \in \mathbb{R}^n; k = 0$
2. Se  $\nabla f(x^k) = 0$ , STOP
3.  $\beta_k = \frac{\nabla f(x_k)^T \nabla f(x_k)}{\nabla f(x_{k-1})^T \nabla f(x_{k-1})}, k \geq 1, \beta_0 = 0$
4.  $p_k = -\nabla f(x_k) + \beta_k p_{k-1} \quad p_0 = -\nabla f(x_0)$
5.  $\alpha_k = r_k^T r_k / p_k^T A p_k$
6.  $x_{k+1} = x_k + \alpha_k p_k$
7.  $k = k + 1$  e ritornare a 2)

Il problema serio nasce al punto 5). La formula descritta in realtà non è necessariamente vera, dato che non possiamo più sfruttare le proprietà del funzionale quadratico. Ciò ci toglie anche la possibilità di effettuare una ricerca esatta per trovare il miglior passo di discesa dato che essa stessa diventerebbe un ulteriore problema di ottimizzazione.

Inoltre non possiamo fare più affidamento alle proprietà necessarie al funzionamento all'algoritmo dimostrate nel capitolo precedente che si basano sul residuo.

## 8.2 Approssimazione del passo

Innanzitutto verifichiamo che la direzione scelta  $p_k$  sia di discesa ( $\nabla f(x_k)^T p_k < 0$ ) e sotto quali condizioni ciò è verificato.

$$\nabla f(x_k)^T p_k = \underbrace{-\|\nabla f(x_k)\|_2^2}_{\text{negativo}} + \underbrace{\beta_k \nabla f(x_k)^T p^{k-1}}_{0 \text{ con ricerca esatta}} < 0$$

Il secondo membro è 0 dato che se la ricerca è esatta, il gradiente è ortogonale alla direzione precedente. Ciò si verifica considerando la funzione monodimensionale di ricerca

$$\varphi(\alpha) = f(x_k + \alpha_k p_k)$$

Cercando i punti di minimo abbiamo:

$$0 = \varphi'(\alpha_k) = \nabla f(\underbrace{x_k + \alpha_k p_k}_{x_{k+1}})^T p_k$$

Quindi abbiamo la ricerca esatta, ma effettuando la minimizzazione di  $\varphi'$ , il che la rende molto costosa e inutilizzabile in pratica.

### 8.3 Convergenza del metodo

Utilizzando le condizioni di Wolfe per ottenere una ricerca inesatta, abbiamo un risultato approssimato che non ci garantisce che il metodo sia di discesa.

Per garantire nuovamente la discesa del metodo, sfruttiamo una delle condizioni di Wolfe, specificatamente quella di curvatura (utilizzata in precedenza per evitare che il passo fosse troppo piccolo).

$$(1) \quad \varphi'(\alpha) \geq c_2 \varphi'(0)$$

Derivata

$$\varphi'(\alpha) = \nabla f(x_k + \alpha p_k)^T p_k \quad \varphi'(0) = \nabla f(x_k)^T p_k < 0 \quad \Rightarrow \text{direzione di discesa}$$

Questa condizione da sola purtroppo non basta. Imponiamo un'altra condizione, più forte, in cui mettiamo i termini in valore assoluto

$$|\varphi'(\alpha_k)| \leq c_2 |\varphi'(0)| \quad \text{CUR1}$$

Questa condizione sostituisce la precedente.

Nella condizione semplice, se  $\varphi'(\alpha_k)$  è positiva, allora basta che la direzione di discesa sia di un valore qualsiasi, mentre nella nuova proprietà forziamo la scelta di un valore di modulo sufficientemente grande.

#### Lemma 8.1

Se il passo  $\alpha_k$  soddisfa (CUR1), allora il metodo genera direzioni  $p_k$  tali che

$$-\frac{1}{1-c_2} \leq \frac{\nabla f(x_k)^T p_k}{\|\nabla f(x_k)\|_2^2} \leq \frac{2c_2-1}{1-c_2} \quad c_2 \in (0, 1)$$

Da notare che il limite maggiore della disuguaglianza è negativo se scegliamo  $c_2 < \frac{1}{2}$ , il che ci garantisce la discesa.

Per fare la ricerca richiede anche la condizione di Armijo per il passo massimo da effettuare.

La proprietà di Armijo (che richiede  $0 < c_1 < c_2 < \frac{1}{2}$ ), insieme alla proprietà di curvatura forte formano le condizioni forti di Wolfe

Riscrivendo il metodo, abbiamo quindi

1. Scegliere  $x_0 \in \mathbb{R}^n; k = 0$
2. Se  $\nabla f(x_k) = 0$ , STOP
3.  $\beta_0 = 0$   
 $\beta_k = \frac{\nabla f(x_k)^T \nabla f(x_k)}{\nabla f(x_{k-1})^T \nabla f(x_{k-1})}, k \geq 1$
4.  $p_0 = -\nabla f(x_0)$   
 $p_k = -\nabla f(x_k) + \beta_k p_{k-1}$
5. Calcolare  $\alpha_k > 0$  che soddisfi Armijo e CUR1,  $0 < c_1 < c_2 < 1/2$



$$6. x_{k+1} = x_k + \alpha_k p_k$$

7.  $k = k + 1$  e ritornare a 2)

Dobbiamo però ora verificare che esista questo  $\alpha_k$ .

Si può dimostrare che, se la funzione è inferiormente limitata, le condizioni forti di Wolfe sono verificate in un intervallo; assunto ciò, dobbiamo verificare la convergenza del metodo per poter anche trarre una valutazione sulle performance di metodo.



### Teorema 8.2 (Convergenza)

Supponiamo che:

1.  $f$  sia inferiormente limitata
2.  $f$  sia differenziabile
3.  $\nabla f$  sia Lipschitziana, ossia

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$$

allora

$$\liminf_{k \rightarrow \infty} \|\nabla f(x_k)\| = 0$$

ciò ci garantisce che

$$\exists \{x_{k_j}\}_j \quad \|\nabla f(x_{k_j})\|_2 \rightarrow_{j \rightarrow +\infty} 0$$

Da notare che nel metodo del gradiente esatto possiamo dire che ogni successione converge a zero, mentre qui possiamo solo garantire l'esistenza di una successione. Comunque questa condizione ci garantisce la successione converge sui gradienti, ma non sulla funzione, bisogna quindi dimostrare che la convergenza del gradiente a 0 implica la convergenza del metodo. La dimostrazione è analoga a quella per la ricerca inesatta e non verrà riportata.

Esiste comunque una differenza nella dimostrazione. La proprietà che si ottiene è:

$$\sum_{k=0}^{\infty} \cos^2 \Theta_k \|\nabla f(x_k)\|_2^2 < +\infty$$

Mentre nella ricerca inesatta potevamo individuare un valore  $\delta > 0$  tale che

$$\cos \Theta_k \leq -\delta$$

La differenza sta nel fatto che in questo metodo le direzioni possono tendere a diventare ortogonali, infatti se verifichiamo il valore del coseno dell'angolo tra il residuo e la direzione scelta al passo  $k$

$$\cos \Theta_k = \frac{-\nabla f(x_k)^T p_k}{\|\nabla f(x_k)\|_2 \|p_k\|_2}$$

Supponendo che al passo  $k - 1$  si ottiene

$$\cos \Theta_{k-1} = \frac{-\nabla f(x_{k-1})^T p_{k-1}}{\|\nabla f(x_{k-1})\|_2 \|p_{k-1}\|_2} \approx 0$$

Allora la direzione scelta tende alla direzione di massima crescita e decrescita, il che porta alla possibilità di avere  $\alpha_{k-1} \approx 0$  il che porterebbe ad avere  $x_k \approx x_{k-1}$  e  $\nabla f(x_k) \approx \nabla f(x_{k-1})$  e inoltre  $\beta_k \approx 1$

Ora è possibile verificare che al passo  $k$  la direzione scelta sarà simile a quella del passo precedente.

Sia

$$\frac{2c_2 - 1}{1 - c_2} = -\delta_1 \quad \delta_1 > 0$$

Per il Lemma 2.1

$$\frac{\nabla f(x_{k-1})^T p_{k-1}}{\|\nabla f(x_{k-1})\|_2^2} \leq \delta_1$$

Moltiplicando entrambi i membri per  $\frac{\|\nabla f(x_{k-1})\|_2}{\|p_{k-1}\|_2}$  otteniamo

$$\frac{\|\nabla f(x_{k-1})\|_2}{\|p_{k-1}\|_2} \frac{\nabla f(x_{k-1})^T p_{k-1}}{\|\nabla f(x_{k-1})\|_2^2} \leq \delta_1 \frac{\|\nabla f(x_{k-1})\|_2}{\|p_{k-1}\|_2}$$

Ma il primo membro è uguale a  $-\cos \Theta_{k-1}$ , quindi:

$$0 \approx \cos \Theta_{k-1} \geq \delta_1 \frac{\|\nabla f(x_{k-1})\|_2}{\|p_{k-1}\|_2} \rightarrow \underbrace{\|\nabla f(x_{k-1})\|_2}_{\approx \|\nabla f(x_k)\|_2} \ll \|p_{k-1}\|_2$$

Considerate queste proprietà, allora  $p_k = -\nabla f(x_k) + \beta_k p_{k-1} \approx \nabla f(x_k) + p_{k-1} \approx p_{k-1}$  dato che  $\beta \approx 1$  e  $\|-\nabla f(x_k)\|_2$  è irrilevante rispetto a  $\|p_{k-1}\|_2$ .

Quindi, avendo due direzioni simili in due passi successivi, allora se all'iterazione precedente è stato usato un passo  $\alpha_{k-1}$  che ha prodotto grandi miglioramenti, allora il passo  $\alpha_k$  sarà molto piccolo.

## 8.4 Variazioni del metodo

Una tecnica per evitare questo fenomeno è il *restart*.

Ogni  $\bar{n}$  iterazioni, viene imposto  $\beta_k = 0$ , eliminando ogni contributo delle direzioni precedenti. Normalmente viene scelto  $\bar{n} = n$  se  $n$  è grande e la sotto successione dei  $\{\beta_{k_j}\}_j$  tali che  $\beta_k = 0$  porta il gradiente a convergere a 0 e sarà tale che soddisfa la condizione sul limite inferiore del gradiente

Un'altra tecnica per risolvere questo problema è avere una variante dell'algorithm.

Fletcher - Reeves

Un altro algorithmo è la Variazione di Polak Ribierre, che impone :

$$\beta_k^{PR} = \frac{r_k^T r_k}{r_{k-1}^T r_{k-1}} = \frac{r_k^T (r_k - r_{k-1})}{r_{k-1}^T r_{k-1}} = \frac{\nabla f(x_k)^T (\nabla f(x_k) - \nabla f(x_{k-1}))}{\nabla f(x_{k-1})^T \nabla f(x_{k+1})}$$

In tal modo abbiamo che:

$$\cos \Theta_{k-1} \approx 0 \rightarrow \beta_k^{PR} \approx 0$$

Ciò ci toglie la necessità di fare restart, ma non esiste un teorema di convergenza e esistono dei controesempi in cui il metodo non converge. Comunque, per la gran parte dei problemi, il metodo si comporta molto bene.

# 9 Metodi per l'ottimizzazione non vincolata senza derivate

Vogliamo definire dei metodi per cui è possibile risolvere il problema di ottimizzazione non vincolata

$$(P) \quad \min\{f(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n\}$$

senza l'uso della derivazione di  $f$ , quindi rinunciando all'uso del gradiente e dei metodi di ricerca esatti basati sui punti stazionari delle derivate. Questi metodi hanno valenza particolare nel caso in cui la funzione non sia derivabile, oppure se la definizione di  $f$  rende difficoltoso, dal punto di vista computazionale, il calcolo delle derivate.

Esistono in letteratura diversi metodi di questo tipo, vediamo alcuni:

## Metodo delle differenze finite

In questo caso si passa dall'uso del limite del rapporto incrementale all'approssimazione della derivata su  $x_i$  tramite il calcolo della variazione di valore della funzione, rispetto ad una perturbazione infinitesimale  $t$  lungo la direzione indicata dal vettore della base canonica  $\mathbf{e}_i$ .

$$\frac{\partial f(\mathbf{x})}{\partial x_i} = \frac{f(\mathbf{x} + t\mathbf{e}_i) - f(\mathbf{x})}{t} + O(t)$$

Oppure si può centrare il calcolo effettuando una stima della variazione al centro dell'intervallo  $t$ , ma in questo caso si possono amplificare eventuali problemi di stabilità a causa della differenza tra numeri simili.

$$\frac{\partial f(\mathbf{x})}{\partial x_i} = \frac{f(\mathbf{x} + t\mathbf{e}_i) - f(\mathbf{x} - t\mathbf{e}_i)}{2t} + O(t^2)$$

Un'altra formula permette l'approssimazione dell'elemento  $(i, j)$  della matrice hessiana di  $f$

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) = \frac{f(\mathbf{x} + t\mathbf{e}_i + t\mathbf{e}_j) - f(\mathbf{x} + t\mathbf{e}_i) - f(\mathbf{x} + t\mathbf{e}_j) + f(\mathbf{x})}{t^2} + O(t^2)$$

## Differenziazione automatica

Questo metodo sfrutta, dove possibile, le proprietà di derivazione di funzioni composte per scomporre la funzione originale e calcolare le derivate delle singole componenti, per poi ricomporle. Questa tipologia di metodo è utilizzata da software per la derivazione automatica di funzioni.

## 9.1 Metodo di ricerca diretta a compasso

Ora descriveremo un metodo per l'ottimizzazione non vincolata di (P) che non usa il calcolo di gradienti. Ad ogni iterazione si cercherà una direzione di discesa lungo una delle direzioni descritte dalle basi canoniche di  $\mathbb{R}^n$  effettuando un passo  $t$  fissato e che viene ridotto ogni volta che non rvengono trovati spostamenti che migliorano l'approssimazione.

Prendendo in considerazione la base canonica

$$\{e_1, \dots, e_n\} \subseteq \mathbb{R}^n$$

Definiamo l'insieme  $D$  come:

$$D = \{d_1, d_2, \dots, d_n, d_{n+1}, \dots, d_{2n}\} \quad d_i = \begin{cases} e_i & 1 \leq i \leq n \\ -e_{i-n} & n+1 \leq i \leq 2n \end{cases}$$

Ogni vettore  $v \in \mathbb{R}^n$  può essere scritto normalmente come combinazione lineare della base canonica  $v = \sum_{i=1}^n \alpha_i e_i$  con  $\alpha_i \in \mathbb{R}$ , quindi può anche essere scritto in funzione dei vettori di  $D$  usando solo il valore assoluto di  $\alpha_i$ :

$$v = \sum_{i=1}^n |\alpha_i| \hat{d}_i \quad \text{con } \hat{d}_i = \begin{cases} e_i = d_i & \alpha_i \geq 0 \\ -e_i = d_{i+n} & \alpha_i < 0 \end{cases}$$

### 9.1.1 Descrizione dell'algoritmo

L'algoritmo è descritto come segue:

1. Scegliere  $x_0 \in \mathbb{R}^n, t_0 > 0, k = 0$
2. Se esiste  $d \in D$  tale che

$$f(x_k + t_k d) < f(x_k)$$

allora

$$x_{k+1} = x_k + t_k d, t_{k+1} = t_k$$

altrimenti

$$x_{k+1} = x_k, t_{k+1} = t_k/2$$

3.  $k = k + 1$ ; ritornare a 2)

C'è da notare che questo algoritmo non dà alcun valore specifico per  $t_0$  dato che non abbiamo a priori alcuna informazione sulla funzione; inoltre non viene esplicitato il criterio di arresto, ma anche se non è possibile utilizzare il gradiente di  $f$ , si usa l'unico parametro che ci dà una stima dell'avanzamento dell'algoritmo, cioè il passo  $t_k$ .

Si sceglie quindi la condizione  $t_k \leq \delta$  ( $\delta > 0$  fissato)

Ora riscriviamo l'algoritmo nella seguente forma per dimostrarne delle proprietà, aggiungendo un contatore  $l$  dei fallimenti nella ricerca della direzione di decrescita e chiamiamo  $\{z^l\}$  la successione dei punti di  $\mathbb{R}^n$  e  $\{t_l\}$  la successione dei passi per cui c'è stato un fallimento dell'algoritmo.

1. Scegliere  $x_0 \in \mathbb{R}^n, t_0 > 0, k = 0, l = 0$
2. Se esiste  $d \in D$  tale che

$$f(x_k + t_k d) < f(x_k)$$

allora

$$x_{k+1} = x_k + t_k d, t_{k+1} = t_k$$

altrimenti

$$x_{k+1} = x_k, t_{k+1} = t_k/2, l = l + 1, z^l = x_k, t_l = t_k$$

3.  $k = k + 1$ ; ritornare a 2)

### 9.1.2 Teorema di convergenza

Ora enunceremo una proprietà dell'algoritmo che ci porta a poter dimostrare la convergenza del metodo

**Lemma 9.1**

Sia

$$L_f(x_0) = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\} \quad \text{compatto (chiuso e limitato)}$$

allora il numero possibile di iterazioni successive senza fallimento è limitato.

*Dimostrazione.* L'insieme dei punti sui quali mi posso spostare ad ogni passo possono essere descritti come:

$$\{x_k + t_k \left( \sum_{i=1}^{2n} u_i d_i \right) : u_i \in \mathbb{Z}_+\} \cap L_f(x_k)$$

Ma se l'insieme è  $L_f$  è compatto, lo è anche l'intersezione, quindi abbiamo un numero finito di punti possibili, quindi le iterazioni con passo  $t_k$  sono limitate.

□

**Teorema 9.2 (Convergenza)**Sia  $f$  differenziabile su  $\mathbb{R}^n$ 

Supponiamo che:

- $L_f(x_0)$  sia compatto
- $\nabla f$  sia Lipschitziana, ossia

$$\exists L > 0 \quad : \quad \|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$$

allora

$$\lim_{l \rightarrow \infty} \|\nabla f(z^l)\|_2 = 0$$

Grazie a questo teorema potremo affermare che, dato  $\{z^l\} \subseteq L_f(x_0)$  allora anche la successione dei passi fallimentari è un insieme compatto. Quindi potremo dire che esiste una sua sottosuccessione  $\{z^{l_j}\}_j$  convergente per  $j \rightarrow \infty$  ad un valore che chiamiamo  $z^*$ . Dunque, dato che  $f$  è Lipschitziana, e quindi continua, e la norma è continua, avremo al limite  $\|\nabla f(z^{l_j})\|_2 \rightarrow \|\nabla f(z^*)\|_2$  che per il teorema sarà uguale a 0

*Dimostrazione.* Sia  $v \in \mathbb{R}^n$  con  $\|v\|_2 = 1$ . Allora può essere scritto come combinazione lineare a coefficienti non negativi dei  $\hat{d}_i$

$$v = \sum_{i=1}^n \gamma_i \hat{d}_i, \quad \gamma_i \geq 0$$

I vettori  $\hat{d}_i$  della base sono tutti ortogonali tra di loro, quindi quando effettuiamo il prodotto scalare avremo per ogni singola componente  $l$ :

$$\hat{d}_l^T v = \gamma_l > 0$$

Ed avendo preso  $v$  unitario possiamo affermare che

$$1 = v^T v = \sum_{i=1}^n \gamma_i^2$$

Dato che la somma delle componenti è uguale a 1, allora esisterà un valore  $k$  tale che

$$\gamma_k \geq \frac{1}{\sqrt{n}}$$

Allora esisterà un vettore del generatore  $D$  tale che  $-d^T v \geq \frac{1}{\sqrt{n}}$ .

Useremo questa proprietà che vale per ogni vettore unitario nella dimostrazione della convergenza.

Possiamo supporre che  $\nabla f(z^l) \neq 0$  dato che altrimenti avremmo già la convergenza alla soluzione. Dato che  $\frac{f(z^l)}{\|\nabla f(z^l)\|}$  è un vettore normalizzato, allora vale la proprietà definita precedentemente, cioè

$$\exists d \in D. - \left( \frac{f(z^l)}{\|\nabla f(z^l)\|} \right)^T d \geq \frac{1}{\sqrt{n}}$$

il che equivale a scrivere  $-\nabla f(z^l)^T d \geq \frac{\|\nabla f(z^l)\|_2}{\sqrt{n}}$ .

Dato che  $z^l$  fa parte di un' iterazione fallimentare, allora non vale la condizione di discesa per ogni direzione di  $D$ , ed in particolare nei confronti di  $d$ . Quindi  $0 \leq f(z^l + t_l d) - f(z^l)$ , il che può essere riscritto per il teorema del valor medio come:

$$0 \leq f(z^l + t_l d) - f(z^l) = \nabla f(z^l + \gamma t_l d)^T d \quad \gamma \in (0, 1)$$

Infine, confrontando le proprietà trovate finora, avremo che:

$$0 < \frac{\|\nabla f(z^l)\|_2}{\sqrt{n}} \leq -\nabla f(z^l)^T d \quad \text{per la prima proprietà}$$

Sommando  $\nabla f(z^l + \gamma t_l d)^T d \geq 0$  e raggruppando per  $d$

$$\leq [\nabla f(z^l + \gamma t_l d) - \nabla f(z^l)]^T d \leq \|\nabla f(z^l + \gamma t_l d) - \nabla f(z^l)\|_2 \underbrace{\|d\|_2}_{=1} \leq L \|\gamma t_l d\|_2 \quad \text{per } f \text{ Lipschitziana}$$

Quindi per  $l \rightarrow \infty$

$$0 \leq \frac{\|\nabla f(z^l)\|_2}{\sqrt{n}} \leq 0 \implies \|\nabla f(z^l)\|_2 \rightarrow 0$$

□

Bisogna notare che non viene fornita alcuna misura della qualità delle soluzioni raggiunte, dato che la condizione di arresto non è legata a  $f(x)$ , ma fa esclusivamente riferimento ad un valore assoluto  $\delta$ .

# 10 Il problema lineare dei minimi quadrati

## 10.1 Metodo delle equazioni normali

Sia

$$Ax = b \quad (10.1)$$

un sistema lineare in cui la matrice  $A \in \mathbb{C}^{m \times n}$  dei coefficienti è tale che  $m \geq n$ . Se  $m > n$ , il sistema (10.1) ha più equazioni che incognite e si dice sovradeterminato. Se il sistema (10.1) non ha soluzione, fissata una norma vettoriale  $\| \cdot \|$ , si ricercano i vettori  $x \in \mathbb{C}^n$  che minimizzano la quantità  $\|Ax - b\|$ . In norma 2, il problema diventa quello di determinare un vettore  $x \in \mathbb{C}^n$  tale che

$$\|Ax - b\|_2 = \min_{y \in \mathbb{C}^n} \|Ay - b\|_2 = \gamma \quad (10.2)$$

Tale problema viene detto problema dei minimi quadrati. Il seguente teorema caratterizza l'insieme  $X$  dei vettori  $x \in \mathbb{C}^n$  che soddisfano alla (10.2).



### Teorema 10.1

Valgono le seguenti proprietà:

1.  $x \in X$  se e solo se

$$A^H Ax = A^H b \quad (10.3)$$

Il sistema (10.3) viene detto sistema delle equazioni normali o sistema normale.

2.  $X$  è un insieme non vuoto, chiuso e convesso.
3. L'insieme  $X$  si riduce ad un solo elemento  $x^*$  se e solo se la matrice  $A$  ha rango massimo.
4. Esiste  $x^* \in X$  tale che

$$\|x^*\|_2 = \min_{x \in X} \|x\|_2 \quad (10.4)$$

Il vettore  $x^*$  è l'unico vettore di  $X$  che appartiene a  $N(A^H A)^\perp$  ed è detto soluzione di minima norma

*Dimostrazione.*

1. Siano

$$S(A) = \{y \in \mathbb{C}^m : y = Ax \text{ per qualche } x \in \mathbb{C}^n\}$$

e

$$S(A)^\perp = \{z \in \mathbb{C}^m : z^H y = 0, \forall y \in S(A)\}$$

il sottospazio di  $\mathbb{C}^m$  immagine di  $A$ , e il sottospazio ortogonale a  $S(A)$  (si vedano i paragrafi 2 e 6 del capitolo 1). Il vettore  $b$  può essere così decomposto

$$b = b_1 + b_2, \quad \text{dove } b_1 \in S(A) \text{ e } b_2 \in S(A)^\perp$$

per cui il residuo

$$r = b_1 - Ax + b_2 = y + b_2, \quad \text{dove } y = b_1 - Ax \in S(A) \text{ e } b_2 \in S(A)^\perp$$

vale

$$\|r\|_2^2 = (y + b_2)^H (y + b_2) = \|y\|_2^2 + \|b_2\|_2^2$$

in quanto  $y^H b_2 = b_2^H y = 0$ . Per minimizzare  $\|r\|_2^2$ , non possiamo agire su  $b_2$  poiché è costante, mentre possiamo porre a zero  $y$  che dipende da  $x$ , il che equivale a porre  $b_1 = Ax$ , che è verificato per qualche  $x$  dato che  $b_1$  appartiene al sottospazio immagine di  $A$ .

A questo punto otteniamo

$$\|r\|_2^2 = \|b_2\|_2^2$$

vero se e solo se il vettore  $r$  appartiene a  $S(A)^\perp$  ed è quindi ortogonale alle colonne di  $A$ , cioè

$$A^H r = A^H(b - Ax) = 0$$

Ne segue quindi che  $x \in X$  se e solo se  $x$  è soluzione di (10.3). Inoltre risulta  $\gamma^2 = \|b_2\|_2^2$

Nel caso di  $\mathbb{R}^2$  con una matrice  $A$  di rango 1 si può dare la seguente interpretazione geometrica, illustrata nella figura 7.1. Il vettore  $b = r - Ax$  risulta decomposto in un sol modo nel vettore  $b_2 = r \in S(A)^\perp$  e nel vettore  $b_1 = Ax \in S(A)$ . Il vettore  $Ax$  è quindi la proiezione ortogonale del vettore  $b$  sul sottospazio generato dalle colonne di  $A$ .

### TODO

Inserire immaginina

2. L'insieme  $X$  non è vuoto dato che, per quanto detto precedentemente,  $Ax = b_1$  ha soluzione per come abbiamo scelto  $b_1$ .

Mostriamo ora che  $X$  è chiuso e convesso:

- a) caso  $\text{rank}(A) = n$   
 $A^H A$  è definita positiva non singolare,  $A^H Ax = A^H b$  ha soluzione unica ed  $X$  è formato da un solo vettore.
- b) caso  $\text{rank}(A) < n$   
 $A^H A$  è semi-definita positiva,  $\det(A^H A) = 0$ , abbiamo infinite soluzioni delle forma

$$X = \{x = x_0 + v\} \quad v \in \text{Ker}(A^H A)$$

e poiché  $N(A^H A)$  è chiuso e convesso,  $X$  un insieme chiuso e convesso.

Ad esempio poniamo che il rango sia 1 in  $\mathbb{R}^2$  con  $n = 2$ , allora  $\dim(\text{Ker}(A^H A)) = 1$  e  $\text{rank}(A^H A) = 1$ , l'insieme delle soluzioni  $X$  è una retta, che è un insieme chiuso e convesso.

3. per quanto detto al secondo punto se  $A$  ha rango massimo,  $X$  contiene un solo elemento. In tal caso l'insieme  $N(A^H A)$  è costituito dal solo elemento nullo.
4. l'esistenza della soluzione di minima norma è ovvia nel caso in cui  $X$  si riduce al solo elemento  $x$ .  
 Invece se  $A$  non ha rango massimo, per quanto detto al secondo punto,  $X$  è formato da infiniti elementi ed esiste  $x^* \in X$  tale che  $\|x^*\| = \min_X \|x\|$  e  $x^*$  è l'unica soluzione appartenente a  $\text{Ker}(A^H A)^\perp$ .

Versione più approfondita dalle dispense:

Se  $X$  non si riduce al solo elemento  $x^*$ , sia  $x_0 \in X$  e si consideri l'insieme

$$B = \{x \in \mathbb{C}^n : \|x\|_2 \leq \|x_0\|_2\}$$

Poiché, se  $x \in X$ , ma  $x \notin B$ , risulta  $\|x\|_2 > \|x_0\|_2$ , allora

$$\min_{x \in X} \|x\|_2 = \min_{x \in X \cap B} \|x\|_2$$

L'insieme  $X \cap B$  è un insieme non vuoto, limitato e chiuso, in quanto intersezione di insiemi chiusi, e quindi compatto; essendo la norma una funzione continua, esiste un  $x^* \in X \cap B$  per cui vale la (10.4).

Inoltre  $x^*$  è l'unico vettore di  $X$  appartenente a  $N(A^H A)^\perp$ . Infatti esistono e sono unici  $y \in N(A^H A)$  e  $z \in N(A^H A)^\perp$  tali che  $x^* = y + z$ . Poiché  $x^*$  è soluzione di (10.3),  $A^H A(y + z) = A^H b$ , da cui  $A^H A z = A^H b$  e quindi  $z$  è soluzione del problema (10.2). Se  $y$  non fosse uguale a 0,  $z$  avrebbe norma 2 minore di  $\|x^*\|$ , ciò è assurdo perché  $x^*$  è la soluzione di minima norma: ne segue che  $x^* = z \in N(A^H A)^\perp$ .

Nel caso di  $\mathbb{R}^2$  con una matrice  $A$  di rango 1, si può dare l'interpretazione geometrica illustrata nella figura in cui è riportata la varietà  $X$ , parallela alla varietà  $N(A^H A)$ . Il punto  $x_0$  un qualunque punto di  $X$ . Il punto  $x^*$  è quello di minima norma e quindi quello più vicino all'origine  $O$  dello spazio  $\mathbb{C}$

□



### Risoluzione tramite Cholesky

Se la matrice  $A$  ha rango massimo, allora la soluzione del problema dei minimi quadrati può essere ottenuta risolvendo il sistema

$$A^H A x = A^H b$$

In tal caso, poiché la matrice  $A^H A$  è definita positiva (1.7), si può utilizzare per la risoluzione il metodo di Cholesky 4.6. Determinata la matrice triangolare inferiore tale che

$$L L^H = A^H A$$

la soluzione  $x^*$  di  $A^H A x = A^H b$  viene calcolata risolvendo successivamente i due sistemi di ordine  $n$  con matrice di coefficienti triangolare

$$L y = A^H b$$

$$L^H x = y$$

Se la matrice  $A$  non ha rango massimo, non si può risolvere il sistema (10.3) con il metodo di Cholesky, ma si può applicare il metodo di Gauss con la variante del massimo pivot.

### Costo

Nel caso di una soluzione:

- Calcolo di  $A^H A$ :  $n^2 m / 2$  per la costruzione della matrice hermitiana
- Cholesky:  $n^3 / 6$  per la risoluzione del sistema

In totale quindi il costo è:  $n^2((m/2) + (n/6))$



### Work in progress

Condizionamento, definito solo per le quadrate, per le rettangolari si passa per SVD,

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

$$\mu(A^H A) = \mu^2(A)$$

Cholesky è stabile

## 10.2 Metodo QR

Si esamina ora un altro procedimento, detto metodo QR, che opera direttamente sulla matrice  $A$  fattorizzandola nella forma QR 4.2, dove  $Q$  è una matrice ortogonale. In particolare, poiché si sta lavorando nel campo complesso,  $Q$  è unitaria.

Si supponga dapprima che la matrice  $A \in \mathbb{C}^{m \times n}$  abbia rango massimo,  $k = n \leq m$ . Si applica il metodo di Householder alla matrice  $A$ , ottenendo una successione di matrici di Householder 4.5.

$$P^{(k)} \in \mathbb{C}^{m \times m}, k = 1, \dots, n$$

Posto  $Q^H = P^{(1)} P^{(2)} \dots P^{(n)}$  risulta

$$A = QR \tag{10.5}$$

dove la matrice  $R \in \mathbb{C}^{m \times n}$  ha la forma

$$R = \left[ \begin{array}{c} R_1 \\ 0 \end{array} \right] \begin{array}{l} \} n \text{ righe} \\ \} m - n \text{ righe} \end{array} \tag{10.6}$$

ed  $R_1$  è una matrice triangolare superiore non singolare in quanto  $A$  ha rango massimo. Dalla (10.5) si ha

$$\|Ax - b\|_2 = \|QRx - b\|_2 \stackrel{*)}{=} \|QRx - QQ^Hb\|_2 = \|Q(Rx - Q^Hb)\|_2 = \|Rx - Q^Hb\|_2 = \|Rx - c\|_2 \quad (10.7)$$

\*) Le matrici unitarie preservano la norma 2 (1.1.2.3)

dove  $c = Q^Hb$ . Partizionando il vettore  $c$  nel modo seguente

$$c = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \begin{array}{l} \text{ } n \text{ componenti} \\ \text{ } m - n \text{ componenti} \end{array}$$

per la (10.6) vale

$$Rx - c = \begin{bmatrix} R_1x - c_1 \\ -c_2 \end{bmatrix}$$

per la (10.7) risulta

$$\min_{x \in \mathbb{C}^n} \|Ax - b\|_2^2 = \min_{x \in \mathbb{C}^n} \|Rx - c\|_2^2 = \min_{x \in \mathbb{C}^n} [\|R_1x - c_1\|_2^2 + \|c_2\|_2^2] = \|c_2\|_2^2 + \min_{x \in \mathbb{C}^n} \|R_1x - c_1\|_2^2$$

Poiché  $R_1$  è non singolare, la soluzione  $x^*$  del sistema lineare

$$R_1x = c_1 \quad (10.8)$$

è tale che

$$\min_{x \in \mathbb{C}^n} \|R_1x - c_1\|_2 = \|R_1x^* - c_1\|_2 = 0$$

ne segue che  $x^*$  è la soluzione del problema

$$\|Ax - b\|_2 = \min_{y \in \mathbb{C}^n} \|Ay - b\|_2 = \gamma$$

e si ottiene

$$\gamma = \|c_2\|_2$$



### Nota

Mi sembra che la parte parte sull'evitare il calcolo delle matrici  $P^{(k)}$  e  $Q$  non l'ha spiegata, ma la scrivo per completezza.

Analogamente al caso della risoluzione dei sistemi lineari, il metodo può essere applicato senza calcolare effettivamente né le matrici  $P^{(k)}$ ,  $k = 1 \dots n$ , né la matrice  $Q$ . Si può procedere infatti nel modo seguente: sia

$$P^{(k)} = I - \beta_k v_k v_k^H, \quad k = 1, \dots, n$$

Consideriamo la matrice

$$T^{(1)} = [A \mid b]$$

e si costruisce le successioni delle matrici  $T^{(k)}$  tali che

$$T^{(k+1)} = P^{(k)}T^{(k)} = T^{(k)} - \beta_k v_k v_k^H T^{(k)}$$

Al termine dopo  $n$  passi si ottiene la matrice

$$T^{(n+1)} = [R \mid c]$$

### 10.2.1 Casi e Costi

**Rango massimo** Nel caso  $\det(R_1) \neq 0 \iff \text{rank}(A) = n$ , abbiamo un'unica soluzione. La complessità computazionale è data da:

- Calcolo QR:  $n^2(m - n/3) \approx n^2m$  (circa il doppio di Cholesky)
- Risolvere  $x = R_1^{-1}C_1$ :  $n^2$  operazioni

In generale quindi QR richiede più operazioni, ma se  $m = n$  i due metodi hanno lo stesso costo.

**Rango < n: metodo del massimo pivot** In questo caso esistono infinite soluzioni, ed è necessario provvedere ad una riformulazione del problema nel seguente modo

$$A\Pi = QR$$

dove  $\Pi$  è una matrice di permutazione delle colonne. Infatti se la matrice  $A$  non ha rango massimo, la matrice  $R_1$  ottenuta ha almeno un elemento diagonale nullo e quindi non è possibile calcolare la soluzione del sistema (??).

Questa difficoltà viene superata applicando il metodo QR con la tecnica *del massimo pivot* (4.5.1) per colonne nel modo seguente: al  $k$ -esimo passo, costruita la matrice  $A^{(k)}$  della forma

$$A^{(k)} = \left[ \begin{array}{cc} C^{(k)} & D^{(k)} \\ 0 & B^{(k)} \end{array} \right] \begin{array}{l} \} k - 1 \text{ righe} \\ \} m - k + 1 \text{ righe} \end{array}$$

si determina la colonna di  $B^{(k)}$  la cui norma 2 è massima. Sia  $j$ ,  $1 \leq j \leq n - k + 1$  l'indice di tale colonna. Se  $j = 1$ , si scambiano fra loro la  $k$ -esima e la  $(k + j - 1)$ -esima colonna della matrice  $A^{(k)}$ . Quindi si applica la matrice elementare  $P^{(k)}$  alla matrice con le colonne così permutate. Se il rango di  $A$  è  $r < m$ , questo procedimento termina dopo  $r$  passi, e si ottiene una decomposizione del tipo

$$A\Pi = QR$$

dove  $\Pi \in \mathbb{R}^{n \times n}$  è una matrice di permutazione,  $Q \in \mathbb{C}^{m \times m}$  è una matrice unitaria ed  $R$  è della forma

$$A^{(k)} = \left[ \begin{array}{cc} R_1 & S \\ 0 & 0 \end{array} \right] \begin{array}{l} \} r \text{ righe} \\ \} m - r \text{ righe} \end{array}$$

in cui  $R_1 \in \mathbb{C}^{r \times r}$  è triangolare superiore non singolare e  $S \in \mathbb{C}^{r \times (n-r)}$ . Gli elementi diagonali di  $R_1$  risultano positivi ed ordinati in ordine modulo non crescente.



#### Nota

Il professore su questo ultimo punto è stato vago, ma sul libro è spiegato tutto per bene.



#### Work in progress

##### Cosa avra mai voluto dire?

Costo:  $(m - k)(n - k) + (m - k)(n - k)$

Sommando su tutti i  $k$

$$2 \sum_{k=1}^n (m - k)(n - k)$$

$$i = n - k \quad i = 1, \dots, n \quad k = n - i$$

$$2 \sum_{i=1}^n (m - n + i)i = 2 \left( \sum_{i=1}^n (m - n)i + i^2 \right) = 2 \left( (m - n) \frac{n^2}{2} + \frac{n^3}{3} \right) = (m - n)n^2 + \frac{2}{3}n^3$$

### 10.3 Norme di matrici non quadrate

Per comprendere i successivi metodi, è necessario introdurre questo concetto. Sia  $A \in \mathbb{C}^{m \times n}$ . Le norme matriciali indotte, come già visto in precedenza, vengono definite per mezzo della relazione

$$\|A\| = \max_{\|x\|=1} \|Ax\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

dove  $\|\cdot\|$  e  $\|\cdot\|'$  sono norme vettoriali rispettivamente su  $\mathbb{C}^n$  e  $\mathbb{C}^m$ .

Estendiamo questo concetto sulla matrici  $m \times n$ , usando la norma 2, usandola in  $\|\cdot\|$  e  $\|\cdot\|'$ . Si dimostra che:

$$\|A\|_2 = \sqrt{\rho(A^H A)} \quad (\rho: \text{raggio spettrale})$$

Utilizzando invece la norma, di Frobenius otteniamo

$$\|A\|_F = \sqrt{\text{tr}(A^H A)} = \sqrt{\sum_{i,j} |a_{ij}|^2} \quad (\text{tr} = \text{traccia})$$

Questa non è una norma indotta.

#### Invarianza della norma

Se  $U \in \mathbb{C}^{m \times m}$  e  $V \in \mathbb{C}^{n \times n}$  sono matrici unitarie, poiché

$$(U^H A V)^H (U^H A V) = V^H A^H A V$$

risulta

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

La stessa cosa vale per la norma di Frobenius.

A lezione Bevilacqua ha scritto:

$$\|AU\|_2^2 = \rho(U^H A^H A U) = \rho(A^H A) = \|A\|_2^2$$

$$\|UA\|_2^2 = \rho(A^H U^H U A) = \rho(A^H A) = \|A\|_2^2$$

### 10.4 SVD: Decomposizione ai valori singolari di una matrice

Vedremo un metodo per risolvere il problema dei minimi quadrati, tramite un nuovo tipo di fattorizzazione: la SVD. Il pregio di questo metodo è che vengono coinvolte matrici unitarie che hanno garanzie di stabilità. Assomiglia ad una trasformazione per similitudine ma non lo è.



#### Teorema 10.2

Sia  $A \in \mathbb{C}^{m \times n}$ . Allora esistono una matrice unitaria  $U \in \mathbb{C}^{m \times m}$  ed una matrice unitaria  $V \in \mathbb{C}^{n \times n}$  tali che

$$A = U \Sigma V^H \quad (10.9)$$

dove la matrice  $\Sigma \in \mathbb{R}^{m \times n}$  ha elementi  $\sigma_{ij}$  nulli per  $i \neq j$  e per  $i = j$  ha elementi  $\sigma_{ii} = \sigma_i$  con

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p \geq 0, \quad p = \min\{m, n\}$$

La decomposizione è detta decomposizione ai valori singolari della matrice  $A$ , mentre i valori  $\sigma_i$ , per  $i = 1, \dots, p$ , sono detti **valori singolari** di  $A$ . Indicate con  $u_i, i = 1, \dots, m$ , e  $v_i, i = 1, \dots, n$ , le colonne rispettivamente di  $U$  e di  $V$ , i vettori  $u_i$  e  $v_i, i = 1, \dots, p$ , sono detti rispettivamente **vettori singolari sinistri** e **vettori singolari destri** di  $A$ . La matrice  $\Sigma$  è univocamente determinata, anche se le matrici  $U$  e  $V$  non lo sono.

*Dimostrazione.* Si considera per semplicità il caso  $m \geq n$  (se  $m < n$ , si sostituisce  $A$  con  $A^H$ ). Si procede dimostrando per induzione su  $n$  che la tesi vale per ogni  $m \geq n$ . Per  $n = 1$  è  $A = a \in \mathbb{C}^m$ . Si pone  $\sigma_1 = \|a\|_2$  e si considera come matrice  $U$  la matrice di Householder tale che  $Ua = \sigma_1 e_1$ . La matrice  $V$  è la matrice  $V = [1]$ . Per  $n > 1$  si dimostra che se la tesi vale per le matrici di  $\mathbb{C}^{k \times (n-1)}$ , con  $k \geq n - 1$ , allora vale per le matrici di  $\mathbb{C}^{k \times n}$  con  $m \geq n$ . Sia  $x \in \mathbb{C}^n$  tale che  $\|x\|_2 = 1$  e  $\|Ax\|_2 = \|A\|_2$ . Si consideri il vettore

$$y = \frac{Ax}{\|Ax\|_2} \in \mathbb{C}^m$$

Allora  $\|y\|_2 = 1$  e  $Ax = \sigma_1 y$ , con  $\sigma_1 = \|A\|_2$ . Siano poi  $V_1 \in \mathbb{C}^{n \times n}$  e  $U_1 \in \mathbb{C}^{m \times m}$  matrici unitarie le cui prime colonne sono uguali rispettivamente a  $x$  e  $y$ . Poiché

$$U_1^H A V_1 e_1 = U_1^H A x = U_1^H \sigma_1 y = \sigma_1 [1; 0; \dots; 0]^T$$

è



**Nota**

Quest'ultimo passaggio si spiega col fatto che  $U_1$  è unitaria, allora anche la sua trasposta coniugata è unitaria, allora  $U_1^H U_1 = I$ . Se al posto di  $U$  prendiamo proprio  $y$ , otteniamo la prima colonna della matrice identità

$$A_1 = U_1^H A V_1 = \begin{bmatrix} \sigma_1 & w^H \\ 0 & B \end{bmatrix} \begin{matrix} \text{]1 riga} \\ \text{]m - 1 righe} \end{matrix}$$

in cui  $w \in \mathbb{C}^{n-1}$ ,  $B \in \mathbb{C}^{(m-1) \times (n-1)}$  e  $\mathbf{0} \in \mathbb{C}^{m-1}$ . Si dimostra ora che  $w = 0$ . Si supponga per assurdo che  $w \neq 0$  e si consideri il vettore  $z = \begin{bmatrix} \sigma_1 \\ w \end{bmatrix} \neq \mathbf{0}$  per cui

$$A_1 z = \begin{bmatrix} \sigma_1^2 + w^H w \\ B w \end{bmatrix}$$

Si ha

$$\|A_1 z\|_2^2 = \|z\|_2^4 + \|B w\|_2^2 \geq \|z\|_2^4$$

da cui, dividendo per  $\|z\|_2^2$  si ottiene

$$\frac{\|A_1 z\|_2^2}{\|z\|_2^2} \geq \|z\|_2^2$$

e quindi

$$\|A_1\|_2^2 \geq \sigma_1^2 + w^H w. \tag{10.10}$$

D'altra parte è  $\|A_1\|_2 = \|A\|_2$  in quanto  $A_1$  è ottenuta da  $A$  con trasformazioni unitarie e quindi

$$\|A_1\|_2 = \sigma_1. \tag{10.11}$$

Dal confronto fra la (10.10) e la (10.11) segue l'assurdo. Quindi

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & \sigma_n \end{pmatrix}$$

$$\sigma_i \in \mathbb{C} \quad \sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0$$

Figura 10.1: Struttura della matrice  $\Sigma$

$$A_1 \begin{bmatrix} \sigma_1 & \mathbf{0}^H \\ \mathbf{0} & B \end{bmatrix}$$

Dalla (10.11) segue che

$$\sigma_1 \geq \|B\|_2 \quad (10.12)$$

Infatti

$$\sigma_1^2 = \|A_1\|_2^2 = \rho(A_1^H A_1) = \rho \left( \begin{bmatrix} \sigma_1^2 & \mathbf{0}^H \\ \mathbf{0} & B^H B \end{bmatrix} \right) = \max[\sigma_1^2, \rho(B^H B)] \geq \rho(B^H B) = \|B\|_2^2.$$

Poiché  $B \in \mathbb{C}^{(m-1) \times (n-1)}$  e  $m-1 \geq n-1$ , per l'ipotesi induttiva si ha

$$U_2^H B V_2 = \Sigma_2$$

dove le matrici  $U_2 \in \mathbb{C}^{(m-1) \times (m-1)}$  e  $V_2 \in \mathbb{C}^{(n-1) \times (n-1)}$  sono unitarie e  $\Sigma_2 \in \mathbb{R}^{(m-1) \times (n-1)}$  ha elementi  $\sigma_2 \geq \dots \geq \sigma_p$ . Poiché  $\sigma_2 = \|B\|_2 \leq \sigma_1$  per la (10.12), la tesi segue con

$$U = U_1 \begin{bmatrix} 1 & \mathbf{0}^H \\ \mathbf{0} & U_2 \end{bmatrix}, \quad V = V_1 \begin{bmatrix} 1 & \mathbf{0}^H \\ \mathbf{0} & V_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_1 & \mathbf{0}^H \\ \mathbf{0} & \Sigma_2 \end{bmatrix}$$

Le matrici  $U$  e  $V$  non sono univocamente determinate: infatti se

$$A = U \Sigma V^H$$

è una decomposizione ai valori singolari di  $A$ , e se  $S \in \mathbb{C}^{n \times n}$  una matrice di fase e  $Z \in \mathbb{C}^{(m-n) \times (m-n)}$  è una matrice unitaria, anche

$$A = U \begin{bmatrix} S & 0 \\ 0 & Z \end{bmatrix} \Sigma S^H V^H$$

è una decomposizione ai valori singolari di  $A$ . Inoltre se  $\sigma_i = \sigma_{i+1} = \dots = \sigma_{i+j}$ , per  $j \geq 1$ , detta  $P$  una qualunque matrice unitaria di ordine  $j+1$  e considerata la matrice diagonale a blocchi

$$Q = \begin{bmatrix} I_{i-1} & O & O \\ O & P & O \\ O & O & I_{n-j-i} \end{bmatrix}$$

si ha che

$$A = U \begin{bmatrix} Q & O \\ O & I_{m-n} \end{bmatrix} \Sigma Q^H V^H$$

è una decomposizione ai valori singolari di  $A$  □

Dal teorema segue che

$$A = U \Sigma V^H = \sum_{i=1}^p \sigma_i u_i v_i^H \quad (10.13)$$

$$\left. \begin{aligned} A v_i &= \sigma_i u_i \\ A^H u_i &= \sigma_i v_i \end{aligned} \right\}, i = 1, \dots, p$$

**Costi**

- Diagonalizzazione:  $A^H A : \frac{1}{2}mn^2$
- $A^H A: O(n^3)$
- Fattorizzazione  $QR$ , costo :  $mn^2$

Normalmente costa  $\frac{1}{2}mn^2$ , usando Golom costa  $2mn^2$  ma è più stabile. (Ricontrollare tutto)

**Teorema 10.3**

Sia  $A \in \mathbb{C}^{m \times n}$  e sia

$$A = U \Sigma V^H$$

la sua decomposizione ai valori singolari, dove

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_k > \sigma_{k+1} = \dots = \sigma_p = 0.$$

Allora valgono le seguenti proprietà

1.

$$A = U_k \Sigma_k V_k^H = \sum_{i=1}^k \sigma_i u_i v_i^H, \quad \text{dove}$$

$U_k \in \mathbb{C}^{m \times k}$  è la matrice le cui colonne sono  $u_1, \dots, u_k$ ,

$V_k \in \mathbb{C}^{n \times k}$  è la matrice le cui colonne sono  $v_1, \dots, v_k$ ,

$\Sigma_k \in \mathbb{R}^{k \times k}$  è la matrice diagonale i cui elementi principali sono

$$\sigma_1, \dots, \sigma_k$$

2. Il nucleo di  $A$  è generato dai vettori  $v_{k+1}, \dots, v_n$ .

3. L'immagine di  $A$  è generata dai vettori  $u_1, \dots, u_k$ , e quindi

$$\text{rank}(A) = k$$

4.  $\sigma_i^2, i = 1, \dots, p$  sono gli autovalori della matrice  $A^H A$  (se  $m < n$  i restanti autovalori sono nulli) e quindi

$$\|A\|_F^2 = \sum_{i=1}^k \sigma_i^2$$

$$\|A\|_2 = \sigma_1$$

5. Se  $m = n$  e  $A$  è normale, allora  $\sigma_i = |\lambda_i|, i = 1, \dots, n$ , dove i  $\lambda_i$  sono gli autovalori di  $A$ , e i vettori singolari destri e sinistri coincidono con gli autovettori di  $A$ .

*Dimostrazione.* Si supponga per semplicità che  $p = n \leq m$  (se  $n > m$  si sostituisce  $A$  a con  $A^H$ ).

1. La matrice  $\Sigma$  ha la forma

$$\begin{bmatrix} \Sigma_k & O \\ O & O \end{bmatrix} \begin{matrix} \text{ } k \text{ righe} \\ \text{ } m - k \text{ righe} \end{matrix}$$

per cui, partizionando le matrici  $U$  e  $V$  nel modo seguente

$$U = [U_k | U'_{m-k}] \quad V = [V_k | V_{n-k}]$$

Dalla (10.13) risulta che

$$A = U_k \Sigma_k V_k^H \quad (10.14)$$

2. Se  $x \in \mathbb{C}^n$ , la condizione  $Ax = 0$  per la (10.13) equivalente alla condizione

$$U \Sigma V^H x = 0$$

e, poiché  $U$  non singolare, è equivalente a

$$\Sigma V^H x = 0 \quad (10.15)$$

Il vettore  $z = \Sigma V^H x$  può essere partizionato nel modo seguente

$$\begin{bmatrix} \Sigma_k V_k^H x \\ O \end{bmatrix} \begin{matrix} \text{ } k \text{ componenti} \\ \text{ } m - k \text{ componenti} \end{matrix} \quad (10.16)$$

per cui la (10.15) può essere scritta come  $\Sigma_k V_k^H x = 0$ , ossia  $V_k^H x = 0$ . Quindi  $Ax = 0$  se e solo se  $x$  è ortogonale alle prime  $k$  colonne di  $V$  ed essendo  $V$  unitaria, se e solo se  $x$  è generato dalle restanti colonne di  $V$ .

3. Dalla (10.14) risulta che

$$y = Ax = U_k \Sigma_k V_k^H x = U_k z \quad (10.17)$$

dove  $z = \Sigma_k V_k^H x \in \mathbb{C}^k$ . Quindi  $y$  è generato dalle colonne di  $U_k$ . Viceversa (20) si ha che, poiché la matrice  $\Sigma_k V_k$  è di rango massimo, per ogni  $x \in \mathbb{C}^n$ ,  $x \neq 0$ , esiste uno  $z \neq 0$  per cui vale la (21).

4. Dalla (10.13) si ha che

$$A^H A = V \Sigma^H \Sigma V^H$$

dove  $\Sigma^H \Sigma \in \mathbb{R}^{n \times n}$  è la matrice diagonale i cui elementi principali sono  $\sigma_1^2, \dots, \sigma_p^2$ . Poiché la traccia e il raggio spettrale di due matrici simili sono uguali, si ha

$$\|A\|_F^2 = \text{tr}(A^H A) = \sum_{i=1}^p \sigma_i^2 = \sum_{i=1}^k \sigma_i^2$$

e

$$\|A\|_2^2 = \rho(A^H A) = \sigma_1^2$$

e poiché  $\sigma_1 > 0$  risulta  $\|A\|_2 = \sigma_1$

5. Se  $A$  è normale, dalla forma normale di Schur di  $A$

$$A = U D U^H$$

, segue che

$$A^H A = U D^H D U^H$$

perciò gli autovalori  $\sigma_i$  di  $A^H A$  sono tali che

$$\sigma_i^2 = \bar{\lambda}_i \lambda_i = |\lambda_i|^2 \quad \text{per } i = 1, \dots, n$$

□

### 10.4.1 Calcolo dei valori e vettori singolari

Dal teorema (10.10) si può ricavare anche un procedimento per calcolare i valori e i vettori singolari di  $A$ . Per semplicità si suppone  $m \geq n$  (se fosse  $m < n$  basta riferirsi alla matrice  $A^H$ ). Questo procedimento si articola nei seguenti passi:

1. si calcolano gli autovalori e gli autovettori, normalizzati in norma 2, della matrice  $A^H A$  e si considera la seguente decomposizione in forma normale di Schur della matrice  $A^H A$

$$A^H A = Q D Q^H, D \in \mathbb{R}^{n \times n}, Q \in \mathbb{C}^{n \times n} \quad (10.18)$$

in cui gli elementi principali di  $D$  sono gli autovalori in ordine non crescente di  $A^H A$  e  $Q$  è la corrispondente matrice degli autovettori ( $Q$  è unitaria) (un metodo stabile per calcolare la decomposizione (10.18) sarà descritto nel paragrafo 9);



2. si calcola la matrice

$$C = AQ \in \mathbb{C}^{m \times n} \quad (10.19)$$

e si determina, utilizzando la tecnica del massimo pivot per colonne, la fattorizzazione  $QR$  della matrice

$$C\Pi = UR = U \begin{bmatrix} R_1 \\ O \end{bmatrix} \quad (10.20)$$

dove  $\Pi \in \mathbb{R}^{n \times n}$  è una matrice di permutazione,  $U \in \mathbb{C}^{m \times m}$  è una matrice unitaria ed  $R_1 \in \mathbb{C}^{m \times m}$  è una matrice triangolare superiore con gli elementi principali reali non negativi e ordinati in modo non crescente. Le condizioni imposte sull'ordinamento degli elementi principali di  $R_1$  rendono unica questa fattorizzazione se gli elementi principali di  $R_1$  sono tutti distinti.

Da (10.19) e (10.20) si ottiene, ricordando che  $Q^{-1} = Q^H$ ,  $C\Pi = UR \rightarrow C = UR\Pi^T$

$$A = UR\Pi^T Q^H, \quad (10.21)$$

da cui

$$A^H A = Q\Pi R^H U^H U R \Pi^T Q^H = Q\Pi R^H R \Pi^T Q^H = Q\Pi R_1^H R_1 \Pi^T Q^H$$

e quindi per la (10.18) risulta

$$R_1^H R_1 = \Pi^T D \Pi$$

Poiché la matrice  $\Pi^T D \Pi$  risulta essere diagonale, ne segue che  $R_1^H R_1$  è diagonale, e quindi  $R_1$  non può che essere diagonale. Inoltre, poiché gli elementi principali di  $R_1$  e di  $D$  sono ordinati in modo non crescente, se gli autovalori di  $A^H A$  sono tutti distinti  $\Pi = I$ . Quindi la (10.21), se si pone  $\Sigma = R$  e  $V = Q\Pi$ , rappresenta la decomposizione ai valori singolari di  $A$ .

## 10.5 Risoluzione del problema dei minimi quadrati con i valori singolari

Utilizzando il teorema è possibile dare una formulazione esplicita e della soluzione  $x^*$  di minima norma del problema dei minimi quadrati e del corrispondente  $\gamma$ , anche nel caso in cui la matrice  $A$  non sia di rango massimo.



### Teorema 10.4

Sia  $A \in \mathbb{C}^{m \times n}$  di rango  $k$ , con  $m \geq n \geq k$ , e sia

$$A = U\Sigma V^H$$

la decomposizione ai valori singolari di  $A$ . Allora la soluzione di minima norma del problema (2) è data da

$$x^* = \sum_{i=1}^k \frac{u_i^H b}{\sigma_i} v_i$$

e

$$\gamma^2 = \sum_{i=k+1}^m |u_i^H b|^2$$

*Dimostrazione.* Poiché la norma 2 è invariante per trasformazioni unitarie, si ha

$$\|Ax - b\|_2^2 = \|U^H(Ax - b)\|_2^2 = \|U^H A V V^H x - U^H b\|_2^2$$

e posto  $y = V^H x$ , si ha

$$\|Ax - b\|_2^2 = \|\Sigma y - U^H b\|_2^2 = \sum_{i=1}^n |\sigma_i y_i - u_i^H b|^2 + \sum_{i=n+1}^m |u_i^H b|^2 \stackrel{(10.5)}{=} \sum_{i=1}^k |\sigma_i y_i - u_i^H b|^2 + \sum_{i=k+1}^m |u_i^H b|^2 \quad (10.23)$$

dove  $y_i, i = 1, \dots, n$ , sono le componenti di  $y$ .  
Il minimo della (10.23) viene raggiunto per

$$y_i = \frac{u_i^H b}{\sigma_i}, \quad i = 1, \dots, k \quad (10.24)$$

Fra tutti i vettori  $y \in \mathbb{C}^n$  per cui vale la (10.24) il vettore di minima norma  $y^*$  è quello per cui

$$y_i^* = \begin{cases} \frac{u_i^H b}{\sigma_i} & \text{per } i = 1, \dots, k, \\ 0 & \text{per } i = k+1, \dots, n \end{cases}$$

Poiché  $x^* = Vy^*$ , è  $\|x^*\|_2 = \|y^*\|_2$  e quindi

$$x^* = Vy^* = \sum_{i=1}^k y_i^* v_i = \sum_{i=1}^k \frac{u_i^H b}{\sigma_i} v_i$$

e dalla (10.23) risulta

$$\gamma^2 = \|Ax - b\|_2^2 = \sum_{i=k+1}^m |u_i^H b|^2$$

□

## 10.6 Pseudoinversa di Moore-Penrose

Se la matrice  $A$  è quadrata e non singolare, la soluzione del sistema  $Ax = b$  e del problema dei minimi quadrati

$$\|Ax - b\|_2 = \min_{y \in \mathbb{C}^n} \|Ay - b\|_2 = \gamma$$

coincidono e possono essere espresse nella forma

$$x^* = A^{-1}b$$

per mezzo della matrice inversa  $A^{-1}$ . Il concetto di matrice inversa può essere esteso anche al caso di matrici  $A$  per cui  $A^{-1}$  non esiste. In questo caso si definisce una matrice pseudoinversa di  $A$ , indicata con il simbolo  $A^+$ , che consente di scrivere la soluzione di minima norma del problema dei minimi quadrati nella forma  $x^* = A^+b$ .

### Definizione 10.5 (Matrice pseudoinversa di Moore-Penrose)

Sia  $A \in \mathbb{C}^{m \times n}$  una matrice di rango  $k$ . La matrice  $A^+ \in \mathbb{C}^{n \times m}$  tale che

$$A^+ = V\Sigma^+U^H$$

dove  $\Sigma^+ \in \mathbb{R}^{n \times m}$  è la matrice che ha elementi  $\sigma_{ij}$  nulli per  $i \neq j$  e per  $i = j$  ha elementi

$$\sigma_{ii}^+ \begin{cases} \frac{1}{\sigma_i} & \text{per } i = 1, \dots, k \\ 0 & \text{per } i = k+1, \dots, p \end{cases}$$

è detta **pseudoinversa di Moore-Penrose** di  $A$

 **Proprietà 10.1**

• La matrice  $X = A^+$  è l'unica matrice di  $\mathbb{C}^{n \times m}$  che soddisfa alle seguenti **equazioni di Moore-Penrose**:

1.  $AXA = A$
2.  $XAX = X$
3.  $(AX)^H = AX$
4.  $(XA)^H = XA$

• Se il rango di  $A$  è massimo, allora

$$\begin{array}{ll} \text{se } m \geq n & A^+ = (A^H A)^{-1} A^H \\ \text{se } m \leq n & A^+ = A^H (A A^H)^{-1} \\ \text{se } m = n = \text{rank}(A) & A^+ = A^{-1} \end{array}$$

## 10.7 Calcolo della soluzione di minima norma con il metodo del gradiente coniugato

Il metodo del gradiente coniugato descritto nel Capitolo 7 per la risoluzione dei sistemi lineari può essere utilizzato anche per calcolare la soluzione di minima norma del problema dei minimi quadrati

$$\min_{y \in \mathbb{R}^n} \|Ay - b\|_2^2, \quad \text{dove } A \in \mathbb{R}^{m \times n} \text{ e } b \in \mathbb{R}^m, \text{ con } m \geq n$$

In questo caso la funzione che si deve minimizzare è

$$(Ay - b)^2 = y^T A^T A y + b^T b - 2b^T A y = 2\left(\frac{1}{2}y^T A^T A y - b^T A y\right) + b^T b, \quad y \in \mathbb{R}^n$$

è soluzione del problema dei minimi quadrati è un vettore  $x$  tale che

$$\Phi(x) = \min_{y \in \mathbb{R}^n} \Phi(y)$$

dove

$$\Phi(y) = \frac{1}{2} \|Ay - b\|_2^2 - b^T b = \frac{1}{2} y^T A^T A y - b^T A y$$

Si può quindi applicare il metodo del gradiente coniugato utilizzando l'algoritmo esposto nel Capitolo 7 con l'ovvia modifica che la direzione del gradiente negativo di  $\Phi(x)$  nel punto  $x_k$  è data da

$$-\nabla \Phi(x_k) = A^T (b - Ax_k) = A^T r_k$$

dove  $r_k$  è il residuo del punto  $x_k$ . L'algoritmo allora risulta il seguente:

1.  $k = 0$ ,  $x_0$  arbitrario,  $r_0 = b - Ax_0$
2. se  $r_k = 0$ , STOP
3. altrimenti si calcoli
  - $s_k = A^T r_k$
  - $\beta_k = \frac{\|s_k\|_2^2}{\|s_{k-1}\|_2^2}$  ( $\beta_0 = 0$ , per  $k = 0$ )
  - $p_k = s_k + \beta_k p_{k-1}$  ( $p_0 = s_0$ , per  $k = 0$ )
  - $\alpha_k = \frac{\|s_k\|_2^2}{\|A p_k\|_2^2}$

- $x_{k+1} = x_k + \alpha_k p_k$
- $r_{k+1} = r_k - \alpha_k A p_k$
- $k = k + 1$  e ritorno al punto 2)

Come condizione di arresto si può usare la stessa condizione di arresto usata nel capitolo 5, cioè

$$\|s_k\|_2 = \epsilon \|b\|_2$$

dove  $\epsilon$  è una tolleranza prefissata

### 10.7.1 Tabella riassuntiva

Sia  $A \in \mathbb{C}^{m \times n}$  ( $m > n$ )

Metodo	Costi	Output	Requisiti	Pregi	Difetti
Cholesky ( $LL^H$ )	Costruzione $A^H A$ : $n^2 m / 2$ . Risoluzione sistema: $n^3 / 6$ . Totale: $\frac{n^2}{2} (m + \frac{n}{3})$	$L$ è una matrice triangolare inferiore con elementi principali uguali ad 1 ed $U$ una matrice triangolare superiore.	$A$ rango massimo $\Rightarrow$ $A$ definita positiva	È unica	Poco stabile
Householder ( $QR$ )	Fattorizzazione: $n^2 (\frac{m-n}{3})$ . Risoluzione sistema triangolare: $\frac{n^2}{2}$ . Totale: $n^2 (m - \frac{n}{3})$	$Q$ è una matrice unitaria ed $R$ è una matrice triangolare superiore.	Sempre applicabile. In caso $A$ non ha rango massimo si usa pivoting	Usa matrici unitarie: stabile	Non unica a meno di matrici di fase. Più lento di Cholesky in caso $n \neq m$ .
SVD ( $U\Sigma V^H$ )					
Gradiente coniugato					

## 10.8 SVD troncata

La decomposizione ai valori singolari di una matrice consente anche di risolvere il seguente problema di minimo: data una matrice  $A \in \mathbb{C}^{m \times n}$  di rango  $k$ , e fissato un intero  $r < k$ , qual'è la matrice  $B \in \mathbb{C}^{m \times n}$  di rango  $r$  e più vicina ad  $A$ ? Vale infatti il seguente

### Teorema 10.6

Sia  $A \in \mathbb{C}^{m \times n}$  e sia

$$A = U\Sigma V^H$$

le decomposizione ai valori singolari di  $A$ , dove

$$\sigma_1 \geq \sigma_2 \geq \dots \sigma_k > \sigma_{k+1} = \dots = \sigma_p = 0$$

e sia  $r$  un intero positivo minore o uguale a  $k$ . Indicando con

$$A_r = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^H$$

e con

$$S = \{B \in \mathbb{C}^{m \times n} : \text{rango di } B = r\}$$

si ha

$$\min_B \in S \|A - B\|_2 = \|A - A_r\|_2 = \sigma_{r+1}$$

*Dimostrazione.* Sia  $\Sigma_r \in \mathbb{R}^{m \times n}$  la matrice i cui elementi sono  $\sigma_{ii} = \sigma_i$  per  $i = 1, \dots, r$  e  $\sigma_{ij} = 0$  altrimenti. Allora vale

$$U^H A_r V = \Sigma_r$$

e quindi per il punto 3) del teorema 10.5 è di rango  $A_r = r$ . Risulta inoltre

$$\|A - A_r\|_2 = \|U^H(A - A_r)V\|_2 = \|\Sigma - \Sigma_r\|_2 = \sigma_{r+1} \quad (10.25)$$

in quanto  $\sigma_{r+1}$  è il massimo degli elementi non nulli di  $\Sigma - \Sigma_r$ . Sia  $B \in S$ . Il nucleo di  $B$  ha dimensione  $n - r$  perché  $B$  ha rango  $r$ . Poiché l'intersezione fra  $N(B)$  e il sottospazio  $T$  di  $\mathbb{C}^n$  generato dai vettori  $\mathbf{v}_1, \dots, \mathbf{v}_r, \mathbf{v}_{r+1}$  non può ridursi al solo vettore nullo, in quanto  $\dim T + \dim N(B) = n + 1$ , esiste un elemento  $\mathbf{z} \in N(B) \cap T$ ,  $\mathbf{z} \neq \mathbf{0}$ . Si supponga che  $\|\mathbf{z}\|_2 = 1$ ; essendo  $\mathbf{z}$  elemento di  $T$ , si può scrivere

$$w\mathbf{z} = \sum_{i=1}^{r+1} \alpha_i \mathbf{v}_i \quad (10.26)$$

per il punto 1) del teorema 10.5 si ha

$$A\mathbf{z} = \sum_{i=1}^k \sigma_i \mathbf{u}_i (\mathbf{v}_i^H \mathbf{z}) = \sum_{i=1}^{r+1} \sigma_i \mathbf{u}_i (\mathbf{v}_i^H \mathbf{z}) \quad (10.27)$$

in quanto, essendo  $\mathbf{z} \in T$  e  $V$  unitaria, è  $\mathbf{v}_i^H \mathbf{z} = 0$  per  $i = r + 2, \dots, k$ . Poiché  $\mathbf{z} \in N(B)$ , è  $B\mathbf{z} = \mathbf{0}$  e si ha

$$\|A - B\|_2^2 \geq \|(A - B)\mathbf{z}\|_2^2 = \|A\mathbf{z}\|_2^2 \quad (10.28)$$

D'altra parte per la (10.27), poiché i vettori  $\mathbf{u}_i, i = 1, \dots, r + 1$  sono ortonormali, si ha

$$\|A\mathbf{z}\|_2^2 = \sum_{i=1}^{r+1} \sigma_i^2 |\mathbf{v}_i^H \mathbf{z}|^2 \quad (10.29)$$

Poiché  $\sigma_i^2 \geq \sigma_{r+1}^2, i = 1, \dots, r + 1$ , si ha

$$\|A\mathbf{z}\|_2^2 \geq \sigma_{r+1}^2 \sum_{i=1}^{r+1} |\mathbf{v}_i^H \mathbf{z}|^2$$

per la 10.26 è

$$\sum_{i=1}^{r+1} |\mathbf{v}_i^H \mathbf{z}|^2 = \sum_{i=1}^{r+1} \left| \mathbf{v}_i^H \sum_{j=1}^{r+1} \alpha_j \mathbf{v}_j \right|^2 = \sum_{i=1}^{r+1} \left| \sum_{j=1}^{r+1} \alpha_j \mathbf{v}_i^H \mathbf{v}_j \right|^2 = \sum_{i=1}^{r+1} |\alpha_i|^2$$

e quindi dalla (10.29) segue

$$\|A\mathbf{z}\|_2 \geq \sigma_{r+1} \quad (10.30)$$

Confrontando la (10.29) e la (10.28) si ha che

$$\|A - B\|_2 \geq \sigma_{r+1}$$

e poiché per la (10.25) è  $\|A - A_r\|_2 = \sigma_{r+1}$ , ne segue che

$$\min_{B \in S} \|A - B\|_2 = \|A - A_r\|_2 = \sigma_{r+1}$$

□

## Golub e Reinsch

Ultima cosa Autovalori ed autovettori  $A^H A$

$$A^H A = Q D Q^H$$

Si può esprimere (algoritmo di Golub-Reinsch)

$$A = P^H B H^H$$

$$P^H, H^H : \text{unitarie}$$

$B$  : bidiagonale superiore

Allora

$$A^H A = H B^H B H^H$$

Si può anche fare in modo che  $B$  sia reale, allora

$$A^H A = H B^T B H^H$$

Il costo di tale procedura è lineare ( $O(n)$ )

# 11 Il problema nonlineare dei minimi quadrati

## 11.1 Minimi quadrati lineari

Analizzeremo il problema dei minimi quadrati lineari dal punto di vista dell'ottimizzazione. Siamo nel caso continuo non vincolato: è un problema quindi che rientra nella classe di problemi che abbiamo studiato. Analizziamone la forma:

$$A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^n \quad \min\{\|Ax - b\|_2 : x \in \mathbb{R}^n\}$$

Per trattare il problema, riscriviamo la funzione obiettivo: possiamo fare la seguente equivalenza che agevolerà meglio i conti (tale riscrittura vale solo nel caso lineare)

$$\min\{\|Ax - b\|_2 : x \in \mathbb{R}^n\} \equiv \min\left\{\frac{1}{2}\|Ax - b\|_2^2 : x \in \mathbb{R}^n\right\}$$

Abbiamo fatto una scalatura ed elevato al quadrato: i minimi in tutte e due le scritte coincidono, in quanto le funzioni che abbiamo applicato sono monotone. Sviluppiamo la funzione obiettivo:

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2 = \frac{1}{2}(Ax - b)^T(Ax - b) = \frac{1}{2}x^T A^T A x - (A^T b)^T x + \frac{1}{2}b^T b$$

Questa è una funzione quadratica, di cui sappiamo calcolare immediatamente il gradiente

$$\nabla f(x) = A^T A x - A^T b$$

mentre la matrice Hessiana è data da :

$$\nabla^2 f(x) = A^T A$$

### Considerazione 1: annullamento del gradiente

Dove si annulla il gradiente?

$$\nabla f(x) = 0 \iff A^T A x = A^T b$$

Quest'ultimo è il sistema delle equazioni normali: i punti stazionari sono le soluzioni delle equazioni normali. Sappiamo dalla teoria che si ha minimo se la funzione è *convessa*, ed è noto inoltre che una funzione è convessa se la matrice Hessiana ( $\nabla^2 f(x)$ ) è semidefinita positiva. Andiamo quindi a controllare se abbiamo questa condizione:

$$x^T A^T A x = (Ax)^T Ax = \|Ax\|_2^2 \geq 0$$

La matrice Hessiana è semidefinita positiva in ogni punto, quindi la funzione è convessa: concludiamo quindi che i punti stazionari coincidono con i punti di minimo. Abbiamo visto quello che ha spiegato Bevilacqua sotto un altro punto di vista, quello dell'ottimizzazione.

### Considerazione 2: matrice A di rango massimo

Se la matrice A è di rango massimo, abbiamo che la funzione è strettamente convessa: infatti

$$\begin{aligned} A \text{ rango massimo} &\implies (Ax = 0 \iff x = 0) \implies x^T \nabla^2 f(x) x > 0 \forall x \neq 0 \\ &\implies f \text{ strettamente convessa} \end{aligned}$$

È noto inoltre che, quando la funzione è strettamente convessa, esiste un solo punto di minimo.

**Considerazione 3**

$$f(x) = \frac{1}{2} \|Ax - b\|_2^2 = \frac{1}{2} \left( (A_1x - b_1)^2 + (A_2x - b_2)^2 + \dots + (A_mx - b_m)^2 \right)$$

$f(x)$  si dice lineare quando gli addendi  $(A_mx - b_m)$  sono lineari.

Questo è il punto di partenza per parlare del problema dei minimi quadrati non lineare.

**11.2 Problema dei minimi quadrati non lineare**

Un problema dei minimi quadrati non lineare ha le seguenti caratteristiche:

$$\min\{f(x) : x \in \mathbb{R}^n\} \quad f(x) = \frac{1}{2} \sum_{j=1}^m r_j^2(x)$$

dove

$$r_j : \mathbb{R}^n \rightarrow \mathbb{R} \text{ non lineari} \quad x \rightarrow A_jx - b_j;$$

Un esempio di problema non lineare era il data fitting, visto nelle prime lezioni. I metodi per risolvere questo problema li avremmo già ma, poiché conosciamo la struttura della funzione obiettivo, vogliamo sfruttarne la struttura per specializzare i metodi che abbiamo visto fino ad adesso.

Supporremo inoltre  $r_j$  differenziabile 2 volte.

Cercheremo di specializzare il metodo di Newton.

**Calcolo del gradiente**

La funzione  $r_j^2$  è la composizione di  $r_j$  con una funzione  $\varphi$  che associa ad un valore il suo quadrato

$$r_j^2(x) = \varphi(r_j(x)) \quad \varphi : \mathbb{R} \rightarrow \mathbb{R} \quad t \rightarrow t^2$$

Tale funzione è derivabile e vale

$$\varphi'(t) = 2t$$

Il gradiente di una una funzione composta è noto per il teorema (6.7) : calcoliamolo in questo caso

$$\nabla r_j^2 = \nabla \varphi(r_j(x)) = \varphi'(r_j(x)) \nabla r_j(x) = 2r_j(x) \nabla r_j(x) \quad (11.1)$$

Quindi, usando la linearità dei gradienti e utilizzando la (11.1) otteniamo

$$\nabla f(x) = \frac{1}{2} \sum_{j=1}^m \nabla r_j^2(x) = \sum_{j=1}^m r_j(x) \nabla r_j(x) = \dots$$

Cerchiamo una rappresentazione più compatta:

$$R = (r_1, \dots, r_m) : \mathbb{R}^n \rightarrow \mathbb{R}^m$$

$$R(x) = (r_1(x), \dots, r_m(x))$$

Matrice Jacobiana, le cui righe sono i gradienti dei componenti

$$J_R(x) = \begin{bmatrix} \nabla r_1(x)^T \\ \vdots \\ \nabla r_m(x)^T \end{bmatrix} \in \mathbb{R}^{m \times n}$$

Il tutto si riduce quindi a:

$$\nabla f(x) = \dots = J_R(x)^T R(x)$$



### Calcolo della matrice hessiana

Calcoliamo le derivate seconde, la matrice hessiana di  $f$

$$[\nabla^2 f(x)]_{kl} = \frac{\partial^2 f(x)}{\partial x_k \partial x_l} = \frac{\partial}{\partial x_k} \left( \frac{\partial f}{\partial x_l}(x) \right) = \frac{\partial}{\partial x_k} \left( \sum_{j=1}^m r_j(x) \frac{\partial r_j}{\partial x_l}(x) \right) \stackrel{*)}{=} \sum_{j=1}^m \underbrace{\frac{\partial r_j}{\partial x_k}(x) \frac{\partial r_j}{\partial x_l}(x)}_{[J_r(x)^T J_R(x)]_{kl}} + \sum_{j=1}^m r_j(x) \underbrace{\frac{\partial^2 r_j}{\partial x_k \partial x_l}}_{[\nabla^2 r_j(x)]_{kl}}$$

#### Domanda aperta

\*) Derivata prodotto?

$$[J_R(x)^T J_R(x)]_{kl} = \begin{bmatrix} \nabla r_1(x)^T & \cdots & \nabla r_m(x)^T \end{bmatrix} \begin{bmatrix} \nabla r_1(x)^T \\ \vdots \\ \nabla r_m(x)^T \end{bmatrix}$$

In forma matriciale possiamo riscrivere tutto come

$$\nabla^2 f(x) = J_R(x)^T J_R(x) + \sum_{j=1}^m r_j(x) \nabla^2 r_j(x)$$

Poiché abbiamo le formule esplicite del gradiente e della matrice hessiana, possiamo sostituirle nei metodi visti specializzarli. Ad esempio nel metodo di Newton 6.7:  $d^k \in \mathbb{R}^n$  soluzione del sistema

$$\nabla^2 f(x^k) d = -\nabla f(x^k)$$

Se  $\nabla^2 f(x^k)$  è definita positiva, allora  $d_N^k$  è una direzione di discesa. Cercheremo un'approssimazione del metodo di Newton per togliere di mezzo i termini del secondo ordine: otterremo il metodo di Gauss-Newton, per evitare il calcolo delle matrici hessiane.

## 11.3 Metodo di Gauss-Newton

Vogliamo cercare di approssimare quindi le matrici hessiane con la parte del primo ordine, che domina la parte del secondo ordine quando i residui sono piccoli.

### Notazioni

Cerchiamo delle notazioni comode per fare le nostre considerazioni

$$J_k = J_R(x^k), \quad R_k = R(x^k) = (r_1(x^k), \dots, r_n(x^k)) \quad \text{Funzioni residuali: } r_j$$

Il gradiente è definito nel seguente modo

$$\nabla f(x^k) = J_k^T r_k,$$

mentre la matrice hessiana viene approssimata come segue

$$\nabla^2 f(x^k) \approx J_k^T J_k$$

Il metodo di Gauss-Newton

$$d_{GN}^k \in \mathbb{R}^n \text{ soluzione di } \underbrace{J_k^T J_k d = -J_k^T r_k}_{\text{sistema lineare}}$$

Mentre nel metodo di Gauss standard per avere una direzione di discesa era necessario che la matrice hessiana fosse definita positiva, qui invece le condizioni sono meno strette. Vale infatti il seguente teorema:

**Teorema 11.1**

Se  $\nabla f(x^k) \neq 0$ , allora  $d_{GN}^k$  è una direzione di discesa

*Dimostrazione.* Facciamo una serie di sostituzioni

$$\begin{aligned} \nabla f(x^k)^T d_{GN}^k &= (J_k^T r_k)^T d_{GN}^k \stackrel{\text{reverse order law}}{=} -(J_k^T J_k d_{GN}^k)^T d_{GN}^k = \\ &\stackrel{\text{reverse order law}}{=} -(J_k^T d_{GN}^k)^T (J_k d_{GN}^k) \stackrel{\text{sistema lineare}}{=} -\|J_k^T d_{GN}^k\|_2^2 \leq 0 \end{aligned}$$

Affinché si abbia una direzione di discesa, la quantità deve essere minore di zero. Ma se

$$J_k^T = d_{GN}^k = 0 \rightarrow J_k^T r_k = 0 \text{ abbiamo } \nabla f(x^k) = 0$$

Quindi quando il gradiente non è zero, abbiamo una direzione di discesa.  $\square$

**Legame tra Gauss Newton e minimi quadrati**

$$J_k^T J_k d = -J_k^T r_k$$

inoltre è il sistema delle equazioni normali associate al sistema  $J_k d + r_k = 0$ , ovvero la direzione di Newton  $d_{GN}^k$  è la soluzione di un problema di minimi quadrati lineare, risolve il problema dei minimi quadrati lineari

$$\min \left\{ \frac{1}{2} \|J_k d + r_k\|_2^2 \quad d \in \mathbb{R}^n \right\}$$

Trovare quindi una direzione di Gauss Newton, corrisponde a risolvere un problema dei minimi quadrati, e con Bevilacqua abbiamo visto 3 metodi che non richiedono il calcolo della matrice hessiana.

Lo sviluppo di Taylor del primo ordine di  $r_j(x^k + d)$  è

$$r_j(x^k + d) \approx r_j(x^k) + \nabla r_j(x^k)^T d$$

$$f(x^k + d) = \frac{1}{2} \sum_{j=1}^n r_j^2(x^k + d) \approx \frac{1}{2} \sum_{j=1}^m [r_j(x^k) + \nabla r_j(x^k)^T d]^2 = \frac{1}{2} \|r_k + J_k d\|_2^2$$

Abbiamo quindi un'approssimazione della funzione obiettivo.

**Domanda aperta**

Non ho capito quest'ultima osservazione fatta sopra su Taylor: perché ci serve?

Potremmo pensare passi unitari in stile Newton, ma in questo caso non sarebbe garantita la convergenza del metodo: bisogna aggiungere una ricerca inesatta del passo.

**Metodo di Gauss-Newton ("damped GN")**

1. si sceglie un punto  $x_0 \in \mathbb{R}^n$ ,  $k = 0$
2. se  $\nabla f(x) = 0$  (il punto trovato è stazionario), allora STOP.
3. altrimenti dobbiamo calcolare  $d_{GN}^k \in \mathbb{R}^n$  soluzione di  $J_k^T J_k d = -J_k^T r_k$
4. calcolare  $t_k > 0$  che soddisfi le condizioni di Wolfe
5.  $x^{k+1} = x^k + t_k d_{GN}^k$
6.  $k = k + 1$  e tornare a 2)

 **Teorema 11.2 (Teorema di convergenza)**

Supponiamo che

- $L_f(x^0) = \{x \in \mathbb{R}^n : f(x) \leq f(x^0)\}$  (curve di sottolivello) sia compatto
- $r_j$  siano differenziabili con continuità per ogni  $j = 1 \dots n$
- $\nabla f$  sia continua di tipo Lipschitziano
- $\exists \gamma > 0$  t.c.  $\|J_k z\|_2 \geq \gamma \|z\|_2 \quad \forall z \in \mathbb{R}^n \quad \forall k$  (iterazioni del metodo) (\*\*\*)

Allora

$$\lim_{k \rightarrow +\infty} \|\nabla f(x^k)\|_2 = 0$$

Vediamo il significato dell'ultima ipotesi: ci garantisce che i valori singolari di  $J_k$  siano uniformemente distanti da 0. Cerchiamo di capire perché: sia  $z = v_i$  vettore singolare destro di  $J_k$ : ha norma 1

$$\|J_k v_i\|_2 = \sigma_i \quad \sigma_i \geq \gamma \text{ per ogni matrice } J_k$$

Quindi per ogni iterazione del metodo tutti i valori singolari saranno maggiori di  $\gamma$ .

( $\sigma_i$ : valori singolari)

*Dimostrazione.*

**TODO**

Fare rivedere a Bigi la dimostrazione

Per ipotesi  $x \rightarrow J_r(x)$  è una funzione continua. Allora  $x \rightarrow \|J_R(x)\|_2$  continua, poiché la norma è continua. La curva di sottolivello  $L_f(x^0)$  per ipotesi è compatta. Allora, per il teorema di Weirstrass esiste un massimo

$$\exists \beta > 0 \quad \|J_R(x)\|_2 \leq \beta \quad \forall x \in L_f(x^0)$$

Sia  $\theta_k$  l'angolo tra  $-\nabla f(x^k)$  e  $d_{GN}^k$ . Abbiamo

$$\cos \theta_k = \frac{-f(x^k)^T d_{GN}^k}{\underbrace{\|\nabla f(x^k)\|_2}_{(*)} \underbrace{\|d_{GN}^k\|_2}_{(*)}} = \frac{\|J_k^T d_{GN}^k\|_2^2}{\underbrace{\|J_k^T J_k d_{GN}^k\|_2}_{(*)} \|d_{GN}^k\|_2} \geq \dots (**)$$

Valgono

- $J_k^T r_k = -J_k^T J_k d_{GN}^k \quad (*)$
- $\|J_k^T J_k d_{GN}^k\|_2 \leq \|J_k\|_2^2 \|d_{GN}^k\|_2 \leq \beta^2 \|d_{GN}^k\|_2 \quad (**)$

$$\dots \geq \frac{\|J_k^T d_{GN}^k\|_2^2}{\beta^2 \|d_{GN}^k\|_2^2} \underset{(***)}{\geq} \frac{\gamma^2 \|d_{GN}^k\|_2^2}{\beta^2 \|d_{GN}^k\|_2^2}$$

Dato Wolfe, Lipschitziana, funzione limite

$$\sum_{k=0}^{\infty} \underbrace{\cos^2 \theta_k}_{\neq 0} \|\nabla f(x^k)\|_2^2 < +\infty \implies \lim_{k \rightarrow +\infty} \|\nabla f(x^k)\|_2 = 0$$

□

In realtà la condizione di Lipschitziana non è necessaria.

## 11.4 Regressioni (non) lineari - data/curve fitting

Contesto: immaginiamo di avere un gruppo di dati, istanti di tempo

$$y \in \mathbb{R} \quad t_j \in \mathbb{R}^p$$

Vediamo se ci sono delle correlazioni tra i  $t_j$  e gli  $y$ . L'idea è avere una black box che prende in input i  $t_j$  e sputa fuori i  $y$

$$y \approx f\left(\underbrace{t_j}_{\text{regressori}}\right)$$

Per cercare di capire il comportamento della black box possiamo fare

$$\min \left\{ \frac{1}{2} \left( \sum_{j=1}^m [y_j - f(t_j)]^2 \right); \quad f \in T \right\}$$

Non abbiamo un vettore come incognita, abbiamo un insieme di funzioni. In  $\mathbb{R}^n$  abbiamo una base finita, nello spazio di funzioni abbiamo una base infinita, le variabili sono delle funzioni. Questo argomento non fa parte del corso. Cerchiamo però di vedere un esempio.

$$T = \{f(x_1, \dots, x_m; \cdot) : \underbrace{x_1, x_2, \dots, x_n}_{\text{parametri}} \in \mathbb{R}\}$$

Assegnato il valore ad un insieme finito di parametri, otteniamo la funzione corrispondente. Abbiamo quindi ristretto l'insieme delle funzioni che è possibile considerare, questo per un discorso di rappresentazione. Basta riprendere l'esempio della prima lezione di curve fitting per avere un esempio concreto. Il nuovo problema è espresso come

$$\min \left\{ \frac{1}{2} \sum_{j=1}^n [y_j - f(x_1, \dots, x_n)]^2 : x_1, x_2, \dots, x_n \in \mathbb{R} \right\}$$

Esempio: Modello di crescita di un nuovo servizio. si vuole ottenere un modello. Vari modelli di crescita:

- Crescita esponenziale:  $eg(t) = x_1 e^{x_2 t}$
- Potenza temporale:  $tp(t) = x_1 t^{x_2}$
- Crescita limitata  $lmg(t) = x_3 - x_1^{-x_2 t}$
- Logistica:  $lgs(t) = x_3 / (1 + x_1 e^{-x_2 t})$

I primi due non hanno un bound superiore, le ultime due sì. Nell'esempio abbiamo gli ultimi 3 modelli ad essere candidati.

$$\min \left\{ \frac{1}{2} \sum_{j=1}^n [y_j - tp(x_1, x_2; t_j)]^2 : x_1, x_2 \in \mathbb{R} \right\}$$

$$tp(x_1, x_2, t) = x_1 t^{x_2}$$

Nella lezione avevamo gli elementi residuali scritti come

$$r_j(x_1, x_2) = y_j - x_1 t_j^{x_2}$$

Abbiamo 36 funzioni residuali candidate, e si può utilizzare Gauss Newton per risolvere il problema. `lsqnonlin` in Matlab. Da notare che nella teoria, come criterio d'arresto abbiamo

$$\nabla f(x^*) = 0$$

nella pratica si utilizza

- $\|\nabla f(x^k)\| \leq \text{tol}$
- $|f(x^{k-1}) - f(x^k)| \leq \text{tol}$

Secondo caso

$$\min \left\{ \frac{1}{2} \sum_{j=1}^n [y_j - \text{lmg}(x_1, x_2; t_j)]^2 : x_1, x_2, x_3 \in \mathbb{R} \right\}$$
$$tp(x_1, x_2, x_3, t) = x_3 - x_1 e^{-x_2 t}$$

Ultimo caso: funzione logistica

$$\min \left\{ \frac{1}{2} \sum_{j=1}^n [y_j - \text{lgs}(x_1, x_2; t_j)]^2 : x_1, x_2, x_3 \in \mathbb{R} \right\}$$
$$lgs(x_1, x_2, x_3, t) = x_3 / (1 + x_1 e^{-x_2 t})$$



## 12 Ottimizzazione vincolata con regione ammissibile convessa

Fino ad adesso abbiamo considerato il caso senza vincoli. Nel caso vincolato il problema diventa:

$$(P) \quad \min \{f(x) : x \in X\} \quad f : \mathbb{R}^n \rightarrow \mathbb{R} \quad X \subseteq \mathbb{R}^n$$

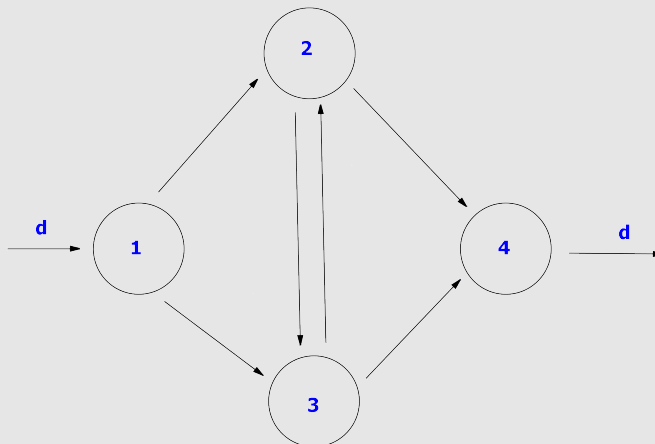
Sono 2 i casi che andremo ad analizzare

- Caso di un  $X$  generico
- Caso di un  $X$  esplicitamente descritto da disuguaglianze e uguaglianze. In questo caso la descrizione della regione ammissibile è la seguente:

$$X = \{x \in \mathbb{R}^n \mid g_i(x) \leq 0, i = 1 \dots n, h_j(x) = 0, j = 1 \dots p\}$$

### Esempio 12.1 (Reti di traffico)

Attraversamento di un arco al variare del flusso. Maggiore è la quantità di traffico attraverso un arco, maggiore è il tempo di percorrenza. Vogliamo minimizzare il tempo di percorrenza nella rete.



Un esempio di funzione che descrive la quantità di tempo necessaria a percorrere l'arco  $ij$  è la seguente:

$$t_{ij}(v) = t_{ij} + \alpha_{ij} \frac{v}{c_{ij} - v} \quad (v < c_{ij}) \quad \begin{cases} t_{ij}, \alpha_{ij} & \text{costanti relative al tipo di strada} \\ c_{ij} & \text{capacità in } ij \\ v & \text{volume di traffico attuale} \end{cases}$$

Un'altra funzione possibile per descrivere il tempo di attraversamento al variare del volume di traffico è

$$t_{ij}(v) = t_{ij} + \alpha_{ij} \frac{v^4}{c_{ij}^4}$$

dove  $c_{ij}$  non è la capacità! La funzione obiettivo da minimizzare è:

$$\min \sum_{(i,j)} x_{ij} t_{ij}(x_{ij})$$

Il problema questa volta è vincolato: abbiamo dei vincoli di conservazione del flusso ossia, la quantità di traffico che entra da un nodo deve essere uguale a quella che esce. La regione ammissibile è descritta da:

$$\begin{cases} x_{12} + x_{13} = d \\ x_{24} + x_{34} = d \\ x_{12} + x_{32} = x_{23} + x_{24} \\ x_{13} + x_{23} = x_{32} + x_{34} \\ x_{ij} \geq 0 \\ (x_{ij} \leq c_{ij}) \quad (\text{opzionale: capacità limitata su un arco}) \end{cases}$$

Il problema può essere descritto più in generale dalla seguente notazione basata su grafi:

$$\sum_{j \in BN(i)} x_{ji} - \sum_{j \in FN(i)} x_{ji} = \begin{cases} -d & i = 0 \\ 0 & i \neq 0, t \\ d & i = t \end{cases}$$

## 12.1 Condizioni di ottimalità con regione ammissibile convessa

Come facciamo a sapere se l'insieme è convesso?

$$X = \{x \in \mathbb{R}^n : g_i(x) \leq 0, h_j(x) = 0\}$$

$$i = 1 \dots n \quad j = 1 \dots p \quad g_i : \mathbb{R}^n \rightarrow \mathbb{R} \quad h_j : \mathbb{R}^n \rightarrow \mathbb{R}$$

Le funzioni  $g_i$  (vincoli di disuguaglianza) e  $h_j$  (vincoli di uguaglianza) definiscono i vincoli applicabili su  $\mathbb{R}^n$  per definire la regione ammissibile. La natura delle funzioni vincolari può garantire la convessità di  $X$ .

### Definizione 12.2 (Funzione affine)

Una funzione  $h_j$  si dice affine se

$$h_j(x) = a_j^T x + b_j \quad a_j \in \mathbb{R}^n, b_j \in \mathbb{R}$$

### Proposizione 12.1

Se le funzioni  $g_i$  sono convesse per  $i = 1 \dots n$  e le funzioni  $h_j$  sono affini per  $j = 1 \dots p$ , allora  $X$  è un insieme convesso

*Dimostrazione.* Si sfrutta la definizione di convessità. Si prendano  $x, y \in X$  e  $\lambda \in [0, 1]$ . Il segmento che unisce i due punti è dato da

$$\lambda x + (1 - \lambda)y$$

- Vincoli di disuguaglianza:

$$g_i(\lambda x + (1 - \lambda)y) \leq \underbrace{\lambda g_i(x)}_{\leq 0} + (1 - \lambda) \underbrace{g_i(y)}_{\leq 0} \leq 0$$



#### Nota

$g_i(x) \leq 0$  e  $g_i(y) \leq 0$  in quanto per ipotesi appartengono ad  $X$ : se non fosse così  $x$  ed  $y$  sarebbero fuori dalla regione ammissibile.



- Vincoli di uguaglianza:

$$\begin{aligned}
 h_j(\lambda x + (1 - \lambda)y) &= \\
 a_j^T(\lambda x + (1 - \lambda)y) + b_j &= \quad [ \text{ponendo } b_j = \lambda b_j + (1 - \lambda)b_j ] \\
 \lambda a_j^T x + (1 - \lambda)a_j^T y + \lambda b_j + (1 - \lambda)b_j &= \\
 \underbrace{\lambda a_j^T x}_{=0} + \underbrace{(1 - \lambda)a_j^T y}_{=0} + \lambda b_j + (1 - \lambda)b_j &= 0
 \end{aligned}$$

**Nota**

$g_i(x) = 0$  e  $g_i(y) = 0$  in quanto per ipotesi appartengono ad  $X$ .

□

Per avere un algoritmo, dobbiamo stabilire le condizioni di ottimalità: minimo locale e globale.

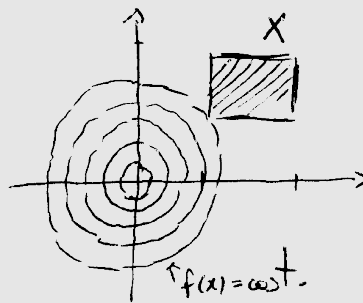
L'ipotesi che assumiamo è che  $f$  sia differenziabile, in modo che si possano utilizzare i gradienti e le matrici hessiane.

Le condizioni di ottimalità che abbiamo visto nel caso non vincolato, valgono anche nel caso vincolato?

**Esempio 12.3**

Vediamo in esempio in cui dimostriamo che le condizioni di ottimalità viste nel caso non vincolato non sono più sufficienti. Vogliamo minimizzare la funzione

$$f(x) = x_1^2 + x_2^2 \quad X = [1, 2] \times [0, 1]$$



Nel caso non vincolato il minimo è in  $(0, 0)$  ed il gradiente di  $f$  è

$$\nabla f(x) = 2x$$

Il gradiente si annulla nel caso del vettore  $x = 0$ , ma  $x \notin X$ .

Questo esempio quindi dimostra che questa non può essere la condizione di ottimalità: bisogna cercare una differente nozione di punto stazionario.

**Definizione 12.4 (Direzione ammissibile)**

Un vettore  $d \in \mathbb{R}^n$  si dice direzione ammissibile per  $X$  a partire da  $\bar{x} \in X$  se  $\exists \bar{t} > 0$  tale che

$$x(t) = \bar{x} + td \in X \quad \forall t \in [0, \bar{t}]$$

Se  $d$  è una direzione ammissibile, anche i suoi multipli sono ancora una direzione ammissibile.

L'insieme in cui appartengono anche i suoi multipli si chiama *cono*.

**Definizione 12.5 (Cono delle direzioni ammissibili)**

$$F(X, \bar{x}) = \{d \in \mathbb{R}^n : \exists \bar{t} > 0 \text{ tale che } \bar{x} + td \in X \quad \forall t \in [0, \bar{t}]\}$$

 **Osservazione 12.6**

Se  $d$  appartiene alla regione ammissibile, anche tutti i suoi multipli ci appartengono.

 **Osservazione 12.7**

Se  $\bar{x}$  è un punto interno a  $X$ , allora

$$F(X, \bar{x}) = \mathbb{R}^n$$

L'idea quindi è quella di testare l'ottimalità lungo tutte le direzioni ammissibili.

 **Teorema 12.8 (Condizione necessaria)**

Sia  $\bar{x} \in X$  un punto di minimo locale per  $(P)$ . Allora

$$\nabla f(\bar{x})^T d \geq 0 \quad \forall d \in F(X, \bar{x}) \quad (CN_F)$$

Che è detta condizione di stazionarietà. ( $F(X, \bar{x})$  sono le direzioni ammissibili)

*Dimostrazione.* Sia  $\epsilon > 0$  tale che  $f(\bar{x}) = \min\{f(x) : x \in X \cap B(\bar{x}, \epsilon)\}$  e sia  $d \in F(X, \bar{x})$  con  $\bar{t} > 0$  fornito dalla definizione. Considerando  $t \leq \min\{\frac{\epsilon}{\|d\|_2}, \bar{t}\}$ , risulta  $\bar{x} + td \in X \cap B(\bar{x}, \epsilon)$  e quindi  $f(\bar{x} + td) - f(\bar{x}) \geq 0$ . A questo punto si procede come nel caso non vincolato, utilizzando lo sviluppo di Taylor

$$0 \leq \frac{f(\bar{x} + td) - f(\bar{x})}{t} = \nabla f(\bar{x})^T d + \frac{r(td)}{t} \xrightarrow{t \downarrow 0} \nabla f(\bar{x})^T d$$

e quindi  $\nabla f(\bar{x})^T d \geq 0$

□

Il teorema contiene al suo interno come caso particolare l'ottimizzazione non vincolata, infatti:

$$X = \mathbb{R}^n \quad \Rightarrow \quad F(X, \bar{x}) = \mathbb{R}^n$$

$$\bar{x} \text{ punto min. loc.} \quad \Leftrightarrow \quad \nabla f(\bar{x}) = 0 \quad \Rightarrow \quad \nabla f(\bar{x})^T d \geq 0$$

Se un punto è stazionario per  $f$  ed appartiene alla regione ammissibile, soddisfa la condizione necessaria di ottimalità.

 **Osservazione 12.9**

1. Se  $\nabla f(\bar{x}) = 0$ , allora  $(CN_F)$  è verificata

2. Se  $x \in \text{int}X$  è un punto di minimo locale di  $(P)$ , allora  $\nabla f(\bar{x}) = 0$

**Proposizione 12.2**

Sia  $X$  un insieme convesso,  $\bar{x} \in X$ . Allora

1.  $x - \bar{x}$  è una direzione ammissibile per  $X$  per ogni  $x \in X$ .
2. Se  $d \in F(X, \bar{x})$ , allora esistono  $\lambda \geq 0$ ,  $x \in X$  tali che  $d = \lambda(x - \bar{x})$

Da 1 e 2 segue che il cono delle direzioni ammissibili è

$$F(X, \bar{x}) = \{\lambda(x - \bar{x}) : \lambda \geq 0, x \in X\}$$

*Dimostrazione.*

1. Sia  $t \in [0, 1]$  :  $\bar{x} + t(x - \bar{x}) = (1 - t)\bar{x} + tx \in X$  poichè  $X$  è convesso. Allora  $(x - \bar{x}) \in F(X, \bar{x})$
2. Sia  $t > 0$  tale che  $\bar{x} + td \in X$ . Posto  $x = \bar{x} + td \in X$ , risulta  $d = \frac{1}{t}(x - \bar{x})$

□

Segue da questa caratterizzazione del teorema 1 una riscrittura

 **Teorema 12.10 (Condizione necessaria (2))**

Sia  $X$  convesso,  $\bar{x} \in X$ . Se  $\bar{x} \in X$  è un punto di minimo locale per  $(P)$ , allora

$$\nabla f(\bar{x})^T (x - \bar{x}) \geq 0 \quad \forall x \in X \quad (CN_X)$$

Viceversa, se  $f$  è convessa,  $\bar{x} \in X$ , e vale la condizione di stazionarietà  $(CN_X)$ , allora  $\bar{x}$  è un punto di minimo (globale) di  $(P)$ .

$(CN_X)$  è una condizione necessaria, se  $f$  è convessa diventa anche sufficiente, questo perchè minimo locale e globale coincidono.

*Dimostrazione.* La prima parte viene da (12.8) (12.2.2), la seconda parte va dimostrata.

$x \in X$ . La funzione  $f$  è convessa: se  $f$  è differenziabile e convessa vale (caratterizzazione delle funzioni convesse)

$$f(x) \underset{f \text{ convessa}}{\geq} f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) \underset{CN_x}{\geq} f(\bar{x})$$

□

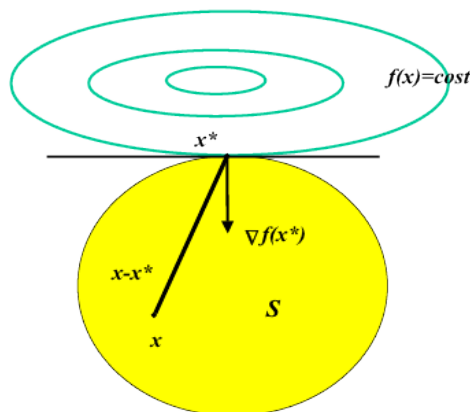



Figura 12.1: Significato geometrico della condizione necessaria di ottimalità

 **Work in progress**

Digressione: compagnie telefoniche (minimi quadrati). Gauss Newton

$$J_k^T J_k + \sum_{j=1}^n r_j(x^k) \nabla^2 r_j(x^k)$$

## 12.2 Metodi risolutivi

Ora che abbiamo il concetto di stazionarietà con vincolo, possiamo ridefinire il punto stazionario.

### Definizione 12.11 (Punto stazionario)

Un punto  $\bar{x} \in X$  si dice stazionario se

$$\nabla f(\bar{x})^T d \geq 0 \quad \forall d \in F(X, \bar{x})$$

L'ipotesi che useremo sempre è che  $X$  sia *convesso*, con la semplificazione 12.2.

Dobbiamo muoverci per punti ammissibili, quindi direzioni ammissibili e di discesa. Quindi:

- Direzione di discesa  $\nabla f(x^k)^T d^k < 0$
- Direzione ammissibile  $d^k \in F(X, x^k)$

### 12.2.1 Metodi delle direzioni ammissibili

Una assunzione che facciamo, oltre alla convessità della regione ammissibile, è che da una direzione, muovendosi di passo unitario, si rimanga nella regione ammissibile.

$$x^k + t_k d^k \in X$$

Questa assunzione non è restrittiva in quanto se con passo unitario si uscisse dalla regione ammissibile, si può sempre scalare accorciando la direzione stessa moltiplicandola per un coefficiente minore di 1.

Questo consente di restringere la ricerca del passo  $t_k \in (0, 1]$ .

Una possibile ricerca del passo di ricerca può essere quella di prendere un passo che soddisfi la condizione di Armijo, la quale ci permette di forzare il valore della funzione obiettivo di una certa quantità nel nuovo punto.

### Definizione 12.12 (Condizione di Armijo (AJO))

$$f(x^k + t d^k) \leq f(x^k) + c_1 t \nabla f(x^k)^T d^k \quad c_1 \in (0, 1)$$

 **Osservazione 12.13**

$t_k \in (0, 1]$  va bene perché

$$X \text{ convesso} \Rightarrow x^k + t_k d^k = (1 - t_k)x^k + t_k(x^k + d^k) \in X$$

Poichè  $\nabla f(x^k)^T d^k$  è negativo, stiamo forzando il valore della funzione obiettivo nel nuovo punto a diminuire di una quantità pari a  $c_1 t \nabla f(x^k)^T d^k$ .

Sappiamo anche che  $\exists \tau \geq 0$  tale che vale (AJO)  $\forall t \in (0, \tau]$

$$(\gamma_k(t) = f(x^k + td^k) \Rightarrow \gamma'_k(0) = \nabla f(x^k)^T d^k < 0)$$

*Dimostrazione.* Consideriamo la funzione di ricerca monodimensionale funzione della sola  $t$

$$\begin{aligned} \varphi(t) &= f(x^k + td^k) \\ \varphi'(t) &= \nabla f(x^k + td^k) d^k \\ \varphi'(0) &= \nabla f(x^k) d^k < 0 \end{aligned}$$

$$0 > \nabla f(x^k) d^k = \varphi'(0) = \lim_{t \rightarrow \infty} \frac{f(x^k + td^k) - f(x^k)}{t}$$

Visto che la derivata in zero è negativa allora vale la condizione (AJO).  $\square$

### Procedura di Armijo

1. Scegliere  $\gamma \in (0, 1)$ ;  $t = 1$
2. Se  $t$  soddisfa (AJO), STOP, ( $t_k = t$ )
3.  $t = \gamma t$  e ritornare a 2.

Alla fine risulta che  $t = \gamma^S$  con S numero di iterazioni effettuate. Quindi  $\gamma^{S-1}$  non soddisfa la condizione (AJO).

Quello che ci servirà sarà (nella dimostrazione della convergenza?) appunto che  $\frac{t_k}{\gamma}$  non sarà uno spostamento corretto.

Riassumiamo la procedura totale

### Procedura per Metodi delle direzioni ammissibili

1. Scegliere  $x^0 \in X$ ;  $k = 0$
2. Se  $\nabla f(x^k)^T d \geq 0 \quad \forall d \in F(X, x^k)$ , allora STOP
3. Scegliere  $d^k \in F(X, x^k)$  tale che  $\nabla f(x^k)^T d^k < 0$  e  $x^k + d^k \in X$
4. Calcolare  $t_k \in (0, 1]$  che soddisfa (AJO) tramite Procedura di Armijo
5.  $x^{k+1} = x^k + t_k d^k$
6.  $k = k + 1$  e ritornare a 2)



#### Osservazione 12.14

Come caso particolare in cui  $X = \mathbb{R}^n$  ritorniamo ai metodi del gradiente visti per l'ottimizzazione non vincolata.

Questa procedura descrive un insieme di metodi che poi si differenziano in

- come verificare la condizione di stazionarietà (ora andrebbe testata su infinite di direzioni)
- come scegliere una direzione ammissibile che sia anche di discesa.

 **Proprietà 12.1**

Supponiamo che  $X$  sia compatto. Allora

$$\lim_{k \rightarrow +\infty} \nabla f(x^k)^T d^k = 0$$

Notare che la condizione di compattezza esclude il caso non vincolato.

*Dimostrazione.* Se vale la condizione di Armijo su  $t_k$  allora

$$\begin{aligned} f(x^k) - f(x^k + t_k d^k) &\geq -c_1 t_k \nabla f(x^k)^T d^k \\ f(x^k) - f(x^{k+1}) &\geq -c_1 t_k \underbrace{\nabla f(x^k)^T d^k}_{\text{negativo}} \\ f(x^k) - f(x^{k+1}) &\geq c_1 t_k |\nabla f(x^k)^T d^k| \end{aligned}$$

$X$  è compatto, quindi  $f$  su  $X$  ammette minimo, quindi  $\{f(x^k)\}$  è una successione decrescente, limitata inferiormente, che converge. Passando al limite  $f(x^k)$  e  $f(x^{k+1})$  convergono allo stesso valore, da cui:

$$\lim_{k \rightarrow +\infty} f(x^k) - f(x^{k+1}) = 0$$

Da cui deduciamo che:

$$\lim_{k \rightarrow +\infty} t_k |\nabla f(x^k)^T d^k| = 0$$

quindi una delle due componenti deve essere nulla.

Supponiamo per assurdo che la seconda componente sia diversa da zero:

$$\lim_{k \rightarrow +\infty} \inf |\nabla f(x^k)^T d^k| \geq \eta > 0 \quad \Rightarrow \quad t_k \rightarrow 0$$

Abbiamo preso il minimo limite, dato che il limite potrebbe non esistere.

Adesso sfruttiamo la compattezza di  $X$ : una successione dentro un compatto ammette una sottosuccessione convergente.

Inoltre

$$\{x^k\} \subseteq X, \quad x^k + d^k \in X \quad \Rightarrow \quad \{d^k\} \text{ è limitata}$$

Quindi entrambe hanno una sottosuccessione convergente:

$$x^k \rightarrow \bar{x} \quad d^k \rightarrow \bar{d}$$

(Abbiamo ommesso gli indici delle sottosuccessioni).

Il prodotto scalare

$$\nabla f(x^k)^T d^k \rightarrow \nabla f(\bar{x})^T \bar{d} \leq -\eta < 0$$

(In quest'ultimo passaggio abbiamo usato la continuità del gradiente e del prodotto scalare).

Importante risultato è che il prodotto scalare fra il gradiente e la direzione calcolato nei punti limite è negativo.

Adesso sfruttiamo la ricerca della condizione di Armijo. Se  $t_k$  soddisfa (AJO) il passo  $\frac{t_k}{\gamma}$  non può soddisfare.

Deve essere:

$$f(x^k + \frac{t_k}{\gamma} d^k) - f(x^k) > c_1 \frac{t_k}{\gamma} \nabla f(x^k)^T d^k$$

Per il teorema del valor medio:

$$f(x^k + \frac{t_k}{\gamma} d^k) - f(x^k) = \nabla f(x^k + \tau_k d^k)^T \frac{t_k}{\gamma} d^k \quad \tau_k \in [0, \frac{t_k}{\gamma}]$$

Riprendendo la disuguaglianza che esprime la non soddisfacibilità di (AJO)

$$\frac{t_k}{\gamma} \nabla f(x^k + \tau_k d^k)^T d^k > \frac{t_k}{\gamma} c_1 \nabla f(x^k)^T d^k$$

Quindi

$$\nabla f(x^k + \tau_k d^k)^T d^k > c_1 \nabla f(x^k)^T d^k$$

Passando ai limiti:

$$\nabla f(\bar{x})^T \bar{d} \geq c_1 \nabla f(\bar{x})^T \bar{d}$$

Quindi deve essere

$$(1 - c_1)\nabla f(\bar{x})^T \geq 0$$

Questo può accadere solo se  $c_1 \geq 1$ , ma la condizione che utilizzavamo è che

$$0 < c_1 < 1$$

quindi abbiamo un *assurdo* e deduciamo che

$$|\nabla f(x^k)^T d^k| = 0$$

□



### Osservazione 12.15

La ricerca di Armijo può essere sostituita dalla Minimizzazione limitata con passo  $t \in [0, 1]$ .

$t_k \in \operatorname{argmin}\{f(x^k + td^k) : t \in [0, 1]\}$  trovato grazie alla ricerca esatta.

Si può riprendere lo stesso dimostrazione del teorema e si usa il  $t_k$  di Armijo.

I punti oscuri nel metodo delle direzioni ammissibili sono:

- Trovare punto ammissibile
- Trovare direzione ammissibile
- Verificare la condizione di stazionarietà

### 12.2.2 Algoritmo di Frank-Wolfe (Gradiente condizionato)

Era stato pensato per le funzioni quadratiche, ma in realtà funziona anche sulle funzioni non quadratiche. È un metodo delle direzioni ammissibili cui si impone direzione

$$d^k = \bar{x}^k - x^k \quad \bar{x}^k \in \operatorname{argmin}\{\nabla f(x^k)(x - x^k) : x \in X\}$$

La verifica della stazionarietà è immediata:

- se  $\nabla f(x^k)^T(\bar{x}^k - x^k) = 0$  (in particolare se  $\bar{x}^k = x^k$ ), allora  $x^k$  è un punto stazionario.
- in caso contrario  $\nabla f(x^k)^T d^k < 0$  e  $x^k + d^k = \bar{x}^k \in X$ , cioè  $d^k$  è la direzione cercata.



### Nota

Si ricordi

$$X \text{ convesso} \Rightarrow F(X, x^k) = \{\lambda(x - x^k) : x \in X, \lambda \geq 0\}$$

Tuttavia, verificare la condizione di stazionarietà è un problema di ottimizzazione non vincolata, se pure più semplice di quello di partenza!

Tuttavia la funzione obiettivo di questo problema risulta *lineare*, infatti poiché

$$\nabla f(x^k)^T(x - x^k) = \underbrace{\nabla f(x^k)^T}_{\text{costante}} x - \underbrace{\nabla f(x^k)^T x^k}_{\text{costante}}$$

risulta

$$\operatorname{argmin}\{\nabla f(x^k)(x - x^k) : x \in X\} \equiv \operatorname{argmin}\{\nabla f(x^k)x : x \in X\}$$

Inoltre se  $X$  è un *poliedro* abbiamo un problema di *programmazione lineare*.

Notare che in questo modo risolviamo sia il passo 2) che il passo 3) dell'algoritmo.


**Teorema 12.16 (Convergenza)**

Supponiamo che  $X$  sia compatto. Allora ogni punto di accumulazione di  $\{x^k\}$  è un punto stazionario di  $(P)$

*Dimostrazione.* Se  $x$  è un punto di accumulazione allora  $x^k \rightarrow x^*$  e dato che siamo in un compatto  $x^k + d^k \in X$  allora  $d^k \rightarrow d^*$ .

Per come abbiamo calcolato  $d_k$  deve essere

$$\nabla f(x^k)^T d^k \leq \nabla f(x^k)^T (x - x^k) \quad \forall x \in X$$

Passiamo al limite membro a membro

$$\nabla f(x^*)^T d^* \leq \nabla f(x^*)^T (x - x^*) \quad \forall x \in X$$

Ma per la proposizione 12.1  $\nabla f(x^*)^T d^* = 0$ , che è la condizione di stazionarietà di  $x^*$ .  $\square$


**Nota**

Dall'esempio visto al calcolatore, si vede che la direzione calcolata andrà dal punto di partenza fino a un vertice del poliedro. Da qui prenderò un punto che sta su questo segmento.

Se il numero dei vertici sono pochi, si ha una convergenza lenta dovuta all'andamento zig-zag fra i vertici.

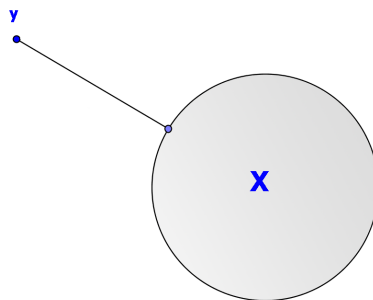
### 12.2.3 Metodi del gradiente proiettato

#### Proiezione di un insieme convesso

$$X \subseteq \mathbb{R}^n \text{ convesso, } y \in \mathbb{R}^n$$

La proiezione del punto  $y$  su  $X$  è:

- se  $y \in X$  la proiezione è  $y$  stesso.
- se  $y \notin X$  la proiezione è il punto più vicino a  $y$  che appartiene a  $X$ .



Per trovare la proiezione dobbiamo risolvere il seguente problema di ottimizzazione:

$$(P_{RX}) \quad \min\{\|y - x\|_2 : x \in X\}$$

Se prendo il quadrato della norma e aggiungo un coefficiente davanti il problema rimane il medesimo:

$$(P_{RX}) \quad \min \left\{ \frac{1}{2} \|y - x\|_2^2 : x \in X \right\}$$



Analizziamo la funzione obiettivo, il gradiente e la matrice hessiana

$$f(x) = \frac{1}{2} \|y - x\|_2^2 = \frac{1}{2} \sum_{i=1}^n (y_i - x_i)^2$$

$$\nabla f(x) = x - y$$

$$\nabla^2 f(x) = I \quad (\text{identità})$$

La matrice hessiana è identica, da questo segue che la funzione è *strettamente* convessa.

Dato che  $X$  è convesso e  $f$  è strettamente convessa esiste un *unico* minimo: la proiezione è unica.

### Definizione 12.17 (Proiezione di $y$ su $X$ )

$P_X : \mathbb{R}^n \rightarrow \mathbb{R}^n$

$$P_X(y) := \operatorname{argmin} \left\{ \frac{1}{2} \|x - y\|_2^2 : x \in X \right\}$$



### Teorema 12.18 (Caratterizzazione delle proiezioni)

Sia  $y \in \mathbb{R}^n$  fissato. Allora

1.  $x^* = P_X(y) \iff (y - x^*)^T (x - x^*) \leq 0 \quad \forall x \in X$
2.  $\|P_X(v) - P_X(z)\|_2 \leq \|v - z\|_2 \quad \forall v, z \in \mathbb{R}^n$



### Osservazione 12.19

La caratterizzazione 1) sta ad indicare che il vettore  $x - p(x)$  deve formare un angolo maggiore o eguale a  $\pi/2$  con ogni vettore  $y - p(x)$ , al variare di  $y$  in  $S$ .

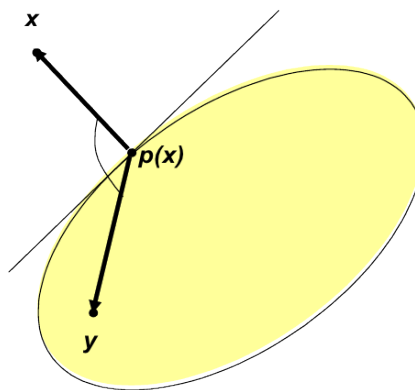


Figura 12.2: Significato geometrico della caratterizzazione 1)



### Osservazione 12.20

Da 2) segue che  $\lim_{v \rightarrow z} P_X(v) = P_X(z)$ , ossia  $P_X$  è continua.

Una funzione così fatta viene anche chiamata *non espansiva*, cioè una funzione così fatta al massimo “accorcia” le distanze fra i punti.

*Dimostrazione.*

1.

$$x^* = P_X(y) \iff x^* \text{ è punto di minimo di } (P_{RX})$$

Quindi dalla definizione di minimo:

$$\nabla f(x^*)^T (x - x^*) \geq 0 \quad \forall x \in X$$

Andando a sostituire il valore del gradiente precedentemente calcolato

$$\begin{aligned} (x^* - y)^T (x - x^*) &\geq 0 \quad \forall x \in X \\ (y - x^*)^T (x - x^*) &\leq 0 \quad \forall x \in X \end{aligned}$$

2. indichiamo con  $v^* = P_X(v)$  e  $z^* = P_X(x)$ .  
Sfruttando la proprietà 1)

$$\begin{aligned} (v - v^*)^T (x - v^*) &\leq 0 \\ (z - z^*)^T (x - z^*) &\leq 0 \end{aligned}$$

Le disuguaglianza sopra valgono  $\forall x \in X$  quindi in particolare valgono per  $x = z^*$  e  $x = v^*$ .

$$\begin{aligned} (v - v^*)^T (z^* - v^*) &\leq 0 \\ (z - z^*)^T (v^* - z^*) &\leq 0 \end{aligned}$$

Portiamo fuori il meno nella seconda

$$\begin{aligned} (v - v^*)^T (z^* - v^*) &\leq 0 \\ -(z - z^*)^T (z^* - v^*) &\leq 0 \end{aligned}$$

Sommiamo membro a membro (distributiva del prodotto scalare)

$$\begin{aligned} (v - v^* - z + z^*)^T (z^* - v^*) &\leq 0 \\ (v - z)^T (z^* - v^*) + \underbrace{(z^* - v^*)^T (z^* - v^*)}_{\|z^* - v^*\|_2^2} &\leq 0 \\ \|z^* - v^*\|_2^2 + (v - z)^T (z^* - v^*) &\leq 0 \\ \|v^* - z^*\|_2^2 \leq (v - z)^T (v^* - z^*) &\leq \|v - z\|_2 \|v^* - z^*\|_2 \end{aligned}$$

\*) per la disuguaglianza di Schwartz

□

Abbiamo introdotto le proiezioni e analizzato le loro proprietà allo scopo di utilizzare per l'ottimizzazione vincolata, un metodo non vincolato, e nel caso si esca dalla regione ammissibile, sia possibile rientrare dentro con una proiezione.

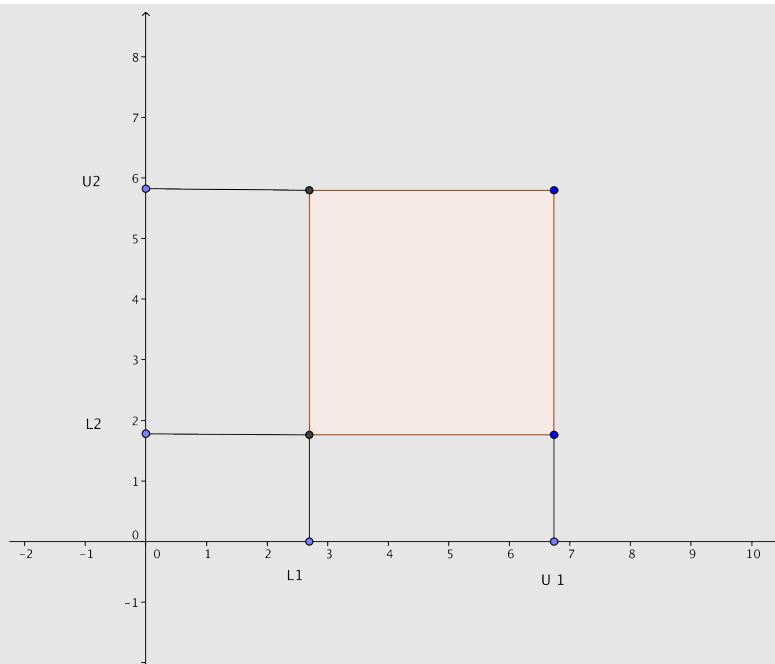
Come abbiamo visto calcolare una proiezione è a sua volta un problema di ottimizzazione vincolata tuttavia esistono proiezioni che si riescono a calcolare facilmente, come nei seguenti casi.

### Esempio 12.21 (Vincoli di scatola)

$$X \in \{x \in \mathbb{R}^n : l_i \leq x_i \leq \mu_i\} \quad i = 1 \dots n$$

Ogni componente  $x_i$  è può variare fra un limite inferiore e un limite superiore.

3 casi: lati esterni, lati superiori, vertici..

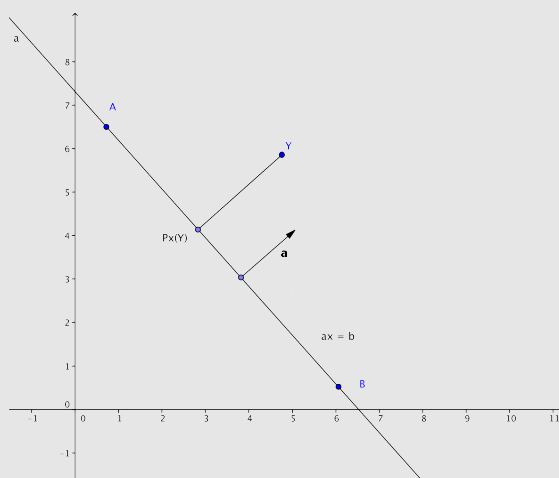


$$[P_X(y)]_i = \begin{cases} \mu_i & \text{se } y_i \geq \mu_i \\ l_i & \text{se } y_i \leq l_i \\ y_i & \text{se } l_i \leq y_i \leq u_i \end{cases}$$

$$\begin{aligned} (y - P_X(y))^T (x - P_X(y)) &= \\ \sum_{i=1}^n (y_i - [P_X(y)]_i)(x_i - [P_X(y)]_i) &= \\ \sum_{i \in I_{>}} (y_i - \mu_i)(x_i - \mu_i) + \sum_{i \in I_{<}} (y_i - l_i)(x_i - l_i) \quad \forall x \in X \quad (I_{>} = \{i \mid y_i > \mu_i\}, I_{<} = \{i \mid y_i < l_i\}) \end{aligned}$$

**Esempio 12.22 (Unico vincolo lineare)**

$$X_1 = \{x \in \mathbb{R}^n : a^T x = b\}$$



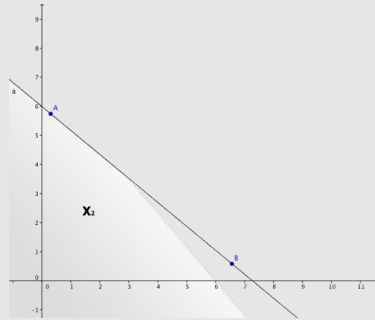
Vettore  $a$  ortogonale alla retta. Mi sposto da  $x$  in una direzione  $\pm a$ .

$$P_{X_1}(y) = y - \lambda_1 a \quad \lambda_1 = \frac{a^T y - b}{a^T a} \quad \lambda_2 = \max\{0, \lambda_1\} = \frac{1}{a^T a} \max\{0, a^T x - b\}$$

$$\begin{aligned}
 & y - [y - \lambda_i a]^T (x - [y - \lambda_i a]) = \\
 & \lambda_i a^T (x - y + \lambda_i a) = \\
 & \lambda_i (a^T x - a^T y) + \lambda_i^2 a^T a = \begin{cases} = \frac{(b - a^T y)(a^T y - b)}{a^T a} + \frac{(a^T y - b)^2}{a^T a} = 0 & \forall x \in X_1 \\ \leq \lambda_1 [(b - a^T y - b) + (a^T y - b)] = 0 & \forall x \in X_2 \end{cases}
 \end{aligned}$$

Nel secondo caso lo spostamento è

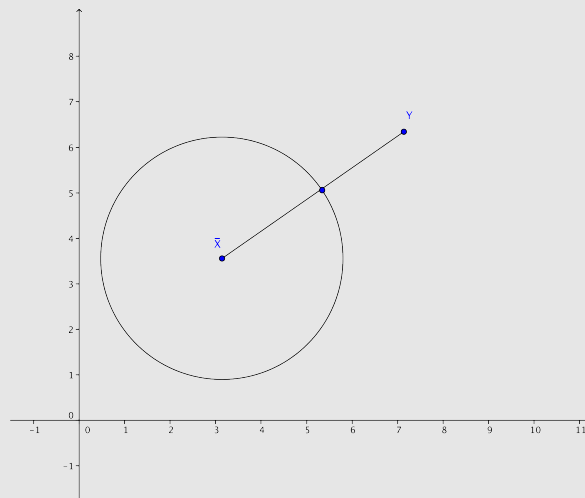
$$X_2 = \{x \in \mathbb{R}^n : a^T x \leq b\}$$



Se sono dentro la regione ammissibile, lo spostamento è 0. Altrimenti se sono fuori scelgo  $\lambda_1$  che sarà  $> 0$  dato che  $a^T y > 0$  Tutto questo discorso vale solo con un vincolo lineare.

**Esempio 12.23 (Sfera)**

$$X = \{x \in \mathbb{R}^n : \|x - \bar{x}\|_2 \leq r\} = B(\bar{x}, r)$$



La proiezione è verso il centro. Se sono dentro non ci si deve spostare.

$$P_X(y) = \bar{x} + \underbrace{(\min\{\|y - \bar{x}\|_2, r\})}_{\lambda} \frac{y - \bar{x}}{\|y - \bar{x}\|_2}$$

Tramite una translazione dall'origine possiamo supporre  $\bar{x} = 0$

$$\begin{aligned}
 \|y\|_2 \left(y - \frac{\lambda y}{\|y\|_2}\right)^T \left(x - \frac{\lambda y}{\|y\|_2}\right) &= (\|y\|_2 - \lambda)(y^T x - \lambda \|y\|_2) \stackrel{\leq}{\leq} \\
 &\leq (\|y\|_2 - \lambda)(\|y\|_2 \|x\|_2 - \lambda \|y\|_2) = \|y\|_2 (\|x\|_2 - \lambda) \stackrel{\leq}{\leq} 0 \quad \text{Schwarz} \\
 & \quad \quad \quad (*)
 \end{aligned}$$

(\*)  $\lambda < r \Rightarrow \lambda = \|y\|_2$  ;  $\lambda = r \Rightarrow \|y\|_2 \geq \lambda$  e  $\|x\|_2 \leq \lambda$

### 12.2.4 Metodo del gradiente proiettato

Il metodo consiste nell'applicare il metodo del gradiente e, nel caso si esca dalla regione ammissibile, di proiettarlo su  $X$ .

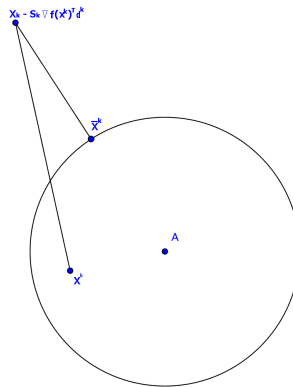
$$(P) \quad \min\{f(x) : x \in X\} \quad X \subseteq \mathbb{R}^n \text{ convesso, } f : \mathbb{R}^n \rightarrow \mathbb{R} \text{ differenziabile con continuità}$$

È un metodo delle direzioni ammissibili in cui  $d^k = \bar{x}^k - x^k$

$$x^{k+1} = x^k + t_k(\bar{x}^k - x^k)$$

$$\bar{x}^k = P_X(x^k - s_k \nabla f(x^k)) \quad \text{con } s_k > 0 \text{ il passo}$$

Poiché  $\bar{x}^k \in X$ , abbiamo  $x^k + d^k = \bar{x}^k \in X$ .



Se  $x^k \neq \bar{x}^k$  (cioè  $d^k \neq 0$ ), allora  $d^k$  è una direzione di discesa.



#### Teorema 12.24

$d^k = \bar{x}^k - x^k$  è una direzione di discesa (in quanto  $\nabla f(x^k)^T d^k < 0$ ).

*Dimostrazione.* Poiché  $\bar{x}^k$  è una proiezione, dalla caratterizzazione 12.18.(1) con  $x = x^k$  abbiamo

$$0 \underset{12.18}{\geq} (x^k - s_k \nabla f(x^k) - \bar{x}^k)^T (x^k - \bar{x}^k) = s_k \nabla f(x^k)^T (\bar{x}^k - x^k) + (x^k - \bar{x}^k)^T (x^k - \bar{x}^k) = s_k \nabla f(x^k)^T d^k + \|x^k - \bar{x}^k\|_2^2$$

da cui

$$\nabla f(x^k)^T d^k \leq -\frac{1}{s_k} \|x^k - \bar{x}^k\|_2^2 < 0$$

□

Se invece  $\bar{x}^k = x^k$  (cioè  $d^k = 0$ ), allora  $x^k$  è un punto stazionario. Infatti:

#### Proposizione 12.3

Sia  $s > 0$  fissato. Allora

$$x^* \in X \text{ è un punto stazionario di } (P) \iff P_X(x^* - s \nabla f(x^*)) = x^*$$

*Dimostrazione.*

$$x^* = P_X(x^* - s\nabla f(x^*)) \iff_{\text{teodiprima}} (x^* - s\nabla f(x^*)) - x^* \perp (x - x^*) \leq 0 \forall x \in X \iff_{12.18.(1)} \underbrace{s\nabla f(x^*)^T (x - x^*)}_{d^k} \geq 0 \forall x \in X$$

□



**Teorema 12.25 (Convergenza)**

Supponiamo che  $X$  sia compatto. Se  $s_k \in [s, s']$  per ogni  $k$  per qualche  $s, s' > 0$ , allora ogni punto di accumulazione  $x^*$  di  $\{x^k\}$  è un punto stazionario di  $(P)$ .

*Dimostrazione.* Dato che  $x^*$  è un punto di accumulazione allora esiste una sottosuccessione di  $x^k \rightarrow x^*$ . Sfruttando la trasformazione usata in una dimostrazione precedente:

$$\|x^k - x^k\|_2 = \|P_X(x^k - s\nabla f(x^k)) - x^k\|_2 \underset{\text{teo12.18}}{\leq} \|x^k - s\nabla f(x^k) - x^k\|_2 \xrightarrow{k \rightarrow \infty 12.1} 0$$

possiamo fare il limite perché la proiezione è continua, quindi

$$\|P_X(x^* - s\nabla f(x^*)) - x^*\|_2 = 0$$

da cui

$$x^* = P_X(x^* - s\nabla f(x^*))$$

e quindi  $x^*$  è stazionario per 12.3.

Abbiamo posto  $s^k = s$  per evitare che andasse all'infinito, basta imporre che sia limitato.

□

1. Scegliere  $x^0 \in X; s > 0; k = 0$
2. Calcolare  $x^{k+1} = P_X(x^k - s\nabla f(x^k))$
3. Se  $x^{k+1} = x^k$ , allora STOP
4.  $k = k + 1$  e ritornare a 2

**Cenni sull'originale metodo del gradiente proiettato**

Si può prendere  $t_k$  identicamente 1. (versione originale dell'algorithmo)

$$(x^{k+1} = x^k = P_X(x^k - s_k \nabla f(x^k)))$$

Ma non è detto che valga la condizione di Armijo, non rientra quindi nell'insieme dei metodi delle direzioni ammissibili e quello che si può dimostrare che otteniamo punti stazionari sotto questa condizione:

- $X$  compatto
- $\nabla f$  lipschitziana  $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$
- $0 < s \leq 2/L$

Risultato che non dimostriamo: questo metodo richiede la conoscenza di  $L$ , e non è banale.

# 13 Ottimizzazione vincolata con regione ammissibile non necessariamente convessa

## 13.1 Condizioni di ottimalità

### 13.1.1 Eliminazione della condizione di convessità

Vogliamo eliminare la condizione di convessità. Il primo passo è ottenere delle nuove condizioni di ottimalità, per poi sfruttare il fatto che la regione ammissibile è espressa tramite disuguaglianze e uguaglianze. Ci troviamo quindi a voler risolvere il solito problema di minimizzazione in un nuovo contesto

$$f : \mathbb{R}^n \rightarrow \mathbb{R} \text{ differenziabile, } X \subseteq \mathbb{R}^n \text{ (non necessariamente convesso)}$$
$$(P) \quad \min\{f(x) : x \in X\}$$

#### Test di ottimalità locale

Il cono delle direzioni ammissibili permette di stare nella regione ammissibile per spostamenti piccoli. Ma muoversi lungo delle direzioni non è equivalente a muoversi a una curva (una successione di punti)! Nel cono delle direzioni ammissibili ci muove per rette (la direzione è fissata), dobbiamo quindi cambiare strategia.

Per testare l'ottimalità locale di  $\bar{x} \in X$  dobbiamo verificare in qualche modo che:

$$f(\bar{x}) \leq f(x_n)$$

per  $n$  sufficientemente grande per ogni successione  $\{x_n\} \subseteq X$  tale che  $x_n \rightarrow \bar{x}$ , con  $x_n \neq \bar{x}$ ,  $x_n = \bar{x} + t_n d_n$ , con

$$t_n \rightarrow 0^+$$

$$t_n = \|x_n - \bar{x}\|_2$$

$$d_n = \frac{x_n - \bar{x}}{\|x_n - \bar{x}\|_2}$$



#### Osservazione 13.1

*I  $d_n$  hanno norma 1, quindi stanno in un compatto e quindi esiste una sottosuccessione convergente. Formalizzando meglio:*

$$d_0 \in B(0, 1) \Rightarrow \exists d \in B(0, 1) \text{ t.c. } d_n \rightarrow d$$

#### Definizione 13.2 (Successione ammissibile)

$\{x_n\}$  si dice successione ammissibile per  $\bar{x} \in X$  se

$$\begin{cases} x_n \in X, x_n \neq \bar{x} & \forall n \\ x_n \rightarrow \bar{x} \end{cases}$$

**Definizione 13.3 (Direzione limite)**

$d \in \mathbb{R}^n$  si dice direzione limite per  $X$  in  $\bar{x}$  se esiste una successione ammissibile per  $\bar{x} \in X$  tale che

$$\frac{x_n - \bar{x}}{\|x_n - \bar{x}\|_2} \rightarrow d$$

(Nota:  $\|d\|_2 = 1$ )

$d_n$  è una direzione limite

Le direzioni limite testano l'ottimalità.

**Definizione 13.4 (Cono tangente (di Bouligand))**

$$T(X, \bar{x}) = \{d \in \mathbb{R}^n \mid \exists t_n \rightarrow 0^+, \exists d_n \rightarrow d \text{ t.c. } \bar{x} + t_n d_n \in X\}$$

Abbiamo tutti le direzioni possibili: il cono tangente contiene le direzioni limite e i loro multipli, cioè il cono generato dalle direzioni limite. Mentre cono delle direzioni ammissibili ci faceva muovere solo lungo delle semirette, qui possiamo muoverci per curve. Vediamone alcune proprietà.

 **Proprietà 13.1 (Proprietà del cono tangente)**

- $\bar{x} \in \text{int } X \Rightarrow T(X, \bar{x}) = \mathbb{R}^n$
- $T(X, \bar{x})$  è un cono chiuso. ( $F$  non è né chiuso né aperto)
- $F(X, \bar{x}) \subseteq T(X, \bar{x})$  [ $d_n \equiv d$ ], dove  $F$  è il cono delle direzioni ammissibili.
- $X$  convesso  $\Rightarrow X \subseteq \bar{x} + T(X, \bar{x})$
- $X$  convesso  $\Rightarrow T(X, \bar{x}) = \underset{\text{chiusura}}{\text{cl}} F(X, \bar{x})$

 **Teorema 13.5 (Condizione necessaria)**

Sia  $\bar{x} \in X$  un punto di minimo locale di  $(P)$ . Allora

$$\nabla f(\bar{x})^T d \geq 0 \quad \forall d \in T(X, \bar{x}) \quad (13.1)$$

*Dimostrazione.* Estensione della vecchia dimostrazione (caso non vincolato).

Sia

$$\exists \epsilon > 0 \quad f(\bar{x}) = \min\{f(x) : x \in X \cap B(\bar{x}, \epsilon)\}. \text{ e } d \in T(X, \bar{x})$$

Per definizione di cono tangente:

$$d \in T(X, \bar{x}) \Rightarrow \exists t_n \rightarrow 0^+, d_n \rightarrow d \text{ t.c. } \bar{x} + t_n d_n \in X$$

Poiché  $\bar{x} + t_n d_n \rightarrow \bar{x}$  risulta  $\bar{x} + t_n d_n \in B(\bar{x}, \epsilon)$  per  $n$  sufficientemente grande. Quindi:

$$0 \leq f(\bar{x} + t_n d_n) - f(\bar{x}) \underbrace{=}_{\text{Taylor}} t_n \nabla f(\bar{x})^T d_n + r(t_n d_n)$$



Dividendo per  $t_n$

$$0 \leq \frac{f(\bar{x} + t_n d_n) - f(\bar{x})}{t_n} = \frac{1}{t_n} \nabla f(\bar{x})^T d_n + \frac{r(t_n d_n)}{t_n} \xrightarrow{n \rightarrow +\infty} \nabla f(\bar{x})^T d \geq 0$$

□

**Osservazione 13.6**

Poiché  $F(X, \bar{x}) \subseteq T(X, \bar{x})$ , se  $X$  è convesso e  $f$  è convessa allora (13.1) è anche condizione sufficiente. Inoltre

$$S\bar{x} \in \text{int } X \implies (13.1) \equiv (\nabla f(\bar{x}) = 0)$$

**13.1.1.1 Caso con soli vincoli di disuguaglianza**

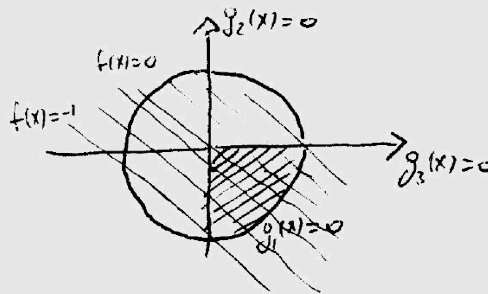
**Esempio 13.7**

$n = 2$

$$f(x) = x_1 + x_2$$

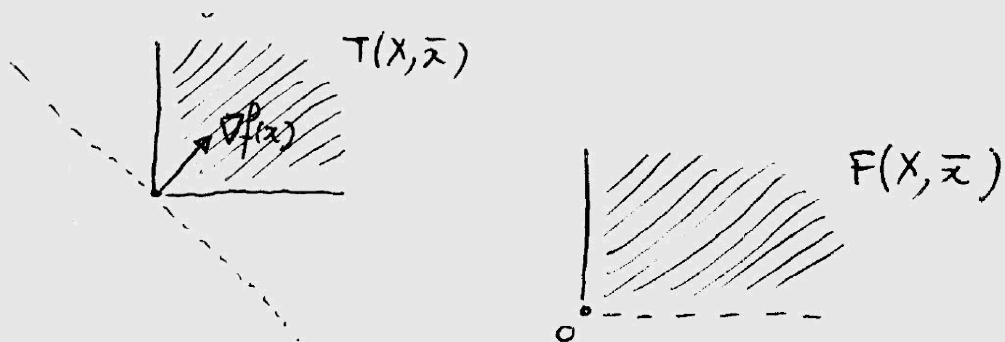
$$g_1(x) = x_1^2 + x_2^2 - 1 \leq 0 \quad g_2(x) = -x_1 \leq 0 \quad g_3(x) = x_2 \leq 0$$

(cerchio)



Il punto di minimo (globale) è  $\bar{x} = (0, -1)$ , mentre  $\nabla f(\bar{x}) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

$$T(X, \bar{x}) \subseteq \{d \in \mathbb{R}^n \mid \underbrace{d_1 + d_2}_{\nabla f(\bar{x})^T d} \geq 0\}$$



Notiamo che nel cono tangente è presente la semiretta orizzontale, che è invece assente nel cono delle direzioni ammissibili: questo è dovuto al fatto che nel cono delle direzioni ammissibili siamo costretti a fissare un  $d$ , e non riusciamo quindi ad arrivare alla semiretta orizzontale. Ci riusciamo invece avendo una sottosuccessione  $d_n$  che tende a  $d$ , dove  $d$  in questo caso è ancora la semiretta orizzontale.

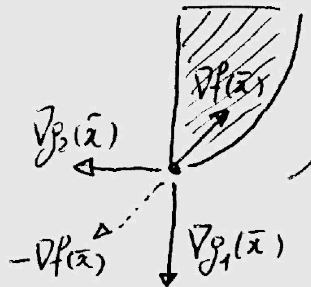
Vincoli attivi:  $g_1, g_2$

$$\bar{x} = (0, 1)$$

$$g_1(\bar{x}) = 0 \quad g_2(\bar{x}) = 0 \quad g_3(\bar{x}) = -1 < 0$$

Calcoliamo i gradienti

$$\nabla g_1(\bar{x}) = \begin{pmatrix} 0 \\ -2 \end{pmatrix} \quad \nabla g_2(\bar{x}) = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$$



Abbiamo che  $-\nabla f(x)$  appartiene al cono tangente  $\{\nabla g_1(\bar{x}), \nabla g_2(\bar{x})\}$  ed è combinazione lineare di  $\nabla g_1$  e  $\nabla g_2$ . Assomiglia al semplice.

$$-\nabla f(\bar{x}) = \lambda_1 \nabla g_1(\bar{x}) + \lambda_2 \nabla g_2(\bar{x})$$

Cerchiamo di formalizzare matematicamente questo concetto. Dobbiamo sfruttare la rappresentazione di  $X$ .

$$X = \{x \in \mathbb{R}^n \mid g_i(x) \leq 0, i = 1 \dots n\} \quad g_i : \mathbb{R}^n \rightarrow \mathbb{R} \text{ diff.}$$

Supponiamo esplicitamente che  $X$  sia espressa in questa forma. Poi introdurremo le uguaglianze. Sostituiamo le formule con il loro sviluppo al prim'ordine. I vincoli non attivi è come se non ci fossero: considereremo i vincoli attivi.

**Linearizzazioni di  $X$  in  $\bar{x} \in X$**

$$I(x) = \{i \mid g_i(\bar{x}) = 0\} \quad \text{indici dei vincoli attivi in } \bar{x}$$

$$g_i(x) \approx \underbrace{g_i(\bar{x})}_{\text{se attivo}} + \nabla g_i(\bar{x})^T \underbrace{(x - \bar{x})}_d$$

Andiamo quindi a inserire in  $L$  le approssimazioni al primo ordine delle  $g_i$  che sono vincoli attivi.

$$L_{<}(X, \bar{x}) = \{d \in \mathbb{R}^n \mid \nabla g_i(\bar{x})^T d < 0, i \in I(\bar{x})\} \text{ (APERTO)}$$

$$L_{\leq}(X, \bar{x}) = \{d \in \mathbb{R}^n \mid \nabla g_i(\bar{x})^T d \leq 0, i \in I(\bar{x})\} \text{ (CHIUSO)}$$

### Proposizione 13.1

Sia  $\bar{x} \in X$ . Allora

$$L_{<}(X, \bar{x}) \subseteq T(X, \bar{x}) \subseteq L_{\leq}(X, \bar{x})$$

(No Dimostrazione)

Abbiamo quindi inscatolato il cono tangente.

### Esempio 13.8 (Continuazione esempio precedente)

Vediamo nell'esempio le linearizzazioni.

$$L_{<}(X, \bar{x}) = \{d \in \mathbb{R}^2 \mid -2d_2 < 0, -d_1 < 0\} = \text{int } \mathbb{R}_+^2$$

$$L_{\leq}(X, \bar{x}) = \{d \in \mathbb{R}^2 \mid -2d_2 \leq 0, -d_1 \leq 0\} = \mathbb{R}_+^2$$

Abbiamo quindi che

$$L_{<}(X, \bar{x}) = T(X, \bar{x})$$

$L_{<}(X, \bar{x})$  è un insieme aperto mentre  $T(X, \bar{x})$  è un insieme chiuso, l'unico caso in cui possono coincidere è che i 2 insiemi sono aperti e chiusi: questo avviene quando tali insiemi sono tutto lo spazio, questo corrisponde a non avere vincoli attivi.

Se  $\bar{x}$  è un punto interno alla regione ammissibile, il cono tangente è tutto  $\mathbb{R}^n$ , perché ci possiamo avvicinare da ogni direzione. È possibile che  $T$  e  $L_{\leq}$  siano diversi, ossia

$$T(X, \bar{x}) \subset L_{\leq}(X, \bar{x})$$

Ad esempio (verificare per casa)

$$X = \{x \in \mathbb{R}^2 \mid x_1^2 - x_2^2 \leq 0\} \quad \bar{x} = (0, 0)$$

La diversità tra  $T(X, \bar{x})$  e  $L_{\leq}(X, \bar{x})$  è quella che ci da fastidio. Vogliamo riscrivere la condizione

$$\nabla f(\bar{x})^T d \geq 0 \quad \forall d \in T(X, \bar{x})$$

Utilizzeremo  $L_{<}$ , grazie infatti all'inclusione  $L_{<}(X, \bar{x}) \subseteq T(X, \bar{x})$  abbiamo:

$$\nabla f(\bar{x})^T d \geq 0 \quad \forall d \in T(X, \bar{x}) \implies \left\{ d \mid \begin{cases} \nabla f(\bar{x})^T d < 0 \\ \nabla g_i(\bar{x})^T d < 0 \quad \forall i \in I(\bar{x}) \end{cases} \right\} = \emptyset$$

ossia il sistema non ammette alcuna soluzione  $d \in \mathbb{R}^n$ . Possiamo dimenticare il cono tangente: abbiamo una condizione necessaria di ottimalità fatta esclusivamente dalla funzione obiettivo e dai vincoli. Il teorema che andremo ad enunciare, risultato dell'algebra lineare, ci fornisce un sistema *duale*, che ammette soluzione quando il sistema di partenza non ne ha, e viceversa.



**Teorema 13.9 (Motzkin)**

Siano  $a_k, b_i, c_j \in \mathbb{R}^n$  con  $k \in I^<, i \in I^{\leq}, j \in I^=$ , con  $I^<, I^{\leq}, I^=$  insiemi finiti di indici con  $I^< \neq \emptyset$ . Allora

$$\begin{cases} a_k^T d < 0, k \in I^< \\ b_i^T d \leq 0, i \in I^{\leq} \\ c_j^T d = 0, j \in I^= \end{cases} \iff \begin{cases} \sum_{k \in I^<} \theta_k a_k + \sum_{i \in I^{\leq}} v_i b_i + \sum_{j \in I^=} \mu_j c_j = 0 \\ \theta_k \geq 0, v_i \geq 0, \mu \in \mathbb{R} \end{cases}$$

non ha soluzione  $d \in \mathbb{R}^n$   ammette soluzione in cui i  $\theta_k$  non sono tutti nulli

Quindi, applicando questo teorema in versione di Gordan al sistema (5) con  $\bar{x}$  punto di minimo locale di (P), e quindi (5) non ammette soluzione, si ottiene che:

$$\begin{cases} \nabla f(\bar{x})^T d < 0 \\ \nabla g_i(\bar{x})^T d < 0 \quad \forall i \in I(\bar{x}) \end{cases} = \emptyset \iff \begin{cases} \exists \theta \geq 0, \lambda_i \geq 0 \text{ con } i \in I(\bar{x}) \text{ non tutti nulli tali che} \\ \theta \nabla f(\bar{x}) + \sum_{i \in I(\bar{x})} \lambda_i \nabla g_i(\bar{x}) = 0 \end{cases}$$

Ponendo

$$\lambda_i = 0 \quad \forall i \notin I(\bar{x})$$

Otteniamo

$$\begin{cases} \theta \nabla f(\bar{x}) + \sum_{i=1}^n \lambda_i \nabla g_i(\bar{x}) = 0 \\ \lambda_i g_i(\bar{x}) = 0 \quad i = 1 \dots n \end{cases}$$

Abbiamo quindi una nuova riformulazione del teorema

**Teorema 13.10 (Fritz John)**

Sia  $\bar{x}$  un punto di minimo locale di  $(P)$ . Allora  $\exists \theta \geq 0, \lambda_i \geq 0$  con  $i = 1, \dots, n$  non tutti nulli (sia  $\theta$  che  $\lambda_i$ ) tali che

$$(FJ) \begin{cases} \theta \nabla f(\bar{x}) + \sum_{i=1}^n \lambda_i \nabla g_i(\bar{x}) = 0 & (FJ_1) \\ \lambda_i g_i(\bar{x}) = 0 & i = 1 \dots n & (FJ_2) \end{cases}$$

Il vincolo  $(FJ_2)$  esprime una condizione di complementarità: nel caso che il vincolo sia attivo in  $\bar{x}$  (cioè  $g(\bar{x}) = 0$ ) può essere  $\lambda_i \geq 0$ . Altrimenti se il vincolo non è attivo ( $g(\bar{x}) < 0$ ) deve essere  $\lambda_i = 0$ .

Per la verifica di ottimalità di  $\bar{x}$  siamo passati da un controllo di infiniti vettori alla risoluzione di un sistema di equazioni che è trattabile. Risolvendo il sistema di Fritz–John si otterranno  $\theta$  e dei  $\lambda_i$ . Questi ultimi varranno 0 per i vincoli non attivi, fornendo un'indicazione su quali siano i vincoli che effettivamente entrano in gioco nella risoluzione del problema di minimizzazione. I  $\lambda_i$  vengono chiamati *moltiplicatori di Lagrange*.

Che fastidio ci da  $\theta = 0$ ? Compaiono esclusivamente i vincoli del problema, non c'è la funzione obiettivo: vorremo che le condizioni valessero per  $\theta \neq 0$ .

Se le condizioni di Fritz John valgono con  $\theta = 0$ , allora  $\{\nabla g_i(\bar{x})\}_{i \in I(\bar{x})}$  sono linearmente dipendenti.

Allora aggiungiamo la negazione come ipotesi. Otteniamo una nuova versione del teorema.

**Osservazione 13.11**

$$(FJ) \text{ valgono con } \theta = 0 \implies \{\nabla g_i(\bar{x})\}_{i \in I(\bar{x})} \text{ linearmente dipendenti}$$

**Teorema 13.12 (Karush-Kuhn-Tucker)**

Sia  $\bar{x}$  un punto di minimo locale di  $(P)$  e supponiamo che  $\{\nabla g_i(\bar{x})\}_{i \in I(\bar{x})}$  siano linearmente indipendenti. Allora  $\exists \lambda_i \geq 0$  con  $i = 1, \dots, n$  tali che

$$\begin{aligned} (KKT_1) \quad \nabla f(\bar{x}) + \sum_{i=1}^n \lambda_i \nabla g_i(\bar{x}) &= 0 \\ (KKT_2) \quad \lambda_i g_i(\bar{x}) &= 0 \quad i = 1 \dots n \end{aligned}$$

Le condizioni di ottimalità KKT sono costituite dal sistema di equazioni e disequazioni (non lineari):

$$\left\{ \begin{array}{ll} (KKT_1) \quad \nabla f(\bar{x}) + \sum_{i=1}^n \lambda_i \nabla g_i(\bar{x}) = 0 & \text{(Ottimalità)} \\ (KKT_2) \quad \lambda_i g_i(\bar{x}) = 0 & i = 1 \dots n \quad \text{(Complementarità)} \\ (KKT_3) \quad g_i(x) \leq 0 & i = 1 \dots n \quad \text{(Ammissibilità (1))} \\ (KKT_4) \quad \lambda_i \geq 0 & i = 1 \dots n \quad \text{(Ammissibilità (2))} \end{array} \right.$$

,dove  $(KKT_3)$  e  $(KKT_4)$  sono le condizioni di ammissibilità, nelle incognite  $x \in \mathbb{R}^n, \lambda = (\lambda_1, \dots, \lambda_n) \in \mathbb{R}^n$

**Nota**

Le condizioni di complementarità fanno in modo che il moltiplicatore di Lagrange associato ad un vincolo possa essere strettamente maggiore di 0 solamente nel caso in cui il vincolo è attivo, infatti:

**Caso 1**

$$\begin{cases} \lambda_i g_i(\bar{x}) = 0 \\ \lambda_i \geq 0 \\ g_i(x) = 0 \end{cases}$$

**Caso 2**

$$\begin{cases} \lambda_i g_i(\bar{x}) = 0 \\ \lambda_i = 0 \\ g_i(x) < 0 \end{cases}$$

**Osservazione 13.13 (I moltiplicatori di Lagrange sono unici)**

Infatti se  $\bar{\lambda} = (\bar{\lambda}_1, \dots, \bar{\lambda}_n)$  e  $\hat{\lambda} = (\hat{\lambda}_1, \dots, \hat{\lambda}_n)$  soddisfano  $(KKT_1)$ , sottraendo un sistema di equazioni dall'altro si ottiene:

$$\sum_{i \in I(\bar{x})} (\bar{\lambda}_i - \hat{\lambda}_i) \nabla g_i(\bar{x}) = 0$$

Allora  $\bar{\lambda}_i = \hat{\lambda}_i \quad \forall i \in [1 \dots n]$  a causa della lineare indipendenza dei gradienti.

Qualifiche dei vincoli:  $\equiv$  condizioni su vincoli per cui (FJ) valgono con  $\theta = 1$ . Altre qualifiche sono:

**Definizione 13.14 (Condizioni di Slater)**

1.  $g_i$  convesse per ogni  $i \in I(\bar{x})$
2.  $\exists \hat{x} \in \mathbb{R}^n$  t.c.  $g_i(\hat{x}) < 0 \quad i = 1 \dots n$

**Definizione 13.15 (Condizioni di Mangasarian-Fromovits)**

$\exists d \in \mathbb{R}^n : \nabla g_i(\hat{x})^T d < 0$  per ogni  $i \in I(\bar{x})$  (cioè  $L_<(X, \hat{x}) \neq \emptyset$ )

Lineare indipendenza si può sostituire con una delle due condizioni

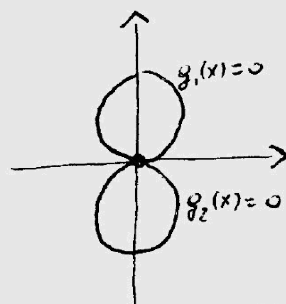
**Esempio 13.16**

È importante notare che le condizioni (KKT) non possono sempre essere utilizzate per dimostrare che un punto è di minimo locale. In questo esempio troveremo un punto di minimo in cui non valgono le ipotesi del teorema KKT

$n = 2, m = 2$

$$f(x) = x_1 + x_2^2$$

$$g_1(x) = x_1^2 + (x_2 - 1)^2 - 1 \quad g_2(x) = x_1^2 + (x_2 + 1)^2 - 1$$



La regione ammissibile è data da un solo punto:

$$X = \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right\}$$

Di conseguenza il punto di minimo può essere soltanto:

$$\bar{x} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \implies I(\bar{x}) = \{1, 2\}$$

Calcolo dei gradienti:

$$\nabla f(\bar{x}) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \nabla g_1(\bar{x}) = \begin{pmatrix} 2\bar{x}_1 \\ 2(\bar{x}_2 - 1) \end{pmatrix} = \begin{pmatrix} 0 \\ -2 \end{pmatrix} \quad \nabla g_2(\bar{x}) = \begin{pmatrix} 2\bar{x}_1 \\ 2(\bar{x}_2 + 1) \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}$$

Calcolo delle condizioni KKT: entrambi i vincoli sono attivi.

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} + \lambda_1 \begin{pmatrix} 0 \\ -2 \end{pmatrix} + \lambda_2 \begin{pmatrix} 0 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2(\lambda_2 - \lambda_1) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

- Il sistema sopra non ha soluzione, quindi il teorema (KKT) non si può applicare in  $\bar{x}$ . Questa condizione si verifica perché non sono verificate le ipotesi di (KKT), dato che  $\nabla g_1(\bar{x})$  e  $\nabla g_2(\bar{x})$  non sono linearmente indipendenti:

$$\begin{pmatrix} 0 \\ 2 \end{pmatrix} = -1 \cdot \begin{pmatrix} 0 \\ -2 \end{pmatrix}$$

- le condizioni (FJ) valgono con per  $\theta = 0 \quad \lambda_2 = \lambda_1 > 0$
- Le qualifiche dei vincoli non valgono:

$$\begin{cases} \nabla g_1(\bar{x}), \nabla g_2(\bar{x}) & \text{linearmente dipendenti} \\ L_{<}(X, \bar{x}) = \emptyset \end{cases}$$

Il cono tangente è più piccolo del linearizzato.

## Vincoli lineari

Esistono dei casi in cui le qualifiche dei vincoli non servono, è il caso dei vincoli lineari. I vincoli lineari non richiedono alcuna qualificazione affinché valgano le condizioni KKT.

I vincoli lineari sono della forma

$$X = \{x \in \mathbb{R}^n \mid Ax \leq b\} \quad A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$$

### Proposizione 13.2

Sia  $\bar{x} \in X$ , allora

$$T(X, \bar{x}) = L_{\leq}(X, \hat{x})$$



#### Nota

No dimostrazione, ma sulla note c'è'.

### Condizioni KKT nella programmazione lineare

Consideriamo la programmazione lineare

$$\max\{c^T x \mid Ax \leq b\} = -\min\{-cx \mid Ax \leq b\}$$

Sappiamo che valgono le condizioni *KKT* sempre.

Come minimo

$$-\min\{-c^T x \mid \hat{x} \in X\}$$

Obiettivo

$$f(x) = -c^T x \quad \nabla f(\bar{x}) = -c$$

Vincoli

$$g_i(x) = A_i x - b_i \quad \rightarrow \quad \nabla g_i(x) = A$$

Condizioni moltiplicatori di Lagrange

$$-c + \sum_{i=1}^n \lambda_i A_i = -c + \lambda^T A = 0$$

Ossia

$$\begin{aligned} \lambda^T A &= c && (KKT_1) \text{ (ammissibilità duale 1)} \\ \lambda &\geq 0 && (KKT_2) \text{ (ammissibilità duale 2)} \\ Ax &\leq b && (KKT_3) \text{ (ammissibilità primale)} \\ \lambda_i \underbrace{(b_i - a_i^T x)}_{g_i} &= 0 && (KKT_4) \text{ (scarti complementari)} \end{aligned}$$

#### 13.1.1.2 Caso con introduzione di vincoli di uguaglianza

Consideriamo adesso il caso in cui  $X$  sia descritto tramite vincoli di disuguaglianza ed uguaglianza.

$$(P) \quad \min\{f(x) : x \in X\} \quad X = \{x \in \mathbb{R}^n \mid g_i(x) \leq 0, h_j(x) = 0 \quad i = 1 \dots n, j = 1 \dots p\}$$

con

$$g_i : \mathbb{R}^n \rightarrow \mathbb{R}, i = 1, \dots, n, \quad h_j : \mathbb{R}^n \rightarrow \mathbb{R}, j = 1, \dots, p \text{ differenziabili}$$



#### Nota

La tentazione sarebbe di ragionare nel metodo più semplice possibile:

$$h_j(x) = 0 \equiv \begin{cases} h_j(x) \leq 0 \\ -h_j(x) \leq 0 \end{cases}$$

ossia considerare un vincolo di uguaglianza come due disuguaglianze. Ma in questo caso si avrebbe

$$L_{<}(X, \bar{x}) = \{d \in \mathbb{R}^n \mid \nabla g_i(\bar{x})^T d < 0 \quad i \in I(\bar{x})\} = \emptyset$$

Quindi dal punto di vista matematico l'equivalenza sopra è corretta, ma noi siamo interessati a trovare un metodo per la verifica dell'ottimalità linearizzando i vincoli e utilizzando  $L_{<}$ , quindi per i nostri fini dobbiamo procedere diversamente.

$$L_{\leq} = \{d \in \mathbb{R}^n \mid \nabla g_i(\bar{x})^T d \leq 0 \quad i \in I(\bar{x}), \quad \nabla h_j(\bar{x})^T d = 0 \quad j = 1 \dots p\}$$

Lo riportiamo a  $L_{<}$

$$L_{<} = \{d \in \mathbb{R}^n \mid \nabla g_i(\bar{x})^T d < 0 \quad i \in I(\bar{x}), \quad \nabla h_j(\bar{x})^T d = 0 \quad j = 1 \dots p\}$$

L'inclusione

$$T(X, \bar{x}) \subseteq (X, \bar{x})$$

si dimostra similmente al caso delle sole disuguaglianze mentre l'inclusione

$$L_{<}(X, \bar{x}) \subseteq T(X, \bar{x})$$

richiede l'ipotesi che  $\{\nabla h_j(\bar{x})\}_{j=1\dots p}$  siano linearmente indipendenti.

A questo punto ripercorriamo la teoria:

$$\begin{aligned} \nabla f(\bar{x})^T d \geq 0 \quad \forall d \in T(X, \bar{x}) \\ \{\nabla h_j(\bar{x})\}_{j=1\dots p} \text{ linearmente indipendenti} \end{aligned} \implies_{L_{<} \subseteq T} \begin{cases} \nabla f(\bar{x})^T d < 0 \\ \nabla g_i(\bar{x})^T d < 0 \quad i \in I(\bar{x}) \\ \nabla h_j(\bar{x})^T d = 0 \quad j = 1 \dots p \end{cases}$$

Non ammette soluzione  $d \in \mathbb{R}^n$

Possiamo applicare il teorema di Motzkin, ottenendo le condizioni di Fritz John, e imponendo la qualifica otteniamo le condizioni KKT.



**Teorema 13.17 (Karush-Kuhn-Tucker con vincoli di uguaglianza e disuguaglianza)**

Sia  $\bar{x} \in X$  un punto di minimo locale di  $(P)$  e supponiamo che i vettori  $\{\nabla g_i(\bar{x})\}_{i \in I(\bar{x})} \cup \{\nabla h_j(\bar{x})\}_{j=1\dots p}$  siano linearmente indipendenti. Allora esistono  $\lambda_i \geq 0$  con  $i = 1 \dots n$ ,  $\mu \in \mathbb{R}$  con  $j = 1 \dots p$  tali che

$$\begin{aligned} \nabla f(\bar{x}) + \sum_{i=1}^n \lambda_i \nabla g_i(\bar{x}) + \sum_{j=1}^p \mu_j \nabla h_j(\bar{x}) = 0 \quad (\overline{KKT}_1) \text{ Moltiplicatori di Lagrange} \\ \lambda_i g_i(\bar{x}) = 0 \quad i = 1 \dots n \quad (\overline{KKT}_2) \text{ Complementarità} \end{aligned}$$

Le condizioni di ottimalità  $\overline{KKT}$  sono quindi costituite dal sistema di equazioni e disequazioni (non lineari):

$$\begin{cases} \nabla f(\bar{x}) + \sum_{i=1}^n \lambda_i \nabla g_i(\bar{x}) + \sum_{j=1}^p \mu_j \nabla h_j(\bar{x}) = 0 & (\overline{KKT}_1) \text{ Moltiplicatori di Lagrange} \\ \lambda_i g_i(\bar{x}) = 0 & i = 1 \dots n \quad (\overline{KKT}_2) \text{ Complementarità} \\ g_i(x) \leq 0 & i = 1 \dots n \quad (\overline{KKT}_3) \text{ Ammissibilità (1)} \\ h_j(x) = 0 & j = 1 \dots p \quad (\overline{KKT}_5) \text{ Ammissibilità (2)} \\ \lambda_i \geq 0 & i = 1 \dots n \quad (\overline{KKT}_4) \text{ Ammissibilità (3)} \end{cases}$$

nelle incognite  $x \in \mathbb{R}^n$ ,  $\lambda = (\lambda_1, \dots, \lambda_n) \in \mathbb{R}^n$ ,  $\mu = (\mu_1, \dots, \mu_p) \in \mathbb{R}^p$

**Definizione 13.18 (Funzione Lagrangiana)**

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^n \lambda_i g_i(x) + \sum_{j=1}^p \mu_j h_j(x)$$

si dice funzione Lagrangiana di

$$(P) \quad \min\{f(x) : g_i(x) \leq 0, i = 1 \dots n, h_j(x) = 0, j = 1 \dots p\}$$



**Osservazione 13.19**

$$(\overline{KKT}_1) \equiv \nabla_x L(x, \lambda, \mu) = 0$$



### Altre qualifiche dei vincoli

#### Slater

- $h_j$  : affini :  $h_j(x) = a_j^T x + b_j, k = 1 \dots p$
- $g_i$  : convesse per ogni  $i \in I(\bar{x})$
- $\{\nabla h_j(\bar{x})\}_{j=1 \dots p}$  linearmente indipendenti
- $\exists \hat{x} \in \mathbb{R}^n$  tale che  $\begin{cases} g_i(\hat{x}) < 0 & i = 1 \dots n \\ h_j(\hat{x}) = 0 & j = 1 \dots p \end{cases}$

#### Mangasarian-Fromovitz

- $\{\nabla h_j(\bar{x})_{j=1 \dots p}\}$  linearmente indipendenti
- $\exists d \in \mathbb{R}^n$  tale che  $\begin{cases} \nabla g_i(\bar{x})^T d \leq 0 & i \in I(\bar{x}) \\ \nabla h_j(\bar{x})^T d = 0 & j = 1 \dots p \end{cases} (\iff L_{<}(X, \bar{x}) \neq \emptyset)$

#### Esercizio 13.1

Possiamo provare a scrivere le condizioni KKT per il problema duale

$$\min\{b^T \lambda \mid \lambda \in X\}$$

$$X = \{\lambda \in \mathbb{R}^n \mid \lambda^T A = c, \lambda \geq 0\}$$

Indicazione, deve venire fuori  $\lambda_i(A_i x - b_i) = 0 \quad i = 1 \dots n$

#### Esercizio 13.2

$$g_i(x) \leq 0 \iff g_i(x) = s_i^2 = 0$$

Trasformiamo il problema in soli vincoli di uguaglianza con le slack  $(x, s) \in \mathbb{R}^{n+m}$ . Scrivere le condizioni KKT in questo caso. Problema: per le uguaglianze serve la lineare indipendenza.

Le condizioni KKT sono necessarie per l'ottimalità e ma vale il seguente teorema:



#### Teorema 13.20

Siano  $f$  e  $g_i$  convesse con  $i \in I(\bar{x})$  per qualche  $\bar{x} \in X$  e siano  $h_j$  affini con  $j = 1 \dots p$  ( $h_j(x) = a_j^T x - b_j$  per opportuni  $a_j \in \mathbb{R}^n, b_j \in \mathbb{R}$ ). Se esistono  $\lambda_i \geq 0$ , con  $i = 1 \dots n$  e  $\mu \in \mathbb{R}$ , con  $j = 1 \dots p$ , tali che valgono  $(KKT_1)$  e  $(KKT_2)$  allora  $\bar{x}$  è un punto di minimo globale.

## 13.2 Metodi per la risoluzione

### 13.2.1 Alcuni approcci alla risoluzione del problema

**Esempio 13.21**

Vogliamo minimizzare la norma, che è una funzione convessa.

$$\min\{x_1^2 + x_2^2\} \quad x_1 + x_2 - 1 = 0$$

Una strada possibile è usare i moltiplicatori KKT e trovare la soluzione.

In realtà alcune variabili possono essere espresse da altre

$$x_2 = 1 - x_1$$

Sostituendo nella funzione obiettivo

$$\min\{x_1^2 + (1 - x_1)^2 \mid x_1 \in \mathbb{R}\}$$

Abbiamo eliminato una variabile: è diventato un problema di ottimizzazione non vincolata: lo possiamo risolvere con i metodi dell'ottimizzazione non vincolata.

Il gradiente è

$$2x_1 - 2(1 - x_1) = 0 \quad \Rightarrow \quad x_1 = \frac{1}{2}$$

**Esercizio 13.3**

Provare a risolvere il problema con le condizioni KKT.

Saremmo tentati ad usare sempre questa strada, ma non è sempre possibile

**Esempio 13.22**

$$\min\{x_1^2 + x_2^2\} \quad x_2^2 - (x_1 - 1)^3 = 0$$

$$x_2^2 = (x_1 - 1)^3$$

$$\min\{x_1^2 + (x_1 - 1)^3 \mid x_1 \in \mathbb{R}\}$$

$x_1 \rightarrow -\infty$  Inferiormente illimitato, ma non è possibile! Cosa abbiamo sbagliato? Il vincolo

$$x_2^2 = (x_1 - 1)^3$$

Sta nascondendo il vincolo  $x_1 \geq 1$ .

L'eliminazione di variabili funziona bene solamente nel caso di vincoli lineari.

**Esempio 13.23**

$$\min\{f(x) \mid Ax = b\} \quad A \in \mathbb{R}^{m \times n} \quad m \leq n \quad A \text{ rango massimo}$$

Se una matrice  $A$  ha rango massimo, si possono riordinare le  $n$  colonne della matrice  $A$  in modo che le prime  $n$  siano linearmente indipendenti

$$A = \left[ \underbrace{A_1}_{n \times n} \mid \underbrace{A_2}_{n \times (n-m)} \right]$$

$$x = \begin{pmatrix} x^1 \\ x^2 \end{pmatrix} \quad x^1 \in \mathbb{R}^m \quad x^2 \in \mathbb{R}^n$$

$$x^1 = A_1^{-1}(b - A_2x^2) \iff A_1x^1 = b - A_2x^2 \iff Ax = b$$

*Problemi: abbiamo una matrice da invertire, potremmo soffrire di problemi di condizionamento, cancellazione, per questo queste tecniche non vengono molto utilizzate.*

Accenno ad un'altra tecnica

### Esempio 13.24

$$\min\{x_1^2 + x_2^2 \mid x_1 + x_2 - 1 = 0\}$$

*Possiamo penalizzare la funzione obiettivo nei punti in cui  $x$  non soddisfa i vincoli. In questo esempio potremmo usare come funzione obiettivo*

$$x_1^2 + x_2^2 + r(x_1 + x_2 - 1)^2$$

*Nuova famiglia di problemi*

$$\min\{x_1^2 + x_2^2 + r(x_1 + x_2 - 1) \mid x_1, x_2 \in \mathbb{R}\}$$

### Esercizio 13.4

*Provare a risolvere il problema per  $r$  fissato.*

## 13.2.2 Excursus di metodi sull'ottimizzazione vincolata

### Condizioni KKT nel caso disuguaglianze e uguaglianze

$$\nabla f(x) + \sum_{i=1}^n \lambda_i \nabla g_i(x) + \sum_{j=1}^p \mu_j \nabla h_j(x) = 0 \iff \nabla_x L(x, \nabla, \mu) = 0$$

$$\lambda_i g_i(x) = 0 \quad i = 1 \dots n$$

$$g_i(x) \leq 0 \quad i = 1 \dots n$$

$$h_j(x) = 0 \quad j = 1 \dots p$$

$$\lambda_i \geq 0 \quad i = 1 \dots r$$

### Lagrangiana

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^n \lambda_i g_i(x) + \sum_{j=1}^p \mu_j h_j(x)$$

$$(P) \min\{f(x) : g_i(x)_{i=1 \dots m} \leq 0, h_j(x)_{j=1 \dots o} = 0\}$$

### 13.2.2.1 Trasformazione di problemi in forma non vincolata: penalizzazione esterna

#### Esempio introduttivo

Consideriamo la minimizzazione della seguente funzione obiettivo convessa:

$$\min\{x_1^2 + x_2^2 : x_1 + x_2 - 1 = 0\}$$

e inseriamo una forma di penalizzazione, portando il vincolo nella funzione con un coefficiente  $r$  di penalizzazione:

$$\min\{x_1^2 + x_2^2 + r(x_1 + x_2 - 1)^2 : x \in \mathbb{R}^2\}$$

$r$  indica di quanto penalizziamo, quando siamo al di fuori della regione ammissibile: il fatto che ci sia un'espressione al quadrato è voluto, in modo che la cosa funzioni sia per quantità negative che positive. Vogliamo capire la relazione che c'è fra  $r$  (parametro di penalizzazione) ed il problema originale. La funzione è convessa, minimo locale implica minimo globale. Deriviamo, le due equazioni sono

$$\text{Derivata rispetto a } x_1 : 2x_1 + 2r(x_1 + x_2 - 1) = 0$$

$$\text{Derivata rispetto a } x_2 : 2x_2 + 2r(x_1 + x_2 - 1) = 0$$

Indichiamo con  $x^r$  il punto di minimo.

$$x_1^r = x_2^r = \frac{r}{1 + 2r}$$

Cosa succede per  $r \rightarrow \infty$ ? Il limite è  $\frac{1}{2}$ .

Scriviamo le condizioni KKT che in questo danno una condizione necessaria è sufficiente all'ottimalità.

$$\begin{cases} 2x_1 + \mu = 0 \\ 2x_2 + \mu = 0 \\ x_1 + x_2 = 1 \quad \text{Ammissibilità} \end{cases}$$

Otteniamo

$$x_1^* = x_2^* = \frac{1}{2} \quad \mu = -1$$

Proviamo a porci in una situazione più generale: sia  $h$  il vincolo di uguaglianza.

$$\min\{f(x) + rh^2(x) \mid x \in \mathbb{R}^n\}$$

e deriviamo rispetto a  $x$

$$\nabla f(x^r) + \underbrace{2rh(x^r)}_{\mu} \nabla h(x^r)$$

dove  $x^r$  è il minimo: abbiamo ottenuto un moltiplicatore

$$2r(x_1^r + x_2^r - 1) = \frac{-2r}{1 + 2r} \xrightarrow{r \rightarrow +\infty} -1$$

Quindi, almeno in questo caso, portare  $r$  ad infinito corrisponde a risolvere il problema originale.

### Teoria della penalizzazione

Consideriamo il problema generico con vincoli di uguaglianza e di disuguaglianza. Introduciamo la *Funzione di penalizzazione*  $p_r$  che tiene conto di entrambi i tipi di vincoli e utilizziamo un parametro  $r > 0$ :

$$p_r(x) = f(x) + r \cdot \left( \sum_{i=1}^n g_i^{+2}(x) + \sum_{j=1}^p h_j^2(x) \right)$$

dove  $g_i^+$  deve ancora essere definita. Come abbiamo visto, per i vincoli di uguaglianza è corretto moltiplicare  $r$  per  $h_j^2(h)$ , ma per quanto riguarda le disuguaglianze non possiamo comportarci allo stesso modo (ovvero prendere  $g_i^+ = g_i$  dato che in questo modo penalizzeremmo anche le soluzioni che si discostano dai vincoli verso l'interno della regione ammissibile).

Prendendo invece

$$g_i^+(x) = \max\{0, g_i(x)\}$$

la penalizzazione  $g_i^+(x)$  vale 0 se il vincolo è rispettato,  $> 0$  altrimenti.

Si pone dunque un nuovo problema di minimizzazione:

$$(P_r) \quad \min\{p_r(x) : x \in \mathbb{R}^n\}$$

**Nota**

$g_i^+(x)$  non è necessariamente differenziabile, ma elevandola al quadrato lo diventa. Tuttavia, non è differenziabile due volte, quindi per minimizzare  $p_r(x)$  non possiamo applicare il metodo di Newton che usa le derivate seconde.

**Osservazioni preliminari**

$$x \in X \Rightarrow p_r(x) = f(x)$$

$$x \notin X \Rightarrow p_r(x) > f(x)$$

Quindi in generale

$$p_r(x) \geq f(x) \quad \forall x \in \mathbb{R}^n, \forall r \geq 0$$

Chiamamo  $\bar{v}$  il valore ottimo del problema originale  $(P)$  e  $\bar{v}_r$  l'ottimo del problema con la penalizzazione  $(P_r)$ , i valori in generale non coincidono!

**Osservazione 13.25**

$\bar{v} = \min\{f(x) : x \in X\} \geq \min\{p_r(x) : x \in \mathbb{R}^n\} = \bar{v}_r \quad \forall r \geq 0$ . Quindi

$$\bar{v} \geq \sup\{\bar{v}_r : r \geq 0\}$$

Inoltre  $r_1 \geq r_2 \Rightarrow \bar{v}_{r_1} \geq \bar{v}_{r_2}$  (cioè  $\min\{p_r(x)\}$  è non decrescente all'aumentare di  $r$ ) e quindi

$$\sup\{\bar{v}_r : r \geq 0\} = \lim_{r \rightarrow \infty} \bar{v}_r$$

**Proposizione 13.3**

Supponiamo che  $x_r$  sia un punto di minimo globale di  $(P_r)$ . Ogni punto di accumulazione di  $\{x^r\}_r$  è un punto di minimo globale di  $(P)$ .

**Osservazione 13.26**

$x^r$  ammissibile, cioè  $x^r \in X$ ,  $\Rightarrow x^r$  punto di minimo globale di  $(P)$ . Infatti

$$\bar{v} \geq \bar{v}_{r_k} = p_{r_k}(x^k) = \inf_{x^k \in X} f(x^k)$$

, da cui  $\bar{v} = f(x^k)$  poiché  $x^k$  è ammissibile.

Quindi in genere,  $\{x^k\}$  è costituita da punti non ammissibili, cioè 'esterni' alla regione ammissibile  $X$  (da cui il nome di *penalizzazione esterna*). Dovremo accontentarci di una soluzione approssimata.

Sappiamo che:

$$x^k \text{ punto di minimo globale di } (P_r) \Rightarrow \nabla p_r(x^r) = 0$$

Calcoliamo esplicitamente il gradiente:

$$0 = \nabla p_r(x^k) = \nabla f(x^k) + \sum_{i=1}^n 2r_k g_i^+(x^k) \nabla g_i(x^k) + \sum_{j=1}^p 2r_k h_j(x^k) \nabla h_j(x^k)$$

$$\lambda_i^k = 2r_k g_i^+(x^k) \quad i = 1 \dots m \quad \text{e} \quad \mu_j^k = 2r_k h_j(x^k), \quad j = 1 \dots p$$

, sono i moltiplicatori associati a  $x^k$  per il problema (P)  
 (Attenzione, le altre condizioni KKT, cioè ammissibilità e complementarità in genere non sono soddisfatte  
 $[g_i(x^k) > 0 \Rightarrow \lambda_i^k > 0]$ )



### Teorema 13.27

Supponiamo che  $f, g_i, h_j$  siano differenziabili con continuità. Siano  $r_k \uparrow +\infty, \tau_k \downarrow 0$  e  $x^k \in \mathbb{R}^n$  un punto tale che  $\|\nabla p_{r_k}(x^k)\|_2 \leq \tau_k$ . Allora ogni punto di accumulazione  $x^*$  di  $\{x^k\}$  tale che

$$\{\nabla h_j(x^*)\}_{j=1 \dots p} \cup \{\nabla g_i(x^*)\}_{i \in I(\bar{x})} \quad \text{linearmente indipendenti}$$

soddisfa le condizioni KKT con i moltiplicatori

$$\lambda_i^*, \mu_j^*$$

dati da

$$\begin{aligned} \lambda_i^* &= \lim_{l \rightarrow +\infty} 2r_{k_l} g_i(x^{k_l}) \quad i = 1 \dots n \\ \mu_j^* &= \lim_{l \rightarrow +\infty} 2r_{k_l} h_j(x^{k_l}) \quad j = 1 \dots p \end{aligned}$$

dove  $\{x^{k_l}\}$  è una sottosuccessione tale che  $x^{k_l} \rightarrow x^*$



### Osservazione 13.28 (Problema mal condizionato)

Fare tendere  $r$  a  $+\infty$  può portare a dei seri problemi di malcondizionamento:  $\nabla_r^2$  può risultare mal condizionato per  $r$  grande. Prendiamo il problema visto nell'esempio iniziale :

$$Pr(x) = x_1^2 + x_2^2 + r(x_1 + x_2 - 1)^2$$

$$\nabla_r^2 p_r(x) = \begin{pmatrix} 2(1+r) & 2r \\ 2r & 2(1+r) \end{pmatrix}$$

Gli autovalori  $\lambda_1^r = 2$ ,  $\lambda_2^r = 2(1+2r)$ .

Gli autovettori corrispondenti sono  $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$ ,  $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ , ma  $\lambda_2^r \xrightarrow{r \rightarrow \infty} +\infty$

#### 13.2.2.2 Metodo dei moltiplicatori

Funziona solo con vincoli di uguaglianza, ma basta introdurre le variabili di slack  $s_i$ , dette anche variabili di scarto, per le disuguaglianze, ossia

$$g_i(x) \leq 0 \quad \longrightarrow \quad g_i(x) + s_i^2 = 0$$

Consideriamo il problema

$$(P_{eq}) \quad \min\{f(x) : h_j(x) = 0 \quad j = 1 \dots p\} \quad (f, h \text{ differenziabili con continuità})$$

Supponiamo di perturbare i vincoli di uguaglianza  $h_j(x)$  di una quantità  $\delta_j$ , a questo punto la funzione di penalizzazione esterna diventa:

$$f(x) + r \sum_{j=1}^p [h_j(x) - \delta_j]^2 = f(x) + \sum_{j=1}^p \underbrace{(-2r\delta_j)}_{\mu_j} h_j(x) + r \sum_{j=1}^p h_j^2(x) + r \underbrace{\sum_{j=1}^p \delta_j^2}_{\text{costante}}$$

Abbiamo ottenuto la Lagrangiana aumentata (dalla penalizzazione)

**Definizione 13.29 (Lagrangiana aumentata per  $(P_{eq})$ )**

$$L_r(x, \mu) = f(x) + \sum_{j=1}^p \mu_j h_j(x) + r \sum_{j=1}^p h_j^2(x)$$

Adesso abbiamo qualcosa in più con cui giocare: i moltiplicatori. Nella precedente funzione obiettivo non erano presenti. Supponiamo che  $\bar{x}$  e  $\bar{\mu}$  soddisfino le condizioni KKT. In questo caso si ottiene che  $\bar{x}$  è un punto stazionario della Lagrangiana aumentata vista come funzione della sola  $x$  dove i moltiplicatori sono fissati.

**Osservazione 13.30**

$L_r(\cdot, \mu)$  è la funzione esterna di

$$(PL_{eq}(\mu)) \quad \min \{ f(x) + \sum_{j=1}^p \mu_j h_j(x) : h_j(x) = 0 \quad j = 1 \dots p \}$$

dove  $\mu \in \mathbb{R}^p$  è fissato, mentre  $L_r$  è la funzione lagrangiana di

$$(P_{eq}(r)) \quad \min \{ f(x) + r \sum_{j=1}^p h_j^2(x) : h_j(x) = 0 \quad j = 1 \dots p \}$$

dove  $r > 0$  è fissato. Sia  $PL_{eq}(\mu)$  che  $(P_{eq}(r))$  sono problemi equivalenti a  $(P_{eq})$ , in quanto gli addendi della penalizzazione, quando ci si trova all'interno della regione ammissibile, si annullano.

**Teorema 13.31**

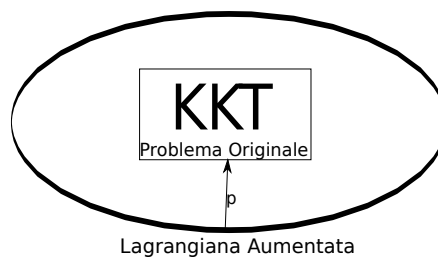
Se  $\bar{x} \in \mathbb{R}^n, \bar{\mu} \in \mathbb{R}^p$  soddisfano le condizioni di KKT per  $(P_{eq})$ , allora  $\bar{x}$  è un punto stazionario di  $L_r(\cdot, \bar{\mu})$

*Dimostrazione.*

$$\nabla_x L_r(\bar{x}, \bar{\mu}) = \nabla f(\bar{x}) + \sum_{j=1}^p \bar{\mu}_j \nabla h_j(\bar{x}) + 2r \sum_{j=1}^p h_j(\bar{x}) \nabla h_j(\bar{x}) \stackrel{KKT \Rightarrow h_j(\bar{x})=0}{=} \nabla f(\bar{x}) + \sum_{j=1}^p \bar{\mu}_j \nabla h_j(\bar{x}) \stackrel{(KKT_1)}{=} 0$$

□

Questo risultato non è in genere vero per la funzione esterna.



In parole povere un punto stazionario del problema della lagrangiana aumentata è candidato ad essere un punto che rispetta le condizioni KKT del problema originario e, per  $p \rightarrow \infty$ , il cerchio, che per l'appunto contiene i punti stazionari della lagrangiana aumentata, tende a restringersi: è questa l'idea che sta alla base del algoritmo che vedremo fra poco.

**Proposizione 13.4**

Siano  $r_k \uparrow +\infty$  (successione che va a infinito),  $\{\mu^k\}_k$  successione limitata e supponiamo che

$$(P_{eq}(x^k)) \quad \min\{L_r(x, \mu^k) : x \in \mathbb{R}^n\}$$

ammetta un punto di minimo globale  $x^k$  per ogni  $k$ . Allora ogni punto di accumulazione di  $\{x^k\}_k$  è un punto di minimo globale di  $(P_{eq})$ . Allora ogni punto di accumulazione della successione  $x^k$  è un punto di minimo globale del problema  $(P_e)$ , quello con i soli vincoli di uguaglianza.

 **Osservazione 13.32**

$$\nabla_x L_r(x, \mu) = \nabla f(x) + \sum_{j=1}^p \frac{(\mu_j + 2rh_j(x))}{2r} \nabla h_j(x)$$

Questa osservazione ci suggerisce una tecnica per aggiornare i moltiplicatori

 **Teorema 13.33**

Siano  $\{\mu^k\} \subseteq \mathbb{R}^p$  limitata,  $r_k \uparrow +\infty$ ,  $\tau_k \downarrow 0$  e  $x^k \in \mathbb{R}^n$  tale che  $\|\nabla L_{r_k}(x^k, \mu^k)\|_2 \leq \tau_k$ . Allora ogni punto di accumulazione  $x^*$  di  $\{x^k\}$  tale che  $\{\nabla h_j(x^*)\}_{j=1 \dots p}$  siano linearmente indipendenti soddisfa le condizioni KKT insieme ai moltiplicatori

$$\mu_j^* = \lim_{l \rightarrow \infty} (\mu_j k^{k_l} + 2r_{k_l} h_j(x^{k_l}))$$

dove  $\{x^{k_l}\}$  è una (qualsiasi) sottosuccessione per cui  $x^{k_l} \xrightarrow{l \rightarrow +\infty} x^*$

**Metodo dei moltiplicatori**

1. Scegliamo  $\delta \in (0, 1)$ ,  $\beta > 1$ ,  $r_0 > 0$ ,  $\mu^0 \in \mathbb{R}^p$ , porre  $viol = +\infty$ ;  $k = 0$
2. Calcolare  $x^k \in \operatorname{argmin}\{L_{r_k}(x, \mu^k) : x \in \mathbb{R}^n\}$
3. Se  $viol(x^k) = \max_{j=1 \dots p} |h_j(x^k)| = 0$  allora STOP,  $(x^k, \mu^k)$  soddisfano KKT
4. Se  $viol(x^k) > \delta viol$  allora il parametro di penalizzazione va aumentato

$$r_{k+1} = \beta r_k \quad \mu^{k+1} = \mu^k \tag{13.2}$$

Altrimenti


$$r_{k+1} = r_k \quad \text{e} \quad \mu_j^{k+1} = \mu_j^k + 2r_k h_j(x^k), \quad j = 1 \dots p$$

5.  $viol = viol(x^k)$ ,  $k = k + 1$  e ritornare a 2

 **Osservazione 13.34**

Se  $viol(x^k) = 0$  (cioè  $x^k$  ammissibile), allora  $\nabla_x L_{r_k}(x^k, \mu^k) = 0 (\approx 0)$  garantisce che  $x^k$  e  $\mu^k$  soddisfano le condizioni KKT per  $(P_{eq})$



 **Osservazione 13.35**

L'aggiornamento (13.2) può verificarsi al più un numero finito di volte consecutive se la condizione di stop viene approssimata da  $\text{viol}(x^k) \leq \epsilon$  per qualche tolleranza  $\epsilon > 0$  (altrimenti si avrebbe  $\mu^k \approx \text{cost}$  definitivamente, ma  $x^k$  non convergerebbe ad una soluzione ammissibile come garantito dai risultati precedenti).

In linea di principio anche qua si può andare a  $+\infty$  ma avremo una buona approssimazione dei moltiplicatori.

### 13.2.2.3 Penalizzazione interna: metodi barriera

Vediamo un altro metodo per la soluzione di problemi con vincoli non (necessariamente) convessi.

$$(P_{in}) \quad \min\{f(x) : x \in X\} \quad X = \{x \in \mathbb{R}^n : g_i(x) \leq 0 \quad i = 1 \dots n\}$$

$f, g_i : \mathbb{R}^n \rightarrow \mathbb{R}$  differenziabili con continuità

Risolviamo dunque solamente problemi in cui  $X$  è formato da vincoli di disuguaglianza. Ci muoviamo inoltre sotto due ipotesi:

1.  $X^0 = \{x \in \mathbb{R}^n : g_i(x) < 0 \quad i = 1 \dots n\} \neq \emptyset$
2.  $\forall x \in X \forall \epsilon > 0 \quad \exists y \in X^0$  tale che  $\|x - y\|_2 \leq \epsilon$

Le idee alla base della penalizzazione interna sono due:

- Aprire la regione ammissibile, ossia eliminarne il bordo. Come vedremo dalle figure, questa operazione non è sempre possibile, ma quando si lavora con sistemi di vincoli, abbiamo la garanzia che questa operazione si possa fare senza particolari problemi

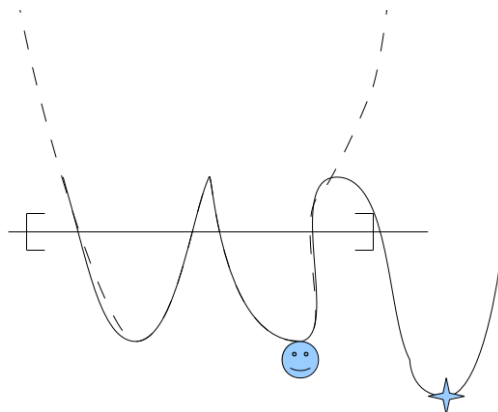


Nel caso della figura qui sopra, sulla parte sinistra abbiamo un chiuso, che può essere aperto.



In questo altro caso, la regione ammissibile è data da un insieme chiuso ed un segmento: non riusciamo a trovare un aperto che contenga tutti i punti senza introdurre nuovi punti della regione ammissibile.

- Dobbiamo introdurre una funzione obiettivo modificata, che faccia schizzare a  $+\infty$  il valore della funzione obiettivo ai bordi, in modo da garantire convergenza.



Vediamo che il minimo globale nell'area non vincolata è diverso dal minimo globale definito nella regione ammissibile: è qui che ci viene in aiuto la funzione di penalizzazione, evidenziata dalla linea tratteggiata.

Il fatto che si utilizzi un aperto e che la nuova funzione  $f'(x)$  ai bordi vada a  $+\infty$  ci garantisce che per trovare i punti stazionari  $\bar{x}$  sia sufficiente verificare che

$$\nabla f'(\bar{x}) = 0$$

Con queste due tecniche possiamo quindi utilizzare i metodi visti per l'ottimizzazione non vincolata.

### Osservazione 13.36

$g_i$  convesse e Ipotesi (1)  $\Rightarrow$  Ipotesi(2)

### Osservazione 13.37

$g_i$  continue  $\Rightarrow X^0$  aperto

L'ipotesi 2) richiede che ogni punto di  $X \setminus X^0$  sia il limite di una opportuna successione di punti di  $X^0$ . Ci riconduciamo ad una ottimizzazione non vincolata. Abbiamo bisogno di una funzione, detta *funzione barriera*

### Definizione 13.38 (Funzione barriera)

Funzione definita su  $X^0$  che tende a  $+\infty$  avvicinandosi a punti di  $X \setminus X^0$

$$B : X^0 \rightarrow \mathbb{R} \quad B(x) \geq 0 \quad \forall x \in X^0, \quad B(x) \rightarrow +\infty \text{ se } x \rightarrow \bar{x} \text{ con } \bar{x} \in X \setminus X^0$$

Si noti che

$$\bar{x} \in X \setminus X^0 \iff \exists i \text{ t.c. } g_i(\bar{x}) = 0$$

### Definizione 13.39 (Barriera inversa)

$$B(x) = - \sum_{i=1}^n \frac{1}{g_i(x)}$$

### Definizione 13.40 (Barriera logaritmica)

$$B(x) = - \sum_{i=1}^n \log(-g_i(x))$$

### Osservazione 13.41

In entrambi i casi:

$$g_i \text{ convesse} \Rightarrow B \text{ è convessa}$$

(fare come esercizio: le funzioni sono monotone, prendere una funzione alla volta)

Le funzioni barriera ci permettono di considerare, al posto del problema di partenza

$$(PB)_\epsilon \quad \min\{f(x) + \epsilon B(x) : x \in X^0\}$$

La funzione barriera è sempre  $> 0$ , quindi:

$$f(x) + \epsilon B(x) > f(x)$$

Avvicinandosi al bordo  $B(x)$  tende a  $+\infty$  e quindi tutto tende a  $+\infty$ .

Poiché l'insieme è aperto,  $(PB)_\epsilon$  è in pratica un problema di ottimizzazione non vincolata, visto che dato un punto interno a  $X^0$  possiamo muoverci in qualunque direzione (con un passo di spostamento adeguato) rimanendo nella regione ammissibile. Inoltre, i punti di minimo soddisfano  $\nabla(f(x) + \epsilon B(x)) = 0$  e utilizzando i metodi di discesa visti nel Capitolo 6 per risolvere i problemi di ottimizzazione non vincolata, partendo da un punto interno ad  $X^0$  verrà generata una sequenza di punti apparenti ad  $X^0$ , ed anche il minimo trovato da questi metodi sarà dentro  $X^0 \subset X$ . Da questo deriva il nome di *metodi del punto interno*.

Ma bisogna fare attenzione: è necessario partire da un punto appartenente a  $X^0$ , e individuarne uno non è sempre una cosa banale (nessun punto è attivo).



#### Osservazione 13.42

Se  $f$  e  $g_i$  sono convesse anche  $B(x)$  è convessa. Quindi cadiamo nell'ottimizzazione convessa.

Idea dei metodi : avere una successione  $\epsilon_k \downarrow 0 \quad x^k \in \operatorname{argmin}\{f(x) + \epsilon_k B(x) : x \in X^0\}$  utilizzando  $x^{k-1}$  come punto iniziale.

Notare che la successione  $x^k$  è contenuta nella regione ammissibile  $X$  (penalizzazione interna).

#### Esempio 13.43

Consideriamo il seguente problema.

$$n = 1$$

$$\min\{x : 1 - x \leq 0\}$$

$\bar{x} = 1$  è l'unico punto di minimo globale.

Proviamo ad applicare questo metodo, utilizzando la barriera logaritmica:

$$\min\{x - \epsilon \log(x - 1) : \underbrace{x > 1}_{X^0}\}$$

per  $x \rightarrow 0$  il rapporto tra  $x$  e  $\epsilon \log(x - 1)$  va a  $+\infty$ .

Calcoliamo il punto in cui il gradiente si annulla

$$\nabla(x - \epsilon \log(x - 1)) = 0 \quad \Rightarrow \quad 1 - \frac{\epsilon}{x - 1} = 0 \quad \Rightarrow \quad x = 1 + \epsilon (> 1)$$

Quindi la soluzione del problema barriera è

$$X(\epsilon) = 1 + \epsilon \xrightarrow{\epsilon \downarrow 0} 1$$

La funzione obiettivo in  $X(\epsilon)$  è

$$X(\epsilon) - \epsilon \log(X(\epsilon) - 1) = 1 + \epsilon - \epsilon \log \epsilon$$

Più  $\epsilon$  tende a 0 più abbiamo un'esplosione.

**Esempio 13.44 (Esempio in due variabili)**

$$(P_{in}) \quad \min\left\{\frac{1}{2}x_1^2 + \frac{1}{2}x_2^2 : 2 - x_1 \leq 0\right\}$$

$\bar{x} = (2, 0)$  risulta essere l'unico punto di minimo globale di  $(P_{in})$ : infatti le condizioni KKT per  $(P_{in})$  risultano essere

$$\begin{cases} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \lambda \begin{pmatrix} -1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\ \lambda_1(2 - x_1) = 0 \\ x_1 \geq 2, \lambda_1 \geq 0 \end{cases} \Rightarrow \begin{cases} x_1 = \lambda_1 \\ x_2 = 0 \\ \lambda_1(2 - x_1) = 0 \\ x_1 \geq 2, \lambda_1 \geq 0 \end{cases} \Rightarrow \begin{cases} x_1 = \lambda_1 \\ x_2 = 0 \\ \lambda_1(2 - x_1) = 0 \\ \lambda_1 \geq 0 \end{cases} \Rightarrow \begin{cases} x_1 = \lambda_1 = 2 \\ x_2 = 0 \end{cases}$$

Cerchiamo di risolvere il problema tramite la tecnica della barriera:

$$(PB_\epsilon) \quad \min\left\{\frac{1}{2}(x_1^2 + x_2^2) - \epsilon \log(x_1 - 2) : x_1 > 2\right\}$$

Il punto di minimo si trova annullando il gradiente

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \nabla\left(\frac{1}{2}(x_1^2 + x_2^2) - \epsilon \log(x_1 - 2)\right) = \begin{cases} x_1 - \frac{\epsilon}{x_1 - 2} \\ x_2 \end{cases} \rightarrow \begin{cases} x_1^2 - 2x_1 - \epsilon = 0 \\ x_2 = 0 \end{cases} \rightarrow x_1 = 1 \pm \sqrt{1 + \epsilon} \xrightarrow{x_1 > 2} x_1 = 1 + \sqrt{1 + \epsilon}$$

$X(\epsilon) = (1 + \sqrt{1 + \epsilon}, 0)$  risulta essere l'unico punto di minimo globale di  $(PB_\epsilon)$  ed inoltre  $X(\epsilon) \xrightarrow{\epsilon \downarrow 0} (2, 0)$

Passiamo ad una formalizzazione di quanto visto negli esempi

**Teorema 13.45**

Ogni punto di accumulazione della successione  $\{x^k\}$  è un punto di minimo globale del problema di partenza  $(P_{in})$

D'ora in avanti considereremo il caso della barriera logaritmica:

$$(PB_\epsilon) \quad \min\{q_\epsilon(x) = f(x) - \epsilon \sum_{i=1}^n \log(-g_i(x)) : x \in X^0\}$$

Nel caso convesso ci possiamo spingere oltre.

**Teorema 13.46**

Supponiamo che

- $f, g_i$  siano convesse
- l'insieme dei punti di minimo (globale) di  $(P_{in})$  sia non vuoto e compatto

Allora:

1. La successione  $\{x^k\}_k$  ammette almeno un punto di accumulazione (non stiamo parlando della regione ammissibile, stiamo usando la barriera)
2.  $f(x^k) \rightarrow f^* \quad q_\epsilon(x^k) \rightarrow f^*$ , dove  $f^*$  è il valore ottimo di  $(P_{in})$

**Legame con KKT**

Cosa possiamo dire sui moltiplicatori di Lagrange? Sia  $x(\epsilon)$  è punto di minimo di  $(PB_\epsilon)$ . Consideriamo il gradiente:

$$0 = \nabla q_\epsilon(x(\epsilon)) = \nabla f(x(\epsilon)) - \epsilon \sum_{i=1}^n \frac{1}{-g_i(x(\epsilon))} \nabla g_i(x(\epsilon)) =$$

$$\nabla f(x(\epsilon)) + \sum_i^n \underbrace{\left( \frac{-\epsilon}{g_i(x(\epsilon))} \right)}_{\substack{\text{moltiplicatori} \\ \text{associati a } x(\epsilon) \\ \text{nel problema} \\ \text{originale } (\lambda_i)} \nabla g_i(x(\epsilon))$$

Sotto opportune ipotesi su un punto di minimo locale  $x^*$  di  $(P_{in})$  (tra cui la lineare indipendenza dei vettori  $\{\nabla g_i(x^*)\}_{i \in I(x^*)}$  [che garantisce l'unicità dei moltiplicatori]), è possibile dimostrare che:

- in un opportuno intorno di  $x^*$  esiste un unico punto di minimo locale  $x(\epsilon)$  di  $(PB_\epsilon)$  per  $\epsilon$  sufficientemente piccolo
- $x(\epsilon) \xrightarrow{\epsilon \downarrow 0} x^*$  e  $\lambda_i(\epsilon) \xrightarrow{\epsilon \downarrow 0} \lambda_i^*$  sono i moltiplicatori associati a  $x^*$ .

**Esempio 13.47**

Tornando all'esempio su 2 dimensioni abbiamo

$$g_1(x) = 2 - x_1$$

e quindi

$$\nabla_i(\epsilon) = \frac{-\epsilon}{g_i(x(\epsilon))} = \frac{-\epsilon}{1 - \sqrt{1 + \epsilon} - 1} \xrightarrow{\epsilon \downarrow 0} 2$$

dove 2 è il moltiplicatore di Lagrange  $\lambda_1$  associato a  $x^* = (2, 0)$ .

$x(\epsilon)$  e  $\lambda_i(\epsilon) = \frac{\epsilon}{-g_i(x(\epsilon))}$  soddisfano:

$$\nabla f(x(\epsilon)) + \sum_{i=1}^n \lambda_i(\epsilon) \nabla g_i(x(\epsilon)) = 0 \quad (\text{per la scelta di } \lambda_i(\epsilon))$$

$$g_i(x(\epsilon)) \leq 0 \quad i = 1 \dots n \quad (\text{poiché } x(\epsilon) \in X^0 [\text{quindi in realtà } g_i(x(\epsilon)) < 0])$$

$$\lambda_i(\epsilon) \geq 0 \quad (\text{conseguenza})$$

Manca la complementarità: infatti in questo caso abbiamo

$$\lambda_i(\epsilon) g_i(x(\epsilon)) = -\epsilon \neq 0$$

ossia

$$\lambda_i(\epsilon) (-g_i(x(\epsilon))) = \epsilon > 0$$

Non è rispettata la condizione di complementarità: se valesse avremmo avuto una soluzione del problema originale. A noi interessa proprio la condizione modificata della complementarità.

**KKT approssimate**

Utilizzando le variabili di scarto (slack)

$$\nabla f(x) + \sum_{i=1}^m \lambda_i \nabla g_i(x) = 0$$

$$g_i(x) + s_i = 0 \quad i = 1 \dots n$$

$$\lambda_i s_i = \epsilon \quad i = 1 \dots n$$

$$\lambda_i s_i \geq 0 \quad i = 1 \dots n$$

 **Osservazione 13.48**

Piccola parentesi sui vincoli di uguaglianza (discorso dell'aperto): 2 possibilità:

- Considerarli come vincoli generici

$$\min\{f(x) + \epsilon B(x) : x \in X^0, h_j(x) = 0 \quad j = 1 \dots p\}$$

- Utilizzare la penalizzazione esterna

$$\min\{f(x) + \epsilon B(X) + r \sum_{j=1}^p h_j^2(x) : x \in X^0\}$$

Chiusa parentesi

### Metodi primali duali del punto interno

Abbiamo  $2m + n$  equazioni +  $i$  vincoli di non negatività. Possiamo cercare di risolvere questo sistema di equazioni adattando Newton Rapsson in modo da soddisfare la condizione

$$\lambda_i s_i \geq 0 \quad i = 1 \dots n$$

L'idea è considerare  $\lambda_i$  ed  $s_i$  positivi

$$\lambda_i, s_i > 0$$

La complementarità non la possiamo avere, dobbiamo avere  $\epsilon > 0$ . Possiamo fare un passo nella direzione di newton mantenendo la positività (accorciamo il passo da unitario a qualcosa di inferiore). Questi sono detti *Metodi primali-duali del punto interno*. Questi metodi funzionano bene con programmazione lineare e quadratica. Nel caso di programmazione lineare sono quelli che hanno rimpiazzato il semplice. Vediamo come si può modificare il metodo nel caso della programmazione lineare

### Metodi primali duali per la programmazione lineare

$$\begin{aligned} \text{(Primale)} \quad & \max\{c^T x : Ax \leq b\} \\ \text{(Duale)} \quad & \min\{b^T \lambda : A^T \lambda = c, \lambda \geq 0\} \end{aligned}$$

Condizioni KKT del duale:

$A^T \lambda = c$	(Ammissibilità duale)
$Ax \leq b$	(Ammissibilità del problema originale)
$\lambda_i (b_i - a_i^T x) = 0 \quad i = 1 \dots n$	
$\lambda_i \geq 0 \quad i = 1 \dots n$	

Possiamo riscrivere in forma equivalente, ponendo  $s = b - Ax$

$$\overline{\text{(Primale)}} \quad \max\{c^T x : Ax + s = b\}$$

$A^T \lambda - c = 0$	(Ammissibilità duale)
$Ax + s - b = 0$	(Ammissibilità del problema originale)
$\lambda_i s_i = 0 \quad i = 1 \dots n$	
$\lambda_i \geq 0, s_i \geq 0 \quad i = 1 \dots n$	

Condizioni KKT approssimate: sostituisce lo 0 del complementarità con un  $\epsilon$ .

$A^T \lambda - c = 0$	(Ammissibilità duale)
$Ax + s - b = 0$	(Ammissibilità del problema originale)
$\lambda_i s_i = \epsilon$	$i = 1 \dots n$
$\lambda_i \geq 0, s_i \geq 0$	$i = 1 \dots n$

Sia  $F : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n$  data da

$$F = (x, s, \lambda) = \begin{pmatrix} A^T \lambda - c \\ Ax + s - b \\ \Lambda S e \end{pmatrix}$$

dove

- $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_1)$  : matrice diagonale con i  $\lambda_i$  sulla diagonale

$$\begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n \end{pmatrix}$$

- $S = \text{diag}(\{s_1, \dots, s_n\})$
- $e = (1, \dots, 1)^T \in \mathbb{R}^n$

Si ha che:

$$\text{Condizioni KKT approximate} \iff F(x, \lambda, s) = \begin{pmatrix} 0 \\ 0 \\ \epsilon e \end{pmatrix} : \lambda, s \geq 0$$

Il metodo di Newton-Raphson per la risoluzione di  $F(x, S, \lambda) = 0$  fornisce la direzione  $d = (d_x, d_s, d_\lambda)$  che risolve il sistema

$$(Nd) \quad JF(x, S, \lambda)d = -F(x, S, \lambda)$$

Se  $(x, s)$  è ammissibile per  $(\bar{P})$  con  $s_i > 0 \ i = 1 \dots n$  e  $\lambda$  è ammissibile per  $(D)$  con  $\lambda_i > 0 \ i = 1 \dots n$ , il sistema  $(Nd)$  diventa

$$(Nd) \quad \begin{pmatrix} 0 & 0 & A^T \\ A & I & 0 \\ 0 & \Lambda & S \end{pmatrix} \begin{pmatrix} d_x \\ d_s \\ d_\lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -\Lambda S e \end{pmatrix}$$

il punto corrente è

$$(X, \lambda S) \rightarrow Ax + s = b \quad A^T \lambda_i > 0 \quad s_i > 0$$

Affinché  $\begin{pmatrix} x' \\ s' \\ \lambda' \end{pmatrix} = \begin{pmatrix} x \\ s \\ \lambda \end{pmatrix} + \alpha \begin{pmatrix} d_x \\ d_s \\ d_\lambda \end{pmatrix}$  soddisfi la richiesta  $\lambda'_i > 0, s'_i > 0 \ i = 1 \dots n$  è molto probabile che risulti  $\alpha \ll 1$ .

L'idea è sostituire  $-\Lambda S e$  con  $-\Lambda S e + \delta \mu e$  dove  $\delta \in [0, 1]$  e  $\mu = \sum_{i=1}^n \frac{\lambda_i s_i}{n}$  in modo che la direzioni di Newton "punti" verso punti  $(x', s', \lambda')$  per cui  $\lambda'_i s'_i \approx \delta \mu > 0$ .



#### Nota

Quello che vuole sapere che con epsilon positivo si può passare tramite Newton-Raphson ai metodi primali duali.

## Metodo Primale Duale

1. Scegliere  $(x^0, S^0, \lambda^0)$  tale che:

$$Ax^0 + s^0 = b, A^T \lambda^0 = c, \lambda_i^0 > 0, s_i^0 > 0 \quad i = 1 \dots n; k = 0$$

2. Risolvere il sistema lineare:

$$\begin{pmatrix} 0 & 0 & A^T \\ A & I & O \\ 0 & \Lambda^k & S^k \end{pmatrix} \begin{pmatrix} d_x^k \\ d_s^k \\ d_\lambda^k \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \Lambda^k S^k e + \delta_k \mu_k e \end{pmatrix}$$

dove

$$\Lambda^k = \text{diag}\{\lambda_1^k, \dots, \lambda_n^k\}, S^k = \text{diag}\{s_1^k, \dots, s_n^k\}, \mu^k = \sum_{i=1}^n \frac{\lambda_i^k s_i^k}{n}, \delta_k \in [0, 1]$$

3. Calcolare  $\alpha^k > 0$  tale che

$$s_i^k + \alpha_k (d_s^k)_i, \lambda_i^k + \alpha + k (d_\lambda^k)_i > 0 \quad i = 1 \dots n$$

4. Porre

$$(x^{k+1}, s^{k+1}, \lambda^{k+1}) = (x^k, s^k, \lambda^k) + \alpha_k (d_x^k, d_s^k, d_\lambda^k) \quad k = k + 1$$

e ritornare a 2)



# 14 Calcolo di autovalori e autovettori

Possibili percorsi

- definizione: radici dell'equazione caratteristica  $p(\lambda) = 0$ . Calcolare i coefficienti di  $p(\lambda)$ : sono stati sviluppati metodi con costo  $O(n^3)$  con il problema che i coefficienti siano malcondizionati: cambiare l'elemento di una matrice può cambiare drasticamente i valori trovati. Per questo non è il caso di usare questo metodo. Per le matrici hermitiane si potrebbe usare comunque, non soffre del problema del condizionamento.
- riduzione di  $A$  (numero finito di passi) per similitudine ad una forma per la quale sia più facile trovare gli autovalori. Con  $O(n^3)$  portiamo la matrice a struttura di Hessemberg ( $i_{ij} = 0$  per  $i \geq j + 1$ ). Nel caso Hermitiano, la forma di Hessemberg può essere portata in forma tridiagonale (zeri anche nella parte superiore). Inoltre le matrici di passaggio sono unitarie. Per le matrici di Hessemberg sono stati sviluppati metodi iterativi.
  - Metodi QR:  $A \rightarrow$  Hessemberg  $\rightarrow$  triangolare superiore
- Metodi iterativi direttamente sulla matrice  $A$ . Prodotto matrice per vettore: convergono a sottoinsiemi di autovalori/autovettori. Indicati per problemi di grandi dimensioni. Metodi possibili: metodo delle potenze, metodi delle iterazioni inverse.

Se gli autovalori sono non reali, come fanno a venire fuori gli autovalori sulla diagonale?

## 14.1 Condizionamento del problema



### Nota

Risultati di localizzazione: non da studiare.

Il Teorema di Bauer–Fike riguarda la perturbazione degli autovalori di una matrice diagonalizzabile a valori complessi. In sostanza, stabilisce un limite superiore alla deviazione di un autovalore calcolato su una matrice perturbata rispetto a quello calcolato sulla matrice esatta. (Informalmente, quello che dice è che la sensibilità degli autovalori è stimata dal condizionamento  $\mu(t)$  della matrice composta dagli autovettori).<sup>1</sup>



### Teorema 14.1 (Teorema di Bauer-Fike)

Sia  $\|\cdot\|$  una norma matriciale indotta che verifichi la seguente proprietà

$$\|D\| = \max_{i=1,\dots,n} |d_{ii}|$$

per ogni matrice diagonale  $D \in \mathbb{C}^{n \times n}$  (una tale norma viene detta norma assoluta, le norme  $\|\cdot\|_1$ ,  $\|\cdot\|_2$  e  $\|\cdot\|_\infty$  sono assolute). Sia  $A \in \mathbb{C}^{n \times n}$  una matrice diagonalizzabile, cioè tale che

$$A = TDT^{-1}$$

con  $D$  diagonale e  $T$  non singolare. Se  $\delta A \in \mathbb{C}^{n \times n}$ , e  $\xi$  è un autovalore di  $A + \delta A$ , allora esiste almeno un autovalore  $\lambda$  di  $A$  tale che

$$|\lambda - \xi| \leq \mu(T) \|\delta A\|,$$

dove  $\mu(T) = \|T\| \|T^{-1}\|$ .

<sup>1</sup>Tradotto da Wikipedia, [http://en.wikipedia.org/wiki/Condition\\_number](http://en.wikipedia.org/wiki/Condition_number)

Condizionamento degli autovalori dipende dagli autovettori. Condizionamento = 1 se  $\|\cdot\|_2$  e  $T$  unitaria allora

$$1 = \mu(T) = \frac{\|T\|_2 \|T^H\|}{1 \cdot 1}$$

(Matrici normali)

Queste sono quantità assolute rispetto a quelle di quelle viste nella risoluzione dei sistemi lineari.

$$\frac{\|\delta A\|}{\|A\|} \approx 10^{-15}$$

Supponiamo che il problema sia non ben condizionato  $\mu(T) = 10^8$   $\|\delta A\| = 10^{-7}$  vupo; dire che non si riesce ad approssimare bene autovalori con  $10^{-7}$ . Gli autovalori di modulo piccolo danno maggiori problemi di approssimazione rispetto a quelli più grandi.

Sono stati trovati risultati relativi ai singoli autovalori, e non rispetto a tutta la matrice

*Dimostrazione.*

- Se  $\xi$  fosse autovalore di  $A$ , la tesi sarebbe verificata. Altrimenti la matrice  $A - \xi I$  risulta non singolare e dalla relazione

$$(A + \delta A)y = \xi y$$

dove  $y$  è autovettore di  $(A + \delta A)y$ , si ha

$$\delta A y = -(A - \xi I)y \implies (A - \xi I)^{-1} \delta A y = -y$$

- Dimostriamo adesso che

$$\|(A - \xi I)^{-1} \delta A\| \geq 1 \quad (14.1)$$

Per definizione di norma abbiamo che

$$\|Z\| = \max_{y \neq 0} \frac{\|Zy\|}{\|y\|}$$

Prendiamo

$$Z = (A - \xi I)^{-1} \delta A$$

Da quanto visto prima abbiamo che

$$\|Zy\| = \|-y\| = \|y\|$$

quindi abbiamo un  $y \neq 0$  ( $y$  è autovettore) per cui

$$\frac{\|Zy\|}{\|y\|} = 1$$

Possiamo affermare che

$$\|Z\| = \max_{y \neq 0} \frac{\|Zy\|}{\|y\|} \geq 1$$

- Facciamo delle trasformazioni algebriche:

$$\begin{aligned} (A - \xi I)^{-1} &= (TDT^{-1} - \xi TT^{-1})^{-1} = ((TD - \xi T)T^{-1})^{-1} = (T(D - \xi I)T^{-1})^{-1} = ((D - \xi I)T^{-1})^{-1}T^{-1} \\ &= T(D - \xi I)^{-1}T^{-1} \end{aligned}$$

- Visto il risultato in (14.1), e poiché le norme sono submoltiplicative, abbiamo

$$1 \leq \|T(D - \xi I)^{-1}T^{-1} \delta A\| \leq \|T\| \|T^{-1}\| \|(D - \xi I)^{-1}\| \|\delta A\|,$$

- Poiché  $\|\cdot\|$  è una norma assoluta, dunque  $\|D\| = \max d_{i,i}$  abbiamo

$$\begin{aligned} 1 &\leq \mu(T) \cdot \max_{i=1, \dots, n} \{|\lambda_i - \xi|^{-1}\} \cdot \|\delta A\| \\ 1 &\leq \mu(T) \frac{1}{\min_{1, \dots, n} |\lambda_i - \xi|} \|\delta A\| \end{aligned} \quad (14.2)$$

in cui i  $\lambda_i, i = 1, \dots, n$  sono gli autovalori di  $A$  e quindi gli elementi principali di  $D$ . Dalla (14.2) segue che

$$\min_{i=1, \dots, n} |\lambda_i - \xi| \leq \mu(T) \|\delta A\|,$$

da cui la tesi.

□

**Osservazione 14.2**

Se  $A$  è una matrice normale, allora  $T$  è unitaria, per cui  $\mu_2(T) = 1$  e dal teorema appena enunciato si ha

$$|\lambda - \epsilon| \leq \|\delta A\|_2$$

ossia il problema del calcolo degli autovalori per matrici normali è ben condizionato per tutti gli autovalori.

## 14.2 Metodo delle potenze

Il *metodo delle potenze* è un classico metodo iterativo per approssimare l'autovalore di modulo massimo di una matrice e il corrispondente autovettore. Sulla base di questo metodo sono stati sviluppati altri metodi che sono particolarmente adatti per approssimare gli autovalori di matrici sparse di grosse dimensioni. È facile dimostrare la convergenza del metodo nel caso che la matrice sia diagonalizzabile e abbia un solo autovalore di modulo massimo.

**Esempio 14.3**

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \quad \lambda_1 = 3 \quad \lambda_2 = 1$$

Se rappresentiamo le direzioni in cui si trovano gli autovettori, sono ortogonali, perchè la matrice è ortogonale. Se si moltiplica  $y_1 = Ay_0$  con  $y_0$  non autovettore. Si calcola a sua volta  $y_2$  sempre con  $y_2 = Ay_1$ . Si converge a  $\lambda_1 = 3$  Vettori di lunghezza che si avvicinano alla bisettrice del primo e del terzo quadrante. Ci si avvicina alla direzione dell'autovettore più grande. Si nota che ad esempio  $y_{10} = (\alpha\beta)$  si vede che

$$\frac{y_1^{(10)}}{y_1^{(9)}} \approx 3 \quad e \quad \frac{y_2^{(10)}}{y_2^{(9)}} \approx 3$$

Sia  $A \in C^{n \times n}$ , con  $n$  autovettori  $x_1, x_2, \dots, x_n$  linearmente indipendenti e autovalori  $\lambda_1, \lambda_2, \dots, \lambda_n$ . Vogliamo costruire una successione di questa forma:

$$y_k = Ay_{k-1} = A \cdot Ay_{k-2} = \dots = A^k y_0$$

Facciamo delle ipotesi su  $A$  e sui suoi autovalori che garantiscono questo tipo di convergenza:

- $A$  diagonalizzabile, quindi vale

$$T^{-1}AT = D$$

dove le colonne di  $T$  sono gli autovettori  $x_i$  linearmente indipendenti (1.21).

- Siano  $|\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_n|$

Il più grande autovalore deve essere distinto, ha quindi molteplicità algebrica 1 e non esistono altri autovalori con lo stesso modulo.

**Nota**

$C$  è un autovalore che attira, per questo introduciamo questa condizione, direzione che prevale sulle altre. Questa ipotesi è problematica perchè dovremmo già sapere la natura degli autovalori.

- Fissiamo un vettore  $y_0$  come base della successione e dato che gli autovettori  $x_1, \dots, x_n$  sono linearmente indipendenti, possiamo esprimerlo come

$$y_0 = \sum_{i=1}^n \alpha_i x_i \quad \alpha_i \neq 0$$

**Nota**

Anche qui abbiamo lo stesso problema: dovremmo già sapere la natura degli autovalori.

**Nota**

Poichè nel calcolo macchina abbiamo un'aritmetica non esatta gli ultimi 2 punti non sono strettamente necessari, si ha comunque convergenza.

Sotto queste ipotesi abbiamo che

$$\lim_{k \rightarrow \infty} \frac{y_j^{k+1}}{y_j^k} = \lambda_1$$

Infatti

$$y_k = A^k y_0 = A^k \left( \sum_{i=1}^n \alpha_i x_i \right) = \left( \sum_{i=1}^n \alpha_i A^k x_i \right) \stackrel{1.17}{=} \sum_{i=1}^n \alpha_i \lambda_i^k x_i = \alpha_1 \lambda_1^k x_1 + \sum_{i=2}^n \alpha_i \lambda_i^k x_i = \lambda_1^k \left[ \alpha_1 x_1 + \sum_{i=2}^n \alpha_i \left( \frac{\lambda_i}{\lambda_1} \right)^k x_i \right] \quad (14.3)$$

Indicate con  $y_r^{(k)}$  e con  $x_r^{(i)}$  le  $r$ -esime componenti dei vettori  $\mathbf{y}_k$  e  $\mathbf{x}_i$ , per gli indici  $j$  per cui  $y_j^{(k)} \neq 0$  e  $x_j^{(1)} \neq 0$ , si ha

$$\frac{y_j^{(k+1)}}{y_j^{(k)}} = \lambda_1 \frac{\alpha_1 x_j^{(1)} + \sum_{i=2}^n \alpha_i \left( \frac{\lambda_i}{\lambda_1} \right)^{k+1} x_j^{(i)}}{\alpha_1 x_j^{(1)} + \sum_{i=2}^n \alpha_i \left( \frac{\lambda_i}{\lambda_1} \right)^k x_j^{(i)}} \quad (14.4)$$

e poiché  $|\lambda_i/\lambda_1| < 1$ , per  $i \geq 2$  si ha

$$\lim_{k \rightarrow \infty} \frac{y_j^{(k+1)}}{y_j^{(k)}} = \lambda_1$$

Quindi da un certo indice  $k$  in poi l'autovalore  $\lambda_1$  può essere approssimato mediante uno dei rapporti  $y_j^{(k+1)}/y_j^{(k)}$

**Calcolo dell'autovettore  $x_1$** 

Con questo metodo si può approssimare anche l'autovettore  $\mathbf{x}_1$ . Dalla (14.3) risulta infatti

$$\lim_{k \rightarrow \infty} \frac{\mathbf{y}_k}{\lambda_1^k} = \alpha_1 \mathbf{x}_1$$

e quindi per  $j = 1, \dots, n$ , è

$$\lim_{k \rightarrow \infty} \frac{y_j^{(k)}}{\lambda_1^k} = \alpha_1 x_j^{(1)}$$

e

$$\lim_{k \rightarrow \infty} \frac{\mathbf{y}_k}{y_j^{(k)}} = \frac{\mathbf{x}_1}{x_j^{(1)}} \quad (14.5)$$

per tutti gli indici  $j$  per cui  $x_j^{(1)} \neq 0$ . Poiché per  $k$  sufficientemente elevato l'indice  $m$  di una componente di massimo modulo di  $\mathbf{y}_k$  rimane costante, la successione  $\mathbf{y}_k/y_m^{(k)}$  converge all'autovettore  $\mathbf{x}_1$  normalizzato in norma  $\infty$ .

Problema: abbiamo un solo autovalore alla volta. Inoltre a convergenza è data dal rapporto tra il primo ed il secondo autovalore.

Questo metodo richiede ad ogni passo il calcolo del prodotto di una matrice  $A$  per un vettore: se  $A$  non è sparsa ogni passo richiede  $n^2$  operazioni moltiplicative.

### 14.2.1 Underflow/Overflow e normalizzazione

Operando in aritmetica finita, dopo pochi passi si possono presentare condizioni di overflow o di underflow. Per evitare che ciò accada è necessario eseguire ad ogni passo una *normalizzazione* del vettore ottenuto, costruendo una successione  $t_k, k = 1, 2, \dots$  così definita

$$\left. \begin{aligned} u_k &= At_{k-1} \\ t_k &= \frac{1}{\beta_k} u_k \end{aligned} \right\} k = 1, 2, \dots \quad (14.6)$$

dove  $\beta_k$  è uno scalare tale che  $\|t_k\| = 1$  per qualche norma vettoriale. Si ha allora

$$t_k = \frac{1}{\beta_1 \beta_2 \cdots \beta_k} y_k = \frac{1}{\beta_1 \beta_2 \cdots \beta_k} A^k t_0$$

il vettore  $y_k$  va normalizzato quindi ad ogni iterazione

Possiamo scegliere  $\|\cdot\|_\infty$  e  $\|\cdot\|_2$ .

#### Norma $\infty$

Utilizzando la norma  $\infty$ , sia  $\|t_0\|_\infty = 1$  e sia  $\beta_k$  una componente di massimo modulo di  $u_k$ , cioè tale che e

$$\beta_k = u_m^{(k)}, \quad \text{con} \quad |u_m^{(k)}| = \max_{j=1, \dots, n} |u_j^{(k)}| = \|u_k\|_\infty$$

I vettori  $t_k$  ottenuti con la (14.6) sono quindi tali che  $t_m^{(k)} = 1$ . Dalla (14.4) risulta

$$u_m^{(k+1)} = \lambda_1 \left( 1 + O\left(\frac{\lambda_2}{\lambda_1}\right)^k \right)$$

Poiché si può assumere che da una certa iterazione in poi l'indice  $m$ , corrispondente a una componente di massimo modulo di  $u_k$ , resti sempre lo stesso, ne segue che la successione dei  $\beta_k$  converge a  $\lambda_1$  e che l'errore che si commette approssimando  $\lambda_1$  con  $\beta_k$  tende a zero come  $|\lambda_2/\lambda_1|^k$ . Inoltre, poiché  $\|t_k\|_\infty = 1$  dalla (14.5) risulta

$$\lim_{k \rightarrow \infty} t_k = \frac{x_1}{x_m^{(1)}}$$

e quindi la successione  $t_k$  converge all'autovettore  $x_1$  normalizzato in norma  $\infty$

#### Caso norma 2 e $A$ ortonormale

Utilizzando la norma 2, sia  $\|t_0\|_2 = 1$  e sia  $\beta_k = \|u_k\|_2$ . Questa scelta di  $\beta_k$  è particolarmente conveniente nel caso che la matrice  $A$  sia normale, perché si ottiene una successione che converge a  $\lambda_1$  più velocemente che nel caso precedente. Infatti, tenendo conto che gli autovettori  $x_1, x_2, \dots, x_n$  di una matrice normale  $A$  possono essere scelti ortonormali, risulta che

$$\begin{aligned} \sigma_k &= t_k^H u_{k+1} = \frac{t_k^H A t_k}{t_k^H t_k} = \frac{(A^k t_0)^H (A^{k+1} t_0)}{(A^k t_0)^H (A^k t_0)} \\ &= \lambda_1 \frac{|\alpha_1|^2 + \sum_{i=2}^n |\alpha_i|^2 \left| \frac{\lambda_i}{\lambda_1} \right|^{2k} \left( \frac{\lambda_i}{\lambda_1} \right)}{|\alpha_1|^2 + \sum_{i=2}^n |\alpha_i|^2 \left| \frac{\lambda_i}{\lambda_1} \right|^{2k}} \\ &= \lambda_1 \left[ 1 + O\left(\left| \frac{\lambda_2}{\lambda_1} \right|^{2k}\right) \right] \end{aligned}$$

La successione dei  $\sigma_k$  converge a  $\lambda_1$  e l'errore che si commette approssimando  $\lambda_1$  con  $\sigma_k$  tende a zero con  $|\lambda_2/\lambda_1|^{2k}$ . Quindi la successione dei  $\sigma_k$  converge più rapidamente della successione dei  $\beta_k$

La velocità di convergenza in questo caso è doppia!

### 14.2.2 Rilassamento delle ipotesi dell'algoritmo

Le ipotesi fin'ora assunte non sono verificabili prima del processo, vediamo se il metodo è ancora valido allentandone alcune.

1. Diagonalizzabilità
2.  $|\lambda_1| > |\lambda_2| \geq \dots$
3.  $t_0 = \sum \alpha_i x_i$  tale che  $\alpha_1 \neq 0$

#### Caso 2a

Supponiamo che  $\lambda_1$  abbia molteplicità algebrica  $r > 1$

$$\lambda_1 = \lambda_2 = \dots = \lambda_r \quad |\lambda_1| = |\lambda_2| = \dots = |\lambda_r| > |\lambda_{r+1}| \geq \dots$$

In questo caso non abbiamo problemi, infatti nella sommatoria basta tirare fuori non solo il primo ma i primi  $r$  autovalori, e la successione converge ugualmente.

Al posto della (14.3) si ha

$$y_k = \lambda_1^k \left[ \sum_{i=1}^r \alpha_i x_i + \sum_{i=r+1}^n \alpha_i \left( \frac{\lambda_i}{\lambda_1} \right)^k x_i \right]$$

L'autovalore  $\lambda_1$  si approssima con la successione dei  $\beta_k$  o dei  $\sigma_k$ , e l'errore dell'approssimazione tende a zero come  $(\lambda_{r+1}/\lambda_1)^k$  o come  $|\lambda_{r+1}/\lambda_1|^{2k}$ . Inoltre

$$\lim_{k \rightarrow \infty} \frac{y_k}{y_j^{(k)}} = \frac{1}{\theta_j} \sum_{i=1}^r \alpha_i x_i \quad \text{dove} \quad \theta_j = \sum_{i=1}^r \alpha_i x_j^{(i)}$$

e quindi la successione  $y_k/y_m^{(k)}$ , dove  $m$  è l'indice di una componente di massimo modulo di  $y_k$ , converge ad un autovettore normalizzato in norma  $\infty$  appartenente allo spazio vettoriale generato da  $x_1, x_2, \dots, x_r$ .

Se invece esistono più autovalori di modulo massimo diversi fra loro, il metodo delle potenze non è convergente.

#### Caso 2b

Supponiamo che  $\lambda_1 = \overline{\lambda_2}$  abbiamo ancora  $|\lambda_1| = |\lambda_2|$ .

Si ha comunque convergenza



#### Nota

a lezione ha detto che non converge, ma una variante converge



#### Nota

Il metodo delle potenze può essere modificato in modo da approssimare o anche autovalori distinti con lo stesso modulo, come nel caso di autovalori complessi coniugati [28] (si veda anche l'esercizio 6.34).

#### Caso 2c

Supponiamo  $\lambda_1 = -\lambda_2$ , il metodo non converge ma se consideriamo solo le iterazioni *pari* di  $A^k$  è come applicare il metodo a  $(A^2)^k$  e sappiamo che  $A^2$  ha autovalori dominanti  $\lambda_1^2 = \lambda_2^2$  e rientriamo nel metodo recuperando la convergenza.

**Caso 1**

Se  $A$  non è diagonalizzabile si ha comunque convergenza, ma più lenta.

**Caso 3**

Se abbiamo  $\alpha_1 = 0$ , nel metodo scompare  $\lambda_1$  ma se  $|\lambda_2| > |\lambda_3| \geq \dots$  c'è convergenza a  $\lambda_2$ .

In pratica però, per la presenza degli errori di arrotondamento, i vettori  $y_k$  effettivamente calcolati avrebbero comunque una componente relativa a  $x_1$  non nulla, come se  $\alpha_1 \neq 0$ . Perciò la successione effettivamente calcolata convergerebbe ugualmente a  $\lambda_1$ , anche se più lentamente.

**14.2.3 Varianti del metodo delle potenze**

Andiamo ad analizzare alcune varianti del metodo delle potenze consentono di calcolare anche gli altri autovalori e i corrispondenti autovettori.

**14.2.3.1 Variante di Wielandt (metodo delle potenze inverse)****Autovalore più piccolo**

Se  $A$  è una matrice non singolare, diagonalizzabile, con autovalori tali che

$$|\lambda_1| \geq \dots \geq |\lambda_{n-1}| > |\lambda_n| > 0,$$

se cerchiamo l'autovalore più piccolo possiamo applicare il metodo delle potenze ad  $A^{-1}$  che ha autovalori  $\frac{1}{\lambda}, i = 1, \dots, n$  tali che

$$\frac{1}{|\lambda_n|} > \frac{1}{|\lambda_{n-1}|} \geq \dots \geq \frac{1}{|\lambda_1|}$$

Per calcolare l'autovalore di modulo minimo di  $A$  si applica il metodo delle potenze alla matrice  $A^{-1}$ . Però sappiamo che calcolare l'inversa di una matrice è molto costoso allora invece di applicare  $y_k = A^{-1}y_{k-1}$  facciamo  $Ay_k = y_{k-1}$ , cioè risolviamo ad ogni passo un sistema lineare.

Il costo è uguale a quello del metodo delle potenze:

- fattorizzazione  $A$ :  $O(n^3)$
- passo  $k$ :  $O(n^2)$

**Autovalore intermedio**

Nel caso ci interessi un autovalore intermedio, basta calcolare l'autovalore massimo di un'altra matrice, così definita:

$$(A - \mu I)^{-1}$$

i cui autovalori sono  $\frac{1}{\lambda_i - \mu}$ , quindi per imporre che il  $k$ -esimo sia il più grande dobbiamo avere

$$\frac{1}{|\lambda_k - \mu|} > \frac{1}{|\lambda_i - \mu|}$$

il che è verificato se  $\mu \approx \lambda_k$ , quindi se conosciamo una stima dell'autovalore che ci interessa, cosa che è possibile ottenere con tecniche di approssimazione come Gerschgorin.

Si possono fare le stesse considerazioni sulla complessità del caso precedente, prendendo in considerazione il sistema

$$(A - \mu I)y_k = y_{k-1}$$

### 14.2.3.2 Variante della deflazione

Sia  $|\lambda_1| > |\lambda_2|$ . Calcolati  $\lambda_1$  e  $x_1$ , di norma 2 unitaria, si considera la matrice di Householder  $P$  tale che  $Px_1 = e_1$ , risulta

$$PAP^H = \begin{bmatrix} \lambda_1 & 0^H \\ 0 & A_1 \end{bmatrix}$$

se  $A$  è hermitiana o

$$PAP^H = \begin{bmatrix} \lambda_1 & a^H \\ 0 & A_1 \end{bmatrix}$$

se  $A$  non lo è.

Si applica il metodo delle potenze alla matrice  $A_1$  di ordine  $n - 1$  (deflazione) e si calcolano  $\lambda_2$  e il corrispondente autovettore  $y_2$  di  $A_1$ . L'autovettore  $x_2$  di  $A$  corrispondente a  $\lambda_2$  è dato da

$$x_1 = P^H \begin{bmatrix} \theta \\ y_2 \end{bmatrix}, \quad \text{con } \theta = \begin{cases} 0 & \text{se } A \text{ è Hermitiana} \\ \frac{a^H y_2}{\lambda_2 - \lambda_1} & \text{se } A \text{ non lo è} \end{cases}$$

Procedendo in questo modo si costruisce la forma di Schur della matrice. Poiché la trasformazione  $A \rightarrow PAP^H$  può distruggere la eventuale struttura o sparsità di  $A$ , questo procedimento può non essere indicato per matrici sparse.

$$PAP^{-1} = \left\| \begin{array}{c|ccc} \lambda & b_2 & \dots & b_n \\ \hline 0 & & & \\ \cdot & & B & \\ \cdot & & & \\ 0 & & & \end{array} \right\|$$

## 14.3 Riduzione di una matrice hermitiana in forma tridiagonale: il metodo di Householder

Il metodo delle potenze è adatto per matrici sparse e si vogliono pochi autovalori.

Nel caso si vogliono ottenere tutti, o quasi tutti, gli autovalori

1. si trasforma  $A$  in una forma  $B$  più condensata (cioè con più zeri) *simile*
2. si calcolano gli autovalori di  $B$  con un procedimento iterativo

### 14.3.1 Trasformazioni

Forma condensate:

- *Hessemberg* (1.3)  $a_{ij} = 0$  per  $i > j + 1$

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

Quest è la forma ottenuta nel caso più generale.



- Tridiagonale

$$T_n = \begin{bmatrix} t_{1,1} & t_{1,2} & 0 & \cdots & 0 \\ t_{2,1} & t_{2,2} & t_{2,3} & \ddots & \vdots \\ 0 & t_{3,2} & t_{3,3} & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & t_{n-1,n} \\ 0 & \cdots & 0 & t_{n,n-1} & t_{n,n} \end{bmatrix}$$

Se la matrice  $A$  è hermitiana, e la trasformazione viene eseguita con matrici unitarie (householder), la matrice  $B$  risulta hermitiana e tridiagonale (rientra in hessemberg).

## 14.4 Il metodo di Householder

La trasformazione per similitudine unitarie della matrice  $A$  nella matrice  $B$  è fatta per passi successivi

$$A^{(k+1)} = T_k^{-1} A^{(k)} T_k, \quad k = 1, 2, \dots, m-1 \quad (14.7)$$

dove

$$A^{(1)} = A \quad \text{e} \quad A^{(m)} = B$$

per cui, posto  $T = T_1 T_2 \dots T_{m-1}$ , risulta  $B = T^{-1} A T$ , e se  $x$  è autovettore di  $B$ ,  $Tx$  è autovettore di  $A$ .

Sia  $A \in \mathbb{C}^{n \times n}$  una matrice hermitiana; si considerino le trasformazioni (14.7), con  $m = n - 1$ , in cui le matrici  $T_k$  siano matrici elementari di Householder (hermitiane e unitarie):

$$T_k = I - \beta_k u_k u_k^H,$$

costruite in modo che nella matrice  $T_k A^{(k)}$  siano nulli tutti gli elementi della  $k$ -esima colonna, con l'indice di riga maggiore di  $k + 1$ . Al primo passo, posto

$$A^{(1)} = A = \begin{bmatrix} a_{11}^{(1)} & a_1^H \\ a_1 & B^{(1)} \end{bmatrix} \begin{array}{l} \text{]1 riga} \\ \text{]n-1 righe} \end{array}$$

si consideri la matrice elementare di Householder  $P^{(1)} \in \mathbb{C}^{(n-1) \times (n-1)}$  tale che

$$P^{(1)} a_1 = \alpha_1 e_1$$

dove  $e_1$  è il primo vettore della base canonica di  $\mathbb{C}^{n-1}$ . La matrice

$$T_1 = \begin{bmatrix} 1 & 0^H \\ 0 & P^{(1)} \end{bmatrix}$$

è tale che nella matrice

$$A^{(2)} = T_1^{-1} A^{(1)} T_1 = T_1 A^{(1)} T_1$$

sono nulli tutti gli elementi della prima colonna con indice di riga maggiore di due e dei simmetrici elementi della prima riga.

Al  $k$ -esimo passo la sottomatrice principale di testa di ordine  $k + 1$  di  $A^{(k)}$  risulta tridiagonale hermitiana e  $A^{(k)}$  ha la forma

$$A^{(k)} = \begin{bmatrix} C^k & b_k & O \\ b_k^H & a_{kk}^{(k)} & a_k^H \\ O & a_k & B^{(k)} \end{bmatrix} \begin{array}{l} \text{]k-1 righe} \\ \text{]1 riga} \\ \text{]n-k righe} \end{array}$$

dove  $C^{(k)} \in \mathbb{C}^{(k-1) \times (k-1)}$  è tridiagonale hermitiana e  $b_k \in \mathbb{C}^{k-1}$  ha nulle le prime  $k-2$  componenti. Sia  $P^{(k)} \in \mathbb{C}^{(n-k) \times (n-k)}$  la matrice di Householder tale che

$$P^{(k)} a_k = \alpha_k e_1$$

dove  $e_1$  è il primo vettore della base canonica di  $C_{n-k}$ . Posto

$$T_k = \begin{bmatrix} I_k & 0^H \\ 0 & P^{(k)} \end{bmatrix}$$

risulta

$$A^{(k+1)} = T_k^{-1} A^{(k)} T_k = T_k A^{(k)} T_k = \begin{bmatrix} C^k & b_k & 0 \\ b_k^H & a_{kk}^{(k)} & a_k^H P^{(k)} \\ 0 & P^{(k)} a_k & P^{(k)} B^{(k)} P^{(k)} \end{bmatrix}$$

Poiché il vettore  $P^{(k)} a_k \in \mathbb{C}^{n-k}$  ha nulle le componenti di indice maggiore o uguale a due, la sottomatrice principale di testa di ordine  $k+2$  della matrice  $A^{(k+1)}$  è tridiagonale hermitiana. Applicando il procedimento  $n-2$  volte si ottiene la matrice  $B = A^{(n-1)}$  tridiagonale hermitiana.

Se  $A$  non è hermitiana si ottiene la forma di Hessemberg.

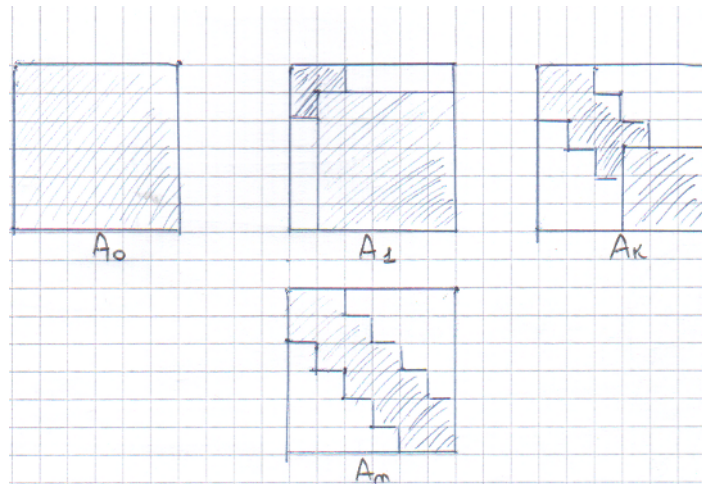


Figura 14.1: Passi della trasformazione di una matrice in forma tridiagonale



### Nota

Nel libro esattamente in questo punto vengono fatti dei conti per stabilire i costi delle varie operazioni, che ometto in quanto il Bevilacqua non li ha nominati a lezione.


La trasformazione  $A^{(k)} \rightarrow A^{(k+1)}$  richiede solo  $2(n-k)^2$  operazioni moltiplicative. Il metodo di Householder per tridiagonalizzare una matrice hermitiana richiede dunque

$$\sum_{k=1}^{n-2} 2(n-k)^2 \approx \frac{2}{3} n^3 \quad \text{operazioni moltiplicative}$$



### Work in progress

Per chi ne ha voglia: aggiungere disegni fatti a lezione

 **Work in progress**

Qua il Bevilacqua ha fatto una digressione sulle proprietà dei minori principali di testa... per chi ha voglia, li metta.

## 14.5 Calcolo degli autovalori delle matrici tridiagonali hermitiane con la successione di Sturm

Per calcolare gli autovalori di una matrice tridiagonale hermitiana conviene utilizzare metodi iterativi che facciano ricorso al polinomio caratteristico solo se il numero degli autovalori che si vogliono determinare è piccolo rispetto alle dimensioni della matrice.

Riassumiamo i passi del metodo:

- utilizziamo lo sviluppo di Laplace per calcolare il polinomio caratteristico della matrice
- la particolare struttura della matrice ci permette di calcolare lo sviluppo in modo iterativo ed efficiente
- utilizziamo il metodo di Newton per calcolare gli zeri del polinomio caratteristico
- per sapere in che intervallo lanciare la ricerca degli zeri usiamo il teorema di Sturm che sfrutta il numero di inversioni di segno

Sia  $B_n \in \mathbb{C}^{n \times n}$  la matrice tridiagonale hermitiana definita da

$$B_n = \begin{bmatrix} \alpha_1 & \bar{\beta}_2 & & & \\ \beta_2 & \alpha_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \bar{\beta}_n \\ & & & \beta_n & \alpha_n \end{bmatrix}$$

Se la matrice  $B_n$  è riducibile, cioè se esiste almeno un indice  $j$ ,  $2 \leq j \leq n$ , tale che  $\beta_j = 0$ , allora il problema del calcolo degli autovalori di  $B_n$  è ricondotto al calcolo degli autovalori di due matrici di ordine inferiore. Infatti si ha

$$B_n = \begin{bmatrix} C_{j-1} & O \\ O & D_{n-j+1} \end{bmatrix}$$

in cui  $C_{j-1} \in \mathbb{C}^{(j-1) \times (j-1)}$ ,  $D_{n-j+1} \in \mathbb{C}^{(n-j+1) \times (n-j+1)}$  e quindi

$$\det(B_n - \lambda I) = \det(C_{j-1} - \lambda I_{j-1}) \det(D_{n-j+1} - \lambda I_{n-j+1}).$$

Se le matrici  $C_{j-1}$  e  $D_{n-j+1}$  sono a loro volta riducibili, si procede in modo analogo.

Si consideri perciò il caso che  $B_n$  sia irriducibile e il suo polinomio caratteristico, calcolato tramite lo sviluppo di Laplace (1.22) rispetto all'ultima riga, risulti

$$P_n(\lambda) = \det(B_n - \lambda I) = (\alpha_n - \lambda)P_{n-1}(\lambda) - |\beta_n|^2 P_{n-2}(\lambda)$$

con  $P_i(\lambda)$  il polinomio caratteristico del minore principale di testa di ordine  $i$ . Questo si può fare per ogni  $i$

$$\begin{aligned} P_0(\lambda) &= 1 \\ P_1(\lambda) &= \alpha_1 - \lambda \\ P_i(\lambda) &= (\alpha_i - \lambda)P_{i-1}(\lambda) - |\beta_i|^2 P_{i-2}(\lambda) \quad i = 2, 3, \dots \end{aligned} \tag{14.8}$$

Quindi è possibile calcolare il valore che il polinomio  $P_n(\lambda)$  assume in un punto con  $2(n-1)$  moltiplicazioni (supponendo di aver già calcolato  $|\beta_i|^2$ ,  $i = 2, 3, \dots, n$ ).

Gli autovalori di  $B_n$  vengono quindi calcolati risolvendo l'equazione caratteristica

$$P_n(\lambda) = 0 \quad (14.9)$$

con un metodo iterativo. Se si utilizza il metodo di Newton, il calcolo di  $P_n(\lambda)$  può essere fatto con le seguenti relazioni ricorrenti, ottenute derivando rispetto a  $\lambda$  entrambi i membri delle (14.8):

$$\begin{aligned} P'_0(\lambda) &= 0 \\ P'_1(\lambda) &= -1 \\ P'_i(\lambda) &= (\alpha_i - \lambda)P'_{i-1}(\lambda) - P_{i-1}(\lambda) - |\beta_i|^2 P'_{i-2}(\lambda), \quad i = 2, 3, \dots, n. \end{aligned}$$

La formula di aggiornamento del metodo di Newton diventa quindi

$$x_{i+1} = x_i - \frac{P_n(x_i)}{P'_n(x_i)}$$

Quindi il rapporto  $P_n(\lambda)/P'_n(\lambda)$ , che interviene ad ogni passo del metodo di Newton, può essere calcolato con  $4(n-1)$  moltiplicazioni e una divisione.

Per separare le radici di (14.9) conviene sfruttare le proprietà delle successioni di Sturm. Infatti nel seguente teorema si dimostra che i polinomi  $P_i(\lambda)$  formano una successione di Sturm.



#### Teorema 14.4

Sia  $A \in \mathbb{C}^{n \times n}$  hermitiana, e sia  $A^k$  la sottomatrice principale di testa di ordine  $k$  di  $A$ . Allora gli autovalori di  $A_k$  separano gli autovalori di  $A_{k+1}$ , per  $k = 1, \dots, n-1$ .



#### Teorema 14.5

Se  $\beta_i \neq 0$ , per  $i = 2, 3, \dots, n$ , la successione dei polinomi  $P_i(\lambda)$ ,  $i = 0, 1, \dots, n$  verifica le seguenti proprietà:

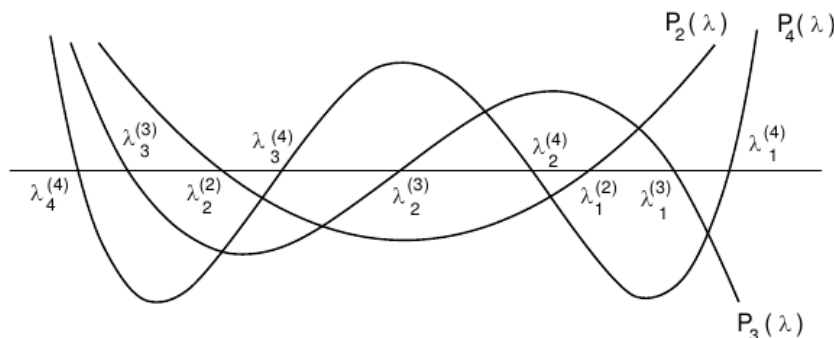
1.  $P_0(\lambda)$  non cambia segno;
2. se  $P_i(\lambda) = 0$ , allora  $P_{i-1}(\lambda)P_{i+1}(\lambda) < 0$ , per  $i = 1, 2, \dots, n-1$
3. se  $P_n(\lambda) = 0$ , allora  $P'_n(\lambda)P_{n-1}(\lambda) < 0$  (e quindi  $P_n(\lambda)$  ha tutti zeri di molteplicità 1).

Una successione di polinomi che verifica le proprietà 1), 2) e 3) è detta successione di Sturm.

*Dimostrazione.* La 1) è ovvia. Per la 2), si osservi che da (14.8) si ha  $P_{i-1}(\lambda)P_{i+1}(\lambda) \leq 0$ . Ma se fosse  $P_{i-1}(\lambda)P_{i+1}(\lambda) = 0$  e  $P_i(\lambda) = 0$ , allora sarebbe  $P_{i-1}(\lambda) = P_{i+1}(\lambda) = 0$ , da cui, per ricorrenza, seguirebbe  $P_0(\lambda) = 0$ , che è assurdo. Da ciò segue anche che gli zeri  $\lambda_j^{(n)}$ ,  $j = 1, 2, \dots, n$ , di  $P_n(\lambda)$  sono distinti dagli zeri  $\lambda_j^{(n-1)}$ ,  $j = 1, 2, \dots, n-1$ , di  $P_{n-1}(\lambda)$  e quindi, per il teorema (14.4), gli zeri  $\lambda_j^{(n-1)}$  separano strettamente gli zeri  $\lambda_j^{(n)}$ , cioè

$$\lambda_{j+1}^{(n)} < \lambda_j^{(n-1)} < \lambda_j^{(n)}, \quad j = 1, 2, \dots, n-1$$

Da questo fatto, tenendo presente che il coefficiente di  $\lambda^i$  in  $P_i(\lambda)$  è  $(-1)^i$ , e quindi  $\lim_{\lambda \rightarrow -\infty} P_i(\lambda) = +\infty$ , segue la 3) (si veda la figura 6.1 per il caso  $\lambda \rightarrow -\infty$ ,  $n = 4$ ).



□

**Teorema 14.6 (Sturm)**

Se  $\{P_i(\lambda)\}, i = 0, 1, \dots, n$ , una successione di Sturm, il numero  $w(b) - w(a)$  è uguale al numero di zeri di  $P_n(\lambda)$  appartenenti all'intervallo  $[a, b)$ .

*Dimostrazione.* Si faccia variare  $\lambda$  con continuità da  $a$  verso  $b$ . Si può avere una variazione nel numero  $w(\lambda)$  solo quando  $\lambda$  incontra uno zero di uno dei polinomi  $P_i(\lambda)$ . Si consideri perciò un  $\lambda^*$  tale che  $P_i(\lambda^*) = 0$  per un indice  $i$ . Per la proprietà 1) del teorema (14.5) deve essere  $i \neq 0$ . Si distinguono allora i due casi:

1.  $i \neq n$ 

In questo caso, per la proprietà 2) del teorema (14.5) si ha

$$P_{i-1}(\lambda^*)P_{i+1}(\lambda^*) < 0.$$

Esiste perciò un numero  $h$  tale che nell'intervallo  $[\lambda^* - h, \lambda^* + h]$  è ancora

$$P_{i-1}(\lambda)P_{i+1}(\lambda) < 0$$

e

$$P_i(\lambda) \neq 0$$

eccetto che nel punto  $\lambda^*$ . Poiché per ogni  $\lambda \in [\lambda^* - h, \lambda^* + h]$  i due polinomi  $P_{i-1}(\lambda)$  e  $P_{i+1}(\lambda)$  hanno segno discorde,  $P_i(\lambda)$  deve avere in questo intervallo segno concorde con uno dei due e discorde con l'altro. Quindi nella sequenza  $P_{i-1}(\lambda), P_i(\lambda), P_{i+1}(\lambda)$  vi è una sola variazione di segno in tutto l'intervallo  $[\lambda^* - h, \lambda^* + h]$ , cioè il fatto che  $P_i(\lambda)$  si annulli in  $\lambda^*$  non comporta variazioni del numero  $w(\lambda)$ .

2.  $i = n$ 

In questo caso, poichè per la proprietà 3) del teorema (14.5) il polinomio  $P_n(\lambda)$  ha radici semplici, la sua derivata  $P'_n(\lambda)$  non si annulla in  $\lambda^*$  ed esiste un numero  $h$  tale che nell'intervallo  $[\lambda^* - h, \lambda^* + h]$   $P'_n(\lambda)$  ha lo stesso segno che  $P_n(\lambda)$  ha in  $\lambda^* + h$  e segno opposto a quello che  $P_n(\lambda)$  ha in  $\lambda^* - h$ . Se  $h$  tale che nell'intervallo  $[\lambda^* - h, \lambda^* + h]$  anche  $P_{n-1}(\lambda)$  non si annulla, poichè per la proprietà 3) del teorema (14.5)  $P_{n-1}(\lambda)$  ha segno opposto a quello di  $P_n(\lambda)$  per  $\lambda \in [\lambda^* - h, \lambda^* + h]$ , la sequenza  $P_{n-1}(\lambda^* + h), P_n(\lambda^* + h)$  presenta una variazione di segno, mentre la sequenza  $P_{n-1}(\lambda^* - h), P_n(\lambda^* - h)$  non presenta alcuna variazione di segno.

Se ne conclude che il numero di variazioni di segno in tutta la sequenza  $P_0(\lambda), P_1(\lambda), \dots, P_n(\lambda)$  può cambiare solo nei punti in cui si annulla  $P_n(\lambda)$ , ed esattamente aumenta di 1 ogni volta che si annulla  $P_n(\lambda)$ . Nella tesi del teorema l'intervallo  $[a, b)$  è aperto a destra perché se fosse

$P_n(b) = 0$ , poichè  $P_n(b)$  viene assegnato lo stesso segno assunto in  $b$  da  $P_{n-1}(\lambda)$ , che diverso da zero in un intorno sinistro di  $b$ ,  $w(\lambda)$  non cambia in tale intorno. Perciò la radice  $b$  non altera il numero di variazioni di segno. □

Risultato serve a metodo di bisezione

$$p_n(a) \cdot p_n(b) < 0$$

$$V(b) - V(a)$$

**14.6 Metodo QR**

L'idea alla base del metodo QR è di generare una successione di matrici  $A_1 \dots A_k$  tutte simili ad  $A$ , ma con  $A_k$  triangolare superiore con gli autovalori sulla diagonale.

Il metodo può sfruttare diversi metodi di fattorizzazione, ma la migliore risulta essere la QR perché

- esiste sempre
- usa trasformazioni unitarie che sono più stabili

### 14.6.1 Metodo $QR$ con trasformazioni di Householder

Nel metodo  $QR$  viene generata una successione  $\{A_k\}$  di matrici nel modo seguente: posto

$$A_1 = A,$$

per  $k = 1, 2, \dots$ , si calcola una fattorizzazione  $QR$  di  $A_k$

$$A_k = Q_k R_k \quad (14.10)$$

dove  $Q_k$  è unitaria e  $R_k$  è triangolare superiore, e si definisce la matrice  $A_{k+1}$  per mezzo della relazione

$$A_{k+1} = R_k Q_k \quad (14.11)$$

Da (14.10) e (14.11) risulta che

$$A_{k+1} = Q_k^H A_k Q_k, \quad (14.12)$$

e quindi le matrici della successione  $\{A_k\}$  sono tutte simili fra di loro, infatti

$$Q_k A_{k+1} Q_k^H = Q_k R_k Q_k Q_k^H = A_k$$

Sotto opportune ipotesi la successione converge ad una matrice triangolare superiore (diagonale se  $A$  è hermitiana) che ha come elementi diagonali gli autovalori di  $A$ .

Il costo dell'algoritmo  $QR$  è  $O(n^3)$ .

#### Proprietà interessanti e miglioramento dell'efficienza

- Se  $A_k$  è Hermitiana, lo è anche  $A_{k+1}$

$$R_k^{-1} A_{k+1} R_k = R_k^{-1} \cancel{R_k} Q_k R_k = A_k$$

- Inoltre se la matrice precedente è di Hessemberg e si moltiplica per una triangolare a sinistra a destra, otteniamo una Hessemberg

$$A_{k+1} = R_k A_k R_k^{-1}$$

Eseguiamo quindi una procedura di bootstrap con

$$A \xrightarrow{\text{Householder}} \text{Hessemberg}$$

questo è un bel vantaggio in termini di costi:  $O(n^3) \rightarrow O(n^2)$ .

Idem per il caso Hermitiano, in questo caso passiamo a costo lineare  $O(n)$

### 14.6.2 Convergenza

Per prima cosa, al fine di dimostrare la convergenza, introduciamo il seguente risultato:



#### Teorema 14.7

Se  $A$  è non singolare e gli  $r_{ii}$  della matrice  $R$  sono reali e positivi, allora le fattorizzazioni  $Q$  e  $R$  sono uniche.

*Dimostrazione.*

$$A = QR$$

Per prima cosa dimostriamo che data una qualsiasi fattorizzazione QR, possiamo riportarci al caso  $Q_1 R_1$  con  $R_1$  avente  $r_{ii}$  tutti positivi. Introduciamo una matrice di fase:

$$A = QR = \underbrace{Q}_{Q_1} \underbrace{S R}_{R_1}$$

Sia infatti

$$R = \begin{pmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{pmatrix} \quad S = \begin{pmatrix} \frac{|r_{11}|}{r_{11}} & 0 \\ 0 & \frac{|r_{22}|}{r_{22}} \end{pmatrix}$$

Allora

$$R_1 = \begin{pmatrix} |r_{11}| & * \\ 0 & |r_{22}| \end{pmatrix}$$

Dimostriamo adesso l'unicità di tale fattorizzazione: per assurdo, siano le seguenti due diverse fattorizzazioni QR:

$$QR = \hat{Q}\hat{R} \quad r_{ii} > 0, \hat{r}_{ii} > 0$$

Ma allora, poichè:

1. il prodotto di matrici unitarie è una matrice unitaria
2. l'inversa di una triangolare superiore, se esiste, è ancora una triangolare superiore. Inoltre l'insieme delle triangolari superiori è chiuso rispetto al prodotto.

Ottiniamo quindi:

$$\underbrace{\hat{Q}Q^{-1}}_{\text{unitaria}} = \underbrace{\hat{R}R^{-1}}_{\text{t.sup}}$$

Ma una matrice  $V$  che sia unitaria e triangolare superiore deve essere necessariamente diagonale in quanto

$$\underbrace{V^H}_{\text{t.inf}} = \underbrace{V^{-1}}_{\text{t.sup}} \rightarrow V \text{ diagonale}$$

In sostanza stiamo lavorando quindi con una matrice *diagonale* e *unitaria*. Sappiamo però

$$1. d_{ii} = \frac{\begin{matrix} >0 \\ \hat{r}_{ii} \\ \end{matrix}}{\begin{matrix} r_{ii} \\ >0 \end{matrix}} > 0$$

2. Essendo la matrice unitaria, affinché  $D = D^H = D^{-1}$  deve essere  $d_{ii} = 1$

Concludiamo che deve essere

$$\hat{R}R^{-1} = I \quad \Rightarrow \quad \hat{R} = R \quad \Rightarrow \quad \hat{Q} = Q$$

□

Adesso possiamo dimostrare il seguente teorema. Purtroppo non è possibile verificarne le ipotesi a priori.



### Teorema 14.8 (Convergenza)

1.

$$|\lambda_1| > |\lambda_2| > \dots > |\lambda_n| > 0$$

, in questo caso la matrice è diagonalizzabile

2. Indicata con  $X$  la matrice degli autovettori di  $A$ , tale che

$$A = XDX^{-1}, \tag{14.13}$$

in cui  $D$  è la matrice diagonale il cui  $i$ -esimo elemento principale è  $\lambda_i$ , si supponga che la matrice  $X^{-1}$  ammetta la fattorizzazione LU.

Allora esistono delle matrici di fase  $S_k$  tali che

$$\lim_{k \rightarrow \infty} S_k^H R_k S_{k-1} = \lim_{k \rightarrow \infty} S_{k-1} A_k S_{k-1} = T \quad (14.14)$$

e

$$\lim_{k \rightarrow \infty} S_{k-1} Q_k S_k = I$$

dove  $T$  è triangolare superiore con gli elementi principali uguali a  $\lambda_1, \lambda_2, \dots, \lambda_n$ . Quindi gli elementi principali di  $A_k$  tendono agli autovalori di  $A$ . Se  $A$  è una matrice hermitiana, allora  $T$  è diagonale.

*Dimostrazione.* Facciamo una semplificazione che sul libro non viene fatta, quindi la dimostrazione è più semplice.

$$A = QR = Q_1 R_1$$

Le due fattorizzazioni possono differire per il prodotto di una matrice di fase, questo significa che

$$Q_1 = QS^H \quad R_1 = SR$$

Utilizziamo il teorema precedente, utilizzando il fatto che la fattorizzazione sia unica.

Il teorema viene dimostrato confrontando due fattorizzazioni  $QR$  della matrice  $A_k$  ottenute in due modi diversi. Due fattorizzazioni  $QR$  di  $A^k$

1. Una prima fattorizzazione è data dalla seguente relazione

$$A^k = \underbrace{H_k}_{\text{tipo Q}} \underbrace{U_k}_{\text{tipo R}}, \quad (14.15)$$

dove

$$H_k = Q_1 Q_2 \dots Q_k$$

è una matrice unitaria e

$$U_k = R_k R_{k-1} \dots R_1$$

$$A^k = \underbrace{Q_1 Q_2 \dots Q_k}_{\text{tipo Q}} \underbrace{R_1 R_2 \dots R_k}_{\text{tipo R}}$$

Per dimostrare la (14.15) si procede per induzione:

**Caso  $k = 1$**

$$A = A_1 = H_1 U_1$$

ovvio

**Caso  $k > 1$**

Supposta valida la (14.15), da (14.11) e (14.12) si ottiene

$$Q_k A_{k+1} = A_k Q_k$$

da cui

$$Q_1 \dots Q_{k-1} Q_k A_{k+1} = Q_1 \dots Q_{k-1} A_k Q_k = \dots = A Q_1 \dots Q_{k-1} Q_k \quad (14.16)$$

e quindi

$$\begin{aligned} H_{k+1} U_{k+1} &= Q_1 \dots Q_k Q_{k+1} R_{k+1} R_k \dots R_1 \\ &= Q_1 \dots Q_{k-1} Q_k A_{k+1} R_k R_{k-1} \dots R_1 \\ &= A Q_1 \dots Q_{k-1} Q_k R_k R_{k-1} \dots R_1 = A H_k U_k = A_{k+1} \end{aligned}$$

cioè  $A_{k+1} = H_{k+1} U_{k+1}$ .

2. Una seconda fattorizzazione  $QR$  della matrice  $A_k$  viene ottenuta dalla relazione (14.13). Sia  $X^{-1} = LU$  la fattorizzazione  $LU$  di  $X^{-1}$ . Allora

$$A^k = X D^k X^{-1} = X D^k L U = X D^k L D^{-k} D^k U$$

Poichè gli elementi della matrice  $D^k L D^{-k}$  sono dati da



$$\begin{cases} l_{ij} \left( \frac{\lambda_i}{\lambda_j} \right)^k & \text{per } i > j, \\ 1 & \text{per } i = j \\ 0 & \text{per } i < j \end{cases} \quad (14.17)$$

e  $|\lambda_i| < |\lambda_j|$  per  $i > j$ , si può porre

$$D^k L D^{-k} = I + E_k,$$

dove

$$\lim_{k \rightarrow \infty} E_k = 0$$

, e quindi è

$$A^k = X(I + E_k)D^k U.$$

Indicata con

$$X = QR$$

una fattorizzazione QR della matrice X, si ha

$$A^k = QR(I + E_k)D^k U = Q(I + RE_k R^{-1})RD^k U,$$

e indicata con

$$I + RE_k R^{-1} = P_k T_k \quad (14.18)$$

una fattorizzazione QR della matrice  $I + RE_k R^{-1}$ , si ha

$$A_k = (QP_k)(T_k R D^k U). \quad (14.19)$$

La (14.19) da una seconda fattorizzazione QR di  $A_k$ : infatti  $QP_k$  è unitaria e  $T_k R D^k U$  è triangolare superiore.

Abbiamo due uguaglianze

### Domanda aperta

qui si sta usando il teorema enunciato e dimostrato prima, ma perchè lo può usare? Non mi è assolutamente chiaro. Dov'è l'ipotesi che i  $r_{ii}$  sono reali? Quest'ultima parte è assolutamente da rivedere.

$$\begin{aligned} QP_k &= Q_1 Q_2 \dots Q_k \\ T_k R D^k U &= R_k \underbrace{R_{k-1} \dots R_1}_{T_{k-1} R D^{k-1} U} \end{aligned}$$

Ottieniamo

$$\begin{aligned} QP_k &= QP_{k-1} Q_k \\ T_k R D^k U &= R_k T_{k-1} R D^{k-1} U \\ Q_k &= P_{k-1}^H P_k \\ R_k &= T_k R D^{k-1} R^{-1} T_{k-1}^{-1} \end{aligned}$$

Ricordiamo che

$$(D^{k-1})^{-1} = D^{1-k}$$

$$A_k = Q_k R_k = \underbrace{P_{k-1}^H}_{\rightarrow \infty I} \underbrace{P_k}_{\rightarrow \infty I} \underbrace{T_k}_{\rightarrow \infty I} R D R^{-1} \underbrace{T_{k-1}^{-1}}_{\rightarrow \infty I}$$

□

### 14.6.3 Indebolimento delle condizioni del teorema

Le condizioni del teorema possono essere indebolite, infatti

1. Esistono

$$\lambda_i, \lambda_j, i \neq j, |\lambda_i| = |\lambda_j| = \begin{cases} (a) & \lambda_i = \lambda_j (\text{multipli}) \\ (b) & \lambda_i \neq \lambda_j \end{cases}$$

nel caso (a) c'è convergenza, nel caso (b), convergenza a  $T$  triangolare superiore a blocchi

2. Davamo per scontata la diagonalizzabilità, supponiamo che non ci sia: c'è lo stesso la convergenza.

3.  $X^{-1}$  non ha fattorizzazione LU, ma non c'è convergenza *ordinata*, autovalori non ordinati.

### 14.6.4 Tecnica di traslazione: QR con shift

La velocità di convergenza del metodo  $QR$  dipende per la (14.17) dai rapporti  $|\lambda_i/\lambda_j|$  per  $i > j$ , e quindi per l'ipotesi

$$|\lambda_1| > |\lambda_2| > \dots > |\lambda_n| > 0$$

dal numero

$$\max_{1 \leq i \leq n-1} \left| \frac{\lambda_{i+1}}{\lambda_i} \right| \quad (14.20)$$

Se tale rapporto è vicino ad 1, la convergenza può essere lenta. In questo caso per accelerare la convergenza si utilizza una tecnica di traslazione dello spettro degli autovalori di  $A$ , detta di *shift*. Sia  $\mu$  un numero che approssima un autovalore  $\lambda$  meglio degli altri autovalori. Le matrici  $Q_k$  e  $R_k$ , generate dal metodo  $QR$  a partire dalla matrice  $A - \mu I$  possono essere costruite anche per mezzo delle seguenti relazioni (metodo  $QR$  con shift)

$$\left. \begin{aligned} A_k - \mu I &= Q_k R_k \\ A_{k+1} &= R_k Q_k + \mu I \end{aligned} \right\} \text{ per } k = 1, 2, \dots$$

e risulta

$$Q_k A_{k+1} = A_k Q_k - \mu Q_k + \mu Q_k = A_k Q_k$$

. Tenendo presente che gli autovalori di  $A - \mu I$  sono  $\lambda_i - \mu$  e che la velocità di convergenza è regolata dalla (14.20), è possibile scegliere un parametro  $\mu$  in modo da accelerare la convergenza del metodo  $QR$  con shift. È conveniente scegliere per  $\mu$  un valore che approssima  $\lambda_n$ . Ciò può essere ottenuto applicando il metodo  $QR$  inizialmente senza shift per un certo numero  $p$  di iterazioni, e scegliendo  $\mu = a_{mn}^{(p)}$  per le successive iterazioni con shift. Poiché  $\mu$  può essere modificato ad ogni iterazione è più conveniente scegliere

$$\mu_k = a_{mn}^{(k)}, \quad k = 1, 2, \dots \quad (14.21)$$

Quando le condizioni di arresto verificata per  $p = n - 1$ , si passa a operare sulla matrice  $B_k$  di ordine  $n - 1$  ottenuta dalla matrice  $A_k$  eliminando l'ultima riga e l'ultima colonna. Per l'approssimazione degli altri autovalori si procede in modo analogo.

### 14.6.5 Matrice di Sylvester



Work in progress

$$S^T S - S S^T \neq 0$$

non è normale, per risultato del condizionamento.

Matlab, quando l'ordine della matrice cresce abbastanza, alcuni autovalori di  $S$  calcolati con il metodo  $QR$  non hanno nessuna cifra significativa.

Varianti del metodo  $QR$  che vanno su sottoninsiemi degli autovalori dominanti in modulo

## 15 DFT: Trasformata discreta di Fourier

Non parleremo di interpolazione trigonometrica. È un'applicazione lineare da  $\mathbb{C}^n$  in  $\mathbb{C}^n$ . È rappresentata quindi da una matrice, la matrice di Fourier  $V \in \mathbb{C}^{n \times n}$ . Questa matrice è particolare: può essere moltiplicata per un vettore  $z \in \mathbb{C}^n$  in  $O(n \log n)$  operazioni aritmetiche.

$$y = Vz$$

Se invece vogliamo risolvere

$$z = V^{-1}y$$

costa  $O(n \log n)$  invece che  $O(n^3)$ .

Un'applicazione è la TAC oppure nel calcolo del gradiente coniugato, dove ad ogni passo di fa un prodotto matrice vettore, il costo passa da  $O(n^2)$  a poco più che lineare.

Notazione: DFT, IDFT (trasformata inversa). Algoritmo FFT: un algoritmo che calcola DFT o IDFT.

### Definizione 15.1 (Radice $n$ -esima)

Sia  $n$  un intero. Si definisce radice  $n$ -esima dell'unità ogni numero complesso  $z$  tale che  $z^n = 1$ . Una radice  $n$ -esima  $\omega$  è detta primitiva se l'insieme  $\{\omega^i, i = 0, \dots, n-1\}$  è costituito da  $n$  elementi distinti. In particolare, indicata con  $i$  l'unità immaginaria ( $i^2 = -1$ ), il numero complesso

$$\omega_n = e^{i2\pi/n}$$

è una radice primitiva  $n$ -esima dell'unità.

Una particolare proprietà delle radici  $n$ -esime è che possono essere generate utilizzando una successione di potenze, infatti:

$$\omega_n^k = e^{ki2\pi/n} = \cos k \frac{2\pi}{n} + i \sin k \frac{2\pi}{n} \quad k = 1, \dots, n$$

$V$  è la matrice di Vandermonde di ordine  $n$ , i cui elementi sono

$$v_{kj} = \omega_n^{kj} \quad k, j = 0, \dots, n-1$$

### Esempio 15.2

Per  $n = 4$  scegliamo  $k = 1$  che corrisponde alle serie di potenze

$$1 \quad i \quad -1 \quad -i$$

La matrice  $V$  contiene tutte le potenze

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{pmatrix}$$

 **Proprietà 15.1**

$V$  è simmetrica (non Hermitiana, perchè sui complessi):

$$V^T = V$$

 **Teorema 15.3**

Vale la relazione di ortogonalità:

$$\sum_{j=0}^{n-1} \omega_n^{kj} = \begin{cases} n & \text{se } k \equiv 0 \pmod{n}, \\ 0 & \text{altrimenti} \end{cases} \quad (15.1)$$

(Ossia le somme per righe o per colonne sono  $n$  sulla prima riga/colonna, 0 sulle altre)

*Dimostrazione.*

- Caso  $k \equiv 0 \pmod{n}$ :

In questo caso esiste un intero  $s$  per cui  $k = sn$  e quindi è  $\omega_n^{jk} = 1 \quad \forall j$

- Caso  $k \not\equiv 0 \pmod{n}$ :

In questo caso è  $\omega_n^k \neq 1$ : ponendo  $x = \omega_n^k$  e utilizzando la nota relazione (serie geometrica)

$$(x + x^2 + \dots + x^{n-1})(1 - x) = 1 - x^n$$

si ottiene

$$\left( \sum_{j=0}^{n-1} \omega_n^{jk} \right) \underbrace{(1 - \omega_n^k)}_{\neq 0} = 1 - \underbrace{\omega_n^{nk}}_1$$

e, ricordando che  $\omega_n^k \neq 1$  e che  $\omega_n^{nk} = 1$ , segue la tesi.

□


 **Proprietà 15.2**

$$V^H V = nI$$

Il risultato si ottiene utilizzando (15.1)

 **Teorema 15.4**

$$V^H V = nI \quad \Rightarrow \quad V^{-1} = \frac{1}{n} V^H$$

 **Work in progress**

A lezione ha spiegato perchè: lascio commentata la dimostrazione che va riscritta.

Una conseguenza è che:

$$z = V^{-1}y \iff z = \frac{1}{n}V^H y$$

**Definizione 15.5 (Trasformata discreta di Fourier)**

L'applicazione

$$z = \frac{1}{n}V^H y$$

è detta trasformata discreta di Fourier e viene generalmente indicata con la sigla *DFT*, mentre il vettore  $z = DFT(y)$  è detto trasformata discreta di Fourier del vettore  $y$  e verifica la relazione

$$z_j = \frac{1}{n} \sum_{k=0}^{n-1} y_k \omega_n^{-jk}, \quad j = 1, \dots, n-1 \quad (15.2)$$

**Definizione 15.6 (Trasformata discreta inversa di Fourier)**

L'applicazione che al vettore  $z$  associa il vettore  $y$  è detta trasformata discreta inversa di Fourier e viene generalmente indicata con la sigla *IDFT*, mentre il vettore  $y = IDFT(z)$  è detto trasformata discreta inversa di Fourier del vettore  $z$  e verifica la relazione

$$y_k = \sum_{j=0}^{n-1} z_j \omega_n^{jk}, \quad k = 0, \dots, n-1 \quad (15.3)$$

Possiamo constatare adesso che possiamo fare il prodotto in  $O(n \log n)$   
 $y = Vz$  prodotto  $Vz$ : IDFT, trasformata inversa.



**Teorema 15.7**

Sia  $n = 2^s$ ; il costo computazionale del calcolo della IDFT di un vettore  $z$  di ordine  $n$  o del calcolo della DFT di un vettore  $y$  di ordine  $n$ , a meno di termini di ordine inferiore, è di  $(n/2) \log_2 n$  moltiplicazioni fra numeri complessi e  $n \log_2 n$  addizioni fra numeri complessi, non contando il calcolo delle  $n$  radici  $n$ -esime dell'unità.

*Dimostrazione.* Posto  $n = 2m$ , per la IDFT( $z$ ) si ha dalla (15.3):

$$y_k = \sum_{j=0}^{n-1} z_j \omega_n^{jk} = \sum_{j \text{ pari}} z_j \omega_n^{jk} + \sum_{j \text{ dispari}} z_j \omega_n^{jk} = \sum_{p=0}^{m-1} z_{2p} \omega_n^{2kp} + \sum_{p=0}^{m-1} z_{2p+1} \omega_n^{k(2p+1)}$$

Ponendo  $z'_p = z_{2p}$  e  $z''_p = z_{2p+1}$ ,  $p = 0, 1, \dots, m-1$ , si ha

$$y_k = \sum_{p=0}^{m-1} z'_p \omega_n^{2kp} + \sum_{p=0}^{m-1} z''_p \omega_n^{k(2p+1)}$$

Tenendo presente che  $\omega_n^{2p} = (\omega_{n/2})^p = \omega_m^p$ , è

$$y_k = \sum_{p=0}^{m-1} z'_p \omega_m^{kp} + \omega_n^k \sum_{p=0}^{m-1} z''_p \omega_m^{kp}, \quad k = 0, \dots, n-1 \quad (15.4)$$

Posto  $y' = IDFT(z')$  e  $y'' = IDFT(z'')$ , cioè

$$y'_q = \sum_{p=0}^{m-1} z'_p \omega_m^{qp}, \quad y''_q = \sum_{p=0}^{m-1} z''_p \omega_m^{qp}, \quad q = 0, \dots, m-1$$

dalla (15.4) segue che i primi  $m$  elementi della trasformata sono dati da

$$y_q = y'_q + \omega_n^q y''_q, \quad q = 0, \dots, m-1 \quad (15.5)$$

Per calcolare i rimanenti  $m$  elementi della trasformata, poiché  $\omega^m = -1$  e  $\omega_m^{q+m} = \omega_m^q$  dalla (15.4) segue

$$\begin{aligned} y_{q+m} &= \sum_{p=0}^{m-1} z'_p \omega_m^{(q+m)p} + \omega_n^{q+m} \sum_{p=0}^{m-1} z''_p \omega_m^{(q+m)p} = \\ &= \sum_{p=0}^{m-1} z'_p \omega_m^{qp} + \omega_n^{q+m} \sum_{p=0}^{m-1} z''_p \omega_m^{qp} = \\ &= y'_q - \omega_n^q y''_q, \quad q = 0, \dots, m-1 \end{aligned} \quad (15.6)$$

La trasformata di ordine  $n$  può quindi essere effettuata con 2 trasformate di ordine  $n/2 = m$  più  $n/2 = m$  moltiplicazioni e  $n$  addizioni.



#### Nota

$\omega_m^{qp}$  è comune sia per i primi  $m$  che per i secondi  $m$  elementi, quindi si calcola una volta sola. Bisogna poi fare 2 volte  $m$  somme, quindi  $n$  somme in totale.

Poiché la trasformata di ordine 1 non richiede operazioni, si possono scrivere le seguenti relazioni di ricorrenza per il numero di addizioni  $A(n)$  e di moltiplicazioni  $M(n)$  occorrenti per il calcolo della trasformata di ordine  $n$

$$\begin{aligned} A(1) &= 0, & A(n) &= 2A(n/2) + n, \\ M(1) &= 0, & M(n) &= 2M(n/2) + n/2. \end{aligned} \quad (15.7)$$

Posto  $t_s = A(n)$ , dove  $s = \log_2 n$ , dalla (15.7) si ottiene l'equazione alle differenze

$$t_0 = 0 \quad t_s = 2t_{s-1} + 2^s$$

la cui soluzione è data da

$$t_s = s2^s$$

da cui  $A(n) = n \log_2 n$ . Analogamente si ottiene  $M(n) = n/2 \log_2 n$ . Si procede nello stesso modo per la  $DFT(y)$ , eseguendo solo al termine le divisioni per  $n$ .

□

### Algoritmo

1. Si calcolano  $\omega_n$  e le sue potenze con esponente per  $k = 0, \dots, n-1$
2. Calcolo di  $y' = IDFT(z')$  di indice pari  
Calcolo di  $y'' = IDFT(z'')$  di indice dispari
- 3.

$$\begin{aligned} y_{m+q} &= y'_q - \omega_n^q y''_q \\ y_p &= y'_q + \omega_n^q y''_q \end{aligned}$$

Algoritmo FFT più famoso Cooley-Tuckey e Sande-Tuckey. Entrambi predispongono l'ordinamento di  $v$ . bit reversal.

$$z_0 z_1 z_2 z_3 z_4 z_5 z_6 z_7$$

$$z_0 z_2 z_4 z_6 \quad z_1 z_3 z_5 z_7$$

Permutazione di indici.

$$z_0 z_4 z_2 z_6 \quad z_1 z_5 z_3 z_7$$

**Applicazione: Prodotto di polinomi**

$$2u(x) = u_0 + u_1x + u_2x^2 + \underbrace{u_3}_0 x^3 + \underbrace{u_4}_0 x^4 \quad 2v(x) = v_0 + v_1x + v_2x^2 + \underbrace{v_3}_0 x^3 + \underbrace{v_4}_0 x^4 \quad 4z(z)$$

$$z(x) = u(x) - v(x) = z_0 + z_1x + z_2x^2 + z_3x^3 + z_4x^4$$

$n = 5$

$$\begin{pmatrix} u(\omega_5^0) \\ u(\omega_5^1) \\ \dots u(\omega_5^4) \end{pmatrix} = V \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} \quad V = \begin{pmatrix} \omega_5^0 & \omega_5^1 & \dots \end{pmatrix}$$

$$t = \begin{pmatrix} z(\omega_5^0) \\ \dots \end{pmatrix} \begin{pmatrix} u(\omega_5^0) = z(\omega_5^0) \\ \dots \end{pmatrix}$$

**Altre applicazioni**

- Prodotto di interi (è come moltiplicare due polinomi)
- FFT e le matrici circolanti: vengono fuori nei modelli che hanno rotazioni (immagini tomografiche)

$$\begin{pmatrix} a_0 & a_1 & a_2 & a_3 \\ a_1 & a_2 & a_3 & a_4 \\ a_2 & a_3 & a_4 & a_1 \\ a_3 & a_4 & a_1 & a_2 \end{pmatrix}$$

Le colonne della matrice di Fourier sono gli autovalori.

- Im(sfocata) = T Im Dove T è matrice di Toeplitz

$$\begin{pmatrix} a & b & c \\ d & a & b \\ e & d & a \end{pmatrix}$$

$$t_{ij} = \alpha^{i-j}$$

la Incorniciamo con una 7x7

$$\begin{pmatrix} a & b & c \\ d & a & b \\ e & d & a \end{pmatrix}$$

Allora diventa circolante. L'aumento di dimensione è compensato dall'uso di DFT.

**Nota**

All'esame vuole sapere l'idea del costo e di come si fanno i prodotti.





# 16 Richiami da Wikipedia

## 16.0.6 Teorema di Lagrange (o valor medio)



### Teorema 16.1 (Teorema di Lagrange (o valor medio))

Sia  $f : [a, b] \rightarrow \mathbb{R}$  una funzione continua in  $[a, b]$ , e derivabile in  $(a, b)$ . Allora

$$\exists c \in (a, b) : f'(c) = \frac{f(b) - f(a)}{b - a}$$

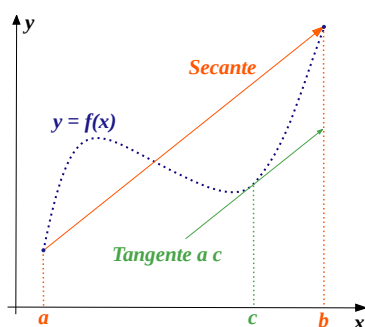


Figura 16.1: Idea intuitiva del teorema

Il teorema rimane valido considerando funzioni definite in  $\mathbb{R}^n$



### Teorema 16.2 (Teorema di Lagrange (o valor medio) su $\mathbb{R}^n$ )

Sia  $f$  una funzione reale derivabile su un aperto  $U \subseteq \mathbb{R}^n$ , siano  $\mathbf{x}, \mathbf{y}$  due punti di  $U$  tali che il segmento  $[\mathbf{x}, \mathbf{y}] = \{t\mathbf{x} + (1-t)\mathbf{y} : t \in [0, 1]\} \subseteq U$ .

Allora esiste  $\xi \in (\mathbf{x}, \mathbf{y})$  tale che

$$f(\mathbf{y}) - f(\mathbf{x}) = Df(\xi)(\mathbf{y} - \mathbf{x})$$

dove con  $Df$  indichiamo il differenziale.

## 16.0.7 Disuguaglianza triangolare

La *disuguaglianza triangolare* è una proprietà matematica. Essa è una delle proprietà caratterizzanti una distanza in uno spazio metrico. Formalmente tale proprietà afferma che, dato lo spazio metrico  $(X, d)$ , allora

$$\forall x, y, z \in X : d(x, y) \leq d(x, z) + d(z, y)$$

Se considerata in un qualsiasi spazio normato  $V$  con la metrica indotta e scegliendo come  $z$  lo 0 di  $V$ , essa dunque ci porta a dire che:

$$\|x + y\| = d(x, -y) \leq d(x, 0) + d(-y, 0) = \|x - 0\| + \|-y - 0\| = \|x\| + \|-y\| = \|x\| + \|y\|$$

Cioè

$$\forall x, y \in V : \|x + y\| \leq \|x\| + \|y\|$$

Dunque, all'interno dei numeri reali con la norma euclidea, assume la nota forma:

 **Proprietà 16.1 (Disuguaglianza triangolare sui moduli)**

$$\forall x, y \in \mathbb{R} : |x + y| \leq |x| + |y|$$