

# **Appunti di Calcolo numerico A.A.2021-2022**

Gabriele Frassi<sup>1</sup> ([g.frassi2@studenti.unipi.it](mailto:g.frassi2@studenti.unipi.it))

16 dicembre 2022

<sup>1</sup>Se questi appunti sono stati utili e vuoi ringraziarmi in qualche modo: <https://www.paypal.com/paypalme/GabrieleFrassi>

## Sommario

La mania di scrivere appunti a modo pure con un mese di tempo (agosto 2022) è degenerata in questa dispensa, che spero possa risultarvi utile nonostante il cambio di professore. Il prof. Massei ha posto come dispensa principale del corso quella del prof. Ghelardoni, su cui si basano pesantemente le diapositive e le lezioni dell'A.A.2021-2022.

Quest'opera è distribuita con licenza Creative Commons “Attribuzione – Non commerciale – Condividi allo stesso modo 4.0 Internazionale”.



# Indice

<b>I Unimap</b>	<b>11</b>
<b>II Lezioni</b>	<b>17</b>
<b>1 Teoria degli errori</b>	<b>18</b>
1.1 Rappresentazione dei numeri . . . . .	19
1.1.1 Teorema di rappresentazione . . . . .	19
1.1.1.1 Esempio . . . . .	20
1.1.1.2 Mantissa del numero . . . . .	20
1.1.2 Rappresentazione dei numeri di macchina . . . . .	21
1.1.2.1 Approssimazione per troncamento . . . . .	21
1.1.2.2 Approssimazione per arrotondamento . . . . .	21
1.1.3 Insieme dei numeri di macchina . . . . .	22
1.1.3.1 <i>overflow</i> . . . . .	22
1.1.3.2 <i>underflow</i> . . . . .	22
1.1.4 Errore assoluto ed errore relativo . . . . .	23
1.1.4.1 Errore assoluto nella rappresentazione del numero reale . . . . .	23
1.1.4.2 Errore relativo nella rappresentazione del numero reale . . . . .	23
1.1.4.3 Errore di macchina . . . . .	23
1.1.5 Operazioni di macchina . . . . .	24
1.1.5.1 Proprietà . . . . .	24
1.1.5.2 Effetto della cancellazione . . . . .	24
1.2 Errori nel calcolo della funzione . . . . .	25
1.2.1 Introduzione . . . . .	25
1.2.2 Insieme di indeterminazione del punto . . . . .	25
1.2.3 Errore assoluto della funzione . . . . .	26
1.2.3.1 Errore assoluto algoritmico . . . . .	26
1.2.3.2 Errore assoluto trasmesso dai dati . . . . .	26
1.2.3.3 Errore assoluto sulla componente $i$ -esima . . . . .	26
1.2.3.4 Coefficiente di amplificazione dell'errore assoluto . . . . .	26
1.2.3.5 Limitazione del modulo dell'errore assoluto . . . . .	26
1.2.4 Errore assoluto - Problema diretto e problema inverso . . . . .	27
1.2.4.1 Problema diretto . . . . .	27
1.2.4.2 Problema inverso . . . . .	28
1.2.4.3 Esempio 1 . . . . .	29
1.2.4.4 Esempio 2 . . . . .	30

1.2.4.5	Esempio 3 . . . . .	31
1.2.4.6	Esempio 4 . . . . .	31
1.2.5	Errore relativo della funzione . . . . .	32
1.2.5.1	Errore relativo algoritmico . . . . .	32
1.2.5.2	Algoritmo stabile e algoritmo instabile . . . . .	32
1.2.5.3	Errore relativo trasmesso dai dati . . . . .	32
1.2.5.4	Errore relativo rispetto alla componente $i$ -esima . . . . .	32
1.2.5.5	Coefficiente di amplificazione dell'errore relativo . . . . .	32
1.2.6	Errore trasmesso dai dati nelle quattro operazioni . . . . .	33
1.2.6.1	Esempio 1 . . . . .	35
1.2.6.2	Esempio 2 . . . . .	35
<b>2</b>	<b>Nozioni di Algebra lineare</b>	<b>36</b>
2.1	Definizioni base sulla matrice . . . . .	36
2.1.1	Matrice / Matrice quadrata / Matrice rettangolare . . . . .	36
2.1.2	Elementi diagonali / Diagonale principale . . . . .	36
2.1.3	Matrice reale . . . . .	36
2.1.4	Matrice identica . . . . .	37
2.2	Osservazioni sulle operazioni tra matrici . . . . .	37
2.3	Definizioni sui vettori . . . . .	37
2.3.1	Vettore . . . . .	37
2.3.2	Vettore reale . . . . .	37
2.3.3	Vettori linearmente indipendenti . . . . .	37
2.3.4	Prodotto scalare . . . . .	38
2.3.5	Vettori ortogonali . . . . .	38
2.4	Definizioni sui determinanti . . . . .	38
2.4.1	Determinante . . . . .	38
2.4.2	Matrice singolare / Matrice non singolare . . . . .	38
2.4.3	Minore di ordine $k$ . . . . .	38
2.4.3.1	Esempio 1 . . . . .	38
2.4.3.2	Esempio 2 . . . . .	39
2.4.4	Minori principali di ordine $k$ . . . . .	40
2.4.5	Minori principali di testa di ordine $k$ . . . . .	40
2.4.6	Rango (Caratteristica) . . . . .	40
2.4.6.1	Esempio 1 . . . . .	40
2.4.6.2	Esempio 2 . . . . .	41
2.4.7	Teorema di Binet-Cauchy (determinante del prodotto matriciale) . . . . .	41
2.4.7.1	Esempio 1 . . . . .	41
2.4.7.2	Esempio 2 . . . . .	42
2.5	Matrici inverse . . . . .	42
2.5.1	Matrice inversa . . . . .	42
2.5.2	Determinante della matrice inversa . . . . .	42
2.5.2.1	Esempio . . . . .	42
2.6	Classificazione delle matrici . . . . .	43
2.6.1	Matrice trasposta . . . . .	43
2.6.2	Determinante di una matrice trasposta . . . . .	43

2.6.2.1	Esempio . . . . .	43
2.6.3	Matrice simmetrica . . . . .	43
2.6.3.1	Esempio . . . . .	43
2.6.4	Matrice anti-simmetrica . . . . .	44
2.6.4.1	Esempi . . . . .	44
2.6.5	Matrice trasposta coniugata . . . . .	44
2.6.5.1	Esempio . . . . .	44
2.6.6	Matrice hermitiana . . . . .	44
2.6.6.1	Esempio . . . . .	45
2.6.7	Matrice anti-hermitiana . . . . .	45
2.6.7.1	Esempi . . . . .	45
2.6.8	Matrice normale . . . . .	46
2.6.8.1	Esempio . . . . .	46
2.6.9	Matrice unitaria . . . . .	46
2.6.9.1	Esempio . . . . .	46
2.6.10	Matrice ortogonale . . . . .	46
2.6.10.1	Esempio . . . . .	47
2.7	Segno della matrice . . . . .	47
2.7.1	Formula per l'individuazione del segno . . . . .	47
2.7.2	Matrice definita positiva . . . . .	47
2.7.3	Matrice definita negativa . . . . .	47
2.7.4	Matrice semidefinita positiva . . . . .	47
2.7.5	Matrice semidefinita negativa . . . . .	47
2.8	Matrici diagonali e triangolari . . . . .	48
2.8.1	Matrice diagonale . . . . .	48
2.8.2	Matrice triangolare inferiore . . . . .	48
2.8.3	Matrice triangolare superiore . . . . .	48
2.8.4	Determinante di una matrice diagonale/triangolare . . . . .	48
2.9	Matrice a predominanza diagonale . . . . .	49
2.9.1	Matrice a predominanza diagonale forte . . . . .	49
2.9.1.1	Esempio . . . . .	49
2.9.2	Matrice a predominanza diagonale debole . . . . .	49
2.9.2.1	Esempio . . . . .	49
2.10	Matrici convergenti e nilpotenti . . . . .	50
2.10.1	Matrice convergente . . . . .	50
2.10.2	Matrice nilpotente . . . . .	50
2.10.2.1	Esempio di matrice nilpotente . . . . .	50
2.11	Trasformaz. per similitudine e matrici di permutazione . . . . .	51
2.11.1	Trasformazioni per similitudine . . . . .	51
2.11.2	Matrice di permutazione . . . . .	51
2.11.2.1	Esempio . . . . .	51
2.11.2.2	Permutazione in simultanea sia su righe che su colonne . . . . .	52
2.11.2.3	Determinante della matrice permutata . . . . .	52
2.12	Partizionamenti a blocchi . . . . .	52
2.12.1	Esempio . . . . .	52

2.12.2	Matrice triangolare a blocchi inferiore . . . . .	53
2.12.2.1	Esempio . . . . .	53
2.12.3	Matrice triangolare a blocchi superiore . . . . .	53
2.12.3.1	Esempio . . . . .	53
2.12.4	Matrice diagonale a blocchi . . . . .	54
2.12.4.1	Esempio . . . . .	54
2.12.5	Determinante di una matrice diagonale/triangolare a blocchi . . . . .	54
2.12.5.1	Esempi . . . . .	54
2.13	Grafi . . . . .	55
2.13.1	Grafo orientato . . . . .	55
2.13.1.1	Esempio . . . . .	55
2.13.2	Grafo fortemente connesso . . . . .	55
2.14	Matrici riducibili . . . . .	56
2.14.1	Matrice riducibile . . . . .	56
2.14.2	Matrice irriducibile . . . . .	56
2.14.3	Forma ridotta . . . . .	56
2.14.4	Verifica dell'irriducibilità della matrice . . . . .	56
2.14.4.1	Esempio 1 . . . . .	57
2.14.4.2	Esempio 2 . . . . .	58
2.14.4.3	Esempio 3 . . . . .	59
2.15	Sistemi lineari . . . . .	59
2.15.1	Sistema lineare . . . . .	59
2.15.2	Teorema di Rouchè-Capelli . . . . .	59
2.15.3	Sistema normale . . . . .	60
2.15.4	Sistema omogeneo . . . . .	60
2.15.5	Risoluzione di un sistema lineare a blocchi . . . . .	60
2.16	Autovalori e autovettori . . . . .	62
2.16.1	Autovalore . . . . .	62
2.16.2	Autovettore destro . . . . .	62
2.16.3	Autovettore sinistro . . . . .	62
2.16.4	Polinomio caratteristico . . . . .	62
2.16.5	Traccia della matrice . . . . .	64
2.16.6	Raggio spettrale . . . . .	64
2.16.7	Teorema sulla matrice convergente . . . . .	64
2.16.8	Autovalori della matrice trasposta . . . . .	64
2.16.9	Teorema sugli autovalori di matrici simili . . . . .	64
2.16.10	Teorema sugli autovalori e autovettori di matrici di potenze . . . . .	65
2.16.10.1	Esempio sul fatto che abbiamo una C.S. . . . .	65
2.16.10.2	Esempio concreto di esercizio . . . . .	66
2.16.11	Teorema sugli autovalori di matrici inverse . . . . .	66
2.16.12	Teorema sugli autovalori di matrici hermitiane . . . . .	67
2.16.13	Molteplicità algebrica . . . . .	67
2.16.14	Molteplicità geometrica . . . . .	67
2.16.15	Teorema su legame tra molteplicità algebrica e geometrica . . . . .	68
2.16.16	Matrice diagonalizzabile . . . . .	69

2.16.17	Spettro . . . . .	69
2.16.18	Teorema sulla traslazione dello spettro . . . . .	69
2.17	Localizzazione degli autovalori . . . . .	70
2.17.1	Cerchi di Gershgorin . . . . .	70
2.17.2	Primo teorema di Gershgorin . . . . .	70
2.17.2.1	Corollario sulla matrice non singolare . . . . .	71
2.17.3	Secondo teorema di Gershgorin . . . . .	71
2.17.3.1	Corollario sui cerchi di Gershgorin . . . . .	72
2.17.4	Terzo teorema di Gershgorin . . . . .	72
2.17.4.1	Corollario sulla matrice non singolare . . . . .	72
2.17.5	Matrice di Frobenius . . . . .	73
2.17.5.1	Esempio . . . . .	73
2.18	Norme vettoriali e norme matriciali . . . . .	76
2.18.1	Norma vettoriale . . . . .	76
2.18.2	Norme classiche . . . . .	76
2.18.3	Teorema di equivalenza fra norme . . . . .	77
2.18.4	Norma matriciale . . . . .	77
2.18.5	Norma matriciale indotta (o naturale) . . . . .	77
2.18.5.1	Esempi . . . . .	78
2.18.6	Norme coerenti (o compatibili) . . . . .	78
2.18.7	Norma di Frobenius . . . . .	78
2.18.8	Teorema di Hirsh (legame tra raggio spettrale e norma matriciale) . . . . .	79
2.18.8.1	Corollario sulla convergenza di una matrice . . . . .	79
2.18.8.2	Corollario sui cerchi . . . . .	79
2.18.8.3	Raggio spettrale nelle matrici hermitiane . . . . .	79
2.18.9	Matrice di rotazione . . . . .	80
<b>3</b>	<b>Sistemi lineari</b>	<b>82</b>
3.1	Sistemi di equazioni lineari . . . . .	82
3.2	Classificazione dei metodi di risoluzione . . . . .	82
3.3	Metodi diretti . . . . .	83
3.3.1	<i>Metodo di Cramer</i> . . . . .	83
3.3.1.1	Spiegazione . . . . .	83
3.3.1.2	Costo computazionale . . . . .	83
3.3.2	<i>Metodo di Gauss (o metodo di eliminazione)</i> . . . . .	83
3.3.2.1	Spiegazione introduttiva . . . . .	83
3.3.2.2	Passi per ottenere il sistema equivalente . . . . .	83
3.3.2.3	Condizione per esecuzione del metodo di Gauss classico . . . . .	85
3.3.2.4	Risoluzione del sistema equivalente . . . . .	85
3.3.2.5	Costo computazionale . . . . .	86
3.3.2.6	Esempio 1 . . . . .	86
3.3.2.7	Risoluzione di $k$ sistemi con matrice $A$ comune: matrice inversa	87
3.3.3	Variante metodo di Gauss: tecniche di pivoting . . . . .	88
3.3.3.1	<i>pivoting parziale</i> . . . . .	88
3.3.3.2	<i>pivoting totale</i> . . . . .	89
3.3.4	Fattorizzazione LR (certe volte detta Fattorizzazione LU) . . . . .	90

3.3.4.1	Dimostrazione pratica . . . . .	90
3.3.4.2	Utilità della fattorizzazione . . . . .	91
3.3.4.3	Esempio di fattorizzazione passando da Gauss . . . . .	92
3.3.4.4	Esempio di fattorizzazione non passando da Gauss . . . . .	93
3.3.4.5	Fattorizzazione LR con pivoting parziale . . . . .	93
3.3.4.6	Determinante della matrice A . . . . .	94
3.3.5	Variante metodo di Gauss: Gauss-Jordan . . . . .	94
3.3.5.1	Esempio: calcolo dell'inversa con Gauss-Jordan . . . . .	94
3.4	Malcondizionamento di un sistema lineare . . . . .	96
3.4.1	Esempi introduttivi . . . . .	96
3.4.2	Definizione di sistema malcondizionato . . . . .	96
3.4.3	Caso particolare: matrice dei coefficienti non perturbata . . . . .	97
3.4.3.1	Numero di condizionamento del sistema con matrice $A$ . . . . .	98
3.4.3.2	Definizione aggiornata di malcondizionamento . . . . .	98
3.4.4	Caso generale con matrice perturbata . . . . .	98
3.4.5	Vettore residuo, errore assoluto e relativo . . . . .	99
3.4.6	Esempio: numero di condizionamento di matrici hermitiane . . . . .	100
3.4.7	Extra: esempio di matrice malcondizionata . . . . .	100
3.5	Metodi iterativi . . . . .	101
3.5.1	Introduzione . . . . .	101
3.5.2	Definizione di metodo convergente . . . . .	101
3.5.3	Costruzione dello schema iterativo . . . . .	101
3.5.4	Teorema di convergenza globale - C.N.S. . . . .	102
3.5.4.1	Errore associato all'iterazione . . . . .	102
3.5.4.2	Dimostrazione . . . . .	103
3.5.5	Condizione sufficiente per la convergenza del metodo iterativo . . . . .	104
3.5.6	Condizione necessaria per la convergenza del metodo iterativo . . . . .	104
3.5.7	Esempio di esercizio sulla convergenza del metodo . . . . .	104
3.5.8	Velocità asintotica di convergenza . . . . .	104
3.5.9	Criterio di arresto . . . . .	106
3.5.9.1	Criterio principale: norma inferiore a errore prefissato . . . . .	106
3.5.9.2	Ulteriore criterio: numero massimo di iterazioni . . . . .	107
3.5.10	Metodi classici . . . . .	107
3.5.10.1	Premesse . . . . .	107
3.5.10.2	Metodo di Jacobi (o metodo delle sostituzioni simultanee) . . . . .	108
3.5.10.3	Metodo di Gauss-Seidel (o metodo delle sostituzioni successive) . . . . .	109
3.5.10.4	Condizioni sufficienti di convergenza . . . . .	111
3.5.10.5	Esempio 1 . . . . .	112
3.5.10.6	Esempio 2 . . . . .	113
3.5.10.7	Esempio 3 . . . . .	114
<b>4</b>	<b>Sistemi non lineari</b>	<b>116</b>
4.1	Introduzione . . . . .	116
4.1.1	Definizione di equazione non lineare . . . . .	116
4.1.2	Cosa vogliamo fare . . . . .	116
4.1.2.1	Esempio di metodo diretto . . . . .	116

4.1.2.2	Rilevanza dei metodi iterativi . . . . .	116
4.1.2.3	Convergenza del metodo . . . . .	117
4.1.3	Metodo stazionario . . . . .	117
4.1.4	Ordine di convergenza e Fattore di convergenza . . . . .	117
4.1.5	Separazione grafica: numero di zeri e intervalli di separazione . . . . .	118
4.1.5.1	Esempio . . . . .	118
4.2	Metodi iterativi a due punti . . . . .	119
4.2.1	Metodo di bisezione (non stazionario) . . . . .	119
4.2.1.1	Spiegazione . . . . .	119
4.2.1.2	Esempio . . . . .	120
4.2.1.3	Criterio di arresto e numero di iterazioni utili . . . . .	121
4.2.1.4	Convergenza lineare . . . . .	122
4.2.2	Metodo delle secanti (stazionario) . . . . .	122
4.3	Metodi iterativi stazionari ad un punto . . . . .	123
4.3.1	Premesse . . . . .	123
4.3.1.1	Cosa vogliamo fare . . . . .	123
4.3.1.2	Schema iterativo . . . . .	123
4.3.1.3	Punto fisso della funzione $\phi(\alpha)$ . . . . .	123
4.3.1.4	Definizione di molteplicità . . . . .	123
4.3.2	Teorema di convergenza locale . . . . .	124
4.3.3	Teorema sull'ordine di convergenza . . . . .	126
4.3.4	Criterio di arresto . . . . .	127
4.3.5	Metodo di Newton (o <i>metodo delle tangenti</i> ) . . . . .	127
4.3.5.1	Introduzione . . . . .	127
4.3.5.2	Interpretazione grafica . . . . .	127
4.3.5.3	Ordine di convergenza e convergenza del metodo . . . . .	128
4.3.5.4	Variante: metodo con convergenza quadratica . . . . .	130
4.3.5.5	Costo computazionale . . . . .	130
4.3.5.6	Primo esempio su Matlab . . . . .	130
4.3.5.7	Secondo esempio su Matlab . . . . .	132
4.3.5.8	Condizioni sufficienti di convergenza . . . . .	132
4.3.6	Primo esempio . . . . .	133
4.4	Metodi iterativi in $\mathbb{R}^n$ . . . . .	135
4.4.1	Introduzione . . . . .	135
4.4.1.1	Funzione su più variabili . . . . .	135
4.4.1.2	Schema iterativo . . . . .	135
4.4.2	Teorema di convergenza locale su $\mathbb{R}^n$ . . . . .	135
4.4.2.1	Matrice jacobiana . . . . .	136
4.4.3	Metodo di Newton-Raphson (Newton su più variabili) . . . . .	136
4.4.3.1	Introduzione . . . . .	136
4.4.3.2	Costo computazionale . . . . .	136
4.4.3.3	Variante: risoluzione senza il calcolo delle inverse . . . . .	136
4.4.3.4	Variante: metodo di Newton semplificato . . . . .	137
4.4.4	Metodo non lineare di Jacobi-Newton . . . . .	138
4.4.4.1	Spiegazione . . . . .	138

4.4.4.2	Costo computazionale . . . . .	138
4.4.4.3	Variante: metodo non lineare di Gauss-Seidel . . . . .	138
4.5	Zeri di polinomi . . . . .	139
4.5.1	Equazioni oggetto di studio . . . . .	139
4.5.2	Successione di Sturm . . . . .	139
4.5.3	Funzione variazione . . . . .	139
4.5.4	Teorema di Sturm . . . . .	140
4.5.4.1	Successione di Sturm completa, corollario sulle radici . . . . .	140
4.5.5	Costruzione della successione di Sturm con Euclide . . . . .	140
4.5.6	Riflessioni sulla molteplicità delle soluzioni . . . . .	142
4.5.7	Uso della successione di Sturm per l'individuazione degli zeri . . . . .	143
<b>5</b>	<b>Calcolo degli autovalori</b>	<b>144</b>
5.1	Premessa: risoluzione dell'equazione caratteristica . . . . .	144
5.2	Metodo delle potenze . . . . .	144
5.2.1	Teorema del metodo delle potenze . . . . .	144
5.2.2	Criterio di arresto . . . . .	146
5.2.3	Metodo delle potenze normalizzato . . . . .	146
5.2.4	Estensione del teorema . . . . .	146
5.2.5	Applicazione del metodo alle matrici normali: deflazione . . . . .	147
5.3	Metodo di Jacobi per matrici reali e simmetriche . . . . .	148
5.3.1	Spiegazione . . . . .	148
5.3.2	Teorema di Jacobi . . . . .	149
5.3.3	Criterio di arresto . . . . .	149
5.3.4	Variante: metodo di Jacobi ciclico . . . . .	149
5.4	Riduzione in forma tridiagonale e di Hessenberg . . . . .	150
5.4.1	Introduzione . . . . .	150
5.4.2	Premessa: matrice tridiagonale . . . . .	150
5.4.3	Metodo di Givens per la tridiagonalizzazione . . . . .	150
5.4.3.1	Applicazione del metodo a matrici simmetriche . . . . .	150
5.4.3.2	Applicazione del metodo a matrici non simmetriche: Hessenberg	151
5.5	Metodo QR . . . . .	151
5.5.1	Fattorizzazione QR . . . . .	151
5.5.2	Fattorizzazione come metodo diretto per risoluzione di sistemi . . . . .	152
5.5.3	Algoritmo del metodo QR . . . . .	153
5.5.4	Teorema di Schur . . . . .	153
5.5.5	Teorema del metodo QR . . . . .	153
5.5.6	Uso della matrice di Hessenberg . . . . .	154
<b>6</b>	<b>Interpolazione e approssimazione di funzioni</b>	<b>155</b>
6.1	Introduzione . . . . .	155
6.1.1	Cosa vogliamo fare . . . . .	155
6.1.2	Funzione interpolante . . . . .	155
6.2	Interpolazione parabolica . . . . .	155
6.2.1	Polinomio di interpolazione . . . . .	155
6.2.2	Matrice di Vandermonde e risoluzione di un sistema . . . . .	156

	6.2.2.1	Grado massimo del polinomio di interpolazione . . . . .	156
	6.2.2.2	Unicità del polinomio di interpolazione . . . . .	156
6.2.3		Interpolazione di Lagrange . . . . .	157
	6.2.3.1	Polinomio fondamentale di interpolazione . . . . .	157
	6.2.3.2	Polinomio di interpolazione di Lagrange . . . . .	157
	6.2.3.3	Esempio . . . . .	157
6.2.4		Interpolazione di Newton . . . . .	158
	6.2.4.1	Premessa: differenze divise di ordine $k$ . . . . .	158
	6.2.4.2	Teorema di espansione per le differenze divise . . . . .	159
	6.2.4.3	Teorema di Newton (Polinomio di interpolazione di Newton) . . . . .	160
	6.2.4.4	Errore nell'interpolazione con le differenze divise . . . . .	161
	6.2.4.5	Quadro di Newton . . . . .	161
	6.2.4.6	Primo esempio . . . . .	162
	6.2.4.7	Secondo esempio . . . . .	162
	6.2.4.8	Terzo esempio . . . . .	162
	6.2.4.9	Quarto esempio . . . . .	163
6.3		Interpolazione osculatoria di Hermite . . . . .	164
	6.3.1	Cosa abbiamo . . . . .	164
	6.3.2	Definizione di interpolazione osculatoria di Hermite . . . . .	164
	6.3.3	Polinomio di interpolazione di Hermite . . . . .	164
	6.3.3.1	Polinomi $h_{0r}(x)$ . . . . .	165
	6.3.3.2	Polinomi $h_{1r}(x)$ . . . . .	165
	6.3.4	Errore nell'interpolazione . . . . .	166
6.4		Interpolazione con funzioni spline . . . . .	167
	6.4.1	Perchè ne parliamo . . . . .	167
	6.4.2	Definizione di funzione spline . . . . .	167
	6.4.3	Esempio: spline cubiche . . . . .	168
	6.4.4	Teorema sull'unicità della funzione spline . . . . .	168
6.5		Metodo dei minimi quadrati nel discreto . . . . .	169
	6.5.1	Spiegazione . . . . .	169
	6.5.2	Sistema delle equazioni normali . . . . .	170
	6.5.3	Sistemi lineari sovradeterminati e minimi quadrati . . . . .	171
	6.5.3.1	Definizione di sistema lineare sovradeterminato . . . . .	171
	6.5.3.2	Metodo dei minimi quadrati . . . . .	171
	6.5.3.3	Primo esempio . . . . .	171
	6.5.3.4	Secondo esempio . . . . .	172
	6.5.3.5	Terzo esempio . . . . .	173
	6.5.3.6	Quarto esempio . . . . .	173
	6.5.3.7	Quinto esempio . . . . .	173
	6.5.3.8	Sesto esempio . . . . .	174
<b>7</b>	<b>Integrazione numerica</b>		<b>175</b>
7.1		Introduzione . . . . .	175
	7.1.1	Promemoria: definizione di primitiva . . . . .	175
	7.1.2	Perchè si parla di integrazione? . . . . .	175
7.2		Grado di precisione ed errore . . . . .	176

7.2.1	Definizione di funzione peso e di momenti . . . . .	176
7.2.2	Approssimazione per mezzo di formula di quadratura . . . . .	176
7.2.3	Errore nell'approssimazione . . . . .	176
7.2.4	Grado di precisione . . . . .	176
7.2.5	Primo esempio (determinare pesi in $J_1$ ) . . . . .	177
7.2.5.1	Conseguenza dell'esempio: formula trapezoidale . . . . .	177
7.2.6	Secondo esempio (determinare pesi in $J_2$ ) . . . . .	178
7.2.6.1	Conseguenza dell'esempio: formula di Simpson . . . . .	179
7.2.7	Terzo esempio (determinare pesi e nodi in $J_1$ ) . . . . .	179
7.2.8	Teorema di Peano per la rappresentazione dell'errore . . . . .	180
7.2.8.1	Semplificazione con nucleo di Peano costante in segno . . . . .	181
7.2.8.2	Errore nella formula trapezoidale . . . . .	181
7.2.8.3	Errore nella formula di Simpson . . . . .	182
7.2.8.4	Esempio di esercizio . . . . .	182
7.3	Formule di tipo interpolatorio . . . . .	183
7.3.1	Dimostrazione di premessa alla definizione . . . . .	183
7.3.2	Definizione di formule di quadratura di tipo interpolatorio . . . . .	183
7.3.3	Unicità della formula di quadratura . . . . .	184
7.3.4	Formule di Newton-Cotes . . . . .	184
7.3.4.1	Caratteristiche . . . . .	184
7.3.4.2	Formula dei trapezi (generalizzazione della formula trapezoidale) . . . . .	184
7.3.4.3	Formula di Cavalieri-Simpson (generalizzaz. della formula di Simpson) . . . . .	185
7.3.4.4	Esempio con confronto tra Trapezi e Cavalieri-Simpson . . . . .	186
7.3.4.5	Tecnica di estrapolazione con la formula dei Trapezi . . . . .	187
<b>III</b>	<b>Appendici</b>	<b>189</b>
<b>A</b>	<b>Calcolo di determinanti</b>	<b>190</b>
<b>B</b>	<b>Differenze col programma di Massei (A.A.22-23 vs A.A.21-22)</b>	<b>192</b>

# Parte I

# Unimap

1. **Mer 29/09/2021 14:00-15:00 (1:0 h)** lezione: Introduzione al corso. Programma, testi consigliati, orario di ricevimento, esercizi d'esame, modalità di svolgimento degli esami. (Paolo Ghelardoni)
2. **Mer 29/09/2021 15:00-16:00 (1:0 h)** lezione: Rappresentazione dei numeri reali in base BETA maggiore di uno: esponente, mantissa e cifre della rappresentazione. Teorema di unicità della rappresentazione. Rappresentazione finita di un numero: troncamento e arrotondamento. Massimo errore assoluto commesso nei due casi. (Paolo Ghelardoni)
3. **Gio 30/09/2021 08:30-10:30 (2:0 h)** lezione: Errore assoluto ed errore relativo nella rappresentazione di un numero reale. Precisione di macchina. Insieme dei numeri di macchina: cardinalità, operazioni, mancanza di alcune proprietà che valgono per le operazioni tra numeri reali. Errore nel calcolo di una funzione di n variabili in un punto assegnato. Errore assoluto: errore assoluto algoritmico ed errore assoluto trasmesso dai dati. Maggiorazione del valore assoluto dell'errore assoluto trasmesso dai dati. Problema diretto: esempio. (Paolo Ghelardoni)
4. **Ven 01/10/2021 08:30-09:30 (1:0 h)** lezione: Errore assoluto: problema inverso con esempio. Errore relativo nel calcolo di una funzione: errore relativo algoritmico, errore relativo trasmesso dai dati. Espressione dell'errore relativo trasmesso dai dati. Errore assoluto ed errore relativo trasmesso dai dati nel caso delle quattro operazioni. (Paolo Ghelardoni)
5. **Mer 06/10/2021 14:00-15:00 (1:0 h)** esercitazione: Calcolo dell'errore relativo nella valutazione di una funzione: esempi con algoritmi di calcolo diversi. (Paolo Ghelardoni)
6. **Mer 06/10/2021 15:00-16:00 (1:0 h)** lezione: Richiami di nozioni di Algebra Lineare: matrici, vettori, matrici diagonali, matrici triangolari, trasposta, hermitiana. Determinante di una matrice. Minori, minori principali e minori principali di testa. Rango o caratteristica di una matrice. Teorema generale di Binet. (Paolo Ghelardoni)
7. **Gio 07/10/2021 08:30-10:30 (2:0 h)** lezione: Richiami di Algebra Lineare: matrice inversa, matrici hermitiane, matrici unitarie e matrici ortogonali. Matrici di permutazione, matrici convergenti, matrici a predominanza forte e a predominanza forte. Sistemi lineari: teorema di Rouche'-Capelli. Matrici partizionate a blocchi. Matrici riducibili: grafo orientato, grafo orientato fortemente connesso. (Paolo Ghelardoni)
8. **Ven 08/10/2021 08:30-09:30 (1:0 h)** esercitazione: Matrici riducibili: costruzione del grafo orientato, studio del grafo e determinazione di una matrice di permutazione che riduce la matrice data. (Paolo Ghelardoni)
9. **Mer 13/10/2021 14:00-15:00 (1:0 h)** lezione: Richiami sulla definizione di autovalore: polinomio caratteristico, equazione caratteristica. Trasformazioni per similitudine. molteplicità algebrica e molteplicità geometrica degli autovalori. Diganalizzabilità di una matrice. (Paolo Ghelardoni)
10. **Mer 13/10/2021 15:00-16:00 (1:0 h)** esercitazione: Applicazione della riducibilità di una matrice alla risoluzione di un sistema lineare. Relazione tra traccia e determinante di una matrice con gli autovalori. Autovalori delle potenze di una matrice. Autovalori di una matrice hermitiana. (Paolo Ghelardoni)
11. **Gio 14/10/2021 08:30-09:30 (1:0 h)** lezione: Traslazione dello spettro di una matrice. Cerchi di Gershgorin, primo Teorema di Gershgorin. Corollario del I° Teorema di Gershgorin. II° e III° Teorema di Gershgorin. Corollario del III° Teorema di Gershgorin. Matrice di Frobenius (Paolo Ghelardoni)

12. **Gio 14/10/2021 09:30-10:03 (1:0 h)** esercitazione: Esempi di calcolo degli autovalori di matrici utilizzando la traccia della matrice e la traslazione dello spettro. (Paolo Ghelardoni)
13. **Ven 15/10/2021 08:30-09:30 (1:0 h)** lezione: Esempio di applicazione della matrice di Frobenius. Norme vettoriali e norme matriciali. Norme classiche. Norma matriciali indotte. Norme vettoriali e norme matriciali coerenti tra loro. Relazione tra raggio spettrale e norma matriciale: teorema di Hirsh. (Paolo Ghelardoni)
14. **Mer 20/10/2021 14:00-15:00 (1:0 h)** esercitazione: Conseguenze del Teorema di Hirsh: condizione sufficiente per matrici convergenti, nuova possibile localizzazione degli autovalori. Matrici di rotazione: struttura, ortogonalità ed autovalori. (Paolo Ghelardoni)
15. **Mer 20/10/2021 15:00-16:00 (1:0 h)** lezione: Sistemi lineari: metodi diretti, metodi iterativi. Confronto del costo computazionale tra il metodo di Cramer e il metodo di Gauss. Algoritmo del metodo di Gauss: moltiplicatori, condizioni di applicabilità senza dover ricorrere a scambi di righe, risoluzione del sistema finale. (Paolo Ghelardoni)
16. **Gio 21/10/2021 08:30-09:30 (1:0 h)** esercitazione: Risoluzione di un sistema lineare applicando il metodo di Gauss. Come si può calcolare la matrice inversa di una matrice data inserendo la matrice identica come matrice di termini noti. Cosa significa applicare la tecnica di pivoting parziale con riferimento alle matrici di permutazione. Tecnica di pivoting totale. (Paolo Ghelardoni)
17. **Gio 21/10/2021 09:30-10:30 (1:0 h)** lezione: Fattorizzazione LR di una matrice: matrici elementari di Gauss, matrice L e matrice R. Utilizzo della fattorizzazione LR per la risoluzione di un sistema lineare. (Paolo Ghelardoni)
18. **Ven 22/10/2021 08:30-09:30 (1:0 h)** esercitazione: Esempi di calcolo della fattorizzazione di matrici quadrate con o senza l'utilizzo dell'algoritmo di Gauss. Metodo di Gauss-Jordan: costo computazionale. Calcolo della matrice inversa seguendo l'algoritmo del metodo di Gauss-Jordan (Paolo Ghelardoni)
19. **Mer 27/10/2021 14:00-15:00 (1:0 h)** lezione: Problema del malcondizionamento nella risoluzione di un sistema lineare: esempi. Caso particolare di perturbazione nulla sulla matrice dei coefficienti. Numero di condizionamento: sua limitazione inferiore. Maggiorazione dell'errore relativo sulla soluzione nel caso generale. (Paolo Ghelardoni)
20. **Mer 27/10/2021 15:00-16:00 (1:0 h)** lezione: Analisi all'indietro dell'errore commesso nella risoluzione di un sistema lineare tenendo conto del condizionamento del problema. Numero di condizionamento in norma 2 di una matrice hermitiana. Metodi iterativi per sistemi lineari: costruzione di un generico metodo, come si genera una successione di vettori. (Paolo Ghelardoni)
21. **Gio 28/10/2021 08:30-10:30 (2:0 h)** lezione: Metodi iterativi per sistemi lineari: costruzione di un generico metodo, come si genera una successione di vettori, teorema di convergenza e le possibili condizioni di convergenza. Velocità asintotica di convergenza di un processo iterativo. Criterio di arresto. Numero massimo di iterazioni. Decomposizione della matrice  $A=D-E-F$ . (Paolo Ghelardoni)
22. **Ven 29/10/2021 08:30-09:30 (1:0 h)** lezione: Metodo iterativo di Jacobi: struttura della matrice di iterazione. Metodo iterativo di Gauss-Seidel: particolarità della matrice di iterazione. Condizioni sufficienti per la convergenza dei metodi di Jacobi e di Gauss-Seidel. (Paolo Ghelardoni)
23. **Mer 03/11/2021 14:00-16:00 (2:0 h)** esercitazione: Versioni per componenti dei metodi di Jacobi e di Gauss-Seidel. Studio della convergenza dei due metodi per la risoluzione di alcuni sistemi lineari. (Paolo Ghelardoni)

24. **Gio 04/11/2021 08:30-09:30 (1:0 h)** esercitazione: Applicazione dei metodi di Jacobi e di Gauss-Seidel mediante codici di calcolo scritti nel linguaggio Matlab. Equazioni non lineari: tecnica di separazione grafica delle soluzioni dell'equazione. (Paolo Ghelardoni)
25. **Gio 04/11/2021 09:30-10:30 (1:0 h)** lezione: Definizione di ordine di convergenza di una successione convergente. Metodo di bisezione: algoritmo, criterio di arresto, numero di iterazioni, ordine e fattore di convergenza, costo computazionale di ogni iterazione. (Paolo Ghelardoni)
26. **Ven 05/11/2021 08:30-09:30 (1:0 h)** lezione: Applicazione del metodo di bisezione utilizzando un codice scritto nel linguaggio Matlab. Metodo delle secanti: interpretazione grafica, ordine di convergenza, costo computazionale. Metodi iterativi stazionari ad un punto: costruzione. Enunciato del Teorema di convergenza locale. (Paolo Ghelardoni)
27. **Mer 10/11/2021 14:00-16:00 (2:0 h)** lezione: Teorema di convergenza locale. Teorema relativo all'ordine di convergenza di un metodo stazionario ad un punto. Criterio di arresto. Metodo di Newton: formulazione, interpretazione grafica e ordine di convergenza nel caso di radici semplici con la certezza dell'esistenza di valori iniziali che lo rendono convergente. (Paolo Ghelardoni)
28. **Gio 11/11/2021 08:30-09:30 (1:0 h)** lezione: Ordine di convergenza del metodo di Newton nel caso di radici multiple. Condizioni sufficienti e scelta del punto iniziale per la convergenza del metodo di Newton. Efficienza dei metodi iterativi stazionari ad un punto. (Paolo Ghelardoni)
29. **Gio 11/11/2021 09:30-10:30 (1:0 h)** esercitazione: Studio di alcune equazioni: intervalli di separazione delle soluzioni, studio della convergenza di alcuni metodi ricavati dall'equazione, studio della convergenza del metodo di Newton. (Paolo Ghelardoni)
30. **Ven 12/11/2021 08:30-09:30 (1:0 h)** lezione: Ultimi esempi di studio di equazioni non lineari. Si stemi di equazioni non lineari: costruzione dei metodi iterativi, estensione del Teorema di convergenza locale. Metodo di Newton-Raphson: struttura, modalità di applicazione. (Paolo Ghelardoni)
31. **Mer 17/11/2021 14:00-16:00 (2:0 h)** lezione: Sistemi non lineari: metodo di Newton semplificato, metodo non lineare di Jacobi-Newton, metodo non lineare di Gauss-Seidel, cenno al metodo delle secanti. Equazioni algebriche: successione di Sturm, successione completa, funzione variazione e Teorema di Sturm. Primo esempio. (Paolo Ghelardoni)
32. **Gio 18/11/2021 08:30-09:30 (1:0 h)** esercitazione: Esempi di applicazione del Teorema di Sturm a particolari equazioni algebriche con parametri tra i coefficienti. (Paolo Ghelardoni)
33. **Gio 18/11/2021 09:30-10:30 (1:0 h)** lezione: Costruzione della successione di Sturm relativa al polinomio caratteristico di una matrice tridiagonale. Metodi numerici per il calcolo degli autovalori. Metodo delle potenze. Teorema di convergenza del metodo delle potenze. (Paolo Ghelardoni)
34. **Ven 19/11/2021 08:30-09:30 (1:0 h)** lezione: Metodo delle potenze classico: criterio di arresto. Metodo delle potenze normalizzato. Tecnica di deflazione nel caso di matrici normali. Matrici di rotazione. Metodo di Jacobi per il calcolo degli autovalori di una matrice reale e simmetrica. (Paolo Ghelardoni)
35. **Mer 24/11/2021 14:00-16:00 (2:0 h)** lezione: Metodo di Jacobi per matrici reali e simmetriche. Metodo di Jacobi ciclico. Metodo di Givens per trasformare una matrice reale e simmetrica in una matrice tridiagonale o trasformare una matrice reale nella forma di Hessenberg superiore. Fattorizzazione QR di una matrice. Algoritmo del metodo QR con teorema di convergenza. (Paolo Ghelardoni)

36. **Gio 25/11/2021 08:30-09:30 (1:0 h)** lezione: Interpolazione parabolica. Esistenza ed unicità del polinomio interpolante: matrice di Vandermonde. Polinomi fondamentali della interpolazione di Lagrange, polinomio interpolante nella forma di Lagrange. Teorema di espansione facendo uso delle differenze divise. (Paolo Ghelardoni)
37. **Gio 25/11/2021 09:30-10:30 (1:0 h)** esercitazione: Calcolo di un polinomio interpolante nella forma di Lagrange. Definizione delle differenze divise di ordine k. Alcune proprietà delle differenze divise. (Paolo Ghelardoni)
38. **Ven 26/11/2021 08:30-09:30 (1:0 h)** lezione: Polinomio di interpolazione nella forma di Newton. Espressione dell'errore commesso nella interpolazione. Quadro delle differenze divise. Primo esempio di calcolo del polinomio interpolante. (Paolo Ghelardoni)
39. **Mer 01/12/2021 14:00-15:00 (1:0 h)** esercitazione: Esempi di calcolo di polinomi di interpolazione con e senza la presenza di parametri reali. (Paolo Ghelardoni)
40. **Mer 01/12/2021 15:00-16:00 (1:0 h)** lezione: Interpolazione osculatoria di Hermite: proprietà del polinomio di interpolazione di Hermite, polinomi di prima e seconda specie, loro struttura. Calcolo dei polinomi di prima specie. (Paolo Ghelardoni)
41. **Gio 02/12/2021 08:30-09:30 (1:0 h)** esercitazione: Calcolo dei polinomi di seconda specie della interpolazione di Hermite. Esempio di applicazione della interpolazione di Hermite per raccordare tra loro due binari ferroviari. (Paolo Ghelardoni)
42. **Gio 02/12/2021 09:30-10:30 (1:0 h)** lezione: Interpolazione mediante funzioni spline: spline lineari e spline cubiche. Cenni sulla determinazione di una spline cubica su un insieme di punti in progressione aritmetica. Introduzione alla approssimazione di funzioni con il metodo dei minimi quadrati. (Paolo Ghelardoni)
43. **Ven 03/12/2021 08:30-09:30 (1:0 h)** lezione: Metodo dei minimi quadrati nel discreto. Sistema delle Equazioni Normali: caso di soluzione unica, proprietà della matrice dei coefficienti. Esempio di approssimazione di una funzione con criterio di scelta tra due o più approssimazioni. (Paolo Ghelardoni)
44. **Gio 09/12/2021 08:30-09:30 (1:0 h)** esercitazione: Sistemi lineari sovradeterminati: soluzione nel senso dei minimi quadrati. Esempi vari. (Paolo Ghelardoni)
45. **Gio 09/12/2021 09:30-10:30 (1:0 h)** lezione: Integrazione numerica: posizione del problema, formule di quadratura. Grado di precisione (algebrico). Formula trapezoidale e formula di Simpson. (Paolo Ghelardoni)
46. **Ven 10/12/2021 08:30-09:30 (1:0 h)** lezione: Calcolo di pesi e nodi di una formula a due punti. Espressione dell'errore nella integrazione numerica: Teorema di Peano, Nucleo di Peano. Espressione dell'errore semplificata nel caso in cui il nucleo di Peano sia di segno costante. (Paolo Ghelardoni)
47. **Mer 15/12/2021 14:00-15:00 (1:0 h)** esercitazione: Calcolo della espressione dell'errore per la formula trapezoidale e per la formula di Simpson. Calcolo dell'errore per una formula di quadratura a due punti. (Paolo Ghelardoni)
48. **Mer 15/12/2021 15:00-16:00 (1:0 h)** lezione: Espressione dell'errore nelle formule di Newton-Cotes. Costruzione delle formule generalizzate: formula dei trapezi e formula di Cavalieri-Simpson, espressione dell'errore. (Paolo Ghelardoni)
49. **Ven 17/12/2021 08:30-09:30 (1:0 h)** esercitazione: Esempio di applicazione della formule di Newton-Cotes generalizzate con valutazione della formula da preferirsi in base al costo computazio-

nale. Estrapolazione di Richardson applicata alla approssimazione di un integrale con la formula dei trapezi: formula di Romberg. (Paolo Ghelardoni)

50. **Lun 20/12/2021 09:30-10:30 (1:0 h)** lezione: Polinomi ortogonali: prodotto scalare che li definisce, proprietà degli zeri. Formule di quadratura gaussiane: scelta dei nodi, segno dei pesi, grado di precisione, espressione dell'errore. Esempi di alcune formule gaussiane. (Paolo Ghelardoni)

# **Parte II**

# **Lezioni**

# Capitolo 1

## Teoria degli errori

## TEORIA DEGLI ERRORI - RAPPRESENTAZIONE DEI NUMERI

### — TEOREMA DI RAPPRESENTAZIONE —

FISSATO UN INTERO  $\beta > 1$ , OGNI NUMERO NON NULLO  $x \in \mathbb{R} - \{0\}$  AMMETTE UNA UNICA RAPPRESENTAZIONE IN BASE  $\beta$ , DATA DA

$$x = \text{sign}(x) \beta^b \sum_{i=1}^{\infty} d_i \beta^{-i}$$

↓                      ↓                      ↓  
FUNZIONE SEGNO    ESPOLENTE b    CIFRE DELLA RAPPRESENTAZIONE  
 $\text{sign}(x) = \begin{cases} 1 & x > 0 \\ -1 & x < 0 \end{cases}$   
 $b \in \mathbb{Z}$ , CON  $\beta^b$  INDICO  
 L'ORDINE DI GRANDEZZA  
 DEL NUMERO  
 TALE CHE  
 1)  $d_1 \neq 0$   
 2)  $\exists k \in \mathbb{N} : d_j = \beta - 1 \quad \forall j > k$

LE DUE CONDIZIONI SONO NECESSARIE AL FINE DI GARANTIRE L'UNICITÀ DELLA RAPPRESENTAZIONE

CON (1) ESCLUO RAPPRESENTAZIONI DEL TIPO

$$\begin{aligned}
 0.123 \cdot 10^5 &\longrightarrow 0.\textcolor{red}{0}123 \cdot 10^6 && \text{NON VA BENE, } d_1 = 0 \\
 &\longrightarrow 0.\textcolor{red}{0}0123 \cdot 10^7 && \text{NON VA BENE, } d_1 = 0 \\
 &\longrightarrow 12300.\textcolor{red}{0} \cdot 10^0 && \text{NON VA BENE, } d_1 = 0
 \end{aligned}$$

CON (2) ESCLUO RAPPRESENTAZIONI IN PERIODO  $\beta - 1$ , DOPPIONI DI ALTRE RAPPRESENTAZIONI!

1.49 NUMERI PERIODICI SONO RAZIONALI, QUINDI POSSONO ESSERE SCRITTI COME RAPPORTO TRA DUE INTERI (FRAZIONE GENERATRICE)

$$x = \frac{149 - 14}{900} = \frac{135}{90} = \frac{15}{10} = 1.5 \quad \text{DOPPIO DI } 1.5 !!$$

AL NUMERATORE IL NUMERO SENZA SEPARATORE È PERIODO, NENO LA PARTE NON PERIODICA.

AL DENOMINATORE

- TANTI 9 QUANTI IL NUMERO DI CIFRE DEL PERIODO
- TANTI 0 QUANTI IL NUMERO DI CIFRE DELL'ANTIPERIODO

ESEMPIO.  $0.123 \cdot 10^5 = (+1) \cdot 10^5 \left( 1 \cdot 10^{-1} + 2 \cdot 10^{-2} + 3 \cdot 10^{-3} \right)$

$\uparrow$  sign(x)       $\uparrow$   $\beta^b$        $\uparrow$   $\sum_{i=1}^m d_i \beta^{-i}$

SI PARLA DI "RAPPRESENTAZIONE DEL NUMERO REALE IN VIRGOLO MOBILE NORMAIZZATA". LA SEGUENTE SOMMATORIA

$$\sum_{i=1}^{\infty} d_i \beta^{-i}$$

É DETTA MANTISSA DEL NUMERO. POSSIAMO DIRE CHE

$$\frac{1}{\beta} \leq \sum_{i=1}^{\infty} d_i \beta^{-i} < 1$$

VALORE MINIMO ACCETTABILE  
 $d_1 = 1, d_j = 0 \quad \forall j > 1$

$$\sum_{i=1}^{\infty} d_i \beta^{-i} = 1 \cdot \beta^{-1} = \frac{1}{\beta}$$

SUPPONIAMO DI AVERE UN NUMERO CON SOLE CIFRE  $\beta-1$ , CIOÈ  
 $d_j = \beta-1 \quad \forall j \geq 1$   
 (COSA ESCLUSA NEL TEOREMA)

OTTENIAMO

$$\sum_{i=1}^{\infty} d_i \beta^{-i} = \sum_{i=1}^{\infty} (\beta-1) \beta^{-i} = (\beta-1) \sum_{i=1}^{\infty} \beta^{-i}$$

SERIE GEOMETRICA

$$\sum_{i=0}^{\infty} q^m$$

DOVE  $q$  È LA RAGIONE

DOVE  $\sum_{i=1}^{\infty} \beta^{-i}$  È UNA SERIE GEOMETRICA

DI RAGIONE  $\frac{1}{\beta}$  E PRIMO TERMINE  $\frac{1}{\beta}$

$0 < \frac{1}{\beta} < 1$ , CONVERGE SICURAMENTE

$$\sum_{i=1}^{\infty} d_i \beta^{-i} < (\beta-1) \frac{1}{\beta} \frac{1}{1 - \left(\frac{1}{\beta}\right)} = \frac{\beta-1}{\beta} \cdot \frac{\beta}{\beta-1} = 1 !!!$$

$\downarrow$  PRIMO TERMINE       $\downarrow$  RAGIONE

## - RAPPRESENTAZIONE DEI NUMERI DI MACCHINA

A FONDAMENTI DI PROGRAMMAZIONE SI È VISTO CHE IL NUMERO DI CIFRE DELLA MANTISSA È  $m \in \mathbb{N}$ , NON È POSSIBILE RAPPRESENTARE UN NUMERO INFINTO DI CIFRE ("PRIMA O POI LO SPAZIO FINISCE", CIT.). POSSIBILI DUE RAPPRESENTAZIONI:

PER TRONCAMENTO

$$tr(x) = \text{sign}(x) \beta^b \sum_{i=1}^m d_i \beta^{-i}$$

PRENDO SOLO LE PRIME  $m$  CIFRE UTILI, SCARTANDO LE SUCCESSIVE.

LA DIFFERENZA TRA IL TRONCAMENTO DEL NUMERO E IL NUMERO STESSO È LA SEGUENTE

$$|tr(x) - x| < \beta^{b-m}$$

PER ARROTONDAMENTO

(RISPETTO ALLA  $m$ -ESIMA CIFRA)

$$rd(x) = \begin{cases} tr(x) & 0 \leq d_{m+1} < \frac{\beta}{2} \\ \text{sign}(x) \beta^b \left[ \sum_{i=1}^m d_i \beta^{-i} + \beta^{-m} \right] & \frac{\beta}{2} \leq d_{m+1} < \beta \end{cases}$$

CON  $0 \leq d_{m+1} < \frac{\beta}{2}$  ABBIANO UN "ARROTONDAMENTO PER DIFETTO"

CON  $\frac{\beta}{2} \leq d_{m+1} < \beta$  ABBIANO UN "ARROTONDAMENTO PER ECCESSO"

- PRENDO LE PRIME  $m$  CIFRE CON  $\sum_{i=1}^m d_i \beta^{-i}$ , MA
- AGGIORNO L'ULTIMA CIFRA CON  $\beta^{-m}$

$$\begin{aligned} \text{ES. } 0.995 \text{ CON } m=2 &\longrightarrow 0.99 + 10^{-2} = 0.99 + \frac{1}{100} = 1.00 \\ &= 0.1 \cdot 10^1 \end{aligned}$$

SI OSSERVI DALL'ESEMPIO CHE POTREBBE ESSERE NECESSARIA UNA NUOVA NORMALIZZAZIONE CON MODIFICA DELL'ESPONENTE  $b$ , E NON SOLO DELLE CIFRE PRECEDENTEMENTE DETERMinate

LA DIFFERENZA TRA L'ARROTONDAMENTO DEL NUMERO E IL NUMERO STESSO È LA SEGUENTE

$$|rd(x) - x| \leq \frac{1}{2} \beta^{b-m}$$

MAGGIORE PRECISIONE RISPETTO AL TRONCAMENTO!

## — INSIEME DEI NUMERI DI MACCHINA —

DEFINIAMO  $M = F(\beta, m, l, u)$  L'INSIEME DEI NUMERI DI MACCHINA, CIOE L'INSIEME DEI NUMERI Z RAPPRESENTABILI ALL'INTERNO DI UN CALCOLATORE

- È UN INSIEME FINITO
- HA LA SEGUENTE CARDINALITÀ, FISSATA LA BASE  $\beta$  E SUPPOSTO  $l \leq b \leq u$  CON  $l, u \in \mathbb{Z}$

$$\text{Card}(M) = 2 (\beta^m - \beta^{m-1}) (u-l+1) + 1$$

↓

SEGNO DEL NUMERO

ZERO, CHE PERO' NON  
GODE DELL'UNICITÀ

PER OGNI MANTISSA TANTI NUMERI  
QUANTI GLI ESPONENTI RAPPRESENTABILI  
 $u-l+1$  (+1 PER GLI ESTREMI INCLUSI)

PRIMA CIFRA:  $0 < d_1 \leq \beta-1 \rightarrow \beta-1$  CIFRE POSSIBILI  
CIFRE SUCCESSIVE:  $0 \leq d_{\beta} \leq \beta-1 \rightarrow \beta$  CIFRE POSSIBILI

POSSIBILI MANTISSE RAPPRESENTABILI:  $(\underbrace{\beta-1}_{m-1 \text{ VOLTE}}) \beta \dots \beta = (\beta-1) \beta^{m-1} = \beta^m - \beta^{m-1}$

**OVERFLOW** PUÒ SUCCEDERE CHE NEL CALCOLO DI  $rd(x)$  SI OTTENGA UN NUMERO "TROPPO GRANDE" NON APPARTENENTE AD  $M$ .

$$F(10, 3, -99, 99) \Rightarrow rd(x) = 0.1 \cdot 10^{100} \notin M$$

$x = 0.9998 \cdot 10^{-99}$

DUE STRADE PER IL CALCOLATORE:

- SEGNALARE L'ERRORE E FERMARSI
- PORRE  $rd(x) = \text{sign}(x) \max_{y \in M} |y|$ , IL VALORE MASSIMO POSSIBILE

**UNDERFLOW** PUÒ SUCCEDERE CHE NEL CALCOLO DI  $rd(x)$  SI OTTENGA UN NUMERO "TROPPO PICCOLO" NON APPARTENENTE AD  $M$ .

$$F(10, 3, -99, 99) \Rightarrow rd(x) = 0.1 \cdot 10^{-100} \notin M$$

$x = 0.01 \cdot 10^{-99}$

DUE STRADE PER IL CALCOLATORE:

- SEGNALARE L'ERRORE (NON LO FANNO TUTTI) E FERMARSI
- PORRE  $rd(x) = 0$

## - ERRORE ASSOLUTO ED ERRORE RELATIVO

DEFINISCO "ERRORE ASSOLUTO DELLA RAPPRESENTAZIONE DEL NUMERO REALE  $x$ " LA SEGUENTE DIFFERENZA (SI OSSERVI CHE PUÒ ESSERE NEGATIVA)

$$\delta_x = \text{rad}(x) - x$$

ES. MISURO UN TAVOLO LUNGO 2m  $\rightarrow x = 1.99 \quad \delta_x = 2.00 - 1.99 = 0.01\text{m}$

$$x = 2.01 \quad \delta_x = 2.00 - 2.01 = -0.01\text{m}$$

É IMMEDIATO  $|\delta_x| \leq \frac{1}{2} \beta^{b-m}$ , IL MASSIMO ERRORE ASSOLUTO DIPENDE DA ESPONENTE  $b$  E NUM. DI CIFRE  $m$

$$\text{ES. } 0.5 \cdot 10^5 \rightarrow 0.2 \cdot 10^5 \quad \delta_x = 0.05 \cdot 10^5 = 0.5 \cdot 10^4 = \frac{1}{2} 10^4$$

5 - 1 ↘

DEFINISCO "ERRORE RELATIVO DELLA RAPPRESENTAZIONE DEL NUMERO REALE  $x$ " IL SEGUENTE RAPPORTO (SI OSSERVI CHE PUÒ ESSERE NEGATIVO)

$$\epsilon_x = \frac{\text{rad}(x) - x}{x} = \frac{\delta_x}{x}$$

SI METTE IN RELAZIONE L'ERRORE COMMESSO CON L'ORDINE DI GRANDEZZA DEL NUMERO.

$$\text{ES. MISURO UN TAVOLO LUNGO 2m} \rightarrow x = 1.99 \quad \delta_x = 2.00 - 1.99 = 0.01\text{m}$$

$$\epsilon_x = \frac{\delta_x}{x} = \frac{0.01\text{m}}{2\text{m}}$$

POSSIAMO DIRE  $|\epsilon_x| < \frac{1}{2} \beta^{1-m}$ , DIPENDENZA SOLO DAL NUMERO DI CIFRE  
QUESTO PERCHÉ DIVIDERE PER  $x$  SIGNIFICA DIVIDERE PER UN FATTORE  $\beta^b$

IL VALORE  $u = \frac{1}{2} \beta^{1-m}$  È DETTO "ERRORE DI MACCHINA"

L'OPERAZIONE È CORRETTA FINO ALLA  $m$ -ESIMA CIFRA SIGNIFICATIVA SE  $\epsilon_x < u$

↑  
IL MASSIMO ERRORE RELATIVO COMPUTATO NEL PASSAGGIO DA  $x$  AD  $\text{rad}(x)$ .

## - OPERAZIONI DI MACCHINA -

**PROPRIETÀ.** LE OPERAZIONI NON SONO CHIUSE : DATI GLI INPUT E M NON È DETTO CHE OTTENGA UN OUTPUT E M. QUESTO VA IN CONTRASTO, AD ESEMPIO, CON L'INSIEME DEI NUMERI REALI DOVE VALE SEMPRE

$$\begin{array}{ccc} a+b=c & & \\ \downarrow & \downarrow & \downarrow \\ \epsilon \text{IR} & \epsilon \text{IR} & \epsilon \text{IR} \end{array}$$

**SEMPRE !!**

addirittura nelle operazioni non valgono automaticamente le quattro proprietà: si consideri il seguente esempio dove si dimostra la non applicabilità della proprietà associativa nella somma:

Sia  $M = F(10, 3, -99, 99)$  e siano  $x = 0.135 \times 10^{-4}$ ,  
 $y = 0.258 \times 10^{-2}$  e  $z = -0.251 \times 10^{-2}$

Indicando con  $\oplus$  l'operazione di addizione tra elementi di  $M$ , si ha

$$\begin{aligned} x \oplus (y \oplus z) &= 0.135 \times 10^{-4} \oplus (0.258 \times 10^{-2} \oplus -0.251 \times 10^{-2}) \\ &= 0.135 \times 10^{-4} \oplus 0.700 \times 10^{-4} \\ &= 0.835 \times 10^{-4}, \end{aligned}$$

mentre

$$\begin{aligned} (x \oplus y) \oplus z &= (0.135 \times 10^{-4} \oplus 0.258 \times 10^{-2}) \oplus -0.251 \times 10^{-2} \\ &= 0.259 \times 10^{-2} \oplus -0.251 \times 10^{-2} \\ &= 0.800 \times 10^{-4}. \end{aligned}$$

**EFFETTO DELLA CANCELLAZIONE.** IN  $(x \oplus y) \oplus z$  ABBIAMO ASSISTITO AL FENOMENO DELLA CANCELLAZIONE. ABBIAMO

- STESSO ESPONENTE  $b$
- MANTISSE DIFFERENTI DI POCO

IL RISULTATO HA UNA PERDITA DI CIFRE SIGNIFICATIVE E UN'AMPLIFICAZIONE DELL'ERRORE RELATIVO.

## TEORIA DEGLI ERRORI - ERRORI NEL CALCOLO DELLA FUNZIONE

### - INTRODUZIONE

SI CONSIDERI LA FUNZIONE  $f: \mathbb{R}^m \rightarrow \mathbb{R}$  ( $m$  VARIABILI, VALORI REALI)

PRESO  $p_0 = (x_1^0, \dots, x_m^0)$  SI VUOLE CALCOLARE IL VALORE DELLA FUNZIONE  $f(p_0)$

PROBLEMA! DOBBIANO RAPPRESENTARE IL TUTTO NEL CALCOLATORE, E ABBIANO GIÀ DETTO PRIMA CHE NON POSSANO LAVORARE CON INFINITE CIFRE.

NON OTTERREMO MAI  $f(p_0)$  PER LE APPROSSIMAZIONI. DUE QUESTIONI.

- RAPPRESENTARE  $p_0$ , CON COMPONENTI IRRAZIONALI. SI PREnda AD ESEMPIO

$$p_0 = (\sqrt{2}, e, \pi)$$

↓      ↓      ↓

1.41...    2.71...    3.14...

NON PORREMO  $p_0 = (\sqrt{2}, e, \pi)$ , MA  $p_1 = (1.41, 2.71, 3.14)$   
OVIAMENTE  $f(p_0) \neq f(p_1)$

- RAPPRESENTARE IL VALORE RESTITUITO DALLA FUNZIONE  $f$ .

ESEMPIO:  $\log(5) \rightarrow$  CALCOLO CON SVILUPPI DI TAYLOR, NON SI PRENDONO TUTTE LE CIFRE

NON SI CALCOLA IL VALORE DELLA FUNZIONE (NELL'ES. IL  $\log$ ), MA SI RICORRE A UN ALGORITMO DI CALCOLO (NELL'ES. SVILUPPI DI TAYLOR).

NON CALCOLEREMO  $f(p_0)$ , MA  $f_2(p_0)$  OVIAMENTE  $f(p_0) \neq f_2(p_0)$

$\Rightarrow$  NON CALCOLEREMO  $f(p_0)$ , MA  $f_2(p_1)$

### - INSIEME DI INDETERMINAZIONE DEL PUNTO

DATO UN PUNTO  $p_0 = (x_1^0, \dots, x_m^0)$  DEFINIAMO "INSIEME DI INDETERMINAZIONE DEL PUNTO" QUANTO SEGUE

$$D = \left\{ p \in \mathbb{R}^m \mid a_i \leq x_i \leq b_i, i = 1, \dots, m \right\}$$

DOVE  $a_i, b_i \in \mathbb{R}, \forall i = 1, \dots, m$ . IN SOSTANZA DEFINIAMO PER OGNI ELEMENTO  $x_i$  UN INTERVALLO A CUI L'ELEMENTO  $x_i$  APPARTIENE

ESEMPIO.  $p_0 = (\sqrt{2}, e) \Rightarrow D = [1, 2] \times [2, 3]$

$$\begin{array}{ll} a_1 = 1 & a_2 = 2 \\ b_1 = 2 & b_2 = 3 \end{array}$$

## — ERRORE ASSOLUTO DELLA FUNZIONE —

SOSTITUIAMO  $f(p_0)$  CON  $f_2(p_1)$ , DOVE  $p_1 \in D$ ,  $p_1 = (x_1^1, \dots, x_m^1)$ .  
DEFINIAMO "ERRORE ASSOLUTO DELLA FUNZIONE" LA DIFFERENZA

$$\delta_f = f_2(p_1) - f(p_0)$$

↓                                  ↓  
VALORE CHE                        VALORE CHE  
DOBBIANO USARE                VOLLEVANO UTILIZZARE

$f(p_0)$  NON È RAPPRESENTABILE  $\rightarrow$  POSSIBILE CALCOLARE SOLO STIME DI  $\delta_f$

PONIAMO

$$\delta_f = f_2(p_1) - f(p_1) + f(p_1) - f(p_0) \implies \delta_f = \delta_a + \delta_d$$

SOTTRAGGO      SOMMO  
↑                      ↑  
 $\delta_a$                      $\delta_d$

$\delta_a$  È L'ERRORE ASSOLUTO ALGORITMICO, DOVUTO ALL'ALGORITMO DI CALCOLO E QUINDI AL SOSTituIRE  $f$  CON  $f_2$ . VALORE DEFINITO E STIMABILE, DATO L'ALGORITMO.

$\delta_d$  È L'ERRORE ASSOLUTO TRASMESSO DAI DATI, DOVUTO ALLA SOSTITUZIONE DI  $p_0$  CON  $p_1$ . POSSIBILE DARE UNA DIMENSIONE SE  $f \in C^1(D)$ , RICORRENDO ALLA FORMULA DI TAYLOR ARRESTATA AL PRIMO TERMINE NEL PUNTO  $x_0$

$$f(p_1) - f(p_0) = \sum_{i=1}^m \frac{\partial f}{\partial x_i} (x_i^{(1)} - x_i^{(0)})$$

↓                              ↓  
 $p_1$                          $p_0$

DERIVATE PARZIALI CALCOLATE  
IN OPPORTUNI PUNTI APPARTENENTI  
AL SEGMENTO CHE CONGIUNGE  $p_0$  E  $p_1$

PONGO L'ERRORE ASSOLUTO SULLA COMPONENTE  $i$ -ESIMA:  $\delta_{x_i} = x_i^1 - x_i^0$

PONGO IL COEFFICIENTE DI AMPLIFICAZIONE DELL'ERRORE,  
SULLA COMPONENTE  $i$ -ESIMA:  $\rho_i = \frac{\partial f}{\partial x_i}$

OTTENIAMO  $\boxed{\delta_d = \sum_{i=1}^m \rho_i \delta_{x_i}}$

VALORE CHE  
ABBIANO                    VALORE CHE  
CI PIACEREbbe  
AUERE

RICORDIAMOCI CHE FACCIAmo SOLO STIME DI  $\delta_d$  E  $\delta_a$ . CALCOLIAMO UNA "LIMITAZIONE DEL MODULO DELL'ERRORE ASSOLUTO"

$$|f_2(p_1) - f(p_0)| \leq E_a + E_d \quad (\text{SOMMA DEI MASSIMI MODULI})$$

$\|\delta_a\|$  È IL MASSIMO MODULO DI  $\delta_a$

$\|\delta_d\|$  È IL MASSIMO MODULO DI  $\delta_d$ , RICORDARSI LA DISUGUAGLIANZA TRIANGOLARE  $|x+y| \leq |x| + |y|$ , INOLTRE È IMMEDIATO  $|xy| = |x||y|$

$$|\delta_d| = \left| \sum_{i=1}^m \frac{\partial f}{\partial x_i} (x_i^{(1)} - x_i^{(0)}) \right| \leq \sum_{i=1}^m \left| \frac{\partial f}{\partial x_i} (x_i^{(1)} - x_i^{(0)}) \right| = \sum_{i=1}^m \left| \frac{\partial f}{\partial x_i} \right| |x_i^{(1)} - x_i^{(0)}|$$

$$\implies \|\delta_d\| = \sum_{i=1}^m A_{x_i} |\delta_{x_i}|$$

$$\text{DOVE } |\delta_{x_i}| = |x_i^{(1)} - x_i^{(0)}| \quad A_{x_i} \geq \sup_{P \in D} \left| \frac{\partial f}{\partial x_i} \right| \quad i = 1, \dots, m$$

### - ERRORE ASSOLUTO - PROBLEMA DIRETTO E PROBLEMA INVERSO -

Se si conosce una stima dell'errore algoritmico e degli errori  $\delta_{x_i}$  nonché le  $A_{x_i}$ , si può stabilire a posteriori un confine superiore per l'errore assoluto con cui si è calcolata la funzione nel punto desiderato; questo problema è detto **problema diretto**

Il **problema inverso** consiste nel richiedere a priori che il valore  $f_a(P_1)$  sia tale che l'errore assoluto  $|f_a(P_1) - f_a(P_0)|$  risulti minore di un valore prefissato, per cui si deve cercare sia un algoritmo  $f_a(P)$  sia un opportuno punto  $P_1$  che soddisfino la richiesta

IL PIÙ  
PRATICO (CIT.)

### ESEMPIO DI PROBLEMA DIRETTO. SI VUOLE CALCOLARE LA FUNZIONE

$$f(x_1, x_2) = \frac{x_1}{x_2}$$

IN UN PUNTO  $P_0 \in D = [1, 3] \times [4, 5]$ . SI SUPPOGA DI ARROTONDARE IL RISULTATO DELL'OPERAZIONE ALLA 2<sup>a</sup> CIFRA DECIMALE E DI INTRODURRE  $(x_1^{(0)}, x_2^{(0)})$  CON ERRORI  $|\delta_{x_1}| < 10^{-2}$ ,  $|\delta_{x_2}| < 10^{-2}$

QUAL È IL MASSIMO  $|\delta_f|$ ?

SE ARROTONDIAMO IL RISULTATO ALLA 2<sup>a</sup> CIFRA DECIMALE POSSIAMO PORRE

$$|\delta_a| \leq \frac{1}{2} \beta^{b-m} \implies |\delta_a| \leq \frac{1}{2} 10^{-2} \quad m=2 \quad b=0$$

CALCOLIAMO LE DERIVATE PARZIALI:  $\frac{\partial f}{\partial x_1} = \frac{1}{x_2} \quad \frac{\partial f}{\partial x_2} = -\frac{x_1}{x_2^2}$

$$\text{CALCOLIAMO GLI } A_{x_i}: \quad A_{x_1} = \max_{P \in D} \left| \frac{1}{x_2} \right| = \frac{1}{4}$$

$$A_{x_2} = \max_{P \in D} \left| -\frac{x_1}{x_2^2} \right| = \frac{3}{16}$$

$$\text{IN CONCLUSIONE: } |\delta_f| \leq |\delta_a| + |\delta_d| = \frac{1}{2} 10^{-2} + \frac{1}{4} 10^{-2} + \frac{3}{16} 10^{-2}$$

**ESEMPIO DI PROBLEMA INVERSO.** SIANO  $x_1^0 \in [1, 2]$ ,  $x_2^0 \in [-2, -1]$ , SI CONSIDERI LA SOLITA FUNZIONE DI PRIMA

$$f(x_1, x_2) = \frac{x_1}{x_2} \quad D = [1, 2] \times [-2, -1]$$

COME SI DEVE ESEGUIRE LA DIVISIONE? QUALE ERRORE AVRANNO  $x_1$  E  $x_2$  AFFINCHÉ  $|8_f| \leq 10^{-2}$ ?

ENTRAMBE LE RICHIESTE RICHIEDONO DI INDICARE UN LIVELLO DI PRECISIONE. SAPPIAMO CHE  $|8_f| \leq |8_a| + |8_d|$

SI PRENDONO TRE CIFRE E SI ARROTONDA ALLA SECONDA PIÙ SIGNIFICATIVA

DUE CONTRIBUTI ALL'ERRORE ASSOLUTO, CHE DIVIDIAMO CONVENZIONALMENTE IN PARTI UGUALI

$$|8_a| \leq \frac{1}{2} 10^{-2} \quad |8_d| \leq \frac{1}{2} 10^{-2}$$

ANCHE  $|8_d|$  È CARATTERIZZATO DA DUE CONTRIBUTI:  $|8_d| \leq A_{x_1} |8_{x_1}| + A_{x_2} |8_{x_2}|$

DIVIDIAMO ANCHE  $|8_d|$  IN PARTI UGUALI:  $A_{x_1} |8_{x_1}| \leq \frac{1}{4} 10^{-2}$ ,  $A_{x_2} |8_{x_2}| \leq \frac{1}{4} 10^{-2}$

CALCOLIAMO LE DERIVATE PARZIALI:  $\frac{\partial f}{\partial x_1} = \frac{1}{x_2}$        $\frac{\partial f}{\partial x_2} = -\frac{x_1}{x_2^2}$

CALCOLIAMO GLI  $A_{x_i}$ :  $A_{x_1} = \max_{P \in D} \left| \frac{1}{x_2} \right| = 1$        $A_{x_2} = \max_{P \in D} \left| -\frac{x_1}{x_2^2} \right| = 2$

QUINDI:  $|8_d| \leq 1 \cdot |8_{x_1}| + 2 \cdot |8_{x_2}|$

$$|8_{x_1}| \leq \frac{1}{4} 10^{-2}$$

$$2 |8_{x_2}| \leq \frac{1}{4} 10^{-2} \rightarrow |8_{x_2}| \leq \frac{1}{8} 10^{-2}$$

PONIAMO  $|8_{x_1}| \leq 10^{-3}$ ,  $|8_{x_2}| \leq 10^{-3}$  IN MODO TALE CHE LE SOGLIE DETTE NON VENGANO SUPERATE (COME INTRODUCO  $x_1$  E  $x_2$ ?)

MORALE DELLA FAVOLA:  $|8_f| \leq |8_a| + |8_d| = \frac{1}{2} 10^{-2} + 10^{-3} + 2 \cdot 10^{-3} = \frac{4}{5} 10^{-2}$

CHE È  $\leq 10^{-2}$ !!!

(CONDIZIONE INIZIALE RISPETTATA)

1) Si vuole calcolare la funzione

$$f(x, y) = \frac{y^2}{x}$$

in un punto  $P_0 \in [-2, -1] \times [2, 3]$ .

Per avere un errore assoluto  $|\delta_f| \leq 10^{-2}$ , quali limitazioni devono soddisfare l'errore assoluto algoritmico  $|\delta_a|$  e gli errori assoluti  $|\delta_x|$  e  $|\delta_y|$ ?

$$|\delta_f| \leq E_a + E_d \leq 10^{-2}$$

$$\max |\delta_a| \quad \max |\delta_a|$$

DIVISO EQUAMENTE L'ERRORE TRA  $E_a$  ED  $E_d$

$$|\delta_a| \leq \frac{1}{2} 10^{-2} \quad |\delta_d| \leq \frac{1}{2} 10^{-2}$$

DIVISO A SUA VOLTA  $|\delta_d|$  EQUAMENTE TRA I DUE TERMINI DELLA SOMMATORIA

$$|\delta_d| = \sum_{i=1}^m \left| \frac{\partial f}{\partial x_i} \delta_{x_i} \right| = \left| \frac{\partial f}{\partial x} \right| \delta_x + \left| \frac{\partial f}{\partial y} \right| \delta_y \implies |\delta_d| \leq \left| \frac{\partial f}{\partial x} \right| |\delta_x| + \left| \frac{\partial f}{\partial y} \right| |\delta_y|$$

$$\left| \frac{\partial f}{\partial x} \right| |\delta_x| \leq \frac{1}{4} 10^{-2} \quad \left| \frac{\partial f}{\partial y} \right| |\delta_y| \leq \frac{1}{4} 10^{-2}$$

$$\left| \frac{\partial f}{\partial x} \right| = y^2 \cdot (-1) x^{-2} = -\frac{y^2}{x^2} \quad A_x = \max_{P \in D} \left| -\frac{y^2}{x^2} \right| = \left| -\frac{3^2}{(-1)^2} \right| = 9$$

$$\left| \frac{\partial f}{\partial y} \right| = \frac{1}{x} 2y = 2 \frac{y}{x} \quad A_y = \max_{P \in D} \left| 2 \frac{y}{x} \right| = \left| 2 \frac{3}{-1} \right| = 6$$

$$9 |\delta_x| \leq \frac{1}{4} 10^{-2} \implies |\delta_x| \leq \frac{1}{36} 10^{-2} > 10^{-4}$$

$$6 |\delta_y| \leq \frac{1}{4} 10^{-2} \implies |\delta_y| \leq \frac{1}{24} 10^{-2} > 10^{-4}$$

FARE PROVE CON  
LA CALCOLATRICE PER  
TOGLIERE IL NUM. ACCANTO  
ALLA POTENZA

IN ENTRAMBI I CASI NON  
POSso DIRE  $< 10^{-3}$

- INTRODUO X TRONCATO ALLA 4<sup>a</sup> CIFRA DECIMALE
- INTRODUO Y TRONCATO ALLA 4<sup>a</sup> CIFRA DECIMALE
- DOPO IL CALCOLO ARROTONDO ALLA 2<sup>a</sup> CIFRA DECIMALE

1) Si vuole calcolare la funzione

$$f(x, y) = x^2 + y^3$$

in un punto  $P_0 \in [0, 1] \times [1, 2]$ .

Per avere un errore assoluto  $|\delta_f| \leq 10^{-2}$ , quali limitazioni devono soddisfare l'errore assoluto algoritmico  $|\delta_a|$  e gli errori assoluti  $|\delta_x|$  e  $|\delta_y|$ ?

$$|\delta_f| \leq E_a + E_d \leq 10^{-2}$$

$$\max |\delta_a| \quad \max |\delta_d|$$

DIVIDO EQUAMENTE L'ERRORE TRA  $E_a$  ED  $E_d$

$$|\delta_a| \leq \frac{1}{2} 10^{-2} \quad |\delta_d| \leq \frac{1}{2} 10^{-2}$$

DIVIDO A SUA VOLTA  $|\delta_d|$  EQUAMENTE TRA I DUE TERMINI DELLA SOMMATORIA

$$\delta_d = \sum_{i=1}^m \frac{\partial f}{\partial x_i} \delta_{x_i} = \frac{\partial f}{\partial x} \delta_x + \frac{\partial f}{\partial y} \delta_y \implies |\delta_d| \leq \left| \frac{\partial f}{\partial x} \right| |\delta_x| + \left| \frac{\partial f}{\partial y} \right| |\delta_y|$$

$$\left| \frac{\partial f}{\partial x} \right| |\delta_x| \leq \frac{1}{4} 10^{-2} \quad \left| \frac{\partial f}{\partial y} \right| |\delta_y| \leq \frac{1}{4} 10^{-2}$$

$$\frac{\partial f}{\partial x} = 2x$$

$$A_x = \max_{P \in D} |2x| = |2 \cdot 1| = 2$$

$$\frac{\partial f}{\partial y} = 3y^2$$

$$A_y = \max_{P \in D} |3y^2| = |3 \cdot 2^2| = 12$$

$$2 |\delta_x| \leq \frac{1}{4} 10^{-2} \implies |\delta_x| \leq \frac{1}{8} 10^{-2} > 10^{-3}$$

$$12 |\delta_y| \leq \frac{1}{4} 10^{-2} \implies |\delta_y| \leq \frac{1}{48} 10^{-2} > 10^{-4}$$

FARE PROVE CON  
LA CALCOLATRICE PER  
TOGLIERE IL NUM. ACCANTO  
ALLA POTENZA

- INTRODUO X TRONCATO ALLA 3<sup>a</sup> CIFRA DECIMALE
- INTRODUO Y TRONCATO ALLA 4<sup>a</sup> CIFRA DECIMALE
- DOPO IL CALCOLO ARROTONDO ALLA 2<sup>a</sup> CIFRA DECIMALE

1) Si vuole calcolare la funzione

$$f(x, y) = \frac{x+2}{y}$$

in un punto  $P_0 \in D = [0, 1] \times [1, 2]$ .

Indicare con quale precisione si devono introdurre i dati e come eseguire le operazioni per avere un errore assoluto che verifichi la limitazione  $|\delta_f| \leq 10^{-3}$ .

$$|\delta_f| \leq |\delta_x| + |\delta_y|$$

$$|\delta_x| \leq \frac{1}{2} \cdot 10^{-3}$$

$$|\delta_y| \leq \frac{1}{2} \cdot 10^{-3}$$

$$\frac{\partial f}{\partial x} = \frac{1}{y}$$

$$\frac{\partial f}{\partial y} = \frac{0 - (x+2) \cdot 1}{y^2} = -\frac{x+2}{y^2}$$

$$|\delta_f| \leq A_1 |\delta_x| + A_2 |\delta_y|$$

$$A_1 = \max_{P \in D} \left| \frac{\partial f}{\partial x} \right| = \max_{P \in D} \left| \frac{1}{y} \right| = 1 \quad A_2 = \max_{P \in D} \left| \frac{\partial f}{\partial y} \right| = \max_{P \in D} \left| \frac{x+2}{y^2} \right| = 3$$

$$|\delta_x| \leq \frac{1}{4} \cdot 10^{-3} > 10^{-4}$$

$$3|\delta_y| \leq \frac{1}{4} \cdot 10^{-3}$$

$$|\delta_y| \leq \frac{1}{12} \cdot 10^{-3} > 10^{-4}$$

1) Si vuole calcolare la funzione

$$f(x, y) = \frac{x}{x+y}$$

in un punto  $P_0 \in D = [0, 1] \times [1, 2]$ . Si suppone di arrotondare il risultato alla 2<sup>a</sup> cifra decimale e di introdurre i valori  $x$  e  $y$  con errori  $|\delta_x| \leq 10^{-2}$  e  $|\delta_y| \leq 10^{-3}$ .

Quale è il massimo di  $|\delta_f|$ ?

$$|\delta_f| \leq |\delta_x| + |\delta_y| = |\delta_x| + A_x |\delta_x| + A_y |\delta_y|$$

$$|\delta_x| \leq \frac{1}{2} \cdot 10^{-2} = \frac{1}{2} \cdot 10^{-2} = \frac{1}{2} \cdot 10^{-2} \rightarrow \text{RISULTATO ARROTONDATO ALLA 2<sup>a</sup> CIFRA}$$

$$\frac{\partial f}{\partial x} = \frac{1(x+y) - x(1)}{(x+y)^2} = \frac{x+y-x}{(x+y)^2} = \frac{y}{(x+y)^2} \quad \frac{\partial f}{\partial y} = \frac{0 - x(y)}{(x+y)^2} = -\frac{x}{(x+y)^2}$$

$$A_x = \max_{P \in D} \left| \frac{\partial f}{\partial x} \right| = \max_{P \in D} \left| \frac{y}{(x+y)^2} \right| = \left| \frac{2}{(0+1)^2} \right| = 2$$

$$A_y = \max_{P \in D} \left| \frac{\partial f}{\partial y} \right| = \max_{P \in D} \left| -\frac{x}{(x+y)^2} \right| = \left| -\frac{1}{(0+1)^2} \right| = 1$$

$$|\delta_f| \leq \frac{1}{2} \cdot 10^{-2} + 2 \cdot 10^{-2} + 1 \cdot 10^{-3} = 2.6 \cdot 10^{-2}$$

## - ERRORE RELATIVO DELLA FUNZIONE

SOSTITUIAMO  $f(p_0)$  CON  $f_2(p_1)$ , DOVE  $p_1 \in D$ ,  $p_1 = (x_1^{-1}, \dots, x_m^{-1})$ .

DEFINIAMO "ERRORE RELATIVO DELLA FUNZIONE" IL RAPPORTO TRA ERRORE ASSOLUTO DELLA FUNZIONE ED  $f(p_0)$

$$\epsilon_f = \frac{f_2(p_1) - f(p_0)}{f(p_0)} = \frac{\delta_f}{f(p_0)}$$

POSSIAMO SCRIVERE

$$\epsilon_f = \frac{f(p_1) - f(p_0)}{f(p_0)} + \frac{f_2(p_1) - f(p_1)}{f(p_1)} \left( 1 + \frac{f(p_1) - f(p_0)}{f(p_0)} \right) = \epsilon_a + \epsilon_d + \epsilon_a \epsilon_d$$

DOVE  $\epsilon_a = \frac{f_2(p_1) - f(p_1)}{f(p_1)}$  È L'ERRORE RELATIVO ALGORITMICO  
(DOVUTO ALL'ALGORITMO DI CALCOLO -  $f_2$  INVECE DI  $f$ )

SE L'ALGORITMO PRODUCE ERRORI "ACCETTABILMENTE LIMITATI" ALLORA SI PARLA DI ALGORITMO STABILE, ALTRIMENTI SI PARLA DI ALGORITMO INSTABILE

$\epsilon_d = \frac{f(p_1) - f(p_0)}{f(p_0)}$  È L'ERRORE RELATIVO TRASMESSO DAI DATI  
(DOVUTO AL PORRE  $p_1$  INVECE DI  $p_0$ )

$\epsilon_a \epsilon_d$  È UN PRODOTTO CHE TRASCURIAMO POICHÉ DI ORDINE INFERIORE  
RISPETTO AGLI ALTRI TERMINI (ES:  $\epsilon_a = \epsilon_d = 10^{-2} \Rightarrow \epsilon_a \epsilon_d = 10^{-4} < 10^{-2}$ )

$$\implies \epsilon_f = \epsilon_a + \epsilon_d$$

COME  $\delta_d$  POSSIAMO SCRIVERE  $\epsilon_d$  RICORRENDO ALLA FORMULA DI TAYLOR

$$\epsilon_d = \frac{f(p_1) - f(p_0)}{f(p_0)} = \sum_{i=1}^m \frac{\partial f}{\partial x_i} \frac{(x_i^{(1)} - x_i^{(0)})}{f(p_0)}$$

DEFINIAMO L'ERRORE RELATIVO RISPETTO ALLA COMPONENTE  $i$ -ESIMA:  $\epsilon_{x_i} = \frac{x_i^{(1)} - x_i^{(0)}}{x_i^{(0)}}$

$$\implies \epsilon_d = \sum_{i=1}^m \frac{x_i^{(0)}}{f(p_0)} \frac{\partial f}{\partial x_i} \epsilon_{x_i}$$

DEFINIAMO IL COEFFICIENTE DI AMPLIFICAZIONE DELL'ERR. RELATIVO:  $\gamma_i = \frac{x_i^{(0)}}{f(p_0)} \frac{\partial f}{\partial x_i}$

$$\implies \boxed{\epsilon_d = \sum_{i=1}^m \gamma_i \epsilon_{x_i}}$$

SE I  $\gamma_i$  SONO TALI CHE  $\epsilon_d$  SARÀ DELLO STESSO ORDINE DEGLI  $\epsilon_{x_i}$  ALLORA SI HA UN PROBLEMA BEN CONDIZIONATO, ALTRIMENTI SI PARLA DI PROBLEMA MAL CONDIZIONATO

## — ERRORE TRASMESSO DAI DATI NEUE QUATTRO OPERAZIONI —

RIEPILOGHIAMO CON LA SEGUENTE TABELLA I COEFFICIENTI DI AMPLIFICAZIONE DELL'ERRORE RELATIVO TRASMESSO DAI DATI

$$\rho_i = \frac{\partial f}{\partial x_i}$$

operazione	$\delta_d$	$\epsilon_d$
$x \oplus y$	$\delta_x + \delta_y$	$\frac{x}{x+y} \epsilon_x + \frac{y}{x+y} \epsilon_y$
$x \ominus y$	$\delta_x - \delta_y$	$\frac{x}{x-y} \epsilon_x - \frac{y}{x-y} \epsilon_y$
$x \otimes y$	$y \delta_x + x \delta_y$	$\epsilon_x + \epsilon_y$
$x \oslash y$	$\frac{1}{y} \delta_x - \frac{x}{y^2} \delta_y$	$\epsilon_x - \epsilon_y$

$$\delta_i = \frac{x_i^0}{f(p_0)} \frac{\partial f}{\partial x_i}$$

### Osservazione 1

Dalla precedente tabella si evidenzia che le operazioni di addizione e sottrazione non danno problemi per quanto riguarda l'errore assoluto, mentre possono rendere grande l'errore relativo nel caso in cui i due termini dell'operazione siano molto vicini in valore assoluto, in quanto può accadere che i denominatori che compaiono nei coefficienti di amplificazione dell'errore relativo siano molto piccoli in valore assoluto avendo così quella che abbiamo chiamato la **cancellazione**

### Osservazione 2

La moltiplicazione non amplifica l'errore relativo e comporta un errore assoluto che dipende dall'ordine di grandezza dei fattori. Anche la divisione non produce amplificazione per quanto riguarda l'errore relativo, mentre l'errore assoluto diminuisce se il divisore aumenta (in valore assoluto)

$$x+y$$

$$\delta_x = \frac{x}{x+y} (1+0) = \frac{x}{x+y}$$

$$\delta_y = \frac{y}{x+y} (0+1) = \frac{y}{x+y}$$

$$\rho_x = \frac{\partial f}{\partial x} = 1+0=1$$

$$\rho_y = \frac{\partial f}{\partial y} = 0+1=1$$

$$x-y$$

$$\delta_x = \frac{x}{x-y} (1+0) = \frac{x}{x-y}$$

$$\delta_y = \frac{y}{x-y} (0-1) = -\frac{y}{x-y}$$

$$\rho_x = \frac{\partial f}{\partial x} = 1-0=1$$

$$\rho_y = \frac{\partial f}{\partial y} = 0-1=-1$$

$$xy$$

$$\delta_x = \frac{x}{xy} y = 1$$

$$\delta_y = \frac{y}{xy} x = 1$$

$$\rho_x = \frac{\partial f}{\partial x} = y$$

$$\rho_y = \frac{\partial f}{\partial y} = x$$

$$\frac{x}{y} \quad \delta_x = \frac{x}{xy} \cdot \frac{1}{y} = -1 \quad \delta_y = \frac{y}{xy} xy^{-2} = 1 \quad \rho_x = \frac{\partial f}{\partial x} = \frac{1}{y} \quad \rho_y = \frac{\partial f}{\partial y} = xy^{-2}$$

- 1) Determinare l'espressione dell'errore relativo nel calcolo della funzione

$$f(x, y) = \frac{x+y}{x-y}.$$

$$\begin{aligned}\pi_1 &= x+y \\ \pi_2 &= x-y \\ \pi_3 &= \pi_1/\pi_2\end{aligned}$$

$$\epsilon_f = \epsilon_{\pi_3} = \epsilon_3 + \epsilon_{\pi_1} - \epsilon_{\pi_2} =$$

$$= \epsilon_3 + \epsilon_1 + \frac{x}{x+y} \epsilon_x + \frac{y}{x+y} \epsilon_y - \epsilon_{\pi_2} =$$

$$= \epsilon_3 + \epsilon_1 + \frac{x}{x+y} \epsilon_x + \frac{y}{x+y} \epsilon_y - \left( \epsilon_2 + \frac{x}{x-y} \epsilon_x - \frac{y}{x-y} \epsilon_y \right) =$$

$$= \epsilon_1 - \epsilon_2 + \epsilon_3 + \left( \frac{x}{x+y} - \frac{x}{x-y} \right) \epsilon_x + \left( \frac{y}{x+y} + \frac{y}{x-y} \right) \epsilon_y =$$

$$= \dots \dots \frac{x(x-y) - x(x+y)}{(x+y)(x-y)} \epsilon_x + \frac{y(x-y) + y(x+y)}{(x+y)(x-y)} \epsilon_y =$$

$$= \dots \dots \frac{x^2 - xy - x^2 - xy}{x^2 - y^2} \epsilon_x + \frac{xy - y^2 + xy + y^2}{x^2 - y^2} \epsilon_y =$$

$$= \epsilon_1 - \epsilon_2 + \epsilon_3 + \left( -\frac{2xy}{x^2 - y^2} \right) \epsilon_x + \left( \frac{2xy}{x^2 - y^2} \right) \epsilon_y$$

operazione	$\delta_d$	$\epsilon_d$
$x \oplus y$	$\delta_x + \delta_y$	$\frac{x}{x+y} \epsilon_x + \frac{y}{x+y} \epsilon_y$
$x \ominus y$	$\delta_x - \delta_y$	$\frac{x}{x-y} \epsilon_x - \frac{y}{x-y} \epsilon_y$
$x \otimes y$	$y \delta_x + x \delta_y$	$\epsilon_x + \epsilon_y$
$x \oslash y$	$\frac{1}{y} \delta_x - \frac{x}{y^2} \delta_y$	$\epsilon_x - \epsilon_y$

- 1) Determinare l'espressione dell'errore relativo nel calcolo della funzione

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

$$\begin{aligned}\pi_1 &= x-y \\ \pi_2 &= xy \\ \pi_3 &= \pi_1/\pi_2\end{aligned}$$

$$\epsilon_f = \epsilon_{\pi_3} = \epsilon_3 + \epsilon_{\pi_1} - \epsilon_{\pi_2} =$$

$$= \epsilon_3 + \epsilon_1 + \frac{x}{x-y} \epsilon_x - \frac{y}{x-y} \epsilon_y - \epsilon_{\pi_2} =$$

$$= \epsilon_3 + \epsilon_1 + \frac{x}{x-y} \epsilon_x - \frac{y}{x-y} \epsilon_y - \epsilon_2 - \epsilon_x - \epsilon_y =$$

$$= \epsilon_1 - \epsilon_2 + \epsilon_3 + \left( \frac{x}{x-y} - 1 \right) \epsilon_x + \left( -\frac{y}{x-y} - 1 \right) \epsilon_y =$$

$$= \epsilon_1 - \epsilon_2 + \epsilon_3 + \left( \frac{x-x+y}{x-y} \right) \epsilon_x + \left( \frac{-y-x+y}{x-y} \right) \epsilon_y =$$

$$= \epsilon_1 - \epsilon_2 + \epsilon_3 + \left( \frac{y}{x-y} \right) \epsilon_x + \left( \frac{-x}{x-y} \right) \epsilon_y$$

operazione	$\delta_d$	$\epsilon_d$
$x \oplus y$	$\delta_x + \delta_y$	$\frac{x}{x+y} \epsilon_x + \frac{y}{x+y} \epsilon_y$
$x \ominus y$	$\delta_x - \delta_y$	$\frac{x}{x-y} \epsilon_x - \frac{y}{x-y} \epsilon_y$
$x \otimes y$	$y \delta_x + x \delta_y$	$\epsilon_x + \epsilon_y$
$x \oslash y$	$\frac{1}{y} \delta_x - \frac{x}{y^2} \delta_y$	$\epsilon_x - \epsilon_y$

1) Si determini l'errore relativo nel calcolo della funzione

$$f(x, y) = \frac{y}{x^2}.$$

$$\eta_1 = x^2$$

$$\eta_2 = y/\eta_1$$

$$\begin{aligned}\epsilon_f - \epsilon_{\eta_2} &= \epsilon_2 + \epsilon_y - \epsilon_{\eta_1} \\ &= \epsilon_2 + \epsilon_y - (\epsilon_1 + \epsilon_x + \epsilon_x) \\ &= \epsilon_2 - \epsilon_1 - 2\epsilon_x + \epsilon_y\end{aligned}$$

operazione	$\delta_d$	$\epsilon_d$
$x \oplus y$	$\delta_x + \delta_y$	$\frac{x}{x+y} \epsilon_x + \frac{y}{x+y} \epsilon_y$
$x \ominus y$	$\delta_x - \delta_y$	$\frac{x}{x-y} \epsilon_x - \frac{y}{x-y} \epsilon_y$
$x \otimes y$	$y \delta_x + x \delta_y$	$\epsilon_x + \epsilon_y$
$x \oslash y$	$\frac{1}{y} \delta_x - \frac{x}{y^2} \delta_y$	$\epsilon_x - \epsilon_y$

## Capitolo 2

# Nozioni di Algebra lineare

### 2.1 Definizioni base sulla matrice

#### 2.1.1 Matrice / Matrice quadrata / Matrice rettangolare

**Definizione.** Con  $A \in \mathbb{C}^{m \times n}$  intendiamo una matrice avente  $m$  righe ed  $n$  colonne ( $m$  ed  $n$  sono dette dimensioni della matrice). Le righe e le colonne sono formate da  $m \times n$  numeri complessi  $a_{ij}$  (con  $i = 1, 2, \dots, m$ ,  $j = 1, 2, \dots, n$ ).

- $i$  è detto *indice di riga*
- $j$  è detto *indice di colonna*

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

Una matrice si dice:

- **quadrata** se  $m = n$  (si parla di *matrice quadrata di ordine  $n$* , hanno come caratteristiche *determinante*, *raggio spettrale* e *norma matriciale*);
- **rettangolare** se  $m \neq n$ .

#### 2.1.2 Elementi diagonali / Diagonale principale

**Definizione.** Data una matrice quadrata  $A \in \mathbb{C}^{n \times n}$  gli elementi  $a_{ij}$  (con  $i, j \in \{1, 2, \dots, n\}$ ) si dicono *elementi diagonali* o *appartenenti alla diagonale principale della matrice* se  $i = j$ .

#### 2.1.3 Matrice reale

**Definizione.** Una matrice  $A \in \mathbb{R}^{m \times n}$  con tutti gli elementi reali è detta *matrice reale*.

## 2.1.4 Matrice identica

**Definizione.** Una matrice quadrata è detta *matrice identica* se è valido quanto segue

$$a_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases} \quad I = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

## 2.2 Osservazioni sulle operazioni tra matrici

A proposito delle operazioni tra matrici ricordiamoci alcuni aspetti.

- **Proprietà valide per l'addizione.**

Nell'addizione tra matrici vale la proprietà commutativa e la proprietà associativa.

- **Proprietà valide nella moltiplicazione.**

Nella moltiplicazioni tra matrici vale la proprietà associativa e la proprietà distributiva, ma non la proprietà commutativa.

$$AB \neq BA$$

In virtù di ciò distinguiamo *premoltiplicazione* da *postmoltiplicazione*

- **premoltiplicazione** quando  $B$  moltiplica  $A$  a sinistra ( $BA$ )
- **postmoltiplicazione** quando  $B$  moltiplica  $A$  a destra ( $AB$ ).

- **Non validità della legge di annullamento del prodotto.**

Non vale neanche la legge di annullamento del prodotto. E' possibile avere  $AB = 0$  con  $A \neq 0$ ,  $B \neq 0$ .

## 2.3 Definizioni sui vettori

### 2.3.1 Vettore

**Definizione.** Con  $a \in \mathbb{C}^m$  si intende un vettore avente  $m$  componenti complesse  $a_i$  (con  $i = 1, 2, \dots, m$ ). Se non si dice nulla il vettore è rappresentato come *vettore colonna*, cioè una matrice avente  $m$  righe ed una colonna  $\mathbb{C}^{m \times 1}$ .

### 2.3.2 Vettore reale

**Definizione.** Un vettore  $a \in \mathbb{R}^m$  con tutti gli elementi reali è detto *vettore reale*.

### 2.3.3 Vettori linearmente indipendenti

**Definizione.** Supponiamo di avere  $k$  vettori  $x^{(1)}, \dots, x^{(k)} \in \mathbb{C}^m$ . Essi si dicono linearmente indipendenti se  $(\alpha_i \in \mathbb{C}, i = 1, 2, \dots, k)$

$$\alpha_1 x^{(1)} + \alpha_2 x^{(2)} + \dots + \alpha_k x^{(k)} = 0$$

solo se  $\alpha_1 = \alpha_2 = \dots = \alpha_k = 0$ .

### 2.3.4 Prodotto scalare

**Definizione.** Definiamo il seguente prodotto tra due vettori  $a, b \in \mathbb{C}^n$  come *prodotto scalare*

$$a^H b = \sum_{i=1}^n \bar{a}_i b_i$$

### 2.3.5 Vettori ortogonali

**Definizione.** Due vettori  $a, b \in \mathbb{C}^n$  si dicono ortogonali se  $a^H b = 0$ .

## 2.4 Definizioni sui determinanti

### 2.4.1 Determinante

**Definizione.** *Definizione non reintrodotta.* Il professore rimanda a quanto studiato ad Algebra lineare. L'unica novità da tenere a mente è che nessuna definizione di determinante, tra quelle riportate sui testi, è utilizzabile con successo. Questo per:

- costo computazionale (numero elevato di operazioni);
- propagazione dell'errore (maggiore è il numero di operazioni, maggiore è la propagazione dell'errore compiuto in un'operazione).

### 2.4.2 Matrice singolare / Matrice non singolare

**Definizione.** Una matrice quadrata  $A \in \mathbb{C}^{n \times n}$  è detta *singolare* (in alcuni testi si dice *degenera*) se  $\det A = 0$ . In caso contrario si parla di matrice *non singolare*.

### 2.4.3 Minore di ordine $k$

**Definizione.** Data una matrice  $A \in \mathbb{C}^{m \times n}$  ed un intero  $k \leq \min\{m, n\}$  si dice *minore di ordine  $k$*  il determinante di una sottomatrice estratta da  $A$  prendendo gli elementi sulla intersezione di  $k$  righe e  $k$  colonne fissate.

#### 2.4.3.1 Esempio 1

Si prenda la seguente matrice

$$A = \begin{pmatrix} 1 & -1 & 2 \\ 2 & 3 & 4 \\ 3 & 2 & 6 \end{pmatrix}$$

il valore  $k$  è  $k = \min\{3, 3\} = 3$ . Seguono

- **Sottomatrici di ordine 3.**

L'unica sottomatrice di ordine 3 è la matrice stessa.

- **Sottomatrici di ordine 2.**

Il numero di sottomatrici di ordine 2 è  $\binom{3}{2} \cdot \binom{3}{2} = 9$

$$\begin{array}{ccc} \begin{pmatrix} 1 & -1 \\ 2 & 3 \end{pmatrix} & \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} & \begin{pmatrix} -1 & 2 \\ 3 & 4 \end{pmatrix} \\ \begin{pmatrix} 1 & -1 \\ 3 & 2 \end{pmatrix} & \begin{pmatrix} 1 & 2 \\ 3 & 6 \end{pmatrix} & \begin{pmatrix} -1 & 2 \\ 2 & 6 \end{pmatrix} \\ \begin{pmatrix} 2 & 3 \\ 3 & 2 \end{pmatrix} & \begin{pmatrix} 2 & 4 \\ 3 & 6 \end{pmatrix} & \begin{pmatrix} 3 & 4 \\ 2 & 6 \end{pmatrix} \end{array}$$

- **Sottomatrici di ordine 1.**

Banalmente il numero di sottomatrici di ordine 1 è pari al numero di elementi della matrice.

$$(1) \quad (-1) \quad (2) \quad (2) \quad (3) \quad (4) \quad (3) \quad (2) \quad (6)$$

#### 2.4.3.2 Esempio 2

Si prenda la seguente matrice

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

il valore  $k$  è  $k = \min\{3, 4\} = 3$ . Seguono

- **Sottomatrici di ordine 3.**

Il numero di sottomatrici di ordine 3 è  $\binom{3}{3} \cdot \binom{4}{3} = 4$

$$\begin{array}{cccc} \begin{pmatrix} 2 & 1 & 3 \\ 1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix} & \begin{pmatrix} 2 & 1 & 4 \\ 1 & 0 & 2 \\ 1 & -1 & 2 \end{pmatrix} & \begin{pmatrix} 2 & 3 & 4 \\ 1 & 1 & 2 \\ 1 & 0 & 2 \end{pmatrix} & \begin{pmatrix} 1 & 3 & 4 \\ 0 & 1 & 2 \\ -1 & 0 & 2 \end{pmatrix} \end{array}$$

- **Sottomatrici di ordine 2.**

Il numero di sottomatrici di ordine 2 è  $\binom{3}{2} \cdot \binom{4}{2} = 18$

$$\begin{array}{ccc} \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix} & \begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix} & \begin{pmatrix} 1 & 0 \\ 1 & -1 \end{pmatrix} \\ \begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix} & \begin{pmatrix} 1 & 3 \\ -1 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \\ \begin{pmatrix} 3 & 4 \\ 1 & 2 \end{pmatrix} & \begin{pmatrix} 3 & 4 \\ 0 & 2 \end{pmatrix} & \begin{pmatrix} 1 & 2 \\ 0 & 2 \end{pmatrix} \\ \begin{pmatrix} 2 & 3 \\ 1 & 1 \end{pmatrix} & \begin{pmatrix} 2 & 3 \\ 1 & 0 \end{pmatrix} & \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \\ \begin{pmatrix} 2 & 4 \\ 1 & 2 \end{pmatrix} & \begin{pmatrix} 2 & 4 \\ 1 & 2 \end{pmatrix} & \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix} \\ \begin{pmatrix} 1 & 4 \\ 0 & 2 \end{pmatrix} & \begin{pmatrix} 1 & 4 \\ -1 & 2 \end{pmatrix} & \begin{pmatrix} 0 & 2 \\ -1 & 2 \end{pmatrix} \end{array}$$

- **Sottomatrici di ordine 1.**

Banalmente il sottomatrici di minori di ordine 1 è pari al numero di elementi della matrice.

$$(2) \quad (1) \quad (3) \quad (4) \quad (1) \quad (0) \quad (1) \quad (2) \quad (1) \quad (-1) \quad (0) \quad (2)$$

#### 2.4.4 Minori principali di ordine $k$

**Definizione.** Data una matrice  $A \in \mathbb{C}^{m \times n}$  ed un intero  $k \leq \min\{m, n\}$  si dicono *minori principali di ordine  $k$*  i determinanti delle sottomatrici di ordine  $k$  estratte da  $A$  e aventi diagonale principale composta da elementi della diagonale principale di  $A$ .

#### 2.4.5 Minori principali di testa di ordine $k$

**Definizione.** Data una matrice  $A \in \mathbb{C}^{m \times n}$  ed un intero  $k \leq \min\{m, n\}$  si dicono *minori principali di testa di ordine  $k$*  i determinanti delle sottomatrici di ordine  $k$  estratte da  $A$  formate dalle prime  $k$  righe e  $k$  colonne di  $A$ .

#### 2.4.6 Rango (Caratteristica)

**Definizione.** Data una matrice  $A \in \mathbb{C}^{m \times n}$  si dice *rango* (o *caratteristica*) il numero  $r(A)$  dato dall'ordine più alto dei suoi minori diversi da zero.

**Cioè?** Considero una matrice  $A$  e prendo tutti i minori possibili, di ordine  $1, 2, \dots, \min\{m, n\}$ . Il rango è la dimensione più grande di una sottomatrice a determinante non nullo che posso estrarre dalla matrice  $A$ .

##### 2.4.6.1 Esempio 1

Si prenda la seguente matrice

$$A = \begin{pmatrix} 1 & -1 & 2 \\ 2 & 3 & 4 \\ 3 & 2 & 6 \end{pmatrix}$$

il valore  $k$  è  $k = \min\{3, 3\} = 3$ . Per ogni ordine  $k$  possibile vediamo se si ha almeno un minore con  $\det \neq 0$ : in quel caso possiamo fermarci e affermare che  $r(A) = k$ , altrimenti decrementiamo  $k$  e continuiamo con gli ordini inferiori.

- **Minori di ordine 3.**

L'unico minore di ordine 3 è la matrice  $A$ .

$$\det(A) = 1(18 - 8) - 1(12 - 12) + 2(4 - 9) = 10 - 10 = 0$$

Sicuramente  $r(A) \neq 3$ .

- **Minori di ordine 2.**

Il numero di minori di ordine 2 è  $\binom{3}{2} \cdot \binom{3}{2} = 9$ . Osserviamo che il determinante della prima matrice è già diverso da zero, e quindi possiamo fermarci

$$\det \begin{pmatrix} 1 & -1 \\ 2 & 3 \end{pmatrix} = 3 + 2 = 5$$

Possiamo fermarci e dire  $r(A) = 2$ .

### 2.4.6.2 Esempio 2

Si prenda la seguente matrice

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

il valore  $k$  è  $k = \min\{3, 4\} = 3$ . Per ogni ordine  $k$  possibile vediamo se si ha almeno un minore con  $\det \neq 0$ : in quel caso possiamo fermarci e affermare che  $r(A) = k$ , altrimenti decrementiamo  $k$  e continuiamo con gli ordini inferiori.

- **Minori di ordine 3.**

Il numero di minori di ordine 3 è  $\binom{3}{3} \cdot \binom{4}{3} = 4$

$$\begin{pmatrix} 2 & 1 & 3 \\ 1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix} \quad \begin{pmatrix} 2 & 1 & 4 \\ 1 & 0 & 2 \\ 1 & -1 & 2 \end{pmatrix} \quad \begin{pmatrix} 2 & 3 & 4 \\ 1 & 1 & 2 \\ 1 & 0 & 2 \end{pmatrix} \quad \begin{pmatrix} 1 & 3 & 4 \\ 0 & 1 & 2 \\ -1 & 0 & 2 \end{pmatrix}$$

Possiamo già fermarci alla prima matrice

$$\det \begin{pmatrix} 2 & 1 & 3 \\ 1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix} = 2(+1) + 1(-1) + 3(-1) = 2 - 1 - 3 = -2 \neq 0$$

Possiamo dire che  $r(A) = 3!$

### 2.4.7 Teorema di Binet-Cauchy (determinante del prodotto matriciale)

**Teorema.** Date due matrici  $A \in \mathbb{C}^{m \times n}$ ,  $B \in \mathbb{C}^{n \times m}$  il determinante della matrice prodotto

$$C = AB \in \mathbb{C}^{m \times m}$$

- è nullo se  $m > n$ ,
- è dato dalla somma dei prodotti di tutti i possibili minori di ordine massimo di  $A$  per i corrispondenti minori di  $B$  (le stesse righe e colonne considerate in  $A$ ).

**Corollario.** Nel caso di un prodotto tra matrici quadrate si ha

$$\det AB = \det A \det B$$

### 2.4.7.1 Esempio 1

Prendiamo il seguente prodotto matriciale

$$\begin{pmatrix} 1 & 2 \\ 2 & 1 \\ 2 & 2 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix} = \begin{pmatrix} 5 & 6 & 5 \\ 4 & 6 & 7 \\ 6 & 8 & 8 \end{pmatrix}$$

Per il teorema il determinante dovrebbe essere nullo! Verifichiamolo

$$\det = 5(6*8 - 8*7) + 6(7*6 - 4*8) + 5(4*8 - 6*6) = 0$$

### 2.4.7.2 Esempio 2

Prendiamo il seguente prodotto matriciale

$$\begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \\ 2 & 2 & 2 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 2 & 1 \\ 2 & 2 \end{pmatrix} = \begin{pmatrix} 11 & 10 \\ 9 & 10 \end{pmatrix}$$

che ha determinante  $\det = 11 * 10 - 10 * 9 = 20$ . Ricalcoliamo il determinante ricorrendo a Binet. L'ordine massimo dei minori in  $A$  è 2, e i possibili sono i seguenti 3

$$\det \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix} = 2 - 6 = -4 \quad \det \begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} = 2 - 6 = -4 \quad \det \begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix} = 1 - 9 = -8$$

Si moltiplica per i minori equivalenti in  $B$

$$\det \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} = 1 - 4 = -3 \quad \det \begin{pmatrix} 2 & 1 \\ 2 & 2 \end{pmatrix} = 4 - 2 = 2 \quad \det \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix} = 2 - 4 = -2$$

cioè

$$\det \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix} \det \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} + \det \begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \det \begin{pmatrix} 2 & 1 \\ 2 & 2 \end{pmatrix} + \det \begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix} \det \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix} = 20$$

## 2.5 Matrici inverse

### 2.5.1 Matrice inversa

**Definizione.** Una matrice non singolare  $A \in \mathbb{C}^{n \times n}$  ha una corrispondente *matrice inversa*  $A^{-1}$  (unica) tale che

$$AA^{-1} = A^{-1}A = I$$

### 2.5.2 Determinante della matrice inversa

**Definizione.** Data una matrice  $A$  non singolare il determinante della matrice inversa  $A^{-1}$  è il seguente

$$\det A^{-1} = \frac{1}{\det A}$$

La cosa è dimostrabile in modo agile ricorrendo al teorema di Binet-Cauchy

$$\det A \det A^{-1} = \det AA^{-1} = \det I = 1$$

#### 2.5.2.1 Esempio

Prendiamo la seguente matrice, di cui calcoliamo l'inversa

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

Calcoliamo i determinanti

$$\det(A) = 2(3 - 2) - 3(1 - 0) + 1(2 - 0) = 2 - 3 + 2 = 1$$

per quanto detto il determinante dell'inversa dovrebbe essere

$$\det(A^{-1}) = \frac{1}{\det(A)} = \frac{1}{1} = 1$$

verifichiamolo

$$\det(A^{-1}) = 1(6 - 4) + 1(-3 + 2) + 0(4 - 4) = 2 - 1 = 1$$

## 2.6 Classificazione delle matrici

### 2.6.1 Matrice trasposta

**Definizione.** Data la matrice  $A \in \mathbb{C}^{m \times n}$ , la matrice  $B \in \mathbb{C}^{n \times m}$  i cui elementi sono  $b_{ij} = a_{ji}$  è detta *matrice trasposta di A*. La si indica con  $A^T$ .

Si tenga a mente la seguente uguaglianza:  $(AB)^T = B^T A^T$

### 2.6.2 Determinante di una matrice trasposta

**Definizione.** Il determinante di una matrice quadrata è uguale al determinante della corrispondente matrice trasposta

$$\det A = \det A^T$$

#### 2.6.2.1 Esempio

Prendiamo la seguente matrice, di cui calcoliamo la trasposta

$$A = \begin{pmatrix} 2 & 3 & 1 \\ 1 & 3 & 1 \\ 0 & 2 & 1 \end{pmatrix} \implies A^T = \begin{pmatrix} 2 & 1 & 0 \\ 3 & 3 & 2 \\ 1 & 1 & 1 \end{pmatrix}$$

Calcoliamo i determinanti

$$\det(A) = 2(3 - 2) - 3(1 - 0) + 1(2 - 0) = 2 - 3 + 2 = 1$$

$$\det(A^T) = 2(3 - 2) - 1(3 - 2) + 0(3 - 3) = 2 - 1 = 1$$

### 2.6.3 Matrice simmetrica

**Definizione.** Una matrice  $A \in \mathbb{C}^{n \times n}$  si dice *simmetrica* se è uguale alla corrispondente trasposta

$$A = A^T$$

#### 2.6.3.1 Esempio

$$\begin{aligned} A_1 &= \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} & \implies A_1^T &= \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \\ A_2 &= \begin{pmatrix} 2+i & 3 & i & -2 \\ 3 & -1 & 7 & 1+i \\ i & 7 & 0 & 0 \\ -2 & 1+i & 0 & 4 \end{pmatrix} & \implies A_2^T &= \begin{pmatrix} 2+i & 3 & i & -2 \\ 3 & -1 & 7 & 1+i \\ i & 7 & 0 & 0 \\ -2 & 1+i & 0 & 4 \end{pmatrix} \end{aligned}$$

## 2.6.4 Matrice anti-simmetrica

**Definizione.** Una matrice  $A \in \mathbb{C}^{n \times n}$  si dice *anti-simmetrica* se è uguale alla corrispondente trasposta, cambiata di segno

$$A = -A^T$$

### 2.6.4.1 Esempi

$$A_1 = \begin{pmatrix} 0 & 2 & 3 \\ -2 & 0 & -5 \\ -3 & 5 & 0 \end{pmatrix} \Rightarrow A_1^T = \begin{pmatrix} 0 & -2 & -3 \\ 2 & 0 & 5 \\ 3 & -5 & 0 \end{pmatrix} \Rightarrow -A_1^T = \begin{pmatrix} 0 & 2 & 3 \\ -2 & 0 & -5 \\ -3 & 5 & 0 \end{pmatrix}$$

$$A_2 = \begin{pmatrix} 0 & 3+i & i & -7 \\ -3-i & 0 & 5 & 2i \\ -i & -5 & 0 & 4 \\ 7 & -2i & -4 & 0 \end{pmatrix} \Rightarrow A_2^T = \begin{pmatrix} 0 & -3-i & -i & 7 \\ 3+i & 0 & -5 & -2i \\ i & 5 & 0 & -4 \\ -7 & 2i & 4 & 0 \end{pmatrix} \Rightarrow -A_2^T = \begin{pmatrix} 0 & 3+i & i & -7 \\ 3-i & 0 & 5 & 2i \\ -i & -5 & 0 & 4 \\ 7 & -2i & -4 & 0 \end{pmatrix}$$

Per costruire una matrice anti-simmetrica:

1. si costruisce una matrice simmetrica;
2. si cambia il segno agli elementi  $a_{ij}$  tali che  $i < j$ , o agli elementi  $a_{ij}$  tali che  $i > j$ .

Possiamo dire con  $B = A^T$  che  $b_{ij} = -a_{ji}$ , con  $i, j = 1, \dots, n$ .

## 2.6.5 Matrice trasposta coniugata

**Definizione.** Data la matrice  $A \in \mathbb{C}^{m \times n}$ , la matrice  $B \in \mathbb{C}^{n \times m}$  i cui elementi sono

$$b_{ij} = \bar{a}_{ji}$$

è detta *matrice trasposta coniugata di A*. La si indica con  $A^H$ . Se la matrice è reale allora le definizioni di matrice trasposta e matrice trasposta coniugata coincidono.

### 2.6.5.1 Esempio

$$A = \begin{pmatrix} i & 2 & 3-i \\ 4 & -5i & 0 \\ 3 & 2 & -2-i \end{pmatrix} \Rightarrow A^T = \begin{pmatrix} i & 4 & 3 \\ 2 & -5i & 2 \\ 3-i & 0 & -2-i \end{pmatrix} \Rightarrow \bar{A}^T = \begin{pmatrix} -i & 4 & 3 \\ 2 & 5i & 2 \\ 3+i & 0 & -2+i \end{pmatrix}$$

## 2.6.6 Matrice hermitiana

**Definizione.** Una matrice  $A \in \mathbb{C}^{n \times n}$  si dice *hermitiana* se è uguale alla corrispondente trasposta coniugata

$$A = A^H$$

questo significa che gli elementi lungo la diagonale principale sono reali! In presenza di una matrice reale ed hermitiana abbiamo una matrice simmetrica (nell'insieme dei numeri reali le definizioni coincidono).

Si tenga a mente la seguente uguaglianza:  $(AB)^H = B^H A^H$

### 2.6.6.1 Esempio

$$A = \begin{pmatrix} 1 & 3+i \\ 3-i & 2 \end{pmatrix} \rightarrow A^T = \begin{pmatrix} 1 & 3-i \\ 3+i & 2 \end{pmatrix} \Rightarrow \bar{A}^T = \begin{pmatrix} 1 & 3+i \\ 3-i & 2 \end{pmatrix}$$

Per costruire una matrice hermitiana si costruisca una matrice:

- con elementi reali lungo la diagonale principale ( $b_{ij} = \bar{a}_{ji}$ );
- simmetrica, ma in presenza di elementi  $\in \mathbb{C}$  uno dei due elementi deve essere il coniugato.

### 2.6.7 Matrice anti-hermitiana

**Definizione.** Una matrice  $A \in \mathbb{C}^{n \times n}$  si dice *anti-hermitiana* se è uguale alla corrispondente trasposta coniugata, cambiata di segno

$$A = -A^H$$

questo significa che gli elementi lungo la diagonale principale non sono solo reali, ma anche nulli!

#### 2.6.7.1 Esempi

**Esempio 1** Si consideri la seguente matrice

$$A = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} \Rightarrow A^T = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} \Rightarrow \bar{A}^T = \begin{pmatrix} 0 & -i \\ -i & 0 \end{pmatrix}$$

che cambiato di segno risulta

$$-\bar{A}^T = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}$$

**Esempio 2** Si consideri la seguente matrice

$$A = \begin{pmatrix} 0 & -1+i \\ 1+i & 0 \end{pmatrix} \Rightarrow A^T = \begin{pmatrix} 0 & 1+i \\ -1+i & 0 \end{pmatrix} \Rightarrow \bar{A}^T = \begin{pmatrix} 0 & 1-i \\ -1-i & 0 \end{pmatrix}$$

che cambiato di segno risulta

$$-\bar{A}^T = \begin{pmatrix} 0 & -1+i \\ 1+i & 0 \end{pmatrix}$$

**Costruzione di una matrice anti-hermitiana** Per costruire una matrice anti-hermitiana si costruisca una matrice:

- con elementi nulli lungo la diagonale principale ( $b_{ij} = \bar{a}_{ji}$ );
- simmetrica, ma in presenza di elementi  $\in \mathbb{C}$  si ha lo stesso modulo di parte reale e parte immaginaria, ma parte reale cambiata di segno.

## 2.6.8 Matrice normale

**Definizione.** Una matrice  $A \in \mathbb{C}^{n \times n}$  si dice *normale* se vale l'uguaglianza

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

### 2.6.8.1 Esempio

Si consideri la seguente matrice, di cui calcoliamo la trasposta coniugata

$$A = \begin{pmatrix} -i & -i & 0 \\ -i & i & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad A^T = \begin{pmatrix} -i & -i & 0 \\ -i & i & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \bar{A}^T = \begin{pmatrix} i & i & 0 \\ i & -i & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Facciamo i prodotti tra matrice  $A$  e trasposta, otteniamo

$$\begin{aligned} AA^H &= \begin{pmatrix} -i & -i & 0 \\ -i & i & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} i & i & 0 \\ i & -i & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix} \\ A^H A &= \begin{pmatrix} i & i & 0 \\ i & -i & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} -i & -i & 0 \\ -i & i & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix} \end{aligned}$$

## 2.6.9 Matrice unitaria

**Definizione.** Una matrice  $A \in \mathbb{C}^{n \times n}$  si dice *unitaria* se il prodotto tra la matrice e la corrispondente trasposta coniugata è uguale alla matrice identica

$$A^H A = AA^H = I$$

Deduciamo dalla definizione che la matrice unitaria consiste in una particolare matrice normale: addirittura la trasposta coniugata è la matrice inversa!

### 2.6.9.1 Esempio

Si consideri la seguente matrice, di cui calcoliamo la trasposta coniugata

$$A = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} \quad A^T = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} \quad \bar{A}^T = \begin{pmatrix} 0 & -i \\ -i & 0 \end{pmatrix}$$

Facciamo i prodotti tra matrice e trasposta, otteniamo

$$AA^H = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} \begin{pmatrix} 0 & -i \\ -i & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad A^H A = \begin{pmatrix} 0 & -i \\ -i & 0 \end{pmatrix} \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

## 2.6.10 Matrice ortogonale

**Definizione.** Una matrice reale e unitaria è detta *ortogonale*. L'inversa della matrice è la trasposta!

### 2.6.10.1 Esempio

Si consideri la seguente matrice, di cui calcoliamo la trasposta

$$A = \frac{1}{3} \begin{pmatrix} 1 & 2 & 2 \\ 2 & 1 & -2 \\ -2 & 2 & -1 \end{pmatrix} \quad A^T = \frac{1}{3} \begin{pmatrix} 1 & 2 & -2 \\ 2 & 1 & 2 \\ 2 & -2 & -1 \end{pmatrix}$$

Facciamo i prodotti tra matrice  $A$  e trasposta, otteniamo

$$\begin{aligned} AA^T &= \frac{1}{9} \begin{pmatrix} 1 & 2 & 2 \\ 2 & 1 & -2 \\ -2 & 2 & -1 \end{pmatrix} \begin{pmatrix} 1 & 2 & -2 \\ 2 & 1 & 2 \\ 2 & -2 & -1 \end{pmatrix} & A^T A &= \frac{1}{9} \begin{pmatrix} 1 & 2 & -2 \\ 2 & 1 & 2 \\ 2 & -2 & -1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 2 \\ 2 & 1 & -2 \\ -2 & 2 & -1 \end{pmatrix} \\ &= \frac{1}{9} \begin{pmatrix} 9 & 0 & 0 \\ 0 & 9 & 0 \\ 0 & 0 & 9 \end{pmatrix} & &= \frac{1}{9} \begin{pmatrix} 9 & 0 & 0 \\ 0 & 9 & 0 \\ 0 & 0 & 9 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{aligned}$$

## 2.7 Segno della matrice

### 2.7.1 Formula per l'individuazione del segno

Data una matrice hermitiana  $A$  ed un vettore  $x \in \mathbb{C}^n$  possiamo affermare che lo scalare  $x^H Ax \in \mathbb{R}$ . Possiamo farlo dimostrando che la formula è uguale alla sua trasposta coniugata (significa avere parte immaginaria nulla)

$$(x^H Ax)^H = x^H A^H x = x^H Ax$$

Gli spostamenti ci portano a mettere  $x$  prima di  $A$  come  $x^H$ ,  $x^H$  dopo  $A$  come  $x$ , e  $A$  come  $A^H$ . Otteniamo lo stesso prodotto iniziale sapendo che  $A^H = A$ .

### 2.7.2 Matrice definita positiva

**Definizione.** Una matrice hermitiana  $A$  si dice *definita positiva* se  $x^H Ax > 0$ .

### 2.7.3 Matrice definita negativa

**Definizione.** Una matrice hermitiana  $A$  si dice *definita negativa* se  $x^H Ax < 0$ .

### 2.7.4 Matrice semidefinita positiva

**Definizione.** Una matrice hermitiana  $A$  si dice *semidefinita positiva* se  $x^H Ax \geq 0$ .

### 2.7.5 Matrice semidefinita negativa

**Definizione.** Una matrice hermitiana  $A$  si dice *semidefinita negativa* se  $x^H Ax \leq 0$ .

## 2.8 Matrici diagonali e triangolari

### 2.8.1 Matrice diagonale

**Definizione.** Una matrice quadrata è detta *matrice diagonale* se tutti gli elementi esterni alla diagonale principale sono nulli

$$D = \begin{pmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & d_n \end{pmatrix}$$

E' possibile scrivere su Matlab una matrice diagonale ponendola nel seguente modo

```
D=diag([d1,d2,...,dn])
```

### 2.8.2 Matrice triangolare inferiore

**Definizione.** Si dice *matrice triangolare inferiore* la matrice quadrata del tipo

$$L = \begin{pmatrix} l_{11} & 0 & \dots & 0 \\ l_{21} & l_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{pmatrix} \quad \text{cioè una matrice dove } l_{ij} = 0, \forall i < j$$

### 2.8.3 Matrice triangolare superiore

**Definizione.** Si dice *matrice triangolare superiore* la matrice quadrata del tipo

$$R = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ 0 & r_{22} & \dots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & r_{nn} \end{pmatrix} \quad \text{cioè una matrice dove } l_{ij} = 0, \forall i > j$$

### 2.8.4 Determinante di una matrice diagonale/triangolare

**Definizione.** Il determinante di una matrice diagonale o triangolare è il prodotto tra le componenti lungo la diagonale principale

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

chiaramente occorre che i blocchi siano quadrati.

## 2.9 Matrice a predominanza diagonale

### 2.9.1 Matrice a predominanza diagonale forte

**Definizione.** Una matrice  $A \in \mathbb{C}^{n \times n}$  si dice a *predominanza diagonale forte* se

$$|a_{ii}| > \sum_{j=1, j \neq i}^n |a_{ij}|, \quad i = 1, 2, \dots, n$$

Data una riga  $i$ -esima il modulo dell'elemento diagonale è maggiore della somma dei moduli degli elementi appartenenti alla stessa riga e diversi dall'elemento diagonale.

#### 2.9.1.1 Esempio

Si consideri il seguente esempio:

$$A = \begin{pmatrix} 3i & -1 & 1 \\ 2 & -7 & 3 \\ 2i & -3 & 7+3i \end{pmatrix} \quad \begin{array}{l} 3 > 1 + 1 \\ 7 > 3 + 2 \\ 58 > 2 + 3 \end{array}$$

la matrice è a predominanza diagonale forte!

### 2.9.2 Matrice a predominanza diagonale debole

**Definizione.** Una matrice  $A \in \mathbb{C}^{n \times n}$  si dice a *predominanza diagonale debole* se

$$|a_{ii}| \geq \sum_{j=1, j \neq i}^n |a_{ij}|, \quad i = 1, 2, \dots, n$$

Inoltre, per almeno un indice  $r$  ( $1 \leq r \leq n$ )

$$|a_{rr}| > \sum_{j=1, j \neq r}^n |a_{rj}|$$

#### 2.9.2.1 Esempio

Si consideri il seguente esempio:

$$A = \begin{pmatrix} -5 & -1 & 1 \\ 2 & -5 & 3 \\ -2i & 4i & 8i \end{pmatrix} \quad \begin{array}{l} 5 \geq 1 + 1 \\ 5 \geq 2 + 3 \\ 8 \geq 2 + 4 \end{array}$$

la matrice è a predominanza diagonale debole!

## 2.10 Matrici convergenti e nilpotenti

### 2.10.1 Matrice convergente

**Definizione.** Una matrice  $A \in \mathbb{C}^{n \times n}$  si dice *convergente* se il limite delle potenze successive della matrice è la matrice nulla  $O$  (avente le stesse dimensioni)

$$\lim_{k \rightarrow \infty} A^k = O$$

### 2.10.2 Matrice nilpotente

**Definizione.** Una matrice  $A \in \mathbb{C}^{n \times n}$  si dice *nilpotente* se esiste un valore  $k$  per cui

$$A^k = O$$

#### 2.10.2.1 Esempio di matrice nilpotente

Si prenda ad esempio la seguente matrice, che è matrice nilpotente con  $k = 3$

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 3 & 0 & 0 \\ -2 & 8 & 0 \end{pmatrix} \quad k = 2 \quad \begin{pmatrix} 0 & 0 & 0 \\ 3 & 0 & 0 \\ -2 & 8 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 3 & 0 & 0 \\ -2 & 8 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 24 & 0 & 0 \end{pmatrix}$$
$$k = 3 \quad \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 24 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 3 & 0 & 0 \\ -2 & 8 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

## 2.11 Trasformaz. per similitudine e matrici di permutazione

### 2.11.1 Trasformazioni per similitudine

**Definizione.** Data una matrice  $A \in \mathbb{C}^{n \times n}$  ed una matrice  $S \in \mathbb{C}^{n \times n}$  non singolare, si dice *trasformata per similitudine della matrice A* la matrice  $B$  tale che

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

Le matrici  $A$  e  $B$  sono dette *matrici simili*.

### 2.11.2 Matrice di permutazione

**Definizione.** Una matrice  $P \in \mathbb{R}^{n \times n}$  è detta *matrice di permutazione* se è ottenuta dalla matrice identica operando su di essa una permutazione di colonne (o di righe). Abbiamo  $n!$  possibili matrici di permutazione.

**Matrice ortogonale** Una matrice di permutazione è una matrice ortogonale, quindi una matrice reale e unitaria. Questo significa che l'inversa di  $P$  è la trasposta  $P^T$

$$P^T P = P P^T = I$$

**Permutazione su una matrice generica  $A$**  Si osservi che il prodotto di una matrice  $A$  per una matrice di permutazione  $P$  produce su  $A$  una permutazione di colonne o righe.

- $AP$  induce sulla matrice  $A$  la permutazione di colonne operata su  $I$  per ottenere  $P$ .
- $PA$  induce sulla matrice  $A$  la permutazione di righe operata su  $I$  per ottenere  $P$ .

**Trasposta della matrice di permutazione** Se  $P$  si ottiene permutando le colonne di  $I$  allora  $P^T$  si ottiene con la stessa permutazione delle righe di  $I$ .

#### 2.11.2.1 Esempio

Prendiamo ad esempio la seguente matrice

$$\begin{aligned} P &= (e^1|e^2|e^3) = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} & PP^T &= \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ P^T &= \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} & AP &= \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 2 & 3 & 1 \\ 5 & 6 & 4 \\ 8 & 9 & 7 \end{pmatrix} \\ A &= \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} & PA &= \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} = \begin{pmatrix} 7 & 8 & 9 \\ 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \end{aligned}$$

### 2.11.2.2 Permutazione in simultanea sia su righe che su colonne

Proviamo a svolgere il seguente calcolo

$$P^T AP = P^T(AP)$$

Il calcolo è letteralmente una particolare *trasformazione per similitudine*: l'insieme in cui si opera è quello dei reali  $\mathbb{R}$ , l'inversa di  $P$  è la sua trasposta  $P^T$  (come già detto). Si riprenda lo scorso esempio: svolgendo i calcoli detti otteniamo

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 5 & 6 & 4 \\ 8 & 3 & 7 \\ 2 & 3 & 1 \end{pmatrix}$$

La matrice ottenuta presenta tutti gli elementi presenti nella matrice  $A$ , ma risulta permuta sia rispetto alle righe sia rispetto alle colonne.

### 2.11.2.3 Determinante della matrice permutata

Qual è il determinante della matrice  $B$  ottenuta per mezzo delle permutazioni indicate con la matrice  $P$ ?

$$B = P^T AP$$

Sviluppiamo il determinante

$$\begin{aligned} \det(B) &= \det(P^T AP) = \det(P^T) \det(A) \det(P) = \det(P^T) \det(P) \det(A) = \\ &= \det(P^T P) \det(A) = \det(I) \det(A) = \det(A) \end{aligned}$$

Il determinante della matrice permutata  $A$  è lo stesso della matrice  $B$ !

## 2.12 Partizionamenti a blocchi

Nelle applicazioni si ricorre spesso al partizionamento di una matrice generica  $A$  (anche non quadrata) in blocchi, cioè da una matrice  $A$  iniziale otteniamo delle sottomatrici. I modi in cui noi possiamo partizionare una matrice sono molteplici: in ogni caso partizioniamo tracciando righe e colonne, che devono andare da un estremo all'altro della matrice.

### 2.12.1 Esempio

Prendiamo il seguente esempio di partizionamento a blocchi

$$\begin{aligned} A &= \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{pmatrix} & A_{11} &= (1) & A_{12} &= (2 \ 3 \ 4) \\ A &= \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} & A_{21} &= \begin{pmatrix} 5 \\ 9 \\ 13 \end{pmatrix} & A_{22} &= \begin{pmatrix} 6 & 7 & 8 \\ 10 & 11 & 12 \\ 14 & 15 & 16 \end{pmatrix} \end{aligned}$$

Rispetta quanto detto in quanto

- $A_{11}$  ha lo stesso numero di righe di  $A_{12}$ ,  $A_{21}$  ha lo stesso numero di righe di  $A_{22}$
- $A_{11}$  ha lo stesso numero di colonne di  $A_{21}$ ,  $A_{12}$  ha lo stesso numero di colonne di  $A_{22}$

## 2.12.2 Matrice triangolare a blocchi inferiore

**Definizione.** Si dice *matrice triangolare a blocchi inferiore* la matrice a blocchi del tipo

$$A = \begin{pmatrix} A_{11} & 0 & \dots & 0 \\ A_{21} & A_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \dots & A_{nn} \end{pmatrix}$$

Si tenga a mente che avere una matrice a blocchi inferiore non implica avere una matrice triangolare inferiore.

### 2.12.2.1 Esempio

Si consideri il seguente esempio

$$\begin{aligned} A &= \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} & A_{11} &= \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} & A_{12} &= \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \\ A &= \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix} & A_{21} &= \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} & A_{22} &= \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \end{aligned}$$

la matrice è triangolare a blocchi inferiore, ma non triangolare inferiore.

## 2.12.3 Matrice triangolare a blocchi superiore

**Definizione.** Si dice *matrice triangolare a blocchi superiore* la matrice a blocchi del tipo

$$A = \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ 0 & A_{22} & \dots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A_{nn} \end{pmatrix}$$

Si tenga a mente che avere una matrice a blocchi superiore non implica avere una matrice triangolare superiore.

### 2.12.3.1 Esempio

Si consideri il seguente esempio

$$\begin{aligned} A &= \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} & A_{11} &= \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} & A_{12} &= \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \\ A &= \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} & A_{21} &= \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} & A_{22} &= \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \end{aligned}$$

la matrice è triangolare a blocchi superiore, ma non triangolare superiore.

## 2.12.4 Matrice diagonale a blocchi

**Definizione.** Una matrice quadrata è detta *matrice blocchi* se tutte le sottomatrici esterne alla diagonale principale sono matrici nulle

$$A = \begin{pmatrix} A_1 & 0 & \dots & 0 \\ 0 & A_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A_n \end{pmatrix}$$

### 2.12.4.1 Esempio

Consideriamo il seguente esempio di matrice diagonale a blocchi

$$A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \quad A_{11} = \begin{pmatrix} 1 & 1 \end{pmatrix} \quad A_{12} = \begin{pmatrix} 0 & 0 \end{pmatrix}$$

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \quad A_{21} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \quad A_{22} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}$$

## 2.12.5 Determinante di una matrice diagonale/triangolare a blocchi

**Definizione.** Il determinante di una matrice diagonale a blocchi o triangolare a blocchi è il prodotto tra i determinanti delle sottomatrici lungo la diagonale principale

$$\det A = \prod_{i=1}^n \det A_{ii}$$

### 2.12.5.1 Esempi

**Primo esempio (scemo)** Consideriamo la seguente matrice triangolare superiore

$$A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

in un certo senso possiamo immaginarci blocchi costituiti da un solo elemento. A quel punto otteniamo il determinante in modo banale

$$\det(A) = \det(A_{11}) \det(A_{22}) = 1 * 1 = 1$$

**Secondo esempio** Consideriamo la seguente matrice triangolare inferiore a blocchi dove  $A_{ij} \in \mathbb{R}^{2 \times 2}$ ,  $i, j = 1, 2$ .

$$A = \begin{pmatrix} 1 & 2 & 0 & 0 \\ 3 & 4 & 0 & 0 \\ 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

Applichiamo quanto detto

$$\det(A) = \det(A_{11}) \det(A_{22}) = [1 * 4 - 3 * 2] [3 * 8 - 4 * 7] = (-2)(-4) = 8$$

## 2.13 Grafi

### 2.13.1 Grafo orientato

**Definizione.** Data una matrice  $A \in \mathbb{C}^{n \times n}$  e fissati  $n$  punti detti nodi  $(N_1, \dots, N_n)$  si dice *grafo orientato* associato ad  $A$  il grafo che si ottiene congiungendo una coppia di nodi  $\langle N_i, N_j \rangle$  per mezzo di un cammino orientato da  $N_i$  ad  $N_j$  (presente solo se  $a_{ij} \neq 0$ ).

#### 2.13.1.1 Esempio

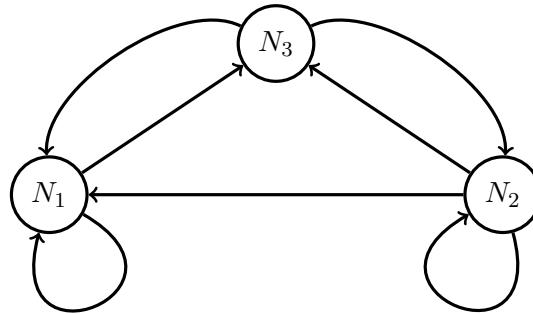
Si consideri la seguente matrice

$$A = \begin{pmatrix} -1 & 0 & 5 \\ 2 & 12 & 3 \\ -4 & 2 & 0 \end{pmatrix}$$

Il grafo orientato associato sarà costituito da 3 nodi.

- $a_{11} = -1$ , arco dal nodo 1 a se stesso
- $a_{12} = 0$ , niente arco orientato dal nodo 1 al nodo 2
- $a_{13} = 5$ , arco orientato dal nodo 1 al nodo 3
- $a_{21} = 2$ , arco orientato dal nodo 2 al nodo 1
- $a_{22} = 12$ , arco dal nodo 2 a se stesso
- $a_{23} = 3$ , arco orientato dal nodo 2 al nodo 3
- $a_{31} = -4$ , arco orientato dal nodo 3 al nodo 1
- $a_{32} = 2$ , arco orientato dal nodo 3 al nodo 2
- $a_{33} = 0$ , niente arco dal nodo 3 a se stesso

Risultato:



### 2.13.2 Grafo fortemente connesso

**Definizione.** Un grafo orientato si dice *fortemente connesso* se da ogni nodo  $N_i$  ( $i = 1, 2, \dots, n$ ) è possibile raggiungere un qualunque altro nodo  $N_j$  ( $j = 1, 2, \dots, n$ ) seguendo un cammino orientato eventualmente passante per altri nodi.

Per gli esempi si veda più avanti l'irriducibilità della matrice.

## 2.14 Matrici riducibili

### 2.14.1 Matrice riducibile

**Definizione.** Una matrice  $A \in \mathbb{C}^{n \times n}$  si dice riducibile se esiste una matrice di permutazione  $P$  tale che la matrice  $P^T AP$  sia partizionabile nella forma

$$B = P^T AP = \begin{pmatrix} B_{11} & 0 \\ B_{21} & B_{22} \end{pmatrix}$$

con blocchi diagonali quadrati.

Nella definizione si ragiona in blocchi  $2 \times 2$ . La cosa rilevante, in generale, è avere una triangolare inferiore a blocchi quadrati (o una triangolare superiore a blocchi, se posso fare la triangolare inferiore potrò fare anche la triangolare superiore): questo può succedere solo in presenza di un numero sufficiente di zeri (si ricordi che non vengono modificati elementi, cambiano solo le posizioni).

### 2.14.2 Matrice irriducibile

**Definizione.** Una matrice che non è riducibile è detta *irriducibile*: non esistono matrici  $P$  che ci permettono di ottenere  $B$  nella forma detta.

### 2.14.3 Forma ridotta

**Definizione.** Una matrice riducibile è detta in *forma ridotta* se nessun blocco diagonale risulta riducibile. Nel seguente partizionamento

$$B = P^T AP = \begin{pmatrix} B_{11} & 0 \\ B_{21} & B_{22} \end{pmatrix}$$

$B_{11}$  e  $B_{22}$  sono non riducibili.

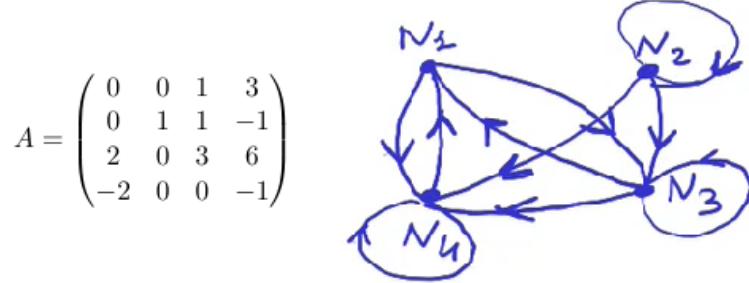
### 2.14.4 Verifica dell'irriducibilità della matrice

**Teorema.** Una matrice  $A \in \mathbb{C}^{n \times n}$  risulta irriducibile se e solo se il grafo orientato ad essa associato risulta fortemente connesso.

Ritorniamo sul discorso del numero di permutazioni possibili:  $n!$  Per verificare se una matrice è riducibile si dovrebbe andare a vedere tutte le  $n!$  permutazioni, cosa non accettabile dal punto di vista computazionale (nella peggiore delle ipotesi le vediamo tutte e nessuna rispetta la definizione di matrice riducibile). Si risolve ricorrendo ai grafi precedentemente introdotti, e al teorema qua sopra.

#### 2.14.4.1 Esempio 1

Consideriamo la seguente matrice e tracciamo il grafo partendo dalla definizione di grafo orientato:



Per ogni nodo vediamo quali sono i nodi collegati (anche indirettamente) e quali scollegati

	Nodi collegati	Nodi scollegati
$N_1$	$N_1, N_3, N_4$	$N_2$
$N_2$	Tutti	
$N_3$	$N_1, N_3, N_4$	$N_2$
$N_4$	$N_1, N_3, N_4$	$N_2$

Se uno vuole sapere solo se la matrice è riducibile può fermarsi alla prima riga: non si ha un grafo fortemente connesso, e quindi la matrice è riducibile. Proseguiamo.

1. Si considera una delle righe della tabella con nodi scollegati.
2. Si costruisce la matrice  $P$  ponendo come colonne prima quelle relative ai nodi collegati, e successivamente quelle relative ai nodi scollegati

$$P = (e^{(1)} | e^{(3)} | e^{(4)} | e^{(2)}) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

L'ordine delle colonne può essere alterato.

3. Si svolge il prodotto matriciale indicato nella definizione di riducibilità

$$B = P^T A P = P^T \begin{pmatrix} 0 & 1 & 3 & 0 \\ 0 & 1 & -1 & 1 \\ 2 & 3 & 6 & 0 \\ -2 & 0 & -1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 3 & 0 \\ 2 & 3 & 6 & 0 \\ -2 & 0 & -1 & 0 \\ 0 & 1 & -1 & 1 \end{pmatrix}$$

Il risultato finale è la seguente partizione

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \quad A_{11} = \begin{pmatrix} 0 & 1 & 3 \\ 2 & 3 & 6 \\ -2 & 0 & -1 \end{pmatrix} \quad A_{12} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

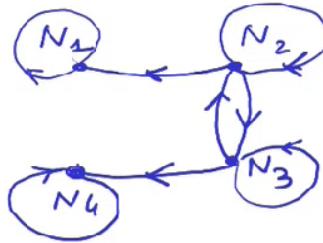
$$A_{21} = (0 \ 1 \ -1) \quad A_{22} = (1)$$

dove i blocchi diagonali  $A_{11}$  e  $A_{22}$  sono quadrati. Si osservi che non è unica (ripetiamo, prendendo le colonne di  $P$  in ordine diverso otteniamo risultati diversi).

#### 2.14.4.2 Esempio 2

Consideriamo la seguente matrice e tracciamo il grafo partendo dalla definizione di grafo orientato:

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 5 & -1 & 2 & 0 \\ 0 & -5 & 11 & 7 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$



Per ogni nodo vediamo quali sono i nodi collegati (anche indirettamente) e quali scollegati

	Nodi collegati	Nodi scollegati
$N_1$	$N_1$	$N_2, N_3, N_4$
$N_2$	Tutti	
$N_3$	Tutti	
$N_4$	$N_4$	$N_1, N_2, N_3$

Se uno vuole sapere solo se la matrice è riducibile può fermarsi alla prima riga: non si ha un grafo fortemente connesso, e quindi la matrice è riducibile. Proseguiamo.

- Considero la **prima riga della tabella**. Otteniamo il seguente  $P$

$$P = (e^{(1)}|e^{(2)}|e^{(3)}|e^{(4)})$$

La matrice  $P$  è la matrice identica, cioè non si modifica nulla della matrice iniziale. La cosa non è strana, se andiamo ad osservare la matrice iniziale osserviamo già un partizionamento a blocchi

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \quad A_{11} = (1) \quad A_{12} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$$A_{21} = \begin{pmatrix} 5 \\ 0 \\ 0 \end{pmatrix} \quad A_{22} = \begin{pmatrix} -1 & 2 & 0 \\ -5 & 11 & 7 \\ 0 & 0 & 1 \end{pmatrix}$$

- Considero la **quarta riga della tabella**. Otteniamo il seguente  $P$

$$P = (e^4|e^1|e^2|e^3)$$

Costruiamo la matrice  $B$

$$B = P^T AP = P^T \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 5 & -1 & 2 \\ 7 & 0 & -5 & 11 \\ 1 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 5 & -1 & 2 \\ 7 & 0 & -5 & 11 \end{pmatrix}$$

Il risultato finale è un partizionamento a blocchi  $3 \times 3$ !

$$A = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix} \quad A_{11} = (1) \quad A_{12} = (0) \quad A_{13} = (0 \ 0)$$

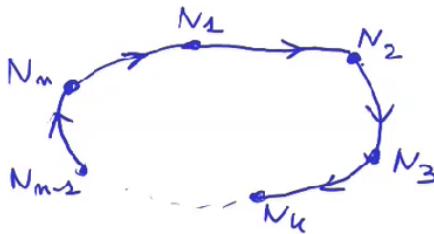
$$A_{21} = (0) \quad A_{22} = (1) \quad A_{23} = (0 \ 0)$$

$$A_{31} = \begin{pmatrix} 0 \\ 7 \end{pmatrix} \quad A_{32} = \begin{pmatrix} 5 \\ 0 \end{pmatrix} \quad A_{33} = \begin{pmatrix} -1 & 2 \\ -5 & 11 \end{pmatrix}$$

### 2.14.4.3 Esempio 3

Precedentemente abbiamo detto che è *necessario un certo numero di zeri affinchè la matrice sia riducibile*. In realtà dobbiamo stare attenti perché sulla riducibilità non influisce solo il numero di zeri, ma anche le "posizioni strategiche" dei numeri. Consideriamo il seguente esempio e tracciamo il relativo grafo

$$A = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & 1 \\ 1 & 0 & 0 & \dots & 0 \end{pmatrix} \in \mathbb{R}^{n \times n}$$



osserviamo che il grafo è un cerchio, quindi ogni coppia di nodi è collegabile. Poichè il grafo è fortemente connesso possiamo concludere che la matrice è irriducibile, pur avendo zeri a volontà.

## 2.15 Sistemi lineari

### 2.15.1 Sistema lineare

**Definizione.** Data una matrice  $A \in \mathbb{C}^{n \times n}$  ed un vettore  $b \in \mathbb{C}^n$ , un sistema di  $n$  equazioni lineari in  $n$  incognite si rappresenta nella forma

$$Ax = b$$

dove  $x \in \mathbb{C}^n$  è il vettore delle incognite. Risolvere un sistema lineare significa trovare gli elementi del vettore  $x$ .

Il sistema si risolve per mezzo del seguente calcolo, premoltiplicando per l'inversa di  $A$

$$A^{-1}Ax = A^{-1}b \longrightarrow x = A^{-1}b$$

Il calcolo non ci piace dal punto di vista computazionale (si pensi al calcolo dell'inversa in matrici di ordine elevato). Nella pratica difficilmente andremo a fare i calcoli delle definizioni matematiche.

### 2.15.2 Teorema di Rouchè-Capelli

**Teorema.** Data una matrice  $A \in \mathbb{C}^{m \times n}$  con  $m \geq n$  ed un vettore  $b \in \mathbb{C}^m$ , un sistema lineare  $Ax = b$  ammette soluzione se e solo se

$$r(A) = r(A|b)$$

dove  $A|b$  è la matrice completa del sistema, costituita da  $m$  righe ed  $n + 1$  colonne. Quindi: se aggiungendo una colonna il rango non aumenta allora sono ammesse soluzioni. Inoltre:

- se  $r(A) = n$  la soluzione è unica;
- se  $r(A) < n$  l'insieme delle soluzioni è un sottospazio di  $\mathbb{C}^n$  avente dimensione  $n - r(A)$ .

Se si considera il caso in cui la matrice è quadrata ( $m = n$ ) l'ultima parte del teorema può essere posta così:

- se  $\det(A) \neq 0$  ( $r(A) = n$ ) la soluzione è unica;
- se  $\det(A) = 0$  ( $r(A) < n$ ) l'insieme delle soluzioni è un sottospazio di  $\mathbb{C}^n$  avente dimensione  $n - r(A)$ .

### 2.15.3 Sistema normale

**Definizione.** Un sistema lineare  $Ax = b$  si dice normale se  $\det A \neq 0$

### 2.15.4 Sistema omogeneo

**Definizione.** Un sistema lineare  $Ax = b$  si dice *omogeneo* se  $b = 0$ .

- Ha sicuramente soluzioni poichè  $r(A) = r(A|b)$  (aggiungere una colonna di zeri non altera sicuramente il rango).
- Se  $\det A \neq 0$  la soluzione nel sistema omogeneo è unica. L'unica soluzione possibile è  $x = 0!!!$
- Se  $\det A = 0$  il sistema omogeneo ha anche soluzioni diverse da  $x = 0$ .

### 2.15.5 Risoluzione di un sistema lineare a blocchi

Precedentemente abbiamo parlato di matrici riducibili, ma abbiamo anche accennato alla risoluzione di sistemi lineari. Sappiamo come il calcolo di una matrice inversa sia operazione estremamente complessa

$$x = A^{-1}b$$

ergo in presenza di un sistema generico  $n \times n$  non andremo mai a calcolare direttamente l'inversa.

**Matrice riducibile e sistema a blocchi** Prendiamo il sistema lineare

$$Ax = b$$

dove  $A \in \mathbb{C}^{n \times n}$ . Supponiamo che la matrice  $A$  del sistema sia riducibile e che si abbia una matrice di permutazione  $P$  tale che

$$B = P^T AP = \begin{pmatrix} B_{11} & 0 & \dots & 0 \\ B_{21} & B_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ B_{k1} & B_{k2} & \dots & B_{kk} \end{pmatrix}$$

dove  $k$  è il numero di blocchi diagonali ottenuti. Vogliamo sostituire nel sistema la matrice  $A$  con la matrice  $B$  appena calcolata.

- Per prima cosa premoltiplico

$$P^T Ax = P^T b$$

- Adesso come procediamo? Non possiamo porre cose come ci pare, considerate le proprietà del prodotto matriciale e le dimensioni delle varie matrici, oltre che del vettore. Quello che noi facciamo è introdurre una matrice identica (che è un po' come moltiplicare uno scalare per 1)

$$P^T A I x = P^T b$$

La matrice identica è uguale al prodotto matriciale tra una matrice e la sua inversa: sostituiamo  $I = P P^T$  (L'inversa della matrice di permutazione  $P$  è la trasposta  $P^T$ ). Risultato:

$$P^T A P P^T x = P^T b \longrightarrow B P^T x = P^T b$$

sostituendo  $y = P^T x$  e  $c = P^T b$ . Ecco il sistema finale

$$\boxed{B y = c}$$

**Risoluzione del sistema a blocchi** Abbiamo ottenuto un sistema a blocchi del tipo

$$\begin{cases} B_{11}y_1 &= c_1 \\ B_{21}y_1 + B_{22}y_2 &= c_2 \\ \vdots &= \vdots \\ B_{k1}y_1 + B_{k2}y_2 + \cdots + B_{kk}y_k &= c_k \end{cases}$$

dove  $y$  e  $c$  sono partizionati adeguatamente in  $k$  blocchi, in modo da rispettare le dimensioni dei blocchi diagonali  $B_{ii}$ . L'algoritmo risolutivo è il seguente

- Risolvo il primo sistema  $B_{11}y_1 = c_1$  ottenendo  $y_1$ . Sostituisco  $y_1$  nei sistemi successivi.
- Risolvo il secondo sistema  $B_{21}y_1 + B_{22}y_2 = c_2$ . Sostituisco  $y_2$  nei sistemi successivi.
- Proseguo fino a quando non avrò calcolato tutti i blocchi  $y_i$  ( $i = 1, \dots, k$ )
- Una volta ottenuto  $y$  ricavo  $x$  premoltiplicando per l'inverso di  $P^T$ , cioè  $P$

$$y = P^T x \longrightarrow P y = P P^T x \longrightarrow \boxed{P y = x}$$

L'algoritmo ci piace perchè migliore dal punto di vista computazionale:

1. da una parte abbiamo la risoluzione di un grande sistema avente dimensioni  $n \times n$ ;
2. dall'altra abbiamo la risoluzione di  $k$  sistemi lineari, tutti di ordine inferiore ad  $n$ .

### Ammissibilità dei calcoli fatti

- Ciò che ci garantisce l'esistenza e l'unicità della soluzione del sistema  $Ax = b$  è il  $\det A \neq 0$  (matrice non singolare).
- Cosa possiamo dire per il sistema  $By = c$ ? Abbiamo già visto che il determinante di  $B$  è lo stesso di  $A$ , quindi

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

L'uguaglianza ci garantisce l'esistenza e l'unicità di tutti i  $k$  sottosistemi da risolvere: il determinante di una matrice a blocchi triangolare si ottiene dal prodotto dei determinanti lungo la diagonale, quindi nessun blocco lungo la diagonale ha determinante nullo.

## 2.16 Autovalori e autovettori

### 2.16.1 Autovalore

**Definizione.** Data una matrice  $A \in \mathbb{C}^{n \times n}$  si dice *autovalore* della matrice  $A$  ogni numero  $\lambda \in \mathbb{C}$  tale che

$$Ax = \lambda x$$

con  $x \in \mathbb{C}^n$  e  $x \neq 0$ .

### 2.16.2 Autovettore destro

**Definizione.** Dato un autovalore  $\lambda$  definiamo *autovettore destro* il vettore  $x \in \mathbb{C}^n$ , ed ogni vettore  $kx$  (con  $k \in \mathbb{C}, k \neq 0$ ), tale che

$$Ax = \lambda x$$

### 2.16.3 Autovettore sinistro

**Definizione.** Dato un autovalore  $\lambda$  definiamo *autovettore sinistro* il vettore  $y \in \mathbb{C}^n$ , ed ogni vettore  $ky$  (con  $k \in \mathbb{C}, k \neq 0$ ), tale che

$$y^T A = \lambda y^T$$

### 2.16.4 Polinomio caratteristico

**Definizione.** Data una matrice  $A \in \mathbb{C}^{n \times n}$  definiamo *polinomio caratteristico* della matrice  $A$  quanto segue

$$P(\lambda) = \det(A - \lambda I)$$

**Come arriviamo al polinomio caratteristico?** Dalla definizione di autovalore sappiamo che è valida l'uguaglianza

$$Ax = \lambda x$$

se spostiamo tutto al primo membro otteniamo

$$(A - \lambda I)x = 0$$

Un sistema omogeneo ha sicuramente soluzioni, ai sensi del teorema di Rouchè-Capelli. Lo stesso teorema, a proposito di un sistema omogeneo, ci permette di dire che avremo soluzioni  $x \neq 0$  solo se  $\det(A - \lambda I) = 0$ . Ecco l'equazione caratteristica, ottenuta ponendo uguale a zero il polinomio caratteristico:

$$\det(A - \lambda I) = 0$$

Si osservi che noi nella pratica non calcoleremo mai l'equazione caratteristica per individuare gli autovalori (si pensi a sistemi  $500 \times 500$ , impossibile fare questi calcoli).

**Formula con cui scriviamo il polinomio caratteristico** Siano  $\lambda_i$  gli autovalori della matrice, cioè le soluzioni dell'equazione caratteristica. Dal teorema di Ruffini scriviamo

$$P(\lambda) = (-1)^n(\lambda - \lambda_1)(\lambda - \lambda_2) \dots (\lambda - \lambda_n)$$

Quanto detto può essere scritto anche nel seguente modo

$$P(\lambda) = (-1)^n \lambda^n + (-1)^{n-1} \sigma_1 \lambda^{n-1} + (-1)^{n-2} \sigma_2 \lambda^{n-2} + \dots + \sigma_{n-2} \lambda^2 - \sigma_{n-1} \lambda + \sigma_n$$

dove i coefficienti  $\sigma_i$  (con  $i = 1, \dots, n$ ) sono, ciascuno, la *somma dei minori principali di ordine  $i$  estratti dalla matrice  $A$* . Poniamo il focus sui seguenti coefficienti

$$\sigma_1 = \sum_{j=1}^n a_{jj} \quad \sigma_n = \det(A)$$

In aggiunta possiamo dire che

$$\sigma_1 = \sum_{j=1}^n \lambda_j \quad \sigma_n = \prod_{j=1}^n \lambda_j$$

- **La somma dei minori principali di ordine 1 è la somma degli autovalori.**

La somma degli autovalori è detta anche *traccia* e si indica con  $\text{tr}(A)$ . Nella scomposizione secondo Ruffini si considerano quei termini dove si moltiplicano tutti i  $\lambda$ , tranne uno, con un  $-\lambda_k$ .

$$P(\lambda) = (-1)^n(\lambda - \lambda_1)(\lambda - \lambda_2) \cdot (\lambda - \lambda_n)$$

$$\begin{aligned} & (-1)^n [\lambda^{n-1}(-\lambda_1) + \lambda^{n-1}(-\lambda_2) + \dots + \lambda^{n-1}(-\lambda_n)] \\ & (-1)^n \lambda^{n-1} (-\lambda_1 - \lambda_2 - \dots - \lambda_n) \\ & (-1)^{n+1} \lambda^{n-1} (\lambda_1 + \lambda_2 + \dots + \lambda_n) \\ & (-1)^{n-1} (-1)(-1) \lambda^{n-1} \sigma_1 \\ & (-1)^{n-1} \lambda^{n-1} \sigma_1 \end{aligned}$$

Abbiamo ottenuto il termine della sommatoria  $(-1)^{n-1} \sigma_1 \lambda^{n-1}$ , mostrando che  $\sigma_1$  è effettivamente la somma degli autovalori.

- **La somma dei minori principali di ordine  $n$ , cioè il determinante della matrice  $A$ , equivale al prodotto degli autovalori.**

Nella scomposizione secondo Ruffini si prende quel termine dove si moltiplicano tutti gli autovalori

$$(-1)^n (-\lambda_1)(-\lambda_2) \cdot (-\lambda_n)$$

Si moltiplica per  $(-1)^n$ : se  $n$  è pari ho il prodotto degli autovalori positivo e  $(-1)^n = 1$ ; se  $n$  è dispari ho il prodotto degli autovalori negativo  $(-1)^n = -1$ , quindi un prodotto sempre positivo.

**Conseguenza.** Se  $\det(A) = 0$  allora avrò almeno un autovalore nullo, se  $\det(A) \neq 0$  non posso avere autovalori nulli!

## 2.16.5 Traccia della matrice

**Definizione.** Data una matrice  $A \in \mathbb{C}^{n \times n}$  definiamo *traccia della matrice A* la somma degli autovalori

$$\text{tr}(A) = \sum_{j=1}^n \lambda_j$$

## 2.16.6 Raggio spettrale

**Definizione.** Definiamo *raggio spettrale* della matrice  $A \in \mathbb{C}^{n \times n}$  il massimo modulo degli autovalori

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

**NB.** Attenzione a non permutare l'ordine delle parole *massimo modulo degli autovalori*: iniziare la definizione dicendo autovalori significa dire qualcosa di sbagliato.

## 2.16.7 Teorema sulla matrice convergente

**Teorema.** Una matrice  $A \in \mathbb{C}^{n \times n}$  è una matrice convergente se e solo se

$$\rho(A) < 1$$

## 2.16.8 Autovalori della matrice trasposta

Si osservi dal seguente calcolo che gli autovalori di una matrice e della corrispondente trasposta sono gli stessi

$$\det(A - \lambda I) = \det[(A - \lambda I)^T] = \det(A^T - \lambda I)$$

ovviamente  $I^T = I$ .

## 2.16.9 Teorema sugli autovalori di matrici simili

**Teorema.** Due matrici simili  $A$  e  $B$  hanno gli stessi autovalori. Inoltre, per ogni autovalore  $\lambda$ , se  $x$  è autovettore di  $A$  allora  $S^{-1}x$  è autovettore di  $B$ .

Il teorema si applica alle trasformazioni per similitudine, quindi anche alle matrici di permutazione.

### 1. Stessi autovalori.

Dimostriamolo per mezzo del seguente calcolo. Sappiamo che  $B = S^{-1}AS$

$$\begin{aligned} \det(B - \lambda I) &= \det(S^{-1}AS - \lambda I) = \\ &= \det(S^{-1}AS - \lambda S^{-1}S) = \\ &= \det(S^{-1}(A - \lambda I)S) = \\ &= \det(S^{-1}) \det(A - \lambda I) \det(S) = \\ &= \det(S^{-1}) \det(S) \det(A - \lambda I) = \det(I) \det(A - \lambda I) = \det(A - \lambda I) \end{aligned}$$

Gli zeri del polinomio caratteristico di  $B$  sono gli zeri del polinomio caratt. di  $A$ .

2. **Dato**  $\lambda$ , se  $x$  è autovettore di  $A$  allora  $S^{-1}x$  è autovettore di  $B$ .

Prendiamo  $By = \lambda y$ . Sostituiammo  $B$

$$S^{-1}ASy = \lambda S^{-1}Sy$$

si osservi che abbiamo posto tra  $\lambda$  ed  $y$  il prodotto  $S^{-1}S$ , che è uguale alla matrice identica. Premoltiplicando per  $S$  otteniamo

$$ASy = \lambda Sy$$

cioè  $x = Sy$ , che invertito è  $y = S^{-1}x$ .

### 2.16.10 Teorema sugli autovalori e autovettori di matrici di potenze

**Teorema.** Se  $\lambda$  è autovalore della matrice  $A$  allora

- $\lambda^k, k \in \mathbb{N}$  è autovalore di  $A^k$ , inoltre
- gli autovettori di  $A$  sono anche autovettori di  $A^k$ .

Si osservi che l'ultima cosa detta è C.S: vale quanto detto solo nella direzione detta, in direzione opposta non è detto (ad esempio se  $x$  è autovettore di  $A^k$  non posso dire automatico che  $x$  è autovettore di  $A$ ).

Dimostriamolo attraverso i seguenti calcoli

$$Ax = \lambda x \longrightarrow AAx = A\lambda x \longrightarrow A^2x = \lambda Ax$$

ma  $Ax = \lambda x$  e quindi

$$A^2x = \lambda^2 x$$

dove  $\lambda^2$  è l'autovalore.

#### 2.16.10.1 Esempio sul fatto che abbiamo una C.S.

Prendiamo la seguente matrice, per quanto riguarda gli autovalori la cosa è pacifica

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad A^2 = I \longrightarrow \lambda^2 = 1 \longrightarrow \lambda_1 = 1, \lambda_2 = -1$$

Entrambe le matrici ( $A$  ed  $I$ ) hanno autovalore comune 1. Quanto segue è banalmente verificato

$$Ix = 1 \cdot x = x$$

qualunque vettore non nullo è autovettore della matrice identica. Prendiamo ad esempio  $x = (1//0)$ , verifichiamo se lo è pure in  $A$

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \neq I \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Chiaramente  $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$  non è autovettore della matrice  $A$ .

### 2.16.10.2 Esempio concreto di esercizio

Si consideri la seguente matrice

$$A = \begin{pmatrix} 1 & -2 & 4 \\ 0 & 5 & 0 \\ 0 & 6 & 1 \end{pmatrix}$$

Determinare il polinomio caratteristico della matrice  $A^3$ .

**Risoluzione** Mi ricordo che il polinomio caratteristico può essere descritto nella seguente forma

$$P(\lambda) = (-1)^3(\lambda - \lambda_1)(\lambda - \lambda_2)(\lambda - \lambda_3)$$

Calcoliamo gli autovalori risolvendo  $\det(A - \lambda I) = 0$ . Ricorriamo allo sviluppo di Laplace

$$\det(A - \lambda I) = (1 - \lambda)[(5 - \lambda)(1 - \lambda)] - 2[0] + 4[0] = (1 - \lambda)(5 - \lambda)(1 - \lambda)$$

Gli autovalori della matrice  $A$  sono  $\alpha_1 = \alpha_2 = 1, \alpha_3 = 5$ . Troviamo gli autovalori di  $A^3$  elevando a 3.

$$\lambda_1 = \lambda_2 = 1^3 \quad \lambda_3 = 5^3 = 125$$

Segue

$$P(\lambda) = -1(\lambda - 1)^2(\lambda - 125)$$

### 2.16.11 Teorema sugli autovalori di matrici inverse

**Teorema.** Il teorema precedente, se la matrice  $A$  è non singolare (quindi autovalori  $\neq 0$ ), può essere esteso a  $k \in \mathbb{Z}$ . In sostanza affermiamo che gli autovalori della matrice inversa  $A^{-1}$  sono i reciproci degli autovalori di  $A$ .

La cosa riguarda gli autovalori delle matrici inverse. Prendiamo

$$Ax = \lambda x$$

premultiplichiamo per  $A^{-1}$ , ottenendo

$$A^{-1}Ax = \lambda A^{-1}x$$

ponendo  $A^{-1}A = I$  e dividendo entrambi i membri per  $\lambda$  otteniamo

$$A^{-1}x = \frac{1}{\lambda}x$$

dove l'autovalore è  $\frac{1}{\lambda}$ .

## 2.16.12 Teorema sugli autovalori di matrici hermitiane

**Teorema.** Gli autovalori di una matrice hermitiana sono tutti reali.

Dire che la matrice è hermitiana significa affermare che  $A = A^H$ , cioè la matrice è uguale alla sua trasposta coniugata. Abbiamo detto precedentemente che il prodotto  $x^H Ax \in \mathbb{R}$ . Premoltiplico per  $x^H$

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

il prodotto  $x^H x$  è il seguente

$$\begin{pmatrix} \bar{x}_1 & \dots & \bar{x}_n \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \bar{x}_1 x_1 + \dots + \bar{x}_n x_n = \sum_{i=1}^n |x_i|^2 > 0$$

La sommatoria può essere nulla solo se tutti gli addendi sono nulli. Riprendiamo l'uguaglianza  $x^H Ax = \lambda x^H x$  e dividiamo per  $x^H x$  (posso farlo perché so che è un numero, ed è un numero  $\neq 0$ ). Otteniamo il cosiddetto *quoziente di Rayleigh*:

$$\lambda = \frac{x^H Ax}{x^H x}$$

Se  $x^H Ax \in \mathbb{R}$  a priori e  $x^H x \in \mathbb{R}$  in quanto appena dimostrato allora  $\lambda \in \mathbb{R}$ . Gli autovalori sono tutti reali.

**Utilità del quoziente di Rayleigh** Il quoziente di Rayleigh è utile perchè ci permette di calcolare l'autovalore  $\lambda$  a partire dall'autovettore  $x$ , data la matrice  $A$ .

$$\lambda = \frac{x^H Ax}{x^H x}$$

## 2.16.13 Molteplicità algebrica

**Definizione.** La *molteplicità algebrica*  $\alpha(\lambda)$  di un autovalore  $\lambda$  è la molteplicità di  $\lambda$  come zero del polinomio caratteristico.

Esempio. Se un polinomio caratteristico ha come zeri tre valori uguali ad 1 allora  $\alpha(1) = 3$ .

## 2.16.14 Molteplicità geometrica

**Definizione.** La *molteplicità geometrica*  $\gamma(\lambda)$  di un autovalore  $\lambda$  è la dimensione dello spazio delle soluzioni del sistema lineare omogeneo

$$(A - \lambda I)x = 0$$

In altre parole è la dimensione dell'autospazio relativo a  $\lambda$ , cioè gli autovettori tali per cui  $Ax = \lambda x$ .

Ricordarsi che con *dimensione di un qualunque spazio vettoriale* intendiamo la cardinalità della base. Si può calcolare la molteplicità geometrica per mezzo della seguente formula

$$\gamma(\lambda) = n - r(A - \lambda I)$$

## 2.16.15 Teorema su legame tra molteplicità algebrica e geometrica

**Teorema.** Per ogni autovalore  $\lambda$  risulta

$$1 \leq \gamma(\lambda) \leq \alpha(\lambda) \leq n$$

- **Molteplicità geometrica.**

Si osservi che  $\gamma(\lambda) \geq 1$ , quindi per ogni autovalore esiste almeno un autovettore associato.

- **Molteplicità algebrica.**

Si osservi che la molteplicità geometrica non può superare la molteplicità algebrica:  $\gamma(\lambda) \leq \alpha(\lambda)$ . Non posso avere, ad esempio, un autovalore che compare 3 volte e 4 autovettori linearmente indipendenti associati.

- **Valore massimo delle due molteplicità.**

Molteplicità geometrica e molteplicità algebrica non superano la dimensione della matrice.

Si prenda ad esempio la seguente matrice

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \implies A - \lambda I = \begin{pmatrix} 1 - \lambda & 0 \\ 0 & 2 - \lambda \end{pmatrix} \implies \det(A - \lambda I) = (1 - \lambda)(2 - \lambda)$$

poniamo  $p(\lambda) = 0$

$$(1 - \lambda)(2 - \lambda) = 0$$

da cui  $\lambda_1 = 1$  e  $\lambda_2 = 2$ . Abbiamo  $\alpha(\lambda_1) = 1$  e  $\alpha(\lambda_2) = 1$ . Per il precedente teorema siamo sicuri che  $\gamma(\lambda_1) = 1$  e  $\gamma(\lambda_2) = 1$ . Si prenda adesso la seguente matrice

$$A = \begin{pmatrix} 5 & 1 & 1 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{pmatrix} \implies A - \lambda I = \begin{pmatrix} 5 - \lambda & 1 & 1 \\ 0 & 5 - \lambda & 0 \\ 0 & 0 & 1 - \lambda \end{pmatrix} \implies \det(A - \lambda I) = (5 - \lambda)(5 - \lambda)(1 - \lambda)$$

Dove abbiamo come autovalori  $\lambda_1 = 5$  e  $\lambda_2 = 1$ . Le molteplicità algebriche sono  $\alpha(\lambda_1) = 2$  e  $\alpha(\lambda_2) = 1$ . Per quanto riguarda le molteplicità geometriche possiamo già dire che  $\gamma(\lambda_2) = 1$  mentre  $\gamma(\lambda_1)$  dobbiamo calcolarla

$$\gamma(\lambda_1) = 3 - r(A - \lambda_1 I) = 3 - r \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & -4 \end{pmatrix} = 3 - 2 = 1$$

Troviamo gli autovettori relativi a  $\lambda_1 = 5$

$$(A - 5I)x = 0 \implies \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & -4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \implies \begin{cases} x_2 + x_3 = 0 \\ -4x_3 = 0 \end{cases}$$

Poniamo un parametro arbitrario  $x_1 = t$

$$x = \begin{pmatrix} t \\ 0 \\ 0 \end{pmatrix} = t \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

### 2.16.16 Matrice diagonalizzabile

**Definizione.** Una matrice  $A$  si dice diagonalizzabile se esiste una matrice  $X$  non singolare tale che

$$X^{-1}AX = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \lambda_n \end{pmatrix}$$

Una matrice è diagonalizzabile se per ogni autovalore  $\lambda$  possiamo dire  $\alpha(\lambda) = \gamma(\lambda)$ .

Si prenda ad esempio una matrice  $A$  con autovalori a due a due distinti: sicuramente la matrice è diagonalizzabile poichè

$$\alpha(\lambda) = \gamma(\lambda) = 1$$

### 2.16.17 Spettro

**Definizione.** Data una matrice  $A \in \mathbb{C}^{n \times n}$  definiamo *spettro della matrice A* l'insieme degli autovalori  $\lambda$ .

### 2.16.18 Teorema sulla traslazione dello spettro

**Teorema.** Sia  $\lambda$  autovalore di  $A$ . Sia  $q \in \mathbb{C}$ . Possiamo dire che la matrice  $B$

$$B = A + qI$$

ha autovalore  $\mu = \lambda + q$  con molteplicità algebrica e geometrica pari a quelle di  $\lambda$ . Gli autovettori di  $B$  sono gli stessi di  $A$ .

- **Sugli autovalori traslati.**

Prendiamo il polinomio caratteristico della matrice  $B$

$$\det(B - \mu I) = \det(A + qI - \mu I) = \det(A - (\mu - q)I)$$

Abbiamo ottenuto  $\lambda = \mu - q \rightarrow \mu = \lambda + q$ .

- **Sugli autovettori invariati.**

Partiamo dalla relazione soddisfatta dagli autovalori della matrice  $B$

$$Bx = \mu x$$

e sostituiamo  $B = A + qI$

$$(A + qI)x = \mu x \rightarrow Ax + qIx = \mu x \rightarrow Ax = (\mu - q)x \rightarrow Ax = \lambda x$$

## 2.17 Localizzazione degli autovalori

Con gli strumenti che introdurremo a breve vogliamo ricercare gli autovalori in una regione limitata di un piano (ovviamente piano complesso).

### 2.17.1 Cerchi di Gershgorin

**Definizione.** Data una matrice  $A \in \mathbb{C}^{n \times n}$  definiamo gli insiemi

$$F_i = \{z \in \mathbb{C} : |z - a_{ii}| \leq \rho_i\} \text{ dove } \rho_i = \sum_{j=1, j \neq i}^n |a_{ij}|$$

con  $i = 1, \dots, n$ , detti *cerchi di Gershgorin*. Insiemi di numeri complessi che hanno, nel piano complesso, una distanza minore o uguale a  $\rho_i$  (raggio del cerchio) rispetto a un punto fisso  $a_{ii}$  (centro del cerchio).

### 2.17.2 Primo teorema di Gershgorin

**Teorema.** Se  $\lambda$  è autovalore di  $A \in \mathbb{C}^{n \times n}$  risulta

$$\lambda \in F = \bigcup_{i=1}^n F_i$$

cioè l'autovalore appartiene all'insieme  $F$  costituito dall'unione di tutti i cerchi di Gershgorin.

**Appartenenza dell'autovalore a un cerchio  $k$ -esimo** Abbiamo detto che  $\lambda$  è autovalore, quindi possiamo dire

$$Ax = \lambda x$$

Prendiamo l'autovettore  $x$  e tra le possibili componenti prendiamo quella  $k$ -esima di modulo massimo.

$$x = (x_1 \ x_2 \ \dots \ x_k \ \dots \ x_n)^T \quad |x_k| \geq |x_i|, i = 1, \dots, n$$

Concentriamoci adesso sulla  $k$ -esima equazione del sistema lineare

$$a_{k1}x_1 + a_{k2}x_2 + \dots + a_{kk}x_k + \dots + a_{kn}x_n = \lambda x_k$$

prendiamo il termine  $x_k$  e portiamolo a seconda membra, e scriviamo il resto della sommatoria in modo compatto

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

Poniamo i moduli dei due membri

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

Per mezzo della diseguaglianza triangolare  $|x + y| < |x| + |y|$  poniamo

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

Sapendo che

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

e considerando la disegualanza precedentemente trovata otteniamo

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

abbiamo ottenuto in entrambi membri un termine comune  $|x_k|$  che sappiamo essere con certezza diverso da zero (un autovettore ha sicuramente almeno una componente diversa da zero, di certo il massimo modulo non è zero). Dividiamo per  $|x_k|$  arrivando alla conclusione

$$\sum_{j=1, j \neq k}^n |a_{kj}| \geq |\lambda - a_{kk}| \longrightarrow |\lambda - a_{kk}| \leq \sum_{j=1, j \neq k}^n |a_{kj}| \longrightarrow |\lambda - a_{kk}| \leq \rho_k$$

che è letteralmente la condizione rispettata dai numeri  $z \in \mathbb{C}$  che appartengono al cerchio di Gershgorin  $k$ -esimo.

**Perchè nel teorema si parla di unione di cerchi?** Nel teorema si afferma che un autovalore  $\lambda$  appartiene a un insieme che consiste nell'unione di tutti i cerchi in quanto non è possibile stabilire a priori a quale cerchio  $k$ -esimo appartiene l'autovalore (noi parliamo di autovalore  $\lambda$  e matrice  $A$ , ma non conosciamo l'autovettore  $x$ ).

### 2.17.2.1 Corollario sulla matrice non singolare

**Corollario.** Una matrice a predominanza diagonale forte è una matrice non singolare.

Dal primo teorema di Gershgorin sappiamo che ogni cerchio ha centro a distanza  $|a_{ii}|$  dall'origine degli assi e raggio  $\rho_i = \sum_{i=1, j \neq i}^n |a_{ij}|$ .

- Se la matrice è a predominanza diagonale forte siamo certi che per ogni cerchio  $i$ -esimo il raggio  $\rho_i$  non sarà maggiore del centro  $a_{ii}$ , quindi nessun cerchio conterrà l'origine.
- Per il primo teorema di Gershgorin possiamo affermare che lo zero non potrà essere autovalore.
- Ricordiamo la scomposizione del polinomio caratteristico, dove

$$\sigma_n = \det(A) = \prod_{i=1}^n \lambda_i$$

se nessun autovalore può essere nullo allora non avremo sicuramente  $\det(A) = 0$ : la matrice è non singolare.

### 2.17.3 Secondo teorema di Gershgorin

**Teorema.** Se  $M_1$  è l'unione di  $k$  cerchi di Gershgorin e  $M_2$  è l'unione dei rimanenti  $n - k$  cerchi, con  $M_1 \cap M_2 = \emptyset$ , allora possiamo dire che

- $k$  autovalori appartengono ad  $M_1$
- $n - k$  autovalori appartengono ad  $M_2$

### 2.17.3.1 Corollario sui cerchi di Gershgorin

**Corollario.** Se una matrice presenta cerchi di Gershgorin a due a due disgiunti allora gli autovalori sono a due a due distinti.

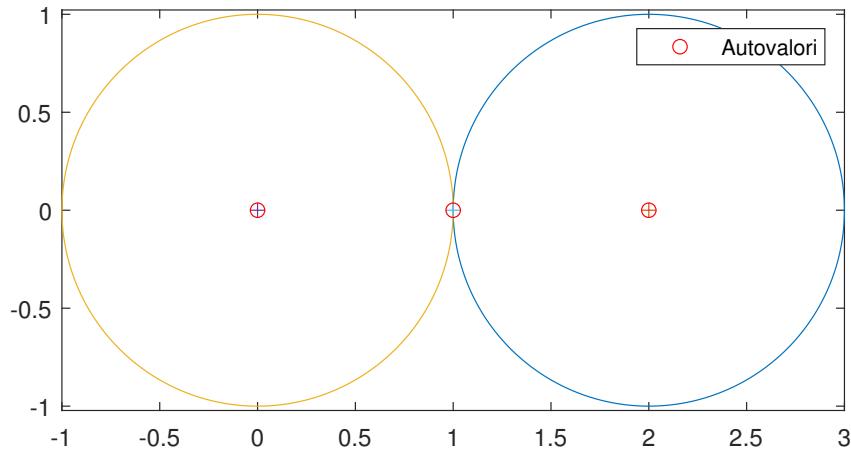
### 2.17.4 Terzo teorema di Gershgorin

**Teorema.** Se la matrice  $A$  è irriducibile allora possiamo dire che se un autovalore appartiene alla frontiera dell'unione dei cerchi di Gershgorin allora l'autovalore  $\lambda$  appartiene alla frontiera di tutti i cerchi costituenti l'insieme  $F$ .

Si consideri come esempio la seguente matrice

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

i cerchi di Gershgorin sono i seguenti



Si osservi che l'autovalore  $\lambda = 1$  sta sulla frontiera dell'unione dei due cerchi, ma affinchè il terzo teorema sia rispettato è necessario che stia anche sulla frontiera dei singoli cerchi.

### 2.17.4.1 Corollario sulla matrice non singolare

**Corollario.** Una matrice a predominanza diagonale debole ed irriducibile è una matrice non singolare.

La predominanza diagonale debole rende possibile per i cerchi avere la circonferenza passante per l'origine (ovviamente la cosa vale per tutti i cerchi tranne uno).

- In queste condizioni lo zero appartiene per forza alla frontiera del cerchio.
- Ma siamo certi che in uno dei cerchi non avremo lo zero tra i possibili autovalori.
- Segue contraddizione col terzo teorema di Gershgorin. La matrice è non singolare, non è possibile avere come autovalore  $\lambda = 0$ .

## 2.17.5 Matrice di Frobenius

**Definizione.** Sia data l'equazione algebrica

$$x^k + a_{k-1}x^{k-1} + \cdots + a_1x + a_0 = 0$$

definiamo *matrice di Frobenius* (o *matrice associata*, o *matrice compagna*) la seguente matrice

$$F = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ -a_0 & -a_1 & -a_2 & \dots & -a_{k-2} & -a_{k-1} \end{pmatrix}$$

Essa è una matrice quadrata di ordine  $k$  dove:

- la codiagonale superiore ha elementi uguali ad 1
- l'ultima riga della matrice presenta i coefficienti dell'equazione, cambiati di segno;
- tutti gli altri elementi sono uguali a zero.

**Forma alternativa** La matrice di Frobenius può essere posta anche nella seguente forma

$$F = \begin{pmatrix} -a_{k-1} & -a_{k-2} & -a_{k-3} & \dots & -a_1 & -a_0 \\ -1 & 0 & 0 & \dots & 0 & 0 \\ 0 & -1 & 0 & \dots & 0 & 0 \\ 0 & 0 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 0 \end{pmatrix}$$

**Utilità** La cosa ci interessa per una ragione fondamentale

L'equazione caratteristica della matrice di Frobenius è l'equazione algebrica!

Questo significa che gli autovalori della matrice  $F$  sono gli zeri dell'equazione algebrica. La matrice di Frobenius trova ampia applicazione nei calcolatori: se dobbiamo risolvere un'equazione algebrica al 99% dei casi la macchina si ricaverà la corrispondente Matrice di Frobenius e individuerà gli autovalori con dei metodi di approssimazione (anticipazione!!).

### 2.17.5.1 Esempio

Si consideri la seguente equazione algebrica

$$2x^3 - 10x^2 + 6 - 2 = 0$$

**Osservazione** L'equazione di partenza è caratterizzata da coefficienti reali. Possiamo dire che

Un'equazione a coefficienti reali ha soluzioni reali o coppie di complessi coniugati.

In presenza di un'equazione di grado dispari avremo almeno una soluzione reale.

**Ricavare la matrice di Frobenius** Riconduciamola al formato detto dividendo per 2 (attenti al coefficiente di  $x^3$ , deve essere 1 per forza)

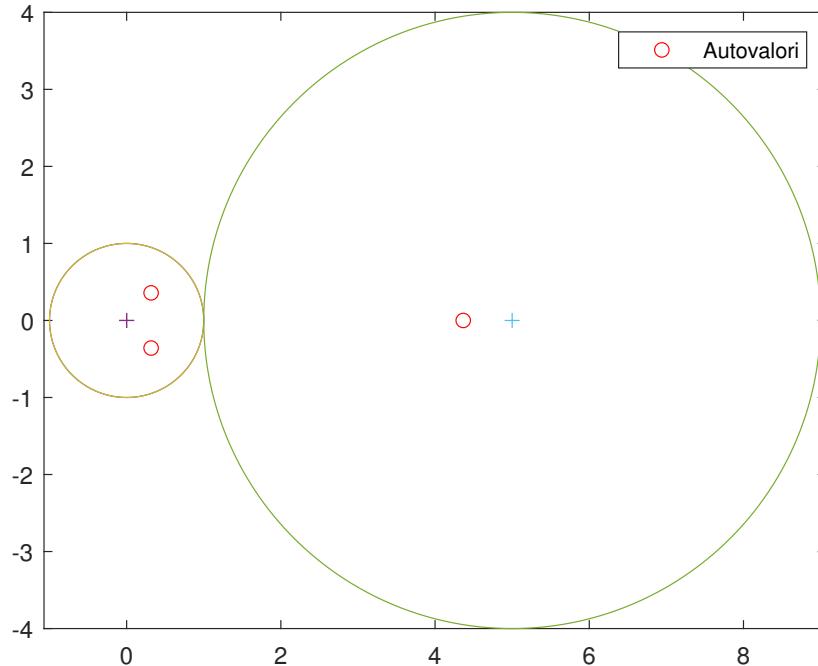
$$x^3 - 5x^2 + 3x - 1 = 0$$

a questo punto ricaviamo la matrice di Frobenius

$$F = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & -3 & 5 \end{pmatrix}$$

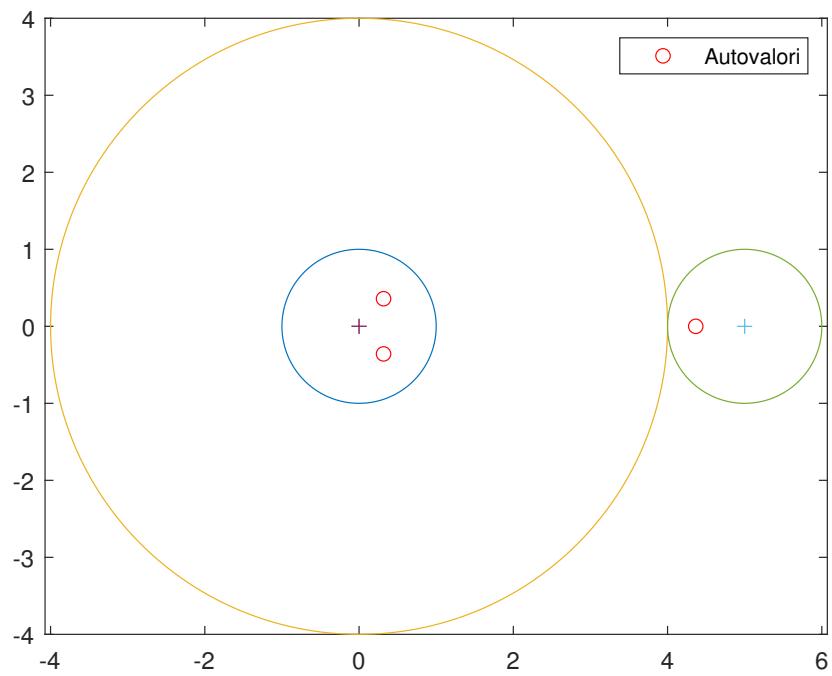
**Applicazione del primo teorema di Gershgorin** Applichiamo Gershgorin (si osservi che in ogni matrice di Frobenius i primi  $n - 1$  cerchi hanno sempre centro 0 e raggio 1)

- Il cerchio 1 ha centro 0 e raggio 1:  $F_1 = \{z \in \mathbb{C} : |z| \leq 1\}$
- Il cerchio 2 ha centro 0 e raggio 1:  $F_2 = \{z \in \mathbb{C} : |z| \leq 1\}$
- Il cerchio 3 ha centro 5 e raggio 4:  $F_3 = \{z \in \mathbb{C} : |z - 5| \leq 4\}$



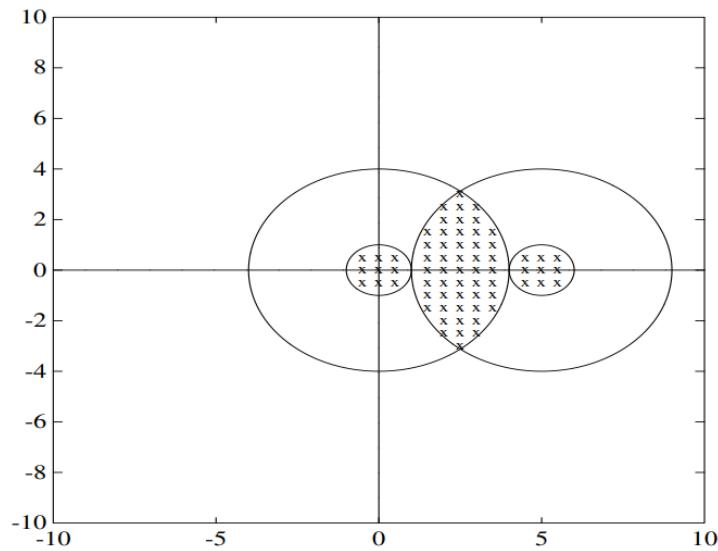
**Riduzione del piano complesso** Chiaramente è nostro interesse ridurre la dimensione del piano complesso, in modo da restringere le scelte possibili. Quello che si può fare è prendere un'altra matrice avente gli stessi autovalori: la trasposta  $F^T$ . Individuare i cerchi di Gershgorin sulla matrice trasposta equivale all'applicazione del primo teorema su  $F$  muovendosi per colonne (e non per righe, come abbiamo fatto prima). Otteniamo

- Il cerchio 1 ha centro 0 e raggio 1:  $G_1 = \{z \in \mathbb{C} : |z| \leq 1\}$
- Il cerchio 2 ha centro 0 e raggio 4:  $G_2 = \{z \in \mathbb{C} : |z| \leq 4\}$
- Il cerchio 3 ha centro 5 e raggio 1:  $G_3 = \{z \in \mathbb{C} : |z - 5| \leq 1\}$



A questo punto intersechiamo le due regioni ottenendo la regione ristretta che volevamo trovare

$$F \cap G$$



## 2.18 Norme vettoriali e norme matriciali

Abbiamo bisogno di strumenti che ci permettano di fare dei confronti tra matrici e vettori.

### 2.18.1 Norma vettoriale

**Definizione.** Si definisce *norma vettoriale* una funzione del tipo

$$\|\cdot\| : \mathbb{C}^n \rightarrow \mathbb{R}_0^+$$

che verifica le seguenti condizioni:

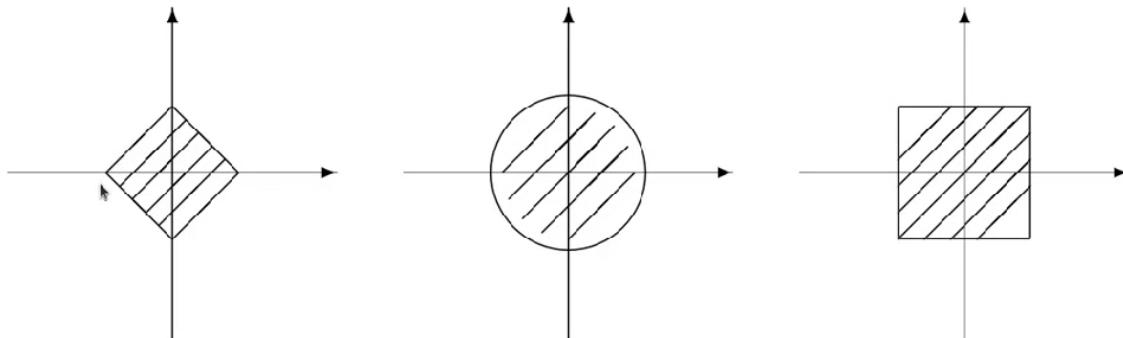
- $\|x\| = 0 \iff x = 0$
- $\|\alpha x\| = |\alpha| \|x\|, \forall x \in \mathbb{C}^n, \forall \alpha \in \mathbb{C}$
- $\|x + y\| \leq \|x\| + \|y\|, \forall x, y \in \mathbb{C}^n$

### 2.18.2 Norme classiche

**Definizione.** Nella definizione di *norma vettoriale* non indichiamo espressamente la funzione. Questo perché la norma può essere definita in modo arbitrario, distinguendo le norme possibili per mezzo di pedici. A noi interessano le cosiddette *norme classiche*:

$$\begin{aligned}\|x\|_1 &= \sum_{i=1}^n |x_i| \\ \|x\|_2 &= \sqrt{\sum_{i=1}^n |x_i|^2} \quad (\text{detta } \textit{norma euclidea}, \text{ quella a cui siamo abituati}) \\ \|x\|_\infty &= \max_{1 \leq i \leq n} |x_i|\end{aligned}$$

Per avere meglio un'idea sulle differenze si consideri l'insieme della *sfera unitaria* in  $\mathbb{R}^2$ . Otteniamo le seguenti rappresentazioni grafiche:



### 2.18.3 Teorema di equivalenza fra nome

**Teorema.** Date due norme vettoriali  $\|x\|_p$  e  $\|x\|_q$  esistono  $\alpha, \beta \in \mathbb{R}^+$  tali che

$$\alpha\|x\|_p \leq \|x\|_q \leq \beta\|x\|_p, \forall x \in \mathbb{C}^n$$

Il teorema ci è utile per studiare una successione di vettori, che converge al vettore nullo: la norma tende a zero, la cosa è valida con qualunque tipologia di norma? Il teorema ci permette di rispondere positivamente: se la norma  $p$  tende a zero allora anche la norma  $q$  tende a zero per il teorema dei carabinieri

$$\alpha\|x\|_p \leq \|x\|_q \leq \beta\|x\|_p, \forall x \in \mathbb{C}^n$$

### 2.18.4 Norma matriciale

**Definizione.** Si definisce *norma vettoriale* una funzione del tipo

$$\|\cdot\| : \mathbb{C}^{n \times n} \rightarrow \mathbb{R}_0^+$$

che verifica le seguenti condizioni:

- $\|A\| = 0 \longleftrightarrow A = O$
- $\|\alpha A\| = |\alpha| \|A\|, \forall a \in \mathbb{C}^{n \times n}, \forall \alpha \in \mathbb{C}$
- $\|A + B\| \leq \|A\| + \|B\|, \forall A, B \in \mathbb{C}^{n \times n}$
- $\|AB\| \leq \|A\| \|B\|, \forall A, B \in \mathbb{C}^{n \times n}$

Le prime tre condizioni sono analoghe, la quarta è la novità.

### 2.18.5 Norma matriciale indotta (o naturale)

**Definizione.** Si dice *norma matriciale indotta* (o *naturale*) la norma matriciale ottenuta partendo dalle norme vettoriali

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

abbiamo il rapporto tra due norme vettoriali. Si dice che la norma è *indotta dalla norma vettoriale*.

Ricordarsi che l'estremo superiore è il più piccolo dei possibili maggioranti (il maggiorante è un valore  $\alpha$  tale che  $\alpha \geq f(x), \forall x$ )

### 2.18.5.1 Esempi

Ci interessano le seguenti *norme matriciali indotte dalle tre norme vettoriali classiche*:

- $\|A\|_1 = \max_j \sum_{i=1}^n |a_{ij}|$  (somme dei moduli di ogni colonna, si prende la più grande)
- $\|A\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|$  (equivalente della precedente ma sulle righe)
- $\|A\|_2 = \sqrt{\rho(A^H A)}$  (detta *norma euclidea*, radice quadrata del raggio spettrale)

### 2.18.6 Norme coerenti (o compatibili)

**Definizione.** Una norma matriciale si dice *coerente* (o *compatibile*) se

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

attenzione, non stiamo ribadendo la quarta proprietà delle norme matriciali (al secondo membro abbiamo il prodotto tra una norma matriciale e una norma vettoriale).

Si osservi che le norme indotte sono coerenti con le rispettive norme vettoriali.

$$\frac{\|Ax\|}{\|x\|} \leq \frac{\|A\| \|x\|}{\|x\|} = \|A\|$$

I calcoli non ci meravigliano visto che  $\|A\|$  è maggiorante.

### 2.18.7 Norma di Frobenius

**Definizione.** La *norma di Frobenius* è una norma matriciale non indotta da norme vettoriali. Essa è definita come segue

$$\|A\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^n |a_{ij}|^2}$$

Non può esistere una norma vettoriale che da la norma di Frobenius come norma indotta. La cosa si verifica agile col caso della matrice identica. Si consideri che con la norma di Frobenius otteniamo

$$\|I\|_F = \sqrt{n}$$

mentre si ha con qualunque norma vettoriale  $\|x\|$ :

$$\|I\| = \sup_{x \neq 0} \frac{\|Ix\|}{\|x\|} = \frac{\|x\|}{\|x\|} = 1$$

## 2.18.8 Teorema di Hirsh (legame tra raggio spettrale e norma matriciale)

**Teorema.** Sia  $A \in \mathbb{C}^{n \times n}$ , per ogni norma matriciale (indotta o no) vale la seguente relazione

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

il raggio spettrale della matrice è inferiore o uguale della norma matriciale.

Dimostriamolo. Partiamo dal sistema che definisce autovalori e autovettori

$$Ax = \lambda x$$

Siamo certi che  $x \neq 0$  visto che è autovettore. Definiamo una matrice  $\mathbb{C}^{n \times n}$

$$B = (x|0|0|\dots|0)$$

Questa matrice ci permette di porre  $AB = \lambda B$ . Calcoliamo la norma di primo e secondo membro:  $\|AB\| = \|\lambda B\|$ . Con la quarta proprietà affermiamo

$$\|AB\| = |\lambda| \|B\| \leq \|A\| \|B\|$$

sicuramente  $\|B\| \neq 0$  poichè la matrice  $B$  è costituita dall'autovettore  $x$ : si ha un numero reale e positivo, quindi possiamo dividere senza cambiare segno. Otteniamo

$$|\lambda| \leq \|A\| \longrightarrow \|A\| \geq |\lambda|$$

Abbiamo dimostrato che la norma matriciale è maggiore o uguale di qualunque modulo degli autovalori, quindi anche di quello che è il massimo tra i moduli degli autovalori. Segue che la norma matriciale è maggiore o uguale del raggio spettrale della matrice

$$\|A\| \geq \rho(A)$$

### 2.18.8.1 Corollario sulla convergenza di una matrice

**Corollario.** Condizione sufficiente affinchè una matrice sia convergente è che una sua norma risulti minore di 1.

Questo deriva dal teorema sulla matrice convergente, dove abbiamo affermato che una matrice quadrata di ordine  $n$  è convergente se e solo se  $\rho(A) < 1$ .

### 2.18.8.2 Corollario sui cerchi

**Corollario.** Se  $A \in \mathbb{C}^{n \times n}$  allora gli autovalori di  $A$  appartengono al cerchio

$$\{z \in \mathbb{C} : |z| \leq \|A\|\}$$

dove  $\|\cdot\|$  è una qualunque norma matriciale.

Ricordarsi che il modulo di un numero complesso è la distanza dall'origine, quindi gli autovalori distano dall'origine al più quanto la norma matriciale  $\|A\|$ .

### 2.18.8.3 Raggio spettrale nelle matrici hermitiane

Nelle matrici hermitiane abbiamo  $A = A^H$ , ricordarsi la norma euclidea matriciale: otteniamo

$$\|A\|_2 = \sqrt{\rho(A^H A)} = \sqrt{\rho(A^2)} = \sqrt{\rho^2(A)} = \rho(A)$$

Ricordarsi che gli autovalori di  $A^2$  sono i quadrati degli autovalori di  $A$ , quindi si considera il quadrato del raggio spettrale di  $A$ .

## 2.18.9 Matrice di rotazione

**Definizione.** Le matrici della forma

$$G_{rt} = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & c & -s \\ & & & & 1 \\ & & & & & \ddots \\ & & & s & c & 1 \\ & & & & & & \ddots \\ & & & & & & & 1 \end{pmatrix} \in \mathbb{R}^{n \times n}$$

con  $c^2 + s^2 = 1$ , si dicono *matrici di rotazione*. Con  $r$  e  $t$  indico, rispettivamente, l'indice di riga (di colonna) in cui colloco  $c$  e  $-s$  (in cui colloco  $c$  ed  $s$ ) e l'indica di riga (di colonna) in cui colloro  $s$  e  $c$  (in cui colloco  $-s$  e  $c$ ).

La condizione  $c^2 + s^2 = 1$  può essere interpretato come  $\cos^2 \alpha + \sin^2 \alpha = 1$ , dove  $\alpha$  è detto *angolo di rotazione*. Si ruota il sistema di riferimento dello spazio  $\mathbb{R}^n$ .

**Matrice riducibile** La matrice  $G_{rt}$  è una matrice riducibile. Poniamo la matrice di permutazione nel seguente modo

$$P = (e^r | e^t | e^1 | \dots | e^n)$$

il risultato è la matrice  $B$

$$B = P^T G_{rt} P = \begin{pmatrix} c & -s & & & \\ s & c & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{pmatrix}$$

questa trasformazione per similitudine ha la proprietà di conservare gli stessi autovalori (questo perchè la matrice di permutazione è una matrice *ortogonale*). La matrice ottenuta è una triangolare a blocchi (addirittura una diagonale a blocchi).

**Autovalori** Abbiamo:

- $n - 2$  autovalori uguali ad 1 relativi al blocco  $B_{22}$

$$\lambda_1 = \lambda_2 = \dots = \lambda_{n-2} = 1$$

- 2 autovalori relativi al blocco  $B_{11}$

$$\lambda^2 - 2c\lambda + c^2 + s^2 = 0 \implies \lambda^2 - 2c\lambda + 1 = 0$$

otteniamo

$$\lambda_{n-1,n} = c \pm \sqrt{c^2 - 1} = c \pm \sqrt{-s^2} = c \pm i\sqrt{s^2}$$

la formula del quadrato del modulo risultante è  $c^2 + s^2$ , che sappiamo essere uguale ad 1 per ipotesi. In conclusione

$$\sqrt{c^2 + s^2} = 1$$

Il modulo degli autovalori è uguale ad 1. I due autovalori sono complessi coniugati di modulo 1.

In conclusione il determinante di una matrice di rotazione è uguale ad 1. Applicando Hirsh otteniamo

$$|\lambda_i| \leq \|G_{rt}\|_2 = \sqrt{\rho(G_{rt}G_{rt}^T)} = 1$$

**Convergente?** No, il raggio spettrale non è strettamente minore di 1.

# Capitolo 3

## Sistemi lineari

### 3.1 Sistemi di equazioni lineari

Ribadiamo cose già viste nel capitolo sulle nozioni di Algebra lineare. Lavoreremo su sistemi lineari del tipo

$$Ax = b$$

dove

- $A$  è la *matrice dei coefficienti* (ci interesseranno soprattutto matrici quadrate di ordine  $n$ , non singolari),
- $b$  è il *vettore dei termini noti*, e
- $x$  è il *vettore delle incognite*

Scriviamo il sistema in notazione non compatta

$$\begin{cases} a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\ \vdots \\ a_{n1}x_1 + \cdots + a_{nn}x_n = b_n \end{cases}$$

Qua diventano evidenti i coefficienti  $a_{ij}$  e i termini noti  $b_i$ . Risolvere un sistema lineare significa individuare un vettore  $x^T = (x_1 \ x_2 \ \dots \ x_n)$  che verifichi tutte le equazioni

### 3.2 Classificazione dei metodi di risoluzione

I metodi di risoluzione di sistemi lineari si classificano in due categorie.

- **Metodi diretti.**

Metodi che con un insieme finito di operazioni sui dati conducono alla soluzione esatta, salvo approssimazioni (quelle viste nel capitolo sulla teoria degli errori).

- **Metodi iterativi.**

Metodi dove abbiamo l'approssimazione della soluzione per mezzo del limite di una successione di vettori. Il numero di operazioni, contrariamente ai metodi diretti, non è determinabile. In alcuni casi questi metodi non funzionano: pensiamo a successioni non convergenti!

### 3.3 Metodi diretti

#### 3.3.1 Metodo di Cramer

Nel capitolo sulle nozioni di Algebra lineare abbiamo introdotto il *metodo di Cramer*.

##### 3.3.1.1 Spiegazione

Dato il sistema lineare  $Ax = b$  la soluzione si ottiene calcolando

$$x_i = \frac{\det(A_i)}{\det(A)}, \quad i = 1, 2, \dots, n$$

dove  $A_i$  è la matrice ottenuta da  $A$  sostituendo la  $i$ -esima colonna col vettore dei termini noti  $b$ . La matrice  $A$  è non singolare.

##### 3.3.1.2 Costo computazionale

Il metodo di Cramer è utile per le dimostrazioni, ma dal punto di vista pratico è inapplicabile. Si vada alla sezione del metodo di Gauss sul costo computazionale per il confronto e la conclusione.

#### 3.3.2 Metodo di Gauss (o metodo di eliminazione)

##### 3.3.2.1 Spiegazione introduttiva

Dato un sistema lineare  $Ax = b$  il metodo di Gauss ne prevede la trasformazione in un sistema equivalente

$$Rx = c$$

dove  $R \in \mathbb{C}^{n \times n}$  è una matrice triangolare superiore. Poichè la matrice  $A$  è non singolare allora lo sarà pure la matrice  $R$ : segue che  $r_{ii} \neq 0$  (ricordarsi come si ottiene il determinante di una matrice triangolare).

**Precisazione** Non stiamo attuando una trasformazione per similitudine: segue che la matrice  $R$  avrà autovalori diversi rispetto alla matrice  $A$ .

##### 3.3.2.2 Passi per ottenere il sistema equivalente

Prendiamo un sistema lineare  $Ax = b$  con  $A \in \mathbb{C}^{3 \times 3}$  e applichiamo il metodo di Gauss.

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

**Promemoria** Ricordarsi che la soluzione di un sistema lineare **NON CAMBIA** se si sostituisce a una qualunque equazione  $i$ -esima una sua combinazione lineare con un'altra equazione del sistema.

##### Passi

- Prendiamo la prima colonna da sinistra, relativa all'incognita  $x_1$ : preso il  $a_{11}$ , nella prima equazione, dobbiamo eliminare le incognite dalla seconda equazione in giù, e quindi i relativi coefficienti  $a_{21}$  e  $a_{31}$ .

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ \mathbf{a_{21}} & a_{22} & a_{23} \\ \mathbf{a_{31}} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

Prendo le equazioni successive alla prima, e sottraggo da ognuna la prima equazione, i cui membri sono stati moltiplicati per un moltiplicatore (ovviamente questi calcoli sono possibili soltanto se  $a_{11} \neq 0$ )

$$a_{ij}^2 = a_{ij} - l_{i1}(a_{1j}) \quad l_{i1} = \frac{a_{i1}}{a_{11}}, \quad i = 2, \dots, n$$

Si ha un particolare moltiplicatore per ogni equazione  $i$ -esima. Facciamo i calcoli

$$\begin{aligned} 2. \quad & (a_{21}x_1 + a_{22}x_2 + a_{23}x_3) - \frac{a_{21}}{a_{11}}(a_{11}x_1 + a_{12}x_2 + a_{13}x_3) = \\ & = 0x_1 + \left(a_{22} - \frac{a_{21}}{a_{11}}a_{12}\right)x_2 + \left(a_{23} - \frac{a_{21}}{a_{11}}a_{13}\right)x_3 = a_{22}^2x_2 + a_{23}^2x_3 \\ 3. \quad & (a_{31}x_1 + a_{32}x_2 + a_{33}x_3) - \frac{a_{31}}{a_{11}}(a_{11}x_1 + a_{12}x_2 + a_{13}x_3) = \\ & = 0x_1 + \left(a_{32} - \frac{a_{31}}{a_{11}}a_{12}\right)x_2 + \left(a_{33} - \frac{a_{31}}{a_{11}}a_{13}\right)x_3 = a_{32}^2x_2 + a_{33}^2x_3 \end{aligned}$$

Si alterano anche i termini noti

$$b_2^2 = b_2 - \frac{a_{21}}{a_{11}}b_1 \quad b_3^2 = b_3 - \frac{a_{31}}{a_{11}}b_1$$

Ecco il risultato

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22}^2 & a_{23}^2 \\ 0 & a_{32}^2 & a_{33}^2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2^2 \\ b_3^2 \end{pmatrix}$$

- Ci spostiamo nella seconda colonna, relativa all'incognita  $x_2$ . Prendo il coefficiente  $a_{22}^2$ , nella seconda equazione: dobbiamo eliminare le incognite dalla terza equazione in giù, nella seconda colonna. In questo caso dobbiamo far sparire il coefficiente  $a_{32}$

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22}^2 & a_{23}^2 \\ 0 & \mathbf{a}_{32}^2 & a_{33}^2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2^2 \\ b_3^2 \end{pmatrix}$$

Lavoriamo solo sulla terza equazione. Definiamo il relativo moltiplicatore

$$l_{32} = \frac{a_{32}^2}{a_{22}^2}$$

e sostituiamo la terza equazione con una sua combinazione lineare.

$$(a_{32}^2x_2 + a_{33}^2x_3) - \frac{a_{32}^2}{a_{22}^2}(a_{22}^2x_2 + a_{23}^2x_3) = 0x_2 + \left(a_{33}^2 - \frac{a_{32}^2}{a_{22}^2}a_{23}^2\right)x_3 = a_{33}^3x_3$$

Modifichiamo anche il termine noto

$$b_3^3 = b_3^2 - \frac{a_{32}^2}{a_{22}^2}b_3^3$$

Il risultato è il sistema  $Rx = c$ !

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22}^2 & a_{23}^2 \\ 0 & 0 & a_{33}^2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2^2 \\ b_3^3 \end{pmatrix}$$

**E se avessi un particolare coefficiente  $a_{ii} = 0$ ?** Cosa succede nel caso in cui un coefficiente lungo la diagonale risulti uguale a zero?

- Scorro le equazioni  $k$ -esime (con  $k > i$ ) lungo la colonna  $i$ -esima per individuare un coefficiente  $a_{ki} \neq 0$
- Individuato il coefficiente scambio di posizione la riga  $i$ -esima e la riga  $k$ -esima. Lecitissimo: l'ordine delle equazioni non altera il risultato del sistema.

Fare queste mosse è un'alterazione del metodo di Gauss classico.

### 3.3.2.3 Condizione per esecuzione del metodo di Gauss classico

In quali condizioni posso eseguire il metodo di Gauss classico, senza fare le mosse appena descritte? Riprendiamo il seguente termine con  $i = 2$

$$a_{22}^2 = a_{22} - \frac{a_{21}}{a_{11}}a_{12} \neq 0$$

Lo imponiamo diverso da zero. Considerando il solo numeratore la condizione diventa

$$a_{11}a_{22} - a_{21}a_{12} \neq 0$$

che significa imporre  $\det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \neq 0$ , cioè un minore principale di testa. Quindi:

- devo trovare  $a_{22}^2$ , che utilizzerò al passo successivo;
- sono certo che  $a_{22}^2$  sarà  $\neq 0$  se il minore principale di testa di ordine 2 sarà diverso da zero.

**Generalizzazione** Tutti gli elementi che incontriamo lungo la diagonale devono essere diversi da zero

$$a_{11} \neq 0, a_{22} \neq 0, \dots, a_{nn} \neq 0$$

e questo è possibile solo se la matrice  $A$  ha tutti i minori principali di testa diversi da zero.

$$a_{11} \neq 0 \longrightarrow \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \neq 0 \longrightarrow \dots \longrightarrow \det(A) \neq 0$$

Se non è possibile fare scambi di riga l'algoritmo si interrompe.

**Problema** Non molte matrici godono di questa proprietà: sicuramente rientrano tra queste le matrici simmetriche e quelle definite positive/negative.

### 3.3.2.4 Risoluzione del sistema equivalente

Un sistema triangolare superiore è banalmente risolubile: l'ultima equazione del sistema ha un'unica incognita  $x_n$ . Scriviamo in modo esteso il sistema  $Rx = c$

$$\begin{aligned} r_{11}x_1 + r_{12}x_2 + \dots + r_{1n}x_n &= c_1 \\ r_{22}x_2 + \dots + r_{2n}x_n &= c_2 \\ &\vdots \\ r_{nn}x_n &= c_n \end{aligned}$$

- Siamo nella equazione  $n$ -esima: calcolo  $x_n$  al volo col rapporto  $x_n = \frac{c_n}{r_{nn}}$ . Prendo  $x_n$  trovato e lo sostituisco nelle  $n - 1$  equazioni precedenti.

- Mi sposto nella equazione  $(n - 1)$ -esima (quella in posizione immediatamente precedente) e trovo  $x_{n-1}$ . Prendo  $x_{n-1}$  trovato e lo sostituisco nelle  $n - 2$  equazioni precedenti.
- Continuiamo allo stesso modo finchè non avremo trovato tutte le componenti  $x_i$  del vettore  $x$ .

I calcoli delle componenti  $x_i$ , a parte  $x_n$ , possono essere rappresentati con la seguente formula

$$x_i = \frac{c_i - \sum_{j=i+1}^n r_{ij}x_j}{r_{ii}}, \quad i = n-1, \dots, 1$$

### 3.3.2.5 Costo computazionale

Il metodo di Gauss è sicuramente più efficiente rispetto al metodo di Cramer. Per prima cosa osserviamo il numero di operazioni necessarie a parità di  $n$  equazioni ed  $n$  incognite.

- **Metodo di Cramer.**  $(n + 1)!(n - 1)$  operazioni.
- **Metodo di Gauss**<sup>1</sup>.  $\frac{n^3}{3} + n^2 - \frac{n}{3}$  operazioni.

Supponendo che una qualunque operazione (moltiplicazioni, divisioni, ...) richieda  $10^{-6}$  secondi poniamo a confronto i tempi di esecuzione tra i due metodi.

$n$	Metodo di Cramer	Metodo di Gauss
12	103 minuti	$7.1 \times 10^{-4}$ secondi
12	24 ore	$8.9 \times 10^{-4}$ secondi
15	15 giorni	$1.1 \times 10^{-3}$ secondi
50	$4.9 \times 10^{52}$ anni	$4.4 \times 10^{-2}$ secondi

Non c'è molto da aggiungere.

### 3.3.2.6 Esempio 1

Prendiamo il seguente sistema lineare

$$A = \begin{pmatrix} -2 & 1 & 1 \\ -1 & 3 & 1 \\ 1 & 1 & 2 \end{pmatrix} \quad b = \begin{pmatrix} -2 \\ -3 \\ 2 \end{pmatrix}$$

1. Considero la colonna 1 e l'elemento diagonale  $a_{11} = -1$ . Vogliamo fare in modo che gli elementi inferiori, nella stessa colonna, siano uguali a zero. Calcoliamo i moltiplicatori, uno per la seconda riga e uno per la terza

$$l_{21} = \frac{a_{21}}{a_{11}} = \frac{1}{-1} = -1 \quad l_{31} = \frac{a_{31}}{a_{11}} = \frac{-2}{-1} = 2$$

Calcoliamo la nuova seconda e terza equazione (sostituite con le seguenti combinazioni lineari)

$$\begin{aligned} -x_1 + 3x_2 + x_3 - l_{21}(-2x_1 + x_2 + x_3) &= 0x_1 + 5/2x_2 + 1/2x_3 & -3 - l_{21}(-2) &= -2 \\ x_1 + x_2 + 2x_3 - l_{31}(-2x_1 + x_2 + x_3) &= 0x_1 + 3/2x_2 + 5/2x_3 & 2 - l_{31}(-2) &= 1 \end{aligned}$$

Abbiamo ottenuto

$$A^{(2)} = \begin{pmatrix} -2 & 1 & 1 \\ 0 & 5/2 & 1/2 \\ 0 & 3/2 & 5/2 \end{pmatrix} \quad b^{(2)} = \begin{pmatrix} -2 \\ -2 \\ 1 \end{pmatrix}$$

---

<sup>1</sup>Basta il primo termine, quelli dopo me li dimentico pure io (cit.).

2. Considero la colonna 2 e l'elemento diagonale  $a_{22}^{(2)} = 5/2$ . Vogliamo fare in modo che l'elemento inferiore, nella stessa colonna, sia uguale a zero. L'unico moltiplicatore da considerare è il seguente

$$l_{32} = \frac{a_{32}}{a_{22}} = \frac{3}{5}$$

Calcoliamo la nuova terza equazione

$$3/2x_2 + 5/2x_3 - l_{32}(5/2x_1 + 1/2x_2) = 0x_1 + 0x_2 + 22/10x_3 \quad 1 - l_{32}(1) = 11/5$$

Abbiamo ottenuto il sistema  $Rx = c$

$$R = \begin{pmatrix} -2 & 1 & 1 \\ 0 & 5/2 & 1/2 \\ 0 & 0 & 22/10 \end{pmatrix} \quad c = \begin{pmatrix} -2 \\ -2 \\ 11/5 \end{pmatrix}$$

3. A questo punto possiamo risolvere il sistema calcolando le componenti

$$x_3 = \frac{11/5}{22/10} = 1 \quad x_2 = \frac{-2 - 1/2}{5/2} = -1 \quad x_1 = \frac{-2 - 1 + 1}{-2} = 1$$

$$x^T = (1 \quad -1 \quad 1)$$

### 3.3.2.7 Risoluzione di $k$ sistemi con matrice $A$ comune: matrice inversa

Supponiamo di avere da risolvere  $k$  sistemi lineari, tutti con la stessa matrice  $A$

$$Ax^{(1)} = b^{(1)} \quad Ax^{(2)} = b^{(2)} \quad \dots \quad Ax^{(k)} = b^{(k)}$$

Possiamo risolvere il tutto in un colpo solo (ovviamente il costo computazionale aumenta) svolgendo il seguente prodotto matriciale

$$AX = B$$

dove  $X = (x^{(1)} | x^{(2)} | \dots | x^{(k)})$  e  $B = (b^{(1)} | b^{(2)} | \dots | b^{(k)})$

**Matrice inversa** Supponiamo di avere come matrice  $B$  la seguente

$$B = (e^{(1)} | e^{(2)} | \dots | e^{(n)}) = I$$

il sistema lineare risolto è il seguente

$$AX = I$$

e noi sappiamo che  $X$  è per forza la matrice inversa di  $A$ !

$$X = A^{-1}$$

Per il calcolo si rimanda alla variante di Gauss-Jordan.

### 3.3.3 Variante metodo di Gauss: tecniche di pivoting

I coefficienti  $a_{11}, a_{22}^2, a_{33}^3, \dots, a_{nn}^n$  sono detti *elementi pivotali*.

Introduciamo le cosiddette *tecniche di pivoting* in modo da ridurre la propagazione degli errori. Queste tecniche sono un'alterazione del metodo di Gauss classico. Distinguiamo *pivoting parziale* e *pivoting totale*.

#### 3.3.3.1 pivoting parziale

Data una colonna  $k$ -esima della matrice  $A$ , prevediamo l'individuazione del massimo modulo tra gli elementi presenti nella colonna  $k$ -esima.

$$\max_{k \leq i \leq n} |a_{ik}| = |a_{rk}|$$

supponiamo che l'elemento sia sulla  $r$ -esima riga. Se  $r \neq k$  effettuiamo scambio di riga tra

- la  $k$ -esima equazione, e
- la  $r$ -esima equazione.

La cosa si applica per  $n - 1$  passi ( $n - 1$  colonne). Gli spostamenti di righe equivalgono a matrici  $A$  premoltiplicate per un'opportuna matrice di permutazione. Se andremo a fare  $k$  spostamenti allora avremo  $k$  matrici di permutazione  $P_1, \dots, P_k$ . Il tutto equivale a risolvere il seguente sistema (se avessimo avuto questo sistema fin da subito non sarebbe stato necessario il pivoting parziale)

$$(P_k P_{k-1} \dots P_2 P_1) A x = (P_k P_{k-1} \dots P_2 P_1) b \longrightarrow PAx = Pb$$

Ricordarsi che il prodotto di  $k$  matrici di permutazione restituisce una nuova matrice di permutazione  $P$ . Per comprendere meglio l'utilità del pivoting parziale prendiamo il seguente esempio, dove alcuni elementi hanno una perturbazione  $\epsilon$  dovuta a calcoli precedenti che non ci interessano.

$$A = \begin{pmatrix} 1 & -1 + \epsilon & 5 \\ 10 & 2 & 4 \\ 1000 & 3 + \epsilon & 7 \end{pmatrix}$$

- **Passo di Gauss (prima colonna) senza pivoting parziale.**

Supponiamo di non applicare il pivoting parziale e di lasciare le equazioni nelle posizioni originarie. Applichiamo Gauss calcolando i moltiplicatori

$$l_{21} = 10 \quad l_{31} = 10^3$$

Facciamo i calcoli

$$\begin{aligned} 10x_1 + 2x_2 + 4x_3 - 10(x_1 + (-1 + \epsilon)x_2 + 5x_3) &= (12 - 10\epsilon)x_2 - x_3 \\ 10^3 x_1 + (3 + \epsilon)x_2 + 7x_3 - 10^3(x_1 + (-1 + \epsilon)x_2 + 5x_3) &= (1003 - 999\epsilon)x_2 - 4993x_3 \end{aligned}$$

Risulta evidente che l'errore  $\epsilon$  è stato amplificato per mezzo dei calcoli fatti.

- **Passo di Gauss (prima colonna) con pivoting parziale.**

Adesso applichiamo il pivoting parziale. Prendiamo la prima colonna, dove

$$\max_{1 \leq i \leq n} |a_{i1}| = 1000$$

effettuiamo uno scambio di posizione tra prima e terza equazione.

$$A_{new} = \begin{pmatrix} 1000 & 3 + \epsilon & 7 \\ 10 & 2 & 4 \\ 1 & -1 + \epsilon & 5 \end{pmatrix}$$

calcoliamo i relativi moltiplicatori

$$l_{21} = 10^{-2} \quad l_{31} = 10^{-3}$$

Ecco l'aspetto rilevante di questa modifica del metodo di Gauss: prendendo come pivot, ogni volta, quello di modulo massimo, otterremo moltiplicatori con modulo  $\in [0, 1]$ . Questa cosa contribuisce a ridurre eventuali perturbazioni, nell'esempio gli  $\epsilon$  (i moltiplicati per moltiplicatori aventi modulo  $\in [0, 1]$ ).

### 3.3.3.2 pivoting totale

Il pivoting totale si basa sulla stessa filosofia del pivoting parziale, ma la ricerca dell'elemento pivotale massimo non avviene solo lungo le colonne, ma anche lungo le righe.

- Supponiamo di dover trovare l'elemento pivotale  $a_{kk}^k$ .
- Troviamo il modulo massimo tra gli elementi  $a_{ij}$  dove  $k \leq i \leq n$  e  $k \leq j \leq n$

$$\max_{\substack{k \leq i \leq n \\ k \leq j \leq n}} |a_{ij}| = |a_{rq}|$$

Al primo passo scorro l'intera matrice, ai passi successivi si scorrono sottomatrici sempre più piccole.

- Effettuo uno scambio di posizione tra l'equazione  $k$ -esima e l'equazione  $r$ -esima, se  $k \neq r$ .
- **Novità.** Effettuo uno scambio di posizione tra la colonna  $k$ -esima e la colonna  $q$ -esima se  $k \neq q$ .

La novità rilevante è che facciamo sia scambi di righe che scambi di colonne:

- gli scambi di righe sono indolore;
- negli scambi di colonne dobbiamo tenere a mente che abbiamo alterato l'ordine delle incognite (dobbiamo ricordarcelo quando avremo ottenuto tutte le componenti del vettore  $x$ ).

### 3.3.4 Fattorizzazione LR (certe volte detta Fattorizzazione LU)

**Teorema.** Se le ipotesi sui minori principali di testa valgono (ergo applicazione del metodo di Gauss classico, senza scambi di righe e/o colonne), l'algoritmo di eliminazione produce la seguente fattorizzazione

$$A = LR$$

dove

- $R$  è la matrice triangolare superiore ottenuta alla fine dell'applicazione dell'algoritmo;
- $L$  è una matrice contenente tutti i moltiplicatori, avente la seguente forma

$$L = \begin{pmatrix} 1 & & & & \\ l_{21} & 1 & & & \\ l_{31} & l_{32} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ \vdots & \vdots & & \ddots & 1 \\ l_{n1} & l_{n2} & \dots & \dots & l_{n,n-1} & 1 \end{pmatrix}$$

#### 3.3.4.1 Dimostrazione pratica

Cerchiamo di dimostrare quanto detto in modo pratico, prendendo una matrice  $A$  quadrata di ordine 4

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix}$$

Prendiamo la prima colonna, otteniamo i seguenti moltiplicatori

$$l_{21} = \frac{a_{21}}{a_{11}} \quad l_{31} = \frac{a_{31}}{a_{11}} \quad l_{41} = \frac{a_{41}}{a_{11}}$$

Utilizziamo i moltiplicatori per scrivere la matrice  $H_1$ , detta *matrice elementare di Gauss*. Con la premoltiplicazione otteniamo l'applicazione del primo passo del metodo di Gauss in formato matriciale

$$H_1 A = \begin{pmatrix} 1 & & & \\ -l_{21} & 1 & & \\ -l_{31} & & 1 & \\ -l_{41} & & & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^2 & a_{23}^2 & a_{24}^2 \\ 0 & a_{32}^2 & a_{33}^2 & a_{34}^2 \\ 0 & a_{42}^2 & a_{43}^2 & a_{44}^2 \end{pmatrix}$$

I calcoli dietro il prodotto matriciale sono letteralmente gli stessi visti nel metodo di Gauss. Passiamo alla seconda colonna, otteniamo i seguenti moltiplicatori

$$l_{32} = \frac{a_{32}}{a_{22}} \quad l_{42} = \frac{a_{42}}{a_{22}}$$

Definiamo la matrice  $H_2$  e facciamo la premoltiplicazione con  $H_1 A$

$$H_2(H_1 A) = \begin{pmatrix} 1 & & & \\ & 1 & & \\ -l_{32} & & 1 & \\ -l_{42} & & & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^2 & a_{23}^2 & a_{24}^2 \\ 0 & a_{32}^2 & a_{33}^2 & a_{34}^2 \\ 0 & a_{42}^2 & a_{43}^2 & a_{44}^2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^2 & a_{23}^2 & a_{24}^2 \\ 0 & 0 & a_{33}^3 & a_{34}^3 \\ 0 & 0 & a_{43}^3 & a_{44}^3 \end{pmatrix}$$

Concludiamo con la terza colonna, da cui otteniamo un solo moltiplicatore

$$l_{43} = \frac{a_{43}}{a_{33}}$$

Definiamo la matrice  $H_3$  e facciamo la premoltiplicazione con  $H_2 H_1 A$

$$H_3(H_2 H_1 A) = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & -l_{43} & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^2 & a_{23}^2 & a_{24}^2 \\ 0 & 0 & a_{33}^3 & a_{34}^3 \\ 0 & 0 & a_{43}^3 & a_{44}^4 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^2 & a_{23}^2 & a_{24}^2 \\ 0 & 0 & a_{33}^3 & a_{34}^3 \\ 0 & 0 & 0 & a_{44}^4 \end{pmatrix} = R$$

Ecco la matrice  $R$ , triangolare superiore! Abbiamo svolto il seguente prodotto

$$(H_3 H_2 H_1) A = R$$

Le matrici  $H_1, H_2, H_3$  sono matrici triangolari, quindi il loro determinante è 1 e sono sicuramente invertibili. Vogliamo trovare  $A$ , poniamo

$$A = H_1^{-1} H_2^{-1} H_3^{-1} R$$

Osserviamo che le inverse sono le matrici stesse coi moltiplicatori cambiati di segno (si togliono i segni negativi)

$$H_1^{-1} = \begin{pmatrix} 1 & & & \\ l_{21} & 1 & & \\ l_{31} & & 1 & \\ l_{41} & & & 1 \end{pmatrix} \quad H_2^{-1} = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & l_{32} & 1 & \\ & l_{42} & & 1 \end{pmatrix} \quad H_3^{-1} = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & l_{43} & 1 \end{pmatrix}$$

Svolgendo il prodotto matriciale  $H_1^{-1} H_2^{-1} H_3^{-1}$  otteniamo

$$\begin{pmatrix} 1 & & & \\ l_{21} & 1 & & \\ l_{31} & l_{32} & 1 & \\ l_{41} & l_{42} & l_{43} & 1 \end{pmatrix} \implies A = LR$$

Il prodotto  $H_1^{-1} H_2^{-1} H_3^{-1}$  è  $L$ ! Abbiamo finito.

### 3.3.4.2 Utilità della fattorizzazione

Supponiamo di voler risolvere il seguente sistema

$$Ax^{(1)} = b^{(1)}$$

Applicando Gauss ho come risultato la matrice  $R$ . Mettendo da parte i vari moltiplicatori otteniamo pure la matrice  $L$  (ovviamente non si devono fare scambi di righe e/o colonne). Sappiamo che  $A = LR$ : possiamo utilizzare la cosa per semplificarci la vita? Scriviamo il secondo sistema nel seguente modo

$$LRx^{(1)} = b^{(1)}$$

da cui otteniamo due sistemi lineari

$$\begin{cases} Rx^{(1)} = y^{(1)} \\ Ly^{(1)} = b^{(1)} \end{cases}$$

**Tutta questa fatica per due sistemi?** Sì, ma entrambi i sistemi sono triangolari:

- $R$  è triangolare superiore;
- $L$  è triangolare inferiore.

In sostanza quello che facciamo è risolvere un sistema per mezzo di due sistemi molto semplici da risolvere.

**Possibile risoluzione in parallelo?** No, purtroppo non possiamo far lavorare più processori in parallelo sui due sistemi. Questo perché dobbiamo calcolare prima

$$Ly^{(1)} = b^{(1)}$$

solo dopo aver trovato  $y^{(1)}$  potremo risolvere il primo sistema  $Rx^{(1)} = y^{(1)}$ .

### 3.3.4.3 Esempio di fattorizzazione passando da Gauss

Calcoliamo la fattorizzazione LR della seguente matrice

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

- **Prima colonna.**

Calcoliamo i moltiplicatori

$$l_{21} = \frac{l_{21}}{l_{11}} = 1 \quad l_{31} = \frac{l_{31}}{l_{11}} = \frac{4}{2} = 2$$

Sostituiamo seconda e terza equazione in modo da far sparire i coefficienti 2 e 4

$$\begin{array}{ll} 2. & 2x_1 + 2x_2 - (2x_1 + x_2 + x_3) = +x_2 - x_3 \\ 3. & 4x_1 + 4x_2 + 3x_3 - 2(2x_1 + x_2 + x_3) = 2x_2 + x_3 \end{array}$$

segue

$$\begin{pmatrix} 2 & 1 & 1 \\ 0 & 1 & -1 \\ 0 & 2 & 1 \end{pmatrix}$$

- **Seconda colonna.**

$$\text{Calcoliamo il moltiplicatore } l_{32} = \frac{l_{32}}{l_{22}} = \frac{4}{2} = 2$$

Sostituiamo terza equazione in modo da far sparire il coefficiente 2

$$2x_2 + x_3 - 2(x_2 - x_3) = -x_3$$

segue la matrice  $R$

$$\begin{pmatrix} 2 & 1 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 3 \end{pmatrix}$$

La matrice  $L$  si ottiene ricordando i coefficienti calcolati nei due passi

$$L = \begin{pmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & 2 & 1 \end{pmatrix}$$

### 3.3.4.4 Esempio di fattorizzazione non passando da Gauss

Calcoliamo la fattorizzazione LR della stessa matrice di prima. ma evitando Gauss

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

- **Termini noti delle matrici.**

Pensiamo alla struttura delle due matrici, per prima cosa poniamo gli elementi noti

$$L = \begin{pmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{pmatrix} \quad R = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \end{pmatrix}$$

- **Prima riga di  $R$ .**

La prima riga della matrice  $R$  è la prima riga della matrice  $A$ .

$$R = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \end{pmatrix} \implies R = \begin{pmatrix} 2 & 1 & 1 \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \end{pmatrix}$$

- **Elementi rimanenti.** Ricapitoliamo

$$\begin{pmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 1 \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \end{pmatrix} = \begin{pmatrix} 2 & 1 & 1 \\ 2 & 2 & 0 \\ 4 & 4 & 3 \end{pmatrix}$$

Ricaviamo dai calcoli del prodotto matriciale alcune equazioni utili per trovare i termini rimanenti:

- $l_{21} * 2 + 1 * 0 + 0 * 0 = 2 \rightarrow 2l_{21} = 2 \rightarrow l_{21} = 1$
- $l_{31} * 2 + l_{32} * 0 + 1 * 0 = 4 \rightarrow 2l_{31} = 4 \rightarrow l_{31} = 2$
- $1 * 1 + 1 * r_{22} + 0 * 0 = 2 \rightarrow 1 + r_{22} = 2 \rightarrow r_{22} = 1$
- $1 * 1 + 1 * r_{23} + 0 * r_{33} = 0 \rightarrow 1 + r_{23} = 0 \rightarrow r_{23} = -1$
- $2 * 1 + l_{32} * 1 + 1 * 0 = 4 \rightarrow 2 + l_{32} = 4 \rightarrow l_{32} = 2$
- $2 * 1 + 2 * (-1) + 1 * r_{33} = 3 \rightarrow r_{33} = 3$

Abbiamo finito!

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & 2 & 1 \end{pmatrix} \quad R = \begin{pmatrix} 2 & 1 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 3 \end{pmatrix}$$

### 3.3.4.5 Fattorizzazione LR con pivoting parziale

Possiamo dimostrare che l'algoritmo di Gauss con pivoting parziale conduce ancora a una fattorizzazione, avente la seguente forma

$$PA = L_p R_p$$

Dove  $P$  è una matrice di permutazione definita dagli scambi di righe richiesti dall'algoritmo.  $R_p$  è triangolare superiore ed  $L_p$  è triangolare inferiore (ed elementi diagonali unitari).

### 3.3.4.6 Determinante della matrice A

**Caso senza pivoting** Dalla fattorizzazione  $A = LR$  deduciamo che il determinante di  $R$  è il seguente

$$\det(A) = \det(L) \det(R) = 1 * \det(R) = \det(R)$$

dove il  $\det(L)$  è il determinante di una triangolare, con diagonale unitaria.

**Caso con pivoting** Per quanto riguarda la fattorizzazione  $PA = L_p R_p$  ragioniamo come segue

$$\det(P) \det(A) = \det(L_p) \det(R_p)$$

la matrice  $L_p$  ha le stesse proprietà della matrice  $L$ , ergo  $\det(L_p) = 1$ . Per quanto riguarda  $\det(P)$  sappiamo che la matrice di permutazione si ottiene a partire dalla matrice identica, permutando righe e colonne: il determinante della matrice identica è 1, e noi sappiamo che per ogni scambio di riga il determinante è lo stesso cambiato di segno.

$$(-1)^s \det(A) = \det(L_p) \det(R_p)$$

dove  $s$  è il numero di scambi di righe effettuati nel pivoting parziale. Concludiamo

$$\det(A) = \frac{1}{(-1)^s} \det(R_p) \implies \det(A) = (-1)^s \det(R_p)$$

### 3.3.5 Variante metodo di Gauss: Gauss-Jordan

Il metodo di Gauss-Jordan è una variante del metodo di Gauss: essa prevede, dato un sistema  $Ax = b$ , l'individuazione di una matrice dei coefficienti  $D$  diagonale!

$$Dx = b'$$

In sostanza non vogliamo soltanto sbarazzarci dei valori sotto la diagonale, ma anche di quelli sopra. Questo significa che ad ogni passo  $i$ -esimo elimineremo l'incognita  $x_i$  da tutte le equazioni, tranne la  $i$ -esima.

**Condizioni di applicabilità** Vale anche qua il discorso dei minori principali di testa, che assicurano l'esecuzione senza necessità di fare scambi.

**Costo computazionale** Risulta ovvio che il costo computazionale di Gauss-Jordan sia maggiore rispetto a Gauss classico. Dato un sistema di  $n$  equazioni in  $n$  incognite otteniamo circa il seguente numero di operazioni

$$\frac{n^3}{2} + n^2 - \frac{n}{2}$$

Si ricordi che il costo computazionale di Gauss classico è nell'ordine di  $O(n^3/3)$ .

#### 3.3.5.1 Esempio: calcolo dell'inversa con Gauss-Jordan

Abbiamo già detto che data una matrice  $A$  di ordine  $n$  non singolare, la sua inversa  $A^{-1}$  è la soluzione del sistema matriciale

$$AX = I$$

Cioè dobbiamo risolvere  $n$  sistemi lineari.

$$Ax^1 = e^1 \quad Ax^2 = e^2 \quad \dots \quad Ax^n = e^n$$

Applicare Gauss-Jordan significa trasformare  $A$  nella matrice identica: nel secondo membro, dove erano presenti i termini noti, troveremo l'inversa!

$$IX = C$$

dove  $C = A^{-1}$ .

**Esempio concreto** Consideriamo la seguente matrice

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

Non andiamo a calcolare a priori il determinante (in generale, soprattutto con matrici quadrate di ordine  $n$  molto grande), ma procediamo subito all'applicazione di Gauss-Jordan: se l'inversa non esiste ce ne accorgeremo dall'impossibilità di portare i conti a termine.

2 3 1	1 0 0	
1 3 1	0 1 0	Si vuole risolvere il sistema $AX = B$ , quindi trovare $X$ .
0 2 1	0 0 1	
2 3 1	1 0 0	Passo uguale al classico metodo di Gauss classico
0 3/2 1/2	-1/2 1 0	Calcolati i moltiplicatori $l_{21} = \frac{a_{21}}{a_{11}}$ , $l_{31} = \frac{a_{31}}{a_{11}}$
0 2 1	0 0 1	
2 0 0	2 -2 0	
0 1 1/3	-1/2 2/3 0	Non abbiamo eliminato solo $a_{32}$ , ma anche $a_{12}$
0 0 1/3	2/3 -4/3 1	
1 0 0	1 -1 0	
0 1 0	-1 2 -1	Non abbiamo eliminato solo $a_{23}$ , ma anche $a_{13}$
0 0 1	2 -4 3	

La matrice inversa è la seguente

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

## 3.4 Malcondizionamento di un sistema lineare

### 3.4.1 Esempi introduttivi

**Termini noti perturbati** Consideriamo i seguenti sistemi lineari  $2 \times 2$

$$\begin{pmatrix} 1 & -1 \\ 1 & -1.000001 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} 1 & -1 \\ 1 & -1.000001 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 0.999999 \\ 1 \end{pmatrix}$$

Il secondo sistema lineare è il primo con un termine noto perturbato, e matrice dei coefficienti invariata. Calcolando le soluzioni esatte troviamo due soluzioni profondamente diverse, con variazioni nell'ordine delle unità:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} -10^{-6} \\ -1 \end{pmatrix}$$

C'è stata un'amplificazione dell'errore: siamo passati da un errore di  $10^{-6}$  a un errore di  $10^0$  (nell'ordine dell'unità).

**Matrice dei coefficienti e termini noti perturbati** Consideriamo un'altra coppia di sistemi lineari  $2 \times 2$

$$\begin{pmatrix} 3 & 4 \\ 3 & 4.00001 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 7 \\ 7.00001 \end{pmatrix}$$

$$\begin{pmatrix} 3 & 4 \\ 3 & 3.99999 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 7 \\ 7.00004 \end{pmatrix}$$

Il secondo sistema lineare è il primo con un elemento della matrice  $A$  perturbato, così come un termine noto perturbato. Calcolando le soluzioni esatte troviamo, anche in questo caso, due soluzioni profondamente diverse:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \frac{23}{3} \\ -4 \end{pmatrix}$$

da errori di  $10^{-5}$  si è passato ad errori di  $10^0$  (nell'ordine dell'unità).

### 3.4.2 Definizione di sistema malcondizionato

**Definizione.** Quando gli errori introdotti nei dati sono amplificati nella soluzione si parla di *sistema malcondizionato*.

- **Matrice di perturbazione.** Se la matrice dei coefficienti di un sistema lineare è perturbata allora andremo a risolvere un sistema con matrice

$$A + \delta A$$

, dove  $\delta A$  è la matrice con la perturbazione dei vari elementi.

- **Vettore di perturbazione.** Se il vettore dei termini noti di un sistema lineare è perturbato allora andremo a risolvere un sistema con vettore dei termini noti

$$b + \delta b$$

dove  $\delta b$  è un vettore con la perturbazione dei vari elementi.

- **Soluzione perturbata.** La conseguenza delle perturbazioni è la perturbazione del vettore  $x$ . Il sistema che si risolve è del tipo

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

dove  $\delta x$  è il vettore soluzione perturbato.

### 3.4.3 Caso particolare: matrice dei coefficienti non perturbata

Supponiamo di non avere perturbazioni sulla matrice  $A$ . Il sistema lineare si presenta nella forma

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

sviluppiamo

$$Ax + A\delta x = b + \delta b$$

se  $Ax = b$  allora  $Ax - b = 0$ , semplifichiamo

$$A\delta x = \delta b$$

la matrice  $A$  ha  $\det \neq 0$ , quindi esiste l'inversa. Premoltiplichiamo con  $A^{-1}$

$$\delta x = A^{-1}\delta b$$

la norma di  $\delta x$  è  $\|\delta x\| = \|A^{-1}\delta b\|$ .

**Prima relazione** Stiamo ragionando con norme matriciali naturali, ergo vale la definizione di *norme coerenti*:

$$\|\delta x\| = \|A^{-1}\delta b\| \leq \|A^{-1}\| \|\delta b\| \longrightarrow \boxed{\|\delta x\| \leq \|A^{-1}\| \|\delta b\|}$$

Lasciamo da parte questa cosa, ci servirà dopo.

**Seconda relazione** Torniamo al sistema  $Ax = b$ . Se si ha equivalenza tra i due membri si ha equivalenza tra le relative norme

$$\|Ax\| = \|b\|$$

se le norme matriciali e vettoriali sono coerenti possiamo dire che

$$\|b\| = \|Ax\| \leq \|A\| \|x\| \implies \|A\| \|x\| \geq \|b\|$$

la norma di  $A$  è nulla solo se si ha la matrice nulla, ma non può esserci perché significherebbe avere  $\det(A) = 0$ . Segue che possiamo dividere per  $\|A\|$ , otteniamo

$$\boxed{\|x\| \geq \frac{\|b\|}{\|A\|}}$$

**Terza relazione** Adesso facciamo la divisione membro tra le due relazioni precedenti.

$$\frac{\|\delta x\|}{\|x\|} \leq \|A^{-1}\| \|\delta b\| \frac{\|A\|}{\|b\|}$$

il segno mantenuto è quello della prima diseguaglianza, il  $\leq$ : se consideriamo i segni delle prime due relazioni osserviamo che nel primo membro si ha una divisione tra termine piccolo e termine grande, nel secondo membro si ha una divisione tra termine grande e termine piccolo.

Per quanto riguarda  $\|b\|$  ci chiediamo se questo possa essere nullo: possibile  $\|b\| = 0$  solo se si ha vettore nullo, ma a quel punto avrei sistema omogeneo con  $\det A \neq 0$ , e l'unica soluzione possibile è  $x = 0$ . La relazione finale, quella che ci interessa, è la seguente

$$\boxed{\frac{\|\delta x\|}{\|x\|} \leq \|A\| \|A^{-1}\| \frac{\|\delta b\|}{\|b\|}}$$

- Il rapporto  $\frac{\|\delta x\|}{\|x\|}$  rappresenta per il vettore  $x$  una misura di quello che è l'errore relativo (*errore relativo nel calcolo della soluzione del sistema perturbato*).
- Il rapporto  $\frac{\|\delta b\|}{\|b\|}$  rappresenta per il vettore  $b$  è l'*errore relativo con cui si introduce il vettore  $b$* .

In sostanza stiamo affermando che l'errore relativo su  $x$  è minore o uguale dell'errore relativo su  $b$  amplificato dal fattore  $\|A\| \|A^{-1}\|$ .

### 3.4.3.1 Numero di condizionamento del sistema con matrice $A$

Sulla base di quanto detto definiamo *numero di condizionamento del sistema con matrice dei coefficienti  $A$*

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

**Valore minimo** Il numero di condizionamento non può risultare minore di 1.

$$\mu(A) = \|A\| \|A^{-1}\| \geq \|AA^{-1}\| = |I| \geq \rho(I) = 1$$

Dal teorema di Hirsh otteniamo la disegualanza col raggio spettrale, che è sicuramente uguale ad 1 per la matrice identica.

### 3.4.3.2 Definizione aggiornata di malcondizionamento

Si parla di *malcondizionamento* quando questo valore è molto maggiore rispetto ad 1

$$\mu(A) \gg 1$$

in generale quando raggiunge l'ordine di  $10^3$ .

### 3.4.4 Caso generale con matrice perturbata

Supponiamo di avere un sistema con perturbazioni anche sulla matrice  $A$ . Il sistema lineare si presenta nella forma

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

Saltiamo la dimostrazione, possiamo affermare che

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\mu(A)}{1 - \frac{\|\delta A\|}{\|A\|} \mu(A)} \left( \frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right)$$

Si osservi che senza perturbazione della matrice  $A$  si ha  $\|\delta A\| = 0$ : annullando i relativi termini si recupera la relazione del caso particolare precedentemente visto.

**Casi esclusi** La relazione detta è valida solo se il denominatore è positivo

$$1 - \frac{\|\delta A\|}{\|A\|} \mu(A) > 0$$

Sostituendo  $\mu(A) = \|A\| \|A^{-1}\|$  si ha una semplificazione

$$1 - \|\delta A\| \|A^{-1}\| > 0$$

ricaviamo che

$$\|\delta A\| \leq \frac{1}{\|A^{-1}\|}$$

La relazione è valida solo se la perturbazione non è così grande (nel senso che non si stravolge completamente la matrice dei coefficienti  $A$ ).

### 3.4.5 Vettore residuo, errore assoluto e relativo

**Definizione.** Sia  $\hat{x}$  la soluzione ottenuta per il sistema  $Ax = b$ , risolto con un qualunque metodo. Abbiamo

$$b - A\hat{x} = r$$

$r$  è il cosiddetto *vettore residuo*.

In presenza di una soluzione esatta il vettore residuo sarà nullo e quindi avremo, come al solito

$$b - Ax = 0$$

**Errore assoluto** Eseguiamo la differenza tra i due sistemi (quello con la soluzione esatta e quello con la soluzione calcolata per mezzo di un metodo), otteniamo

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

dove l'ultima formula si è ottenuta premoltiplicando i membri per  $A^{-1}$  cosa possibile visto che  $A$  è matrice non singolare. Con questi calcoli abbiamo trovato l'errore assoluto nel calcolo della soluzione del sistema perturbato.

**Errore relativo** Lavoriamo sulle norme possiamo dire (se si ha coerenza tra le norme)

$$\|\hat{x} - x\| = \|A^{-1}r\| \leq \|A^{-1}\| \|r\|$$

Si riprenda la seconda relazione del caso particolare

$$\|x\| \geq \frac{\|b\|}{\|A\|}$$

anche qua dividiamo (segno non cambia per le stesse ragioni nel caso particolare)

$$\frac{\|\hat{x} - x\|}{\|x\|} \leq \mu(A) \frac{\|r\|}{\|b\|}$$

Abbiamo trovato l'errore relativo nel calcolo della soluzione del sistema perturbato. Vettore dei residui piccoli non significa avere una buona approssimazione della soluzione ( $\|b\|$  è al denominatore!!)

### 3.4.6 Esempio: numero di condizionamento di matrici hermitiane

Si consideri una matrice hermitiana  $A \in \mathbb{C}^{n \times n}$ . Si vuole determinare il numero di condizionamento  $\mu_2(A)$ , calcolato in norma 2. Ricordiamoci che la norma matriciale 2 è la seguente

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

In presenza di una matrice Hermitiana abbiamo  $A = A^H$  otteniamo

$$\|A\|_2 = \rho(A)$$

Sviluppiamo  $\mu_2(A)$

$$\mu_2(A) = \|A\|_2 \|A^{-1}\|_2 = \rho(A) \rho(A^{-1})$$

Ricordarsi che  $A^{-1}$  è a sua volta una matrice Hermitiana, ricordarsi che gli autovalori dell'inversa di una matrice  $A$  sono i reciproci degli autovalori della matrice  $A$ . Concludiamo

$$\mu_2(A) = \max_{1 \leq i \leq n} |\lambda_i| \frac{1}{\min_{1 \leq i \leq n} |\lambda_i|} = \frac{\max_{1 \leq i \leq n} |\lambda_i|}{\min_{1 \leq i \leq n} |\lambda_i|}$$

### 3.4.7 Extra: esempio di matrice malcondizionata

Il professore ha presentato come esempio per eccellenza di matrice malcondizionata la cosiddetta *matrice hilbertiana*, che ha come componenti

$$a_{ij} = \frac{1}{i+j-1}, \quad i, j = 1, \dots, n$$

che avrà componenti razionali! Possibile costruire matrici hilbertiane su Matlab per mezzo del seguente comando

**hilb(k)**

dove  $k$  è l'ordine della matrice hilbertiana. Per mezzo del seguente codice

```
r=[];
for j=2:20;
    c=cond(hilb(j),1);
    r=[r; j c];
end;
r
```

calcoliamo il numero di condizionamento della matrice di Hilbert al crescere dell'ordine della matrice, da  $j = 2$  a  $j = 20$ . Il seguente output mostra che abbiamo una crescita rapida del malcondizionamento, già alla matrice hilbertiana di ordine 4

<b>r =</b>	I
2.0000e+00	2.7000e+01
3.0000e+00	7.4800e+02
4.0000e+00	2.8375e+04
5.0000e+00	9.4366e+05
6.0000e+00	2.9070e+07
7.0000e+00	9.8519e+08
8.0000e+00	3.3873e+10
9.0000e+00	1.0996e+12
1.0000e+01	3.5352e+13
1.1000e+01	1.2295e+15
1.2000e+01	3.8420e+16
1.3000e+01	7.4907e+17
1.4000e+01	1.5679e+18
1.5000e+01	1.2092e+18
1.6000e+01	7.1626e+19
1.7000e+01	1.2269e+19
1.8000e+01	4.4141e+18
1.9000e+01	4.7506e+18
2.0000e+01	1.0480e+19

Il comportamento è lo stesso anche in norma 2 e norma di Frobenius.

## 3.5 Metodi iterativi

### 3.5.1 Introduzione

I metodi iterativi sono metodi che costruiscono una successione di vettori. Se il metodo funziona la successione di vettori ha come limite la soluzione del problema  $a \in \mathbb{C}^n$

$$\lim_{k \rightarrow +\infty} x^{(k)} = a \quad \boxed{Aa = b}$$

come tutti i limiti è difficile che il valore venga raggiunto, quindi si approssima la soluzione.

#### Contesti suggeriti

- I metodi iterativi sono suggeriti in presenza di una matrice *A sparsa*, cioè una matrice con pochi elementi diversi da zero.
- Questo non vieta un uso anche con matrici *A dense* (o *piene*).

**Costo di un'iterazione** I metodi iterativi sono basati su iterazioni, ed ogni iterazione è data da un prodotto tra matrice  $H$  e vettore. Quindi il costo dell'iterazione è il costo del prodotto matrice-vettore.

### 3.5.2 Definizione di metodo convergente

**Definizione.** Un metodo iterativo si dice *convergente* se la successione  $\{x^{(k)}\}$  converge alla soluzione del sistema  $a$ .

### 3.5.3 Costruzione dello schema iterativo

Presentiamo due strategie per costruire lo schema iterativo da noi adottato nei vari metodi.

**Primo metodo di costruzione** Si consideri un sistema  $Ax = b$  con  $A \in \mathbb{C}^{n \times n}$ . Si prenda una matrice  $G \in \mathbb{C}^{n \times n}$  non singolare. Spostiamo  $b$  nel primo membro e premoltiplichiamo per  $G$

$$Ax - b = 0 \longrightarrow G(Ax - b) = 0$$

Il sistema lineare ottenuto è omogeneo. Se  $G$  è matrice non singolare allora l'unica soluzione si ha con  $Ax - b = 0$ , cioè è verificato il sistema di partenza  $Ax = b$ .

$$GAx - Gb = 0$$

aggiungiamo  $x$  a primo e secondo membro

$$x + GAx - Gb = x$$

isolo la  $x$  a primo membro e raccolgo a secondo membro ( $x$  va a destra, si postmoltiplica)

$$x = x - GAx + Gb \longrightarrow x = (I - GA)x + Gb$$

se indichiamo  $H = I - GA$  e  $c = Gb$  otteniamo un sistema equivalente a quello di partenza.

$$x = Hx + c$$

La matrice  $H$  è detta *matrice di iterazione*. La scrittura ha l'incognita uguale a un'espressione dipendente dall'incognita stessa, e suggerisce il seguente schema iterativo (con  $k = 1, 2, 3, \dots$ )

$$\boxed{x^{(k)} = Hx^{(k-1)} + c}$$

**Secondo metodo di costruzione** Si consideri un sistema  $Ax = b$  con  $A \in \mathbb{C}^{n \times n}$ . Si prenda una decomposizione della matrice  $A$ , supponendo  $\det(M) \neq 0$

$$A = M - N$$

Sostituisco la  $x$  nel sistema lineare

$$(M - N)x = b \longrightarrow Mx - Nx = b$$

il fatto di avere  $M$  non singolare ci permette di calcolare l'inversa, e quindi possiamo premoltiplicare per  $M^{-1}$

$$x = M^{-1}Nx + M^{-1}b$$

ponendo  $H = M^{-1}N$  e  $c = M^{-1}b$  otteniamo lo stesso schema di prima

$$x = Hx + c$$

### 3.5.4 Teorema di convergenza globale - C.N.S.

**Teorema.** Condizione necessarie e sufficiente affinché un metodo iterativo della forma

$$x^{(k+1)} = Hx^{(k)} + c, k = 0, 1, \dots$$

sia convergente per qualunque vettore iniziale  $x^{(0)}$  è che la sua matrice di iterazione  $H$  sia convergente!

Prendiamo l'uguaglianza da cui partono i nostri ragionamenti:  $x = Hx + c$ . Se  $a \in \mathbb{C}^n$  è la soluzione esatta del sistema lineare allora potremo dire anche

$$a = Ha + c$$

#### 3.5.4.1 Errore associato all'iterazione

Prendiamo lo schema iterativo  $x^{(k+1)} = Hx^{(k)} + c, k = 0, 1, 2, \dots$ , e sottraiamo  $a$  ad entrambi i membri

$$x^{(k+1)} - a = Hx^{(k)} + c - a = Hx^{(k)} - Ha \longrightarrow \boxed{x^{(k+1)} - a = H(x^{(k)} - a)}$$

Abbiamo ottenuto la differenza tra la  $k$ -esima iterazione e la soluzione esatta del sistema lineare. Sostituiammo  $e^{(k+1)} = x^{(k+1)} - a$

$$e^{(k+1)} = He^{(k)}$$

La relazione lega l'errore commesso al passo  $k$ -esimo con l'errore compiuto al passo  $(k+1)$ -esimo. E se volessi legare l'errore compiuto al passo  $(k+1)$ -esimo con l'errore iniziale? Si osservi che

$$e^{(k+1)} = H \left( He^{(k-1)} \right) = H^2 E^{(k-1)} = H^3 E^{(k-2)} = \dots = H^{k+1} e^{(0)}$$

Abbiamo ottenuto

$$\boxed{e^{(k+1)} = H^{k+1} e^{(0)}}$$

### 3.5.4.2 Dimostrazione

Visto che parliamo di una condizione necessaria e sufficiente dobbiamo dimostrare "in entrambi i sensi".

- **Matrice  $H$  convergente implica metodo convergente.**

Quello che a noi piacerebbe, nei nostri calcoli, è avere

$$\lim_{k \rightarrow +\infty} x^{(k)} = a \implies \lim_{k \rightarrow +\infty} (x^{(k)} - a) = 0 \implies \lim_{k \rightarrow +\infty} e^{(k)} = 0$$

Se  $e^{(0)}$  è costante allora la condizione che vogliamo ottenere è la seguente

$$\lim_{k \rightarrow +\infty} H^k = O$$

se la matrice  $H$  è convergente allora lo schema iterativo converge e! Il valore iniziale, costante, non influenza in alcun modo la convergenza (lo si porta fuori dal limite).

- **Metodo convergente implica matrice  $H$  convergente.**

Tra tutti i possibili vettori iniziali  $x^{(0)}$  ci saranno gli autovettori di  $H$ . Se  $e^{(0)}$  è autovettore allora potremo dire

$$He^{(0)} = \lambda e^{(0)}$$

Ricordandoci che gli autovalori di  $H^k$  sono potenze  $k$ -esime degli autovalori di  $H$  otteniamo

$$e^{(k+1)} = H^{(k+1)}e^{(0)} = \lambda^{k+1}e^{(0)}$$

Se il metodo converge allora  $e^{(k+1)} \rightarrow 0$ , quindi  $\lambda^{k+1}e^{(0)} \rightarrow 0$ . Ciò è possibile solo se  $\lambda^{k+1} \rightarrow 0$ , cosa possibile se  $|\lambda| < 1$ . Se tutti gli autovalori avranno modulo minore di 1 avremo raggio spettrale minore di 1

$$\rho(H) < 1$$

Avevamo già visto un teorema dove si afferma che una matrice generica  $A \in \mathbb{C}^{n \times n}$  è convergente se  $\rho(A) < 1$ .

**Quindi.**

$$H \text{ convergente} \iff x^{(k)} \rightarrow a$$

- Se la matrice  $H$  è convergente il metodo converge, al di là della scelta del vettore iniziale.
- Se il metodo converge indipendentemente dalla scelta del vettore iniziale allora la matrice  $H$  è convergente.

Il teorema ci dà le condizioni, ma è poco utile sul piano pratico (troppi conti, costo computazionale elevato).

**Osservazione** Non è necessario avere  $\det H \neq 0$  per la convergenza del metodo.

**Corollario 1** Per la convergenza del metodo iterativo è necessario e sufficiente che

$$\rho(H) < 1$$

### 3.5.5 Condizione sufficiente per la convergenza del metodo iterativo

Si prenda teorema di Hirsh e condizione necessaria e sufficiente per avere convergenza del metodo iterativo

$$\rho(H) \leq \|H\| \quad \rho(H) < 1$$

Se io potessi dire (unendo le due relazioni)

$$\|H\| < 1$$

allora  $\rho(A) < 1$ : convergenza dello schema iterativo alla soluzione del sistema

$$x^{(k)} \rightarrow a$$

La condizione è sufficiente, ma non necessaria: potrei avere una norma che vale 1.0001 e raggio spettrale 0.0009 (cit.).

### 3.5.6 Condizione necessaria per la convergenza del metodo iterativo

Condizione necessaria per la convergenza del metodo iterativo è

$$|\det(H)| < 1$$

Ricordarsi del termine  $\sigma_n$  nel polinomio caratteristico: il prodotto degli autovalori. Se questo prodotto è maggiore di 1 è automatico che ci sia qualche autovalore  $> 1$  (e quindi  $\rho(H) > 1$  sicuramente).

**Non è condizione sufficiente** Potrei avere, ad esempio, gli autovalori 0.4 e 2: il prodotto è uguale a 0.8, ma il raggio spettrale è 2!

### 3.5.7 Esempio di esercizio sulla convergenza del metodo

Un processo iterativo presenta la seguente matrice di iterazione  $H$

$$H = \frac{1}{20} \begin{pmatrix} 2 & -1 & 4 \\ 3 & 1 & 7 \\ -10 & 2 & 3 \end{pmatrix}$$

Il metodo risulta convergente?

**Risoluzione** Per la risoluzione ricorriamo alle condizioni introdotte precedentemente. Tra queste la più utile è la condizione sufficiente. Prendiamo, tra le norme matriciali, quella più facile da calcolare: la norma  $\infty$

$$\|H\|_\infty = \frac{1}{20} \max\{7, 11, 15\} = \frac{15}{20} = \frac{3}{4} < 1$$

il metodo è convergente!

### 3.5.8 Velocità asintotica di convergenza

Per qualunque norma matriciale naturale si ha

$$\boxed{\lim_{k \rightarrow +\infty} \sqrt[k]{\|H^k\|} = \rho(H)}$$

Da questo possiamo immaginare che per  $k$  abbastanza grande, risulta

$$\sqrt[k]{\|H^k\|} \simeq \rho(H)$$

Riprendiamo dall'errore al passo  $k$ -esimo

$$e^{(k)} = H^k e^{(0)}$$

Se i due vettori sono uguali lo sono anche le norme, e se le norme sono coerenti allora vale la diseguaglianza

$$\|e^{(k)}\| = \|H^k e^{(0)}\| \leq \|H^k\| \|e^{(0)}\|$$

Possiamo dividere per  $\|e^{(0)}\|$ ? La norma è nulla se il vettore  $e^{(0)}$  è nullo: questo significa scegliere come vettore  $x^{(0)}$  la soluzione stessa! Il caso è molto raro e renderebbe inutili i calcoli su cui stiamo discutendo. Supponiamo che  $x^{(0)}$  non sia la soluzione e dividiamo

$$\frac{\|e^{(k)}\|}{\|e^{(0)}\|} \leq \|H^k\|$$

Abbiamo messo in relazione l'errore iniziale (legato alla scelta di  $x^{(0)}$ ) con l'errore al passo  $k$ -esimo.

**Calcolo-esercizio** Supponiamo che dopo  $k$  iterazioni si voglia ottenere un rapporto almeno nell'ordine  $10^{-m}$ , dove  $m \in \mathbb{N}$ .

$$\frac{\|e^{(k)}\|}{\|e^{(0)}\|} \leq \|H^k\| \simeq 10^{-m}$$

Che relazione c'è tra  $k$  ed  $m$ ? Riprendiamo la relazione iniziale: se essa è vera allora potremo dire

$$\|H^k\| \simeq \rho^k(H)$$

Per trovare la relazione poniamo  $\rho^k(H) = 10^{-m}$ .

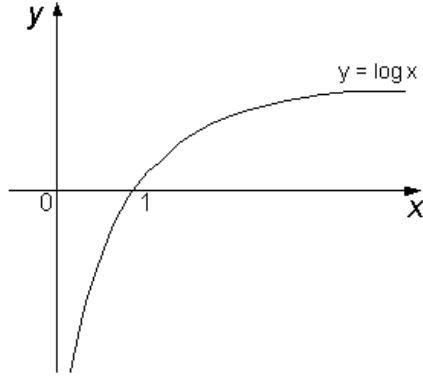
$$\begin{aligned} \rho^k(H) &= 10^{-m} \\ \text{Log}(\rho^k(H)) &= -m \\ k \text{Log}(\rho(H)) &= -m \end{aligned}$$

Si consideri che il professore rappresenta il logaritmo in base dieci scrivendo log con la  $L$  maiuscola, e si osservi che non è necessario il valore assoluto nel logaritmo (il raggio spettrale avrà sicuramente valore  $> 0$ ). Dividendo entrambi i membri per  $k$  e cambiando il segno concludiamo

$v = \frac{m}{k} = -\text{Log}(\rho(H))$

Fissato  $m$  trovo il numero di iterazioni per abbattere l'errore di un fattore  $10^{-m}$ . La formula trovata è la *velocità asintotica di convergenza del metodo* (avente matrice di iterazione  $H$ ).

**Osservazione** Si tenga a mente il grafico di  $\log_a(x)$  con  $a > 1$



Se  $\rho(H) < 1$  allora  $\text{Log}(\rho(H)) < 0$ , ergo  $-\text{Log}(\rho(H)) > 0$ . Il metodo più efficiente è quello che impiega meno iterazioni per abbattere l'errore, quindi quello che ha velocità  $v = -\text{Log}(\rho(H))$  maggiore. Conclusione

Più piccolo è  $\rho(H)$ , più veloce sarà l'abbattimento dell'errore

### 3.5.9 Criterio di arresto

#### 3.5.9.1 Criterio principale: norma inferiore a errore prefissato

Nei metodi iterativi lavoriamo col limite di una successione di vettori: raramente si raggiunge il valore  $a$

$$\lim_{k \rightarrow +\infty} x^{(k)} = a$$

Segue la necessità di introdurre dei *criteri di arresto*. Recuperiamo lo schema

$$x^{(k+1)} = Hx^{(k)} + c$$

sottraggo a entrambi i membri  $x^{(k)}$

$$x^{(k+1)} - x^{(k)} = Hx^{(k)} - x^{(k)} + c$$

Se  $a$  è la soluzione del sistema possiamo dire  $a = Ha + c$ , e quanto segue

$$a = Ha + c \longrightarrow c = a - Ha \longrightarrow c = (I - H)a \longrightarrow c = -(H - I)a$$

sostituiamo  $c$  nella relazione precedente

$$x^{(k+1)} - x^{(k)} = Hx^{(k)} - x^{(k)} - (H - I)a$$

raccogliamo

$$\begin{aligned} x^{(k+1)} - x^{(k)} &= (H - I)x^{(k)} - (H - I)a \\ x^{(k+1)} - x^{(k)} &= (H - I)(x^{(k)} - a) \end{aligned}$$

Premoltiplichiamo per l'inverso di  $H - I$

$$x^k - a = (H - I)^{-1}(x^{(k+1)} - x^{(k)})$$

Ricordarsi la traslazione dello spettro a proposito della matrice  $(H - I)$ : gli autovalori di  $H - I$  sono gli autovalori di  $H$  diminuiti di 1. Questi autovalori non possono essere nulli

perchè abbiamo  $\rho(H) < 1$ : segue che esiste l'inversa della matrice  $H - I$ . Come al solito in presenza di norme coerenti possiamo dire

$$\|x^{(k)} - A\| = \|(H - I)^{-1}(x^{(k+1)} - x^{(k)})\| \leq \|(H - I)^{-1}\| \|x^{(k+1)} - x^{(k)}\|$$

il fattore  $\|(H - I)^{-1}\|$  è un coefficiente (che può dar fastidio se molto grande), mentre  $\|x^{(k+1)} - x^{(k)}\|$  varia. L'idea è di dire che il modulo di questa differenza di vettori deve essere minore di un certo errore  $E$  prefissato (gli ingegneri preferiscono chiamarla *tolleranza*, cit.)

$$\|x^{(k+1)} - x^{(k)}\| < E$$

Se la successione converge il modulo tende a 0. Abbiamo trovato il criterio di arresto più utilizzato: ci si ferma non appena la norma è inferiore al valore prefissato.

**Raggio spettrale nullo** Sappiamo che  $\text{Log}(\rho(H))$  deve avere per forza  $\rho(H) > 0$ , ma esiste un caso in cui  $\rho(H)$  è nullo. Se lo è la matrice  $H$  è nilpotente (ricordare la definizione). In quel caso tutti i discorsi sul criterio di arresto non servirebbero, perchè abbiamo un numero di iterazioni  $k$  dopo le quali abbiamo  $a$  sicuramente!

### 3.5.9.2 Ulteriore criterio: numero massimo di iterazioni

Supponiamo che dopo  $k$  iterazioni la condizione di arresto non si sia verificata. Due opzioni:

- il metodo sta divergendo;
- le  $k$  iterazioni non sono sufficienti, e la velocità asintotica di convergenza è lenta.

In entrambi i casi è necessario approfondire continuando con ulteriori step. Per garantire che l'algoritmo termini per forza introduciamo un numero massimo di iterazioni. Il calcolo si arresta se

$$k \geq N$$

Il criterio è affiancato a quello principale.

## 3.5.10 Metodi classici

### 3.5.10.1 Premesse

Questi due metodi si basano sulla scomposizione di  $A$  in tre matrici

$$A = D - E - F$$

- La matrice  $D$  è una matrice diagonale che ha come elementi quelli della diagonale principale di  $A$ .

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

- La matrice  $-E$  è una matrice triangolare inferiore stretta (zeri lungo la diagonale principale) che ha come elementi quelli sotto la diagonale principale di  $A$ .
- La matrice  $-F$  è una matrice triangolare superiore stretta che ha come elementi quelli sopra la diagonale principale di  $A$ .

Se si pone  $E$  ed  $F$  invece di  $-E$  e  $-F$  si hanno i relativi elementi della matrice  $A$ , ma cambiati di segno.

### Esempio

$$\begin{aligned} \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 9 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 4 & 0 & 0 \\ 7 & 8 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 2 & 3 \\ 0 & 0 & 6 \\ 0 & 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 9 \end{pmatrix} - \begin{pmatrix} 0 & 0 & 0 \\ -4 & 0 & 0 \\ -7 & -8 & 0 \end{pmatrix} - \begin{pmatrix} 0 & -2 & -3 \\ 0 & 0 & -6 \\ 0 & 0 & 0 \end{pmatrix} \end{aligned}$$

#### 3.5.10.2 Metodo di Jacobi (o metodo delle sostituzioni simultanee)

Dato il sistema lineare  $Ax = b$ , sappiamo che è equivalente scrivere

$$x = Hx + c$$

per arrivare a questa equazione avevamo preso una matrice  $G$  non singolare, e una matrice  $H = I - GA$ , con  $c = Gb$ . Decomponiamo la matrice  $A$

$$A = D - E - F$$

in questo contesto poniamo  $G = D^{-1}$ : segue che per poter applicare il metodo dobbiamo imporre  $a_{ii} \neq 0$  ( $i = 1, \dots, n$ ), altrimenti non esisterebbe l'inversa (determinante della matrice diagonale uguale al prodotto degli elementi lungo la diagonale principale, necessario avere matrice  $G$  non singolare). Sviluppiamo  $H$

$$H = I - GA = I - D^{-1}(D - E - F) = I - I + D^{-1}(E + F) = D^{-1}(E + F)$$

Abbiamo trovato la matrice di iterazione del *metodo di Jacobi*

$$H_J = D^{-1}(E + F)$$

con  $c_J = D^{-1}b$

$$x^{(k+1)} = H_J x^{(k)} + c_J, \quad k = 0, 1, 2, \dots$$

**Calcoli nella versione matriciale** La matrice  $H_J = D^{-1}(E + F)$  può essere ricavata banalmente:

- l'inversa della matrice  $D$  è una matrice dove si pongono come elementi diagonali i reciproci degli elementi diagonali di  $D$  (proprietà di qualunque matrice diagonale)

$$D^{-1} = \begin{pmatrix} \frac{1}{a_{11}} & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{a_{22}} & 0 & \dots & 0 \\ 0 & 0 & \frac{1}{a_{33}} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{1}{a_{nn}} \end{pmatrix}$$

- la matrice  $E + F$  è la matrice  $A$  senza gli elementi diagonali, e i rimanenti elementi cambiati di segno

$$E + F = \begin{pmatrix} 0 & -a_{12} & -a_{13} & \dots & -a_{1n} \\ -a_{21} & 0 & -a_{23} & \dots & -a_{2n} \\ -a_{31} & -a_{32} & 0 & \dots & -a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -a_{n1} & -a_{n2} & -a_{n3} & \dots & 0 \end{pmatrix} \quad -(E + F) = \begin{pmatrix} 0 & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & 0 & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & 0 & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & 0 \end{pmatrix}$$

otteniamo

$$\begin{aligned}
H_J = D^{-1}(E + F) &= - \begin{pmatrix} \frac{1}{a_{11}} & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{a_{22}} & 0 & \dots & 0 \\ 0 & 0 & \frac{1}{a_{33}} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{1}{a_{nn}} \end{pmatrix} \begin{pmatrix} 0 & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & 0 & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & 0 & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & 0 \end{pmatrix} = \\
&= \begin{pmatrix} 0 & a_{12}/a_{11} & a_{13}/a_{11} & \dots & a_{1n}/a_{11} \\ a_{21}/a_{22} & 0 & a_{23}/a_{22} & \dots & a_{2n}/a_{22} \\ a_{31}/a_{33} & a_{32}/a_{33} & 0 & \dots & a_{3n}/a_{33} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1}/a_{nn} & a_{n2}/a_{nn} & a_{n3}/a_{nn} & \dots & 0 \end{pmatrix}
\end{aligned}$$

**Versione per componenti** Ricaviamo la *versione per componenti*.

$$x^{(k+1)} = H_J x^{(k)} + c_J \quad H_J = D^{-1}(E + F) \quad c_J = D^{-1}b$$

$$x^{(k+1)} = D^{-1}(E + F)x^{(k)} + D^{-1}b$$

premultiplichiamo per  $D$

$$Dx^{(k+1)} = (E + F)x^{(k)} + b$$

esplicitiamo le matrici

$$\begin{pmatrix} a_{11} & 0 & 0 & \dots & 0 \\ 0 & a_{22} & 0 & \dots & 0 \\ 0 & 0 & a_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & a_{nn} \end{pmatrix} x^{(k+1)} = - \begin{pmatrix} 0 & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & 0 & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & 0 & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & 0 \end{pmatrix} x^{(k)} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

Quali sono le componenti del vettore?

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

con  $i = 1, 2, \dots, n$ ,  $k = 0, 1, \dots$ . Per il sottraendo si è posta letteralmente parte della definizione di prodotto matriciale ( $i$ -esima riga nota per gli elementi della  $j$ -esima colonna)

**Metodo delle sostituzioni simultanee** Supponiamo di voler implementare al calcolatore il metodo di Jacobi. Per la tipologia di calcolo non è possibile sostituire immediata la componente  $j$ -esima di  $x^{(k)}$  con la componente  $j$ -esima di  $x^{(k+1)}$ . Si parla di metodo delle *sostituzioni simultanee* in quanto è possibile solo sostituire l'intero vettore  $x^{(k)}$  con l'intero vettore  $x^{(k+1)}$  (necessario calcolare tutte le componenti di  $x^{(k+1)}$ ).

### 3.5.10.3 Metodo di Gauss-Seidel (o metodo delle sostituzioni successive)

Sostituiamo  $A = D - E - F$  nel sistema  $Ax = b$

$$(D - E - F)x = b$$

spostiamo  $Fx$  nel secondo membro ottenendo

$$(D - E)x = Fx + b$$

premultiplichiamo per l'inversa di  $D - E$ , Possiamo farlo perchè  $D - E$  è una triangolare inferiore, inoltre imponiamo  $a_{ii} \neq 0, i = 1, 2, \dots, n$  (come nel metodo di Jacobi)

$$x = (D - E)^{-1}x + (D - E)^{-1}b, \quad k = 0, 1, \dots$$

Abbiamo ottenuto la matrice di iterazione del *metodo di Gauss-Seidel* e relativo schema iterativo!

$$H_{GS} = (D - E)^{-1}F$$

**A proposito della condizione** Abbiamo detto che  $a_{ii} \neq 0$  ( $i = 1, 2, \dots, n$ ). Se  $\det(A) \neq 0$  e la condizione non è soddisfatta possiamo rimediare riordinando le equazioni ed eventualmente anche le incognite.

**Calcoli nella versione matriciale** I calcoli per ottenere la matrice di iterazione  $H_{GS}$  non sono immediati come nel metodo di Jacobi. Si osservi che

- dobbiamo calcolare l'inversa della triangolare inferiore  $(D - E)^{-1}$ ;
- $F$  è una matrice triangolare superiore stretta, quindi gli elementi lungo la diagonale principale sono nulli, questo significa che la prima colonna della matrice  $F$  è una colonna nulla;
- per il punto precedente segue che **la prima colonna della matrice  $H_{GS}$  sarà una colonna nulla** (unica certezza che abbiamo, non è un problema visto che  $H$  non è ls matrice  $A$  di un sistema lineare).

Da questo otteniamo che  $\det(H_{GS}) = 0$  sempre!

**Versione per componenti** Ricaviamo la *versione per componenti*.

$$x^{(k+1)} = H_{GS}x^{(k)} + c_{GS} \quad H_{GS} = (D - E)^{-1}F \quad c_{GS} = (D - E)^{-1}b$$

$$x^{(k+1)} = (D - E)^{-1}Fx^{(k)} + (D - E)^{-1}b$$

premultiplichiamo per  $D - E$

$$(D - E)x^{(k+1)} = Fx^{(k)} + b$$

che diventa

$$Dx^{(k+1)} = Ex^{(k+1)} + Fx^{(k)} + b$$

esplicitiamo le matrici

$$\begin{pmatrix} a_{11} & 0 & 0 & \dots & 0 \\ 0 & a_{22} & 0 & \dots & 0 \\ 0 & 0 & a_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & a_{nn} \end{pmatrix} x^{(k+1)} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{pmatrix} - \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ a_{21} & 0 & 0 & \dots & 0 \\ a_{31} & a_{32} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & 0 \end{pmatrix} x^{(k+1)} - \begin{pmatrix} 0 & a_{12} & a_{13} & \dots & a_{1n} \\ 0 & 0 & a_{23} & \dots & a_{2n} \\ 0 & 0 & 0 & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix} x^{(k)}$$

Quali sono le componenti del vettore?

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

concludiamo

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

**Metodo delle sostituzioni successive** Supponiamo di voler implementare il metodo al calcolatore. Contrariamente al metodo di Jacobi è possibile sostituire di volta in volta la componente  $x_i^k$  con la componente  $x_i^{k+1}$ . Si osservino le sommatorie: la prima sommatoria riguarda i termini precedentemente sostituiti, la seconda riguarda i termini ancora da sostituire. Per queste proprietà si parla di *metodo delle sostituzioni successive*.

### 3.5.10.4 Condizioni sufficienti di convergenza

Per quanto riguarda il *metodo di Jacobi* e il *metodo di Gauss-Seidel* possiamo porre delle condizioni sufficienti per la convergenza relative alla matrice  $A$ .

- **Teorema 1.**

Se  $A$  è una matrice a predominanza diagonale forte **allora** il metodo di Jacobi e quello di Gauss-Seidel sono convergenti.

- **Teorema 2.**

Se  $A$  è una matrice irriducibile e a predominanza diagonale debole **allora** il metodo di Jacobi e quello di Gauss-Seidel sono convergenti.

**Osservazione** La convergenza di uno dei due metodi classici non implica la convergenza con l'altro metodo.

**Verifica del primo teorema col metodo di Jacobi** Consideriamo la matrice  $H_J$  e calcoliamo la norma matriciale  $\infty$

$$\|H_J\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1, j \neq i}^n \left| \frac{a_{ij}}{a_{ii}} \right|$$

dall'ipotesi di predominanza diagonale forte della matrice  $A$  affermiamo che

$$\sum_{j=1, j \neq i}^n |a_{ij}| < |a_{ii}|, \quad i = 1, 2, \dots, n$$

dividendo entrambi i membri per  $a_{ii}$  (lo possiamo fare) otteniamo

$$\sum_{j=1, j \neq i}^n \left| \frac{a_{ij}}{a_{ii}} \right| < 1, \quad i = 1, 2, \dots, n$$

che è la stessa sommatoria posta nella norma matriciale  $\infty$ . Dire che  $\|H_J\|_\infty < 1$  è condizione sufficiente per la convergenza del metodo.

### 3.5.10.5 Esempio 1

Un sistema lineare ha come matrice dei coefficienti la matrice

$$A = \begin{pmatrix} 1 & \alpha & \alpha \\ \alpha & 1 & \alpha \\ \alpha & \alpha & 1 \end{pmatrix} \quad \alpha \in \mathbb{C}$$

Determinare l'insieme dei valori complessi  $\alpha$  per i quali il *metodo di Jacobi* converge.

**Risoluzione** Scriviamo la matrice di iterazione del metodo di Jacobi

$$H_J = D^{-1}(E + F) = I(E + F) = I \begin{pmatrix} 0 & -\alpha & -\alpha \\ -\alpha & 0 & -\alpha \\ -\alpha & -\alpha & 0 \end{pmatrix} = \begin{pmatrix} 0 & -\alpha & -\alpha \\ -\alpha & 0 & -\alpha \\ -\alpha & -\alpha & 0 \end{pmatrix}$$

- Attuiamo una traslazione dello spettro, diminuendo tutti gli autovalori di  $\alpha$

$$H_j - \alpha I = - \begin{pmatrix} \alpha & \alpha & \alpha \\ \alpha & \alpha & \alpha \\ \alpha & \alpha & \alpha \end{pmatrix} = \dots$$

ricordarsi che per trovare gli autovalori di  $H_j$  dovremo sommare  $\alpha$ .

- Raccogliamo

$$\dots = -\alpha \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

La matrice  $\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$  è sicuramente singolare perché ho colonne uguali. Tre autovalori: due nulli e un terzo. Il terzo è 3: lo otteniamo dal fatto che la traccia, cioè la sommatoria degli autovalori, è uguale alla somma delle componenti lungo la diagonale principale.

- Si moltiplichì per  $-\alpha$ . Gli autovalori saranno

$$\sigma_1 = 0 \quad \sigma_2 = 0 \quad \sigma_3 = -3\alpha$$

- Traslazione dello spettro, troviamo i seguenti autovalori

$$\lambda_1 = \alpha \quad \lambda_2 = \alpha \quad \lambda_3 = -2\alpha$$

A questo punto troviamo il raggio spettrale

$$\rho(H_j) = 2|\alpha|$$

Poniamo il modulo visto che si parla anche di elementi complessi. Ricordiamoci che CNS per la convergenza è  $\rho(H_j) < 1$ . Abbiamo finito!

$$|\alpha| < \frac{1}{2}$$

### 3.5.10.6 Esempio 2

Un sistema lineare ha come matrice dei coefficienti la matrice

$$A = \begin{pmatrix} 2 & 0 & 0 & \alpha \\ 1 & 2 & 0 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 1 & 2 \end{pmatrix} \quad \alpha \in \mathbb{C}$$

Determinare l'insieme dei valori complessi  $\alpha$  per i quali il *metodo di Jacobi* converge e per quali  $\alpha$  converge il *metodo di Gauss-Seidel*.

#### Risoluzione

- **Metodo di Jacobi.**

Calcoliamo la matrice di iterazione di Jacobi

$$H_J = D^{-1}(E + F) = -\frac{1}{2} \begin{pmatrix} 0 & 0 & 0 & \alpha \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

La matrice ottenuta è sul modello della matrice di Frobenius, che ha come equazione caratteristica

$$\lambda^4 - \alpha = 0 \longrightarrow \lambda = \sqrt[4]{\alpha} \longrightarrow |\lambda| = \sqrt[4]{|\alpha|}$$

Si tenga a mente che c'è  $1/2$ , quindi gli autovalori risultano moltiplicati. Segue il raggio spettrale

$$\rho(H_J) = \frac{1}{2} \sqrt[4]{|\alpha|}$$

Imponiamo per avere la convergenza

$$\frac{1}{2} \sqrt[4]{|\alpha|} < 1 \longrightarrow \sqrt[4]{|\alpha|} < 2 \longrightarrow |\alpha| < 2^4$$

Si osservi che in questo non abbiamo la predominanza diagonale forte: se avessi studiato la convergenza usando la predominanza diagonale forte avrei ottenuto un insieme di valori  $\alpha$  molto più piccolo.

- **Metodo di Gauss-Seidel.**

Calcoliamo la matrice di iterazione di Gauss-Seidel

$$H_{GS} = (D - E)^{-1}F =$$

Possiamo dire al volo

$$D - E = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 1 & 2 \end{pmatrix} \quad F = \begin{pmatrix} 0 & 0 & 0 & -\alpha \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Dobbiamo calcolarci l'inversa per forza ricorrendo a Gauss Jordan: otteniamo

$$(D - E)^{-1} = \begin{pmatrix} 1/2 & 0 & 0 & 0 \\ -1/4 & 1/2 & 0 & 0 \\ 1/8 & -1/4 & 1/2 & 0 \\ -1/16 & 1/8 & -1/4 & 1/2 \end{pmatrix}$$

Concludiamo il calcolo della matrice di iterazione

$$H_{GS} = \begin{pmatrix} 1/2 & 0 & 0 & 0 \\ -1/4 & 1/2 & 0 & 0 \\ 1/8 & -1/4 & 1/2 & 0 \\ -1/16 & 1/8 & -1/4 & 1/2 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & -\alpha \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & -1/2\alpha \\ 0 & 0 & 0 & 1/4\alpha \\ 0 & 0 & 0 & -1/8\alpha \\ 0 & 0 & 0 & 1/16\alpha \end{pmatrix}$$

Dove gli autovalori sono  $\lambda_1 = \lambda_2 = \lambda_3 = 0$  e  $\lambda_4 = \frac{1}{16}\alpha$ . Il raggio spettrale sarà

$$\rho(H_{GS}) = \frac{1}{16}|\alpha|$$

se imponiamo  $\rho(H_{GS}) < 1$  concludiamo

$$|\alpha| < 16$$

abbiamo trovato la stessa condizione del caso precedente (col metodo di Jacobi).

Si osservi che  $\rho(H_{GS}) = \rho^4(H_J)$ . Se si fa un confronto a livello di velocità troviamo che

$$-\text{Log}(\rho(H_{GS})) = -4\text{Log}(\rho(H_J))$$

cioè Gauss-Seidel è quattro volte più veloce di Jacobi, questo significa muoversi con un numero di passi minore e quindi ridurre possibilità di propagazione dell'errore.

### 3.5.10.7 Esempio 3

Un sistema lineare ha come matrice dei coefficienti la matrice

$$A = \begin{pmatrix} 1 & \alpha & 0 \\ 0 & 1 & \alpha \\ \alpha & 0 & 1 \end{pmatrix} \quad \alpha \in \mathbb{C}$$

Determinare l'insieme dei valori complessi  $\alpha$  per i quali il *metodo di Jacobi* converge e per quali  $\alpha$  converge il *metodo di Gauss-Seidel*.

#### Risoluzione

- **Metodo di Jacobi.**

Costruiamo la matrice di iterazione

$$H_J = D^{-1}(E + F) - \begin{pmatrix} 0 & \alpha & 0 \\ 0 & 0 & \alpha \\ \alpha & 0 & 0 \end{pmatrix} = -\alpha \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

Abbiamo di nuovo una matrice di Frobenius, questa volta nella forma con cui l'abbiamo definita. L'equazione caratteristica è

$$\lambda^3 - 1 = 0 \longrightarrow \lambda\lambda = \sqrt[3]{1} \longrightarrow |\lambda| = 1$$

Gli autovalori sono tutti di modulo 1. Si consideri il  $-\alpha$ , troviamo  $|\lambda| = |\alpha|$ . Segue il raggio spettrale

$$\rho(H_J) = |\alpha|$$

la condizione è  $|\alpha| < 1$ .

- **Metodo di Gauss-Seidel.**

Costruiamo la matrice di iterazione

$$H_{GS} = (D - E)^{-1}F$$

Possiamo già dire

$$D - E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \alpha & 0 & 1 \end{pmatrix} \quad F = \begin{pmatrix} 0 & -\alpha & 0 \\ 0 & 0 & -\alpha \\ 0 & 0 & 0 \end{pmatrix}$$

La matrice  $D - E$  è catalogabile come matrice elementare di Gauss: segue al volo l'inversa, senza passare da Gauss-Jordan

$$(D - E)^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\alpha & 0 & 1 \end{pmatrix}$$

Otteniamo

$$H_{GS} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\alpha & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & -\alpha & 0 \\ 0 & 0 & -\alpha \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -\alpha & 0 \\ 0 & 0 & -\alpha \\ 0 & \alpha^2 & 0 \end{pmatrix}$$

Ho almeno un autovalore nullo. Si osservi che possiamo trattare  $H_{GS}$  come una matrice a blocchi

$$\begin{pmatrix} (0) & (-\alpha & 0) \\ \begin{pmatrix} 0 \\ 0 \end{pmatrix} & \begin{pmatrix} 0 & -\alpha \\ \alpha^2 & 0 \end{pmatrix} \end{pmatrix}$$

Dal determinante della seconda matrice della diagonale a blocchi otteniamo i rimanenti autovalori

$$\lambda^2 + \alpha^3 = 0 \rightarrow \lambda_{4,3} = \pm \sqrt{\alpha^3} \lambda_{4,3} = \alpha^{3/2}$$

Segue il raggio spettrale

$$\rho(H_{GS}) = |\alpha|^{3/2} < 1$$

A cui segue  $|\alpha| < 1$  visto che  $3/2 > 1$  (proprietà della funzione, roba da superiori che si danno per acquisite - cit.).

# Capitolo 4

## Sistemi non lineari

### 4.1 Introduzione

#### 4.1.1 Definizione di equazione non lineare

**Definizione.** Sia  $f : A \rightarrow \mathbb{R}$ , con  $A \subseteq \mathbb{R}$ , una funzione continua almeno su un certo intervallo e si supponga che  $f(x)$  non sia della seguente forma

$$f(x) = a_1x + a_0$$

con  $a_0, a_1$  costanti. La relazione  $f(x) = 0$  è l'*equazione non lineare* nell'incognita  $x$ .

#### 4.1.2 Cosa vogliamo fare

Vogliamo *determinare/approssimare* gli zeri di  $f(x)$ , cioè i valori  $\bar{x}$  tali che  $f(\bar{x}) = 0$ . Si consideri che non abbiamo a disposizione metodi diretti (cioè formule che con conti corretti ci restituiscano i risultati esatti), salvo casi molto particolari.

##### 4.1.2.1 Esempio di metodo diretto

Un esempio tra questi è l'*equazione algebrica di grado m*

$$a_m x^m + a_{m-1} x^{m-1} + \cdots + a_1 x + a_0 = 0$$

con  $a_m \neq 0$  ed  $m$  intero  $\geq 2$ . L'equazione possiede  $m$  radici nel campo complesso: queste si possono trovare con metodo diretto soltanto con  $m \leq 4$ .

##### 4.1.2.2 Rilevanza dei metodi iterativi

Abbiamo capito che in buona parte dei casi ricorreremo a metodi iterativi. Per individuare una radice  $\alpha$  procediamo all'applicazione ripetuta di una *funzione di iterazione*

$$x_{n+1} = \phi_n(x_n, x_{n-1}, \dots, x_{n-k+1})$$

dove la funzione  $\phi_n$  dipende dalle  $k$  valutazioni precedenti ( $k \geq 1$ ): si parla per questo di *metodo iterativo a k punti*. La funzione dipende anche da  $f(x)$ , e può cambiare di forma al variare di  $n$ .

#### 4.1.2.3 Convergenza del metodo

La buona riuscita dei nostri calcoli dipende:

- dalla funzione di iterazione scelta;
- dalle prime  $k$  approssimazioni iniziali  $x_0, x_1, \dots, x_{k+1}$  (ricordarsi che la funzione di iterazione ha bisogno di  $k$  valori precedenti per poter restituire una nuova approssimazione).

In caso di buona riuscita avremo la successione  $\{x_n\}$  convergente alla radice  $\alpha$ .

$$\lim_{n \rightarrow +\infty} x_n = \alpha$$

Il calcolo viene arrestato al verificarsi di un opportuno *criterio di arresto*.

#### 4.1.3 Metodo stazionario

**Definizione.** Se la forma di  $\phi_n$  non varia al variare di  $n$  il metodo si dice *stazionario*.

#### 4.1.4 Ordine di convergenza e Fattore di convergenza

**Definizione.** Sia data una successione convergente  $\{x_n\}$ , cioè

$$\lim_{n \rightarrow +\infty} x_n = \alpha$$

poniamo  $e_n = x_n - \alpha$ . Se esiste un valore  $\mathbb{R} \ni P \geq 1$  tale che il seguente limite è finito e diverso da zero

$$\lim_{n \rightarrow +\infty} \frac{|e_{n+1}|}{|e_n|^P} = c$$

allora definiamo  $P$  *ordine di convergenza* e  $c$  *fattore di convergenza*.

#### Considerazioni

1. Il rapporto è tra due moduli, quindi il valore  $c$ , se esiste, è sicuramente positivo.
2. Se la successione  $\{x_n\}$  converge ad  $\alpha$  allora  $\{e_n\}$  converge a zero.
3. Per il punto (2) individuiamo che si ha un rapporto fra due infinitesimi.
4. Se il limite ha valore finito e diverso da zero allora i due infinitesimi sono dello stesso ordine (tendono a zero con la stessa velocità).
5. Se il limite è uguale a zero allora l'infinitesimo al numeratore tende a zero più velocemente di quello al denominatore.
6. Se il limite va a infinito allora l'infinitesimo al denominatore tende a zero più velocemente di quello al numeratore.
7. Maggiore è  $P$ , maggiore è la velocità di convergenza.

**Convergenza lineare** Se  $p = 1$  si parla di *convergenza lineare*. Nel caso in cui si abbia una successione  $\{x_n\}$  convergente possiamo dire sicuro che  $c < 1$ : se avessi, ad esempio,  $c = 2$  significherebbe avere  $e_{n+1} > e_n$ , cioè avere un'amplificazione dell'errore (quindi non stiamo convergendo ad  $\alpha$ ).

#### 4.1.5 Separazione grafica: numero di zeri e intervalli di separazione

**Definizione.** L'intervallo di separazione di uno zero è un intervallo a cui appartiene esclusivamente uno zero della funzione.

Introduciamo la cosiddetta *separazione grafica* per:

- individuare il numero di zeri;
- definire per ogni zero un intervallo di separazione

Si tenga a mente che questo metodo non ha sempre successo: in alcuni casi potrebbe diventare una complicazione esistenziale (per esempio necessità di fare lo studio di funzione visto ad Analisi I), in altri la rappresentazione grafica potrebbe impedirci di apprezzare gli zeri della funzione. Quello che facciamo è porre

$$f(x) = g(x) - h(x)$$

Se  $f(x) = 0$  allora

$$g(x) - h(x) = 0 \implies \boxed{g(x) = h(x)}$$

$$\begin{cases} y = g(x) \\ y = h(x) \end{cases}$$

Risolviamo il sistema appena posto individuando le intersezioni dei due grafici, che sono coppie ordinate  $(x, y)$ . Il tutto è legato ad  $f(x) = 0$  con le componenti  $x$  delle soluzioni del sistema, che coincidono con le soluzioni della equazione  $f(x) = 0$ . Ricapitoliamo.

1. Trovo  $g(x)$  ed  $h(x)$
2. Rappresento graficamente  $g(x)$  e  $h(x)$ .
3. Individuo le intersezioni tra i due grafici, cioè le coppie ordinate  $(x, y)$ .
4. Dalla rappresentazione grafica sono in grado di individuare il numero di zeri, e per ogni zero il relativo intervallo di separazione. Si tenga a mente che le ascisse dei punti di intersezione sono le soluzioni dell'equazione  $f(x) = 0$ .

##### 4.1.5.1 Esempio

Si consideri la seguente equazione

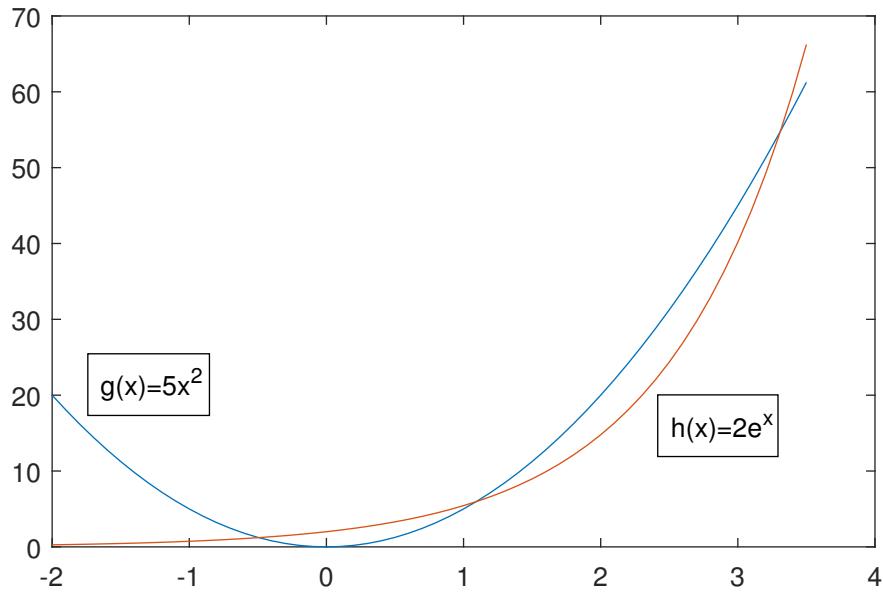
$$5x^2 - 2e^x = 0$$

La separazione più ovvia è  $g(x) = 5x^2$  e  $h(x) = 2e^x$ . Tracciando i grafici individuiamo tre zeri e tre intervalli di separazione

$$\alpha_1 \in [-1, 0] \quad \alpha_2 \in [0.5, 1.5] \quad \alpha_3 \in [3, 3.5]$$

Questi sono intervalli da noi scelti, ma non abbiamo unicità. L'aspetto più controverso dell'esercizio sta nel numero di zeri: tracciando un grafico limitato ai valori  $x \in [0, 2]$  una persona potrebbe essere tentata dal dire che ci sono solo due zeri. Per non sbagliare basta seguire la logica.

- $e^x$  tende a infinito molto più velocemente di  $x^2$ .
- Segue che per questa proprietà l'esponenziale dovrà intersecarsi nuovamente con la parabola, per forza!



La morale dell'esercizio è che bisogna tenere a mente le nozioni di Analisi I, e riflettere sul comportamento delle funzioni.

## 4.2 Metodi iterativi a due punti

### 4.2.1 Metodo di bisezione (non stazionario)

#### 4.2.1.1 Spiegazione

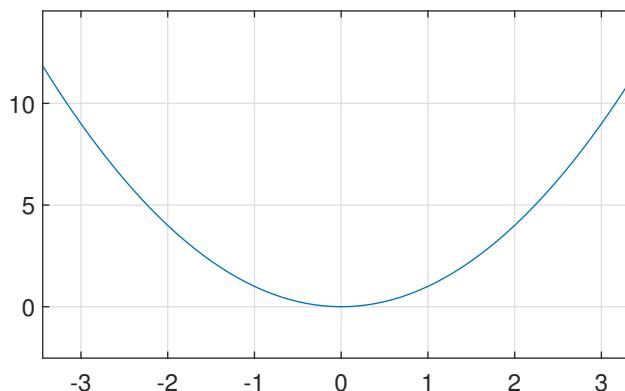
Vogliamo trovare, come al solito, gli zeri della funzione  $f(x)$ , cioè i valori  $\bar{x}$  tali che  $f(\bar{x}) = 0$

**Condizioni** Il metodo è applicabile solo se sono rispettate le seguenti condizioni:

1. continuità della funzione  $f$  nell'intervallo  $[a, b]$ :  $f \in C^0([a, b])$
2. esistenza di uno zero della funzione  $f$  all'interno dell'intervallo  $[a, b]$

$$f(a)f(b) = f(x_0)f(x_1) < 0$$

La condizione (2) è necessaria in quanto il metodo si regge sulla variazione di segno della funzione. Si prenda ad esempio la seguente parabola



Il metodo non funziona perchè lo zero sta sull'ascisse, e non si ha variazione di segno. Il problema non emerge se la soluzione è di molteplicità dispari.

## Passaggi

- Definiamo un intervallo iniziale  $[a, b]$  che contiene almeno uno zero della funzione. Per comodità supporremo che ce ne sia uno solo<sup>1</sup>
- Si assume come approssimazione dello zero l'ascissa del punto medio dell'intervallo  $[a, b]$ . Con questo intendiamo i seguenti calcoli

$$f(x) = 0 \quad \alpha \in [a, b] \quad x_0 = a, x_1 = b \quad x_2 = \frac{a+b}{2} = \frac{x_0+x_1}{2}$$

$x_2$  sarà sicuramente una migliore approssimazione di  $\alpha$  rispetto a  $x_1$  e  $x_0$ .

- I calcoli precedenti ci portano a definire due intervalli:  $[x_0, x_2]$  e  $[x_2, x_1]$ . Per mezzo del seguente prodotto vogliamo determinare in quale dei due intervalli si trova lo zero

$$f(x_1)f(x_2)$$

- Se  $f(x_1)f(x_2) < 0$  allora lo zero si trova nell'intervallo  $[x_2, x_1]$ .
- Se  $f(x_1)f(x_2) > 0$  allora lo zero si trova nell'intervallo  $[x_0, x_2]$ .

In sostanza individuiamo che l'approssimazione  $x_3$  potrà essere una tra le seguenti

$$x_3 = \begin{cases} \frac{x_2+x_1}{2} & f(x_2)f(x_1) < 0 \text{ (siamo nell'intervallo } [x_2, x_1]) \\ \frac{x_2+x_0}{2} & f(x_2)f(x_1) > 0 \text{ (siamo nell'intervallo } [x_0, x_2]) \end{cases}$$

**Detto in modo compatto.** Determiniamo una successione di approssimazioni  $x_1, x_2, x_3, x_4, \dots$  secondo la formula

$$x_{n+1} = \frac{x_n + \hat{x}_n}{2} \quad n = 1, 2, \dots$$

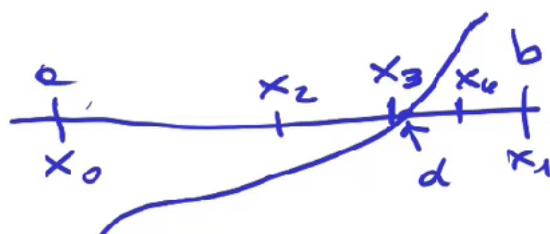
con  $n = 1$  abbiamo  $\hat{x}_1 = x_0$ , mentre con  $n > 1$  poniamo

$$\hat{x}_n = \begin{cases} x_{n-1} & f(x_n)f(x_{n-1}) < 0 \\ \hat{x}_{n-1} & \text{altrimenti} \end{cases}$$

Deduciamo che siamo in presenza di un *metodo non stazionario*, visto che la funzione non è la stessa ad ogni passo (non abbiamo una formula fissa, ma una selezione di uno dei due termini da fare ogni volta).

### 4.2.1.2 Esempio

Prendiamo ad esempio la seguente curva



Vediamo i primi passi del metodo

<sup>1</sup>In presenza di più zeri il metodo privilegerà un particolare zero tra i presenti.

1. Calcoliamo

$$x_2 = \frac{x_1 + \hat{x}_1}{2} = \frac{x_1 + x_0}{2}$$

Abbiamo posto  $\hat{x}_1 = x_0$ , visto che siamo alla prima iterazione.

2. Osserviamo che  $f(x_2)f(x_1) < 0$ , quindi  $\alpha \in [x_2, x_1]$ . Calcoliamo

$$x_3 = \frac{x_2 + \hat{x}_2}{2} = \frac{x_2 + x_1}{2}$$

Abbiamo posto  $\hat{x}_2 = x_1$  visto che  $f(x_2)f(x_1) < 0$ .

$$\hat{x}_2 = \begin{cases} x_1 & f(x_2)f(x_1) < 0 \\ \hat{x}_1 & \text{altrimenti} \end{cases} = \begin{cases} x_1 & f(x_2)f(x_1) < 0 \\ x_0 & \text{altrimenti} \end{cases}$$

3. Osserviamo che  $f(x_3)f(x_2) > 0$ , quindi  $\alpha \in [x_3, x_1]$ . Calcoliamo

$$x_4 = \frac{x_3 + \hat{x}_3}{2} = \frac{x_3 + x_1}{2}$$

Abbiamo posto  $\hat{x}_3 = x_1$  visto che  $f(x_3)f(x_2) > 0$ .

$$\hat{x}_3 = \begin{cases} x_2 & f(x_3)f(x_2) < 0 \\ \hat{x}_2 & \text{altrimenti} \end{cases} = \begin{cases} x_2 & f(x_3)f(x_2) < 0 \\ x_1 & \text{altrimenti} \end{cases}$$

4. Osserviamo che  $f(x_4)f(x_3) < 0$ , quindi  $\alpha \in [x_3, x_4]$ . Calcoliamo

$$x_5 = \frac{x_4 + \hat{x}_4}{2} = \frac{x_4 + x_3}{2}$$

Abbiamo posto  $\hat{x}_4 = x_3$  visto che  $f(x_4)f(x_3) < 0$ .

$$\hat{x}_4 = \begin{cases} x_3 & f(x_4)f(x_3) < 0 \\ \hat{x}_3 & \text{altrimenti} \end{cases} = \begin{cases} x_3 & f(x_4)f(x_3) < 0 \\ x_1 & \text{altrimenti} \end{cases}$$

#### 4.2.1.3 Criterio di arresto e numero di iterazioni utili

Quando ci fermiamo? Quando otterremo il livello di precisione desiderato. Stabiliamo un errore  $E$  e affermiamo che

$$|\hat{x} - \alpha| < E$$

Abbiamo stabilito il limite alla bontà dell'approssimazione. La cosa molto interessante di questo metodo è che possiamo determinare prima di iniziare il numero di iterazioni necessarie per ottenere il livello di precisione stabilito.

$$\frac{1}{2^n}(b - a) < E \rightarrow \frac{b - a}{E} < 2^n \rightarrow \log_2 \frac{b - a}{E} < n$$

segue  $n = \lceil \log_2 \frac{b-a}{E} \rceil$ . La formula iniziale è stata ottenuta ricordando che ad ogni iterazione l'intervallo considerato si dimezza

$$b - a \rightarrow \frac{1}{2}(b - a) \rightarrow \frac{1}{2^2}(b - a) \rightarrow \frac{1}{2^3}(b - a) \rightarrow \dots \rightarrow \frac{1}{2^n}(b - a)$$

#### 4.2.1.4 Convergenza lineare

Il metodo di bisezione ha *convergenza lineare*.

$$\frac{|e_{n+1}|}{|e_n|} \simeq \frac{|x_{n+1} - x_n|}{|x_n - x_{n-1}|} = \frac{1}{2}$$

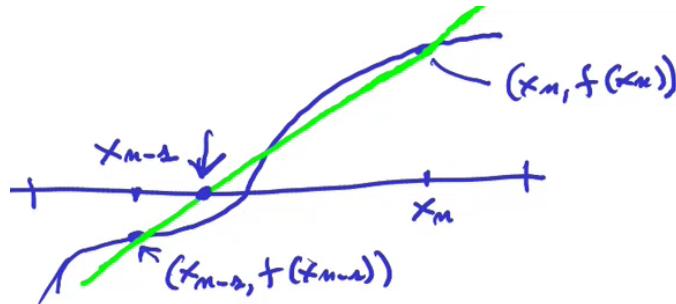
con fattore di convergenza  $\frac{1}{2}$ . Convergenza lenta, di solito si usa questo metodo per ottenere una prima approssimazione (riduzione di un intervallo inizialmente molto ampio) e successivamente passare ad algoritmi con velocità di convergenza maggiore.

#### 4.2.2 Metodo delle secanti (stazionario)

Introduciamo il *metodo delle secanti*, che si basa su tutt'altra strategia.

**Condizioni** Si scelgono due approssimazioni iniziali di  $\alpha$  senza particolari condizioni. La convergenza, in generale, si realizza se  $x_0$  e  $x_1$  sono "abbastanza vicine" alla radice  $\alpha$ .

**Spiegazione** Il metodo si basa sull'uso di due approssimazioni precedenti  $x_n, x_{n-1}$  per individuarne una nuova  $x_{n+1}$ : geometricamente parlando andremo a tracciare la secante al grafico di  $f(x)$  passante per i punti  $(x_n, f(x_n))$  e  $(x_{n-1}, f(x_{n-1}))$



Ricordiamoci la formula per costruire l'equazione di una retta passante per due punti generici  $(x_1, y_1)$  e  $(x_2, y_2)$

$$\frac{x - x_1}{x_2 - x_1} = \frac{y - y_1}{y_2 - y_1}$$

Applichiamo la formula coi punti detti, e imponiamo  $y = 0$  per cercare gli zeri

$$\begin{cases} \frac{x - x_n}{x_{n-1} - x_n} = \frac{y - f(x_n)}{f(x_{n-1}) - f(x_n)} \\ y = 0 \end{cases}$$

otteniamo

$$x - x_n = -f(x_n) \frac{x_{n-1} - x_n}{f(x_{n-1}) - f(x_n)}$$

spostando  $x_n$  a secondo membro e ponendo  $x = x_{n+1}$  concludiamo

$$x_{n+1} = x_n - f(x_n) \frac{x_{n-1} - x_n}{f(x_{n-1}) - f(x_n)}$$

Si osservi che il rapporto  $\frac{x_{n-1} - x_n}{f(x_{n-1}) - f(x_n)}$  è il reciproco del rapporto incrementale (non della derivata, la derivabilità non è richiesta).

**Metodo stazionario** Il metodo è *stazionario*, la formula è sempre la stessa e le uniche cose che cambiano sono le approssimazioni precedenti considerate (i parametri della funzione).

**Ordine di convergenza** L'ordine di convergenza del metodo delle secanti è irrazionale, e superiore rispetto all'ordine del metodo di bisezione.

$$p = \frac{1 + \sqrt{5}}{2} = 1.618\dots$$

## 4.3 Metodi iterativi stazionari ad un punto

### 4.3.1 Premesse

#### 4.3.1.1 Cosa vogliamo fare

Data l'equazione  $f(x) = 0$  si può costruire una funzione  $\phi(x)$  tale che l'equazione data sia equivalente ad una equazione della forma

$$x = \phi(x)$$

#### 4.3.1.2 Schema iterativo

Supponiamo di avere  $f(x) = 0$  e di lavorare su un certo intervallo  $[a, b]$ . Consideriamo una funzione continua  $g(x) \neq 0, \forall x \in [a, b]$  (contenente gli zeri)<sup>2</sup>. Se mi muovo sull'intervallo  $[a, b]$  il seguente prodotto risulterà nullo

$$g(x)f(x) = 0$$

La funzione  $g$  non si annulla, quindi le eventuali soluzioni sono quelle che danno  $f(x) = 0$ . Sommo  $x$  ad entrambi i membri

$$x + g(x)f(x) = x$$

isolo la  $x$  a primo membro

$$x = x - g(x)f(x)$$

il secondo membro è  $\phi(x)$ ! Per ogni scelta della funzione  $\phi(x)$  si può considerare un metodo iterativo stazionario ad un punto, della forma  $x_{n+1} = \phi(x_n)$ , dove  $n = 0, 1, 2, \dots$

#### 4.3.1.3 Punto fisso della funzione $\phi(\alpha)$

Supponiamo di avere una  $x = \alpha$  tale che  $f(\alpha) = 0$ , avremo

$$x = \alpha - g(\alpha)f(\alpha)$$

se  $f(\alpha) = 0$  allora  $x = \alpha$ , cioè  $\alpha = \phi(\alpha)$ . Definiamo  $\alpha$  *punto fisso della funzione  $\phi(\alpha)$* .

#### 4.3.1.4 Definizione di molteplicità

1. Sia  $\alpha$  una soluzione dell'equazione  $f(\alpha) = 0$ . Se abbiamo quanto segue

$$f(\alpha) = 0 \quad f'(\alpha) = 0 \quad f''(\alpha) = 0 \quad \dots \quad f^{(k)}(\alpha) = 0 \quad f^{(k+1)}(\alpha) \neq 0$$

allora possiamo dire che la molteplicità della soluzione  $\alpha$  è  $k + 1$ . Se la molteplicità è pari allora  $\alpha$  è un punto di tangenza senza cambio di segno.

2. Se  $f(x)$  può essere scritto come  $f(x) = (x - \alpha)^{k+1}g(x)$  e

$$\lim_{x \rightarrow \alpha} g(x) = c \neq 0$$

allora possiamo dire che la molteplicità della soluzione  $\alpha$  è  $k + 1$ .

---

<sup>2</sup>Condizioni simili a quelle viste per i sistemi lineari, ma al posto delle funzioni abbiamo le matrici e invece di una funzione diversa da zero abbiamo la matrice non singolare.

### 4.3.2 Teorema di convergenza locale

**Teorema.** Supponiamo di avere l'equazione  $x = \phi(x)$ , supponiamo che  $\alpha$  sia punto fisso (cioè  $\alpha = \phi(\alpha)$ ), supponiamo che  $\alpha \in I$  (intervallo  $[a, b]$ ), supponiamo che  $\phi \in C^1(I)$  (derivabile con continuità). Supponiamo anche di avere la successione

$$x_{n+1} = \phi(x_n), \quad n = 0, 1, 2, \dots$$

se (C.S.)  $\exists \rho, k \in \mathbb{R}^+$ , con  $k < 1$ , tali che ( $I_\rho$  intervallo di raggio  $\rho$  e centro  $\alpha$ )

$$|\phi'(x)| \leq k, \quad x \in I_\rho = ]\alpha - \rho, \alpha + \rho[ \cap I$$

allora

1.  $x_0 \in I_\rho \implies x_n \in I_\rho, \forall n \in \mathbb{N}$  (conv. locale, se esco da  $I_\rho$  non ho più certezze)
2.  $x_0 \in I_\rho \implies \lim_{n \rightarrow +\infty} x_n = \alpha$
3.  $\alpha$  unico punto fisso in  $I_\rho$  (non ho più di un punto fisso in un intervallo  $\rho$ ).

Emerge la differenza fondamentale rispetto ai metodi iterativi di sistemi lineari, dove in caso di convergenza non influisce il punto iniziale: nei sistemi lineari non è più così!

Convergenza globale (sistemi lineari) vs Convergenza locale (sistemi non lineari)

Dimostriamo il teorema punto per punto

1.  $\mathbf{x}_0 \in I_\rho \implies \mathbf{x}_n \in I_\rho, \forall n \in \mathbb{N}$

Dimostriamo per induzione. Nell'induttiva la proposizione che si vuole dimostrare deve essere valida per un certo  $n$  segnato: dalle ipotesi affermiamo che

$$x_0 \in I_\rho$$

Vogliamo dimostrare che

$$x_n \in I_\rho \implies x_{n+1} \in I_\rho$$

il che equivale a dire (con  $I_\rho = ]\alpha - \rho, \alpha + \rho[ \cap I$ )

$$|x_n - \alpha| < \rho \implies |x_{n+1} - \alpha| < \rho$$

Consideriamo adesso la seguente differenza, e applichiamo il teorema di Lagrange rispetto a un intervallo avente estremi  $x_n$  ed  $\alpha$  (non si può dire appartiene all'intervallo, perchè non sappiamo se  $x_n < \alpha$  o  $x_n > \alpha$ )

$$x_{n+1} - \alpha = \phi(x_n) - \phi(\alpha) = \phi'(\xi_n)(x_n - \alpha)$$

dove  $\xi_n$  appartiene all'intervallo con gli estremi detti. Poniamo i valori assoluti

$$|x_{n+1} - \alpha| = |\phi'(\xi_n)| |x_n - \alpha|$$

Se  $\xi_n \in I_\rho$  allora (con  $k < 1$  posto tra le ipotesi, tale che  $|\phi'(x)| \leq k$ )

$$|x_{n+1} - \alpha| = |\phi'(\xi_n)| |x_n - \alpha| \leq k |x_n - \alpha|$$

Nell'ipotesi induttiva abbiamo detto che si ha una valore strettamente minore di  $\rho$ , inoltre ribadiamo che  $k < 1$ . Segue

$$|x_n - \alpha| < k\rho < \rho$$

Unendo tutto abbiamo

$$|x_{n+1} - \alpha| < \rho$$

Abbiamo dimostrato che quanto valido al valore  $n$ -esimo è valido anche al valore  $(n + 1)$ -esimo.

2.  $\mathbf{x_0} \in I_\rho \implies \lim_{n \rightarrow +\infty} \mathbf{x_n} = \alpha$

Dimostrazione per ricorrenza. Possiamo dire

$$\lim_{n \rightarrow +\infty} x_n = \alpha \implies \lim_{n \rightarrow +\infty} (x_n - \alpha) = 0$$

Nella dimostrazione della prima tesi abbiamo detto che

$$|x_{n+1} - \alpha| \leq k|x_n - \alpha|$$

se decrementiamo  $n$  possiamo dire

$$|x_n - \alpha| \leq k|x_{n-1} - \alpha|$$

unendo le due relazioni otteniamo

$$0 \leq |x_{n+1} - \alpha| \leq k|x_n - \alpha| \leq k^2|x_{n-1} - \alpha| \leq \dots \leq k^{n+1}|x_0 - \alpha|$$

$|x_0 - \alpha|$  è un valore costante, che non cambia all'aumentare di  $n$ , mentre

$$\lim_{n \rightarrow +\infty} k^{n+1} = 0$$

poichè  $k < 1$ . Per il teorema dei carabinieri abbiamo dimostrato

$$\lim_{n \rightarrow +\infty} (x_n - \alpha) = 0$$

3.  $\alpha$  **unico punto fisso in  $I_\rho$ .**

Dimostrazione per assurdo. Supponiamo di avere un altro punto fisso  $\beta$  ( $\beta \neq \alpha$ ) tale che

$$\phi(\beta) = \beta$$

se applichiamo (anche qua) il teorema di Lagrange otteniamo (intervallo avente estremi  $\alpha$  e  $\beta$ )

$$\alpha - \beta = \phi(\alpha) - \phi(\beta) = \phi'(\xi)(\alpha - \beta)$$

visto che  $\xi \in I_\rho$  possiamo dire

$$|\alpha - \beta| = |\phi(\alpha) - \phi(\beta)| = |\phi'(\xi)| |\alpha - \beta| \leq k |\alpha - \beta|$$

visto che  $k < 1$  allora

$$k |\alpha - \beta| < |\alpha - \beta|$$

abbiamo trovato l'assurdo

$$|\alpha - \beta| < |\alpha - \beta|$$

che ci permette di stabilire l'unicità del punto fisso.

### 4.3.3 Teorema sull'ordine di convergenza

**Teorema.** Consideriamo uno schema iterativo convergente

$$x_{n+1} = \phi(x_n), \quad n = 0, 1, 2, \dots$$

e un punto fisso  $\alpha$  ( $\alpha = \phi(\alpha)$ ). Se possiamo dire che  $\phi \in C^P(I)$ , e inoltre

$$\phi(\alpha) = \alpha \quad \phi'(\alpha) = 0 \quad \phi''(\alpha) = 0 \quad \dots \quad \phi^{(P-1)}(\alpha) = 0 \quad \phi^{(P)}(\alpha) \neq 0$$

allora  $P \in \mathbb{Z}$  è l'ordine di convergenza. **Vale anche il contrario** (dato  $P$  ordine di convergenza individuiamo il comportamento precedentemente descritto).

Prendiamo la differenza  $x_{n+1} - \alpha = \phi(x_n) - \phi(\alpha)$ . Prima di proseguire dobbiamo reintrodurre lo *sviluppo di Taylor* visto ad Analisi I

$$\phi(x_n) = \phi(\alpha) + \phi'(\alpha)(x_n - \alpha) + \phi''(\alpha) \frac{(x_n - \alpha)^2}{2!} + \dots + \phi^{(P-1)}(\alpha) \frac{(x_n - \alpha)^{P-1}}{(P-1)!} + \phi^{(P)}(\sigma_n) \frac{(x_n - \alpha)^P}{P!}$$

dove si pone il *resto nella forma di Lagrange*:  $\phi^{(P)}(\sigma_n) \frac{(x_n - \alpha)^P}{P!}$ . Spostiamo  $\phi(\alpha)$  nel primo membro

$$\phi(x_n) - \phi(\alpha) = \phi'(\alpha)(x_n - \alpha) + \phi''(\alpha) \frac{(x_n - \alpha)^2}{2!} + \dots + \phi^{(P-1)}(\alpha) \frac{(x_n - \alpha)^{P-1}}{(P-1)!} + \phi^{(P)}(\sigma_n) \frac{(x_n - \alpha)^P}{P!}$$

Sostituiamo nell'equazione iniziale

$$x_{n+1} - \alpha = \phi'(\alpha)(x_n - \alpha) + \phi''(\alpha) \frac{(x_n - \alpha)^2}{2!} + \dots + \phi^{(P-1)}(\alpha) \frac{(x_n - \alpha)^{P-1}}{(P-1)!} + \phi^{(P)}(\sigma_n) \frac{(x_n - \alpha)^P}{P!}$$

Ipotizziamo che  $\phi'(\alpha) = 0, \phi''(\alpha) = 0, \dots, \phi^{(P-1)}(\alpha) = 0$  e  $\phi^{(P)}(\alpha) \neq 0$ . Otterremo

$$x_{n+1} - \alpha = \phi^{(P)}(\sigma_n) \frac{(x_n - \alpha)^P}{P!}$$

Per avere l'ordine di convergenza dobbiamo verificare il limite della definizione

$$\lim_{n \rightarrow +\infty} \frac{|x_{n+1} - \alpha|}{|x_n - \alpha|^P} = c \neq 0$$

riconduciamo lo sviluppo di Taylor semplificato a una forma col rapporto nel limite

$$\frac{x_{n+1} - \alpha}{(x_n - \alpha)^P} = \phi^{(P)}(\sigma_n) \frac{1}{P!}$$

poniamo i valori assoluti e calcoliamo il limite

$$\lim_{n \rightarrow +\infty} \frac{|x_{n+1} - \alpha|}{|x_n - \alpha|^P} = \lim_{n \rightarrow +\infty} \frac{\phi^{(P)}(\sigma_n)}{P!}$$

$\sigma_n$  è compreso tra  $\alpha$  ed  $x_n$  e sicuramente converge ad  $\alpha$ : se la derivata  $P$ -esima è continua allora potremo dire che

$$\lim_{n \rightarrow +\infty} \frac{\phi^{(P)}(\sigma_n)}{P!} = \frac{\phi^{(P)}(\alpha)}{P!} \neq 0$$

**Vale anche il contrario** Prendiamo lo sviluppo di Taylor e supponiamo di avere a un certo punto  $\phi^{(i)}(\alpha) \neq 0$ , con  $1 \leq i < P$ . Se succede questo allora l'ordine di convergenza non è  $P$ , ma  $i$ !!!

#### 4.3.4 Criterio di arresto

Come criterio di arresto dell'algoritmo iterativo si assume la condizione

$$|x_{n+1} - x_n| \leq E$$

#### 4.3.5 Metodo di Newton (o *metodo delle tangenti*)

##### 4.3.5.1 Introduzione

Il metodo iterativo ad un punto principale è il cosiddetto *metodo di Newton*. Si consideri un punto fisso  $x$  ( $f(x) = 0$ ). Sappiamo che

$$x = \phi(x) \quad \phi(x) = x - g(x)f(x)$$

dove  $g(x) \neq 0, \forall x \in I$ . Poniamo

$$g(x) = \frac{1}{f'(x)} \longrightarrow \phi(x) = x - \frac{f(x)}{f'(x)}$$

a cui segue lo schema del metodo

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

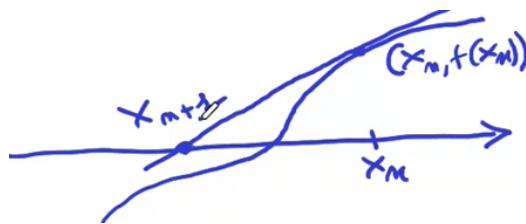
**Confronto col metodo delle secanti** Ricordarsi lo schema del metodo delle secanti

$$x_{n+1} = x_n - f(x_n) \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}$$

Si osservi che qua abbiamo il reciproco del rapporto incrementale, mentre nel metodo di Newton abbiamo il reciproco della derivata (che ricordiamo essere il limite del rapporto incrementale).

##### 4.3.5.2 Interpretazione grafica

Consideriamo il grafico della funzione  $f(x)$



L'equazione per ottenere la retta passante per un punto è la seguente

$$y - y_1 = m(x - x_1)$$

dove  $(x_1, y_1)$  è il punto e  $m$  è il coefficiente angolare. Tra le rette ci interessa quella tangente alla funzione  $f(x)$  nel punto  $(x_n, f(x_n))$ .

$$y - f(x_n) = f'(x_n)(x - x_n)$$

siamo alla ricerca degli zeri, quindi imponiamo a sistema  $y = 0$

$$\begin{cases} y - f(x_n) = f'(x_n)(x - x_n) \\ y = 0 \end{cases}$$

otteniamo

$$-f(x_n) = f'(x_n)(x - x_n) \longrightarrow f'(x_n)x = f'(x_n)x_n - f(x_n)$$

dividendo per  $f'(x_n)$  otteniamo lo schema del metodo

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

ovviamente  $x = x_{n+1}$ . Per l'interpretazione grafica è noto anche come *metodo delle tangenti*.

#### 4.3.5.3 Ordine di convergenza e convergenza del metodo

Abbiamo detto che

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

supponiamo che  $\alpha = \phi(\alpha) \iff f(\alpha) = 0$ . Due casi da affrontare:

- **Radici  $\alpha$  di molteplicità 1.**

Ricordarsi la definizione di molteplicità per mezzo delle derivate (ovviamente la funzione deve essere derivabile). In questo caso, con molteplicità di 1, otterremo

$$f(\alpha) = 0 \quad f'(\alpha) \neq 0$$

Calcoliamo la derivata

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

chiaramente la funzione deve essere derivabile due volte. Calcoliamo  $\phi'(\alpha)$

$$\phi'(\alpha) = \frac{f(\alpha)f''(\alpha)}{[f'(\alpha)]^2} = 0$$

visto che il numeratore è nullo per  $f(\alpha) = 0$ , inoltre  $f'(\alpha) \neq 0$ .

Per quanto riguarda l'ordine di convergenza, individuiamo, ai sensi del teorema appena dimostrato, che questo sarà minimo 2. Calcoliamo la derivata seconda

$$\phi''(x) = \frac{(f'f'' + ff'')(f')^2 - ff''2f'f''}{(f')^4}$$

poniamo

$$\phi''(\alpha) = \frac{[f'(\alpha)]^3 f''(\alpha)}{[f'(\alpha)]^4} = \frac{f''(\alpha)}{f'(\alpha)}$$

non possiamo dire se  $f''(\alpha) = 0$  o no, quindi ci fermiamo. Segue (ribadiamo)

$$P \geq 2$$

Un'ordine di convergenza superiore al metodo di bisezione e a quello delle secanti.

**Ma il metodo converge con radici semplici?** Ricordiamo la condizione centrale del teorema di convergenza locale

$$|\phi'(x)| \leq k < 1$$

Questo deve valere in un intervallo  $I_\rho = [\alpha - \rho, \alpha + \rho]$ . Sappiamo che  $\phi'(\alpha) = 0$ , ma sappiamo anche che  $\phi \in C^1(I_\rho)$ : non è possibile ottenere un modulo  $|\phi'(x)| > 1$  subito dopo essersi discostati da  $\alpha$ . L'intorno richiesto dal teorema di convergenza locale esiste, e quindi il teorema è applicabile: esiste un punto  $x_0$  appartenente all'intervallo  $I_\rho$  tale che si ha convergenza ad  $\alpha$ . Questo  $x_0$  può essere trovato applicando inizialmente il metodo di bisezione.

• **Radici  $\alpha$  di molteplicità  $\geq 1^3$ .**

Abbiamo già detto in passato che se  $\alpha$  è uno zero di  $f(x)$  avente molteplicità  $s \geq 1$  allora

$$f(x) = g(x)(x - \alpha)^s$$

dove  $g(x) = \frac{f(x)}{(x - \alpha)^s}$  e  $g(\alpha) = \lim_{x \rightarrow \alpha} g(x) \neq 0$  (altrimenti avrei una molteplicità diversa, ricordarsi la definizione di molteplicità col limite). Il metodo di Newton prevede il calcolo della derivata prima, procediamo

$$f'(x) = g'(x)(x - \alpha)^s + g(x)s(x - \alpha)^{s-1} = (x - \alpha)^{s-1} [g'(x)(x - \alpha) + sg(x)]$$

Sostituiamo in  $\phi(x)$

$$\phi(x) = x - \frac{f(x)}{f'(x)} = x - \frac{g(x)(x - \alpha)^s}{(x - \alpha)^{s-1} [g'(x)(x - \alpha) + sg(x)]} = x - \frac{g(x)(x - \alpha)}{g'(x)(x - \alpha) + sg(x)}$$

segue la derivata

$$\phi'(x) = 1 - \frac{[g'(x)(x - \alpha) + g(x)][g'(x)(x - \alpha) + sg(x)] - g(x)(x - \alpha)[\dots]}{[g'(x)(x - \alpha) + sg(x)]^2}$$

se imponiamo  $x = \alpha$  otteniamo

$$\phi'(\alpha) = 1 - \frac{g(\alpha)sg(\alpha)}{s^2g^2(\alpha)} = 1 - \frac{1}{s}$$

Ecco il fattore di convergenza, dipendente dalla molteplicità  $s$  della radice!

- Se  $s = 1$  otteniamo  $\phi'(\alpha) = 0$ , quindi si ribadisce quanto visto nella pagina precedente sulle radici di molteplicità 1.
- Se  $s > 1$  otteniamo  $\phi'(\alpha) > 0$ , ma anche  $\phi'(\alpha) < 1$  (visto che  $\frac{1}{s}$  si riduce all'aumentare di  $s$ ).

Segue che con  $s > 1$  l'ordine di convergenza è sicuramente 1

$$\boxed{\phi'(\alpha) \neq 0 \implies P = 1}$$

L'ordine di convergenza è 1! Si ricordi che i metodi precedenti vanno in crisi con soluzioni aventi particolari molteplicità (per esempio con radici di molteplicità pari), mentre il metodo di Newton funziona con radici di qualunque molteplicità. Tuttavia la molteplicità uguale ad 1 per radici di molteplicità  $> 1$  ci segnala una maggiore lentezza del metodo nell'individuare queste radici.

**Ma si ha convergenza con radici di molteplicità  $> 1$ ?** Consideriamo  $s > 1$

- Sappiamo che  $\phi'(\alpha) > 0$  (se  $s \neq 1$ ), ma anche che  $\phi'(\alpha) < 1$ .
- Se ci discostiamo da  $\alpha$  non sarà possibile avere immediatamente  $|\phi'(x)| > 1$  in virtù della continuità della funzione.
- Seguono le stesse conclusioni dette per le radici di molteplicità 1.

---

<sup>3</sup>Mettiamo  $\geq 1$  perchè rientra in questa categoria anche il caso precedente, che abbiamo affrontato a parte semplificandoci la vita.

#### 4.3.5.4 Variante: metodo con convergenza quadratica

Se si conosce la molteplicità  $s$  della radice  $\alpha$  e questa è  $> 1$  allora è possibile alterare il metodo di Newton nel seguente modo

$$x_{n+1} = x_n - s \frac{f(x_n)}{f'(x_n)}, \quad n = 0, 1, \dots$$

**Cosa cambia nei calcoli fatti?** Nel calcolo di  $\phi'(\alpha)$  si otterebbe

$$\phi'(\alpha) = 1 - s \frac{g(\alpha)sg(\alpha)}{s^2 g^2(\alpha)} = 1 - s \frac{1}{s} = 0$$

quindi derivata prima in  $\alpha$  nulla. Si avrà, con radici di molteplicità  $> 1$ , un ordine di convergenza  $P \geq 2$ .

#### 4.3.5.5 Costo computazionale

Se prendiamo la formula

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

osserviamo che dobbiamo valutare due funzioni ad ogni iterazione:  $f(x), f'(x)$ . Questo va in contrasto coi metodi precedenti, dove l'iterazione chiedeva di valutare una sola funzione.

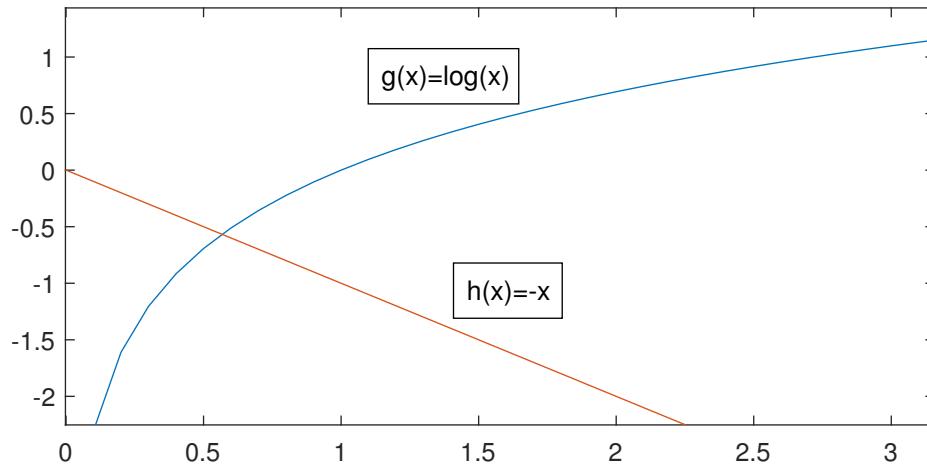
Coperta corta, da una parte si converge più velocemente, dall'altra iterazione più costosa.

#### 4.3.5.6 Primo esempio su Matlab

Separiamo le radici dell'equazione seguente e applichiamo il metodo di Newton ricorrendo a una funzione su Matlab

$$x + \log x = 0$$

Otteniamo come formule separate  $g(x) = \log(x)$  e  $h(x) = -x$ .



Per mezzo della rappresentazione grafica affermiamo che la radice è compresa tra 0.5 e 0.6. Calcoliamo la radice applicando il metodo di Newton, ricorrendo a una funzione Matlab fornita dal professore

```

function [X,i,diff]=newt(f,f1,x0,tol,Nmax)
X=x0;
i=1;
diff=2*tol;
while(diff>tol & i<Nmax)
    X(i+1)=X(i)-f(X(i))/f1(X(i));
    diff=abs(X(i+1)-X(i));
    i=i+1;
end;

```

dove:

- $f$  è la funzione (in questo caso  $x + \log(x)$ );
- $f1$  è la derivata di  $f$  (in questo caso  $1 + \frac{1}{x}$ );
- $x0$  è il punto iniziale
- $tol$  è la tolleranza dell'errore;
- $Nmax$  è un numero di passi massimo da noi imposto.

La funzione restituisce:

- un array  $X$  con le varie approssimazioni calcolate;
- uno scalare  $i$  indicante il numero di passi compiuti;
- uno scalare  $diff$  che consiste nella differenza tra le ultime due approssimazioni.

Poniamo

```

f=@x x+log(x)
f1=@x 1+1./x
[X,i,diff]=newt(f,f1,0,1e-6,80)

```

Otterremo

```

>> [X,i,diff]=newt(f,f1,0.5,1e-6,80)

X =
0.5000    0.5644    0.5671    0.5671    0.5671

i =
5

diff =
1.0415e-11

```

Se io ponessi come  $x0$  il numero 1, invece di 0.5, otterrei un risultato simile ma con passi  $i = 6$ .

#### 4.3.5.7 Secondo esempio su Matlab

Adesso applichiamo il metodo di Newton sulla seguente equazione

$$x^3 - x^2 - 5x - 3 = 0$$

Poniamo

```
f=@x x.^3-x.^2-5*x-3
f1=@x 3*x.^2-2*x-5
[X,i,diff]=newt(f,f1,0,1e-6,80)
```

Con punto iniziale  $x_0 = 0$  otteniamo  $-1$  per mezzo di  $i = 21$  iterazioni. La cosa ci inquieta! Poniamo adesso  $x_0 = 4$ : otteniamo  $3$  per mezzo di  $i = 6$  iterazioni. Perchè?

$-1$  è una radice di molteplicità  $s > 1$ , mentre  $3$  è radice semplice

La radice  $-1$  ha molteplicità 2: sfruttiamola per alterare la funzione

```
function [X,i,diff]=newt(f,f1,x0,tol,Nmax)
X=x0;
i=1;
diff=2*tol;
while(diff>tol & i<Nmax)
    X(i+1)=X(i)-f(X(i))/f1(X(i));
    diff=abs(X(i+1)-X(i));
    i=i+1;
end;
```

Se proviamo a ricalcolare la radice  $-1$  otterremo il risultato di prima con  $i = 6$  iterazioni. Questo perchè l'ordine di convergenza è almeno  $P \geq 2$  (grazie alla modifica fatta).

#### 4.3.5.8 Condizioni sufficienti di convergenza

Data l'equazione  $f(x) = 0$  sia  $\alpha$  una soluzione di molteplicità dispari con  $[a, b]$  intervallo di separazione per  $\alpha$ , inoltre sia  $f \in C^2([a, b])$ . Se  $f'(x)$  e  $f''(x)$  sono di segno costante in  $[a, b]$  allora il metodo di Newton converge scegliendo come punto iniziale un valore  $x_0 \in [a, b]$  tale che

$$f(x_0)f''(x_0) > 0$$

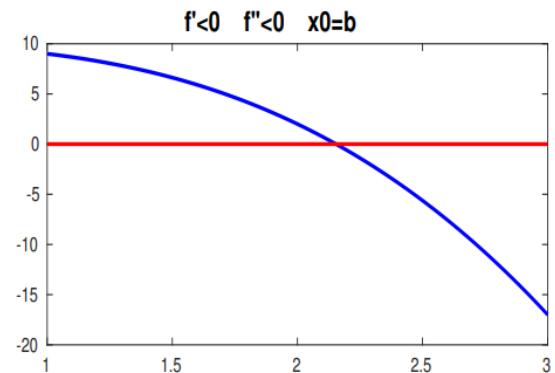
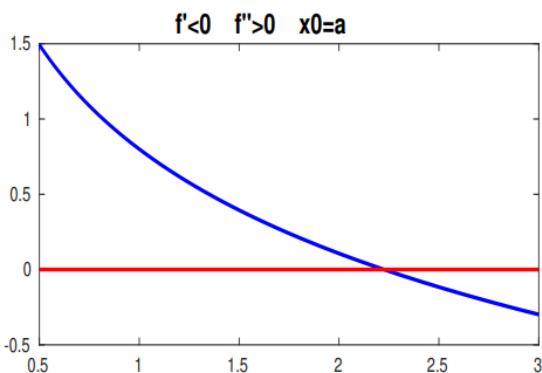
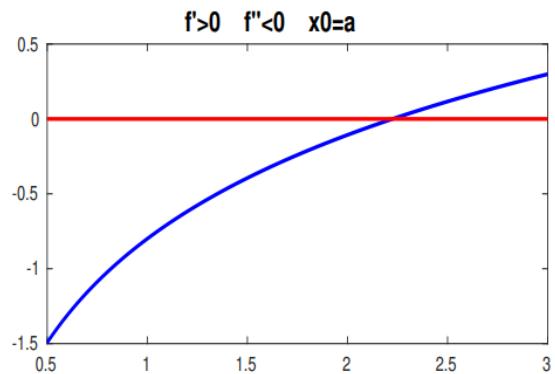
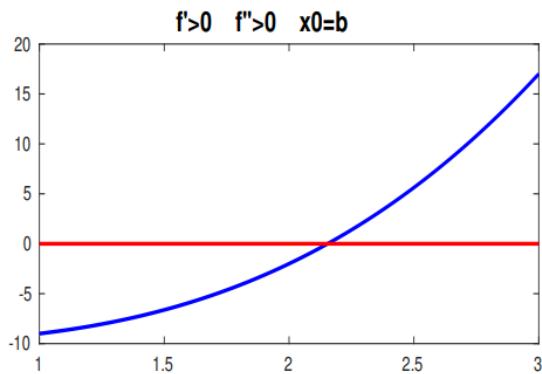
La cosa può essere dimostrata "graficamente". Ricordarsi che:

- la derivata prima positiva segnala andamento crescente della funzione, la derivata prima negativa segnala andamento decrescente;
- la derivata seconda positiva segnala concavità della funzione rivolta verso l'alto, la derivata seconda negativa segnala concavità rivolta verso il basso.

I casi possibili sono i seguenti (ci limitiamo a dire che derivata prima e seconda devono avere segno costante nell'intervallo  $[a, b]$ , nient'altro):

- |                               |                               |
|-------------------------------|-------------------------------|
| 1. $f'(x) > 0$ e $f''(x) > 0$ | 3. $f'(x) < 0$ e $f''(x) > 0$ |
| 2. $f'(x) > 0$ e $f''(x) < 0$ | 4. $f'(x) < 0$ e $f''(x) < 0$ |

Tracciamo una funzione  $f$  per ognuno di questi casi:



Per ogni caso abbiamo un punto  $x_0$  scelto in modo che  $f(x_0)f''(x_0) > 0$ . Quando la condizione è soddisfatta il metodo delle tangenti genera una successione monotonica

- decrescente quando si sceglie  $x_0 = b$
- crescente quando si sceglie  $x_0 = a$

Le successioni sono limitate inferiormente o superiormente da  $\alpha$ . Se scegliamo un altro punto (ad esempio l'altro estremo) corriamo il rischio di intersecare l'asse delle ascisse fuori dall'intervallo  $[a, b]$ , su cui non abbiamo certezze.

#### 4.3.6 Primo esempio

Riprendiamo l'equazione (ricordarsi che per il prof log è in base 2, Log è in base 10)

$$x + \log x = 0$$

Studiamo la convergenza dei metodi iterativi seguenti:

1.  $x_{n+1} = -\log(x_n);$
2. metodo di Newton.

**Risoluzione** Ricordarsi che abbiamo solo condizioni sufficienti: questo significa che se le condizioni da noi conosciute non sono verificate allora non abbiamo idee chiare, e non è conveniente adottare il metodo.

## 1. Convergenza del primo metodo.

Il primo metodo da dove viene fuori? Si osservi che questo metodo è stato ricavato banalmente, isolando il termine  $x$

$$x + \log x = 0 \longrightarrow x = -\log x \longrightarrow x_{n+1} = -\log(x_n)$$

La funzione di iterazione  $\phi(x)$  è la seguente

$$\phi(x) = -\log(x)$$

questo significa che avremo come derivata

$$\phi'(x) = -\frac{1}{x}$$

Abbiamo detto (dalla separazione grafica vista precedentemente) che  $\alpha \in ]0.5, 0.6[$ .

Ricorriamo al teorema di convergenza locale, sappiamo che

$$|\phi'(x)| \leq k < 1, \quad x \in ]\alpha - \delta, \alpha + \delta[$$

Il modulo della derivata, ovviamente, è

$$|\phi'(x)| = \frac{1}{x}$$

Si osservi che non abbiamo valori  $x$  per cui  $\frac{1}{x} < 1$ . Conclusione: non abbiamo certezze sulla convergenza, per sicurezza non adotteremo questo metodo (ricordiamo che il teorema di convergenza locale è solo una C.S)

## 2. Convergenza del metodo di Newton.

L'unica soluzione individuata ha logicamente molteplicità dispari. Possiamo applicare la condizione sufficiente di convergenza vista per il metodo di Newton

$$f(x) = x + \log(x) \longrightarrow f'(x) = 1 + \frac{1}{x} \longrightarrow f''(x) = -\frac{1}{x^2}$$

Occhio al segno di  $f(x)$  ed  $f''(x)$  nell'intervallo, dobbiamo porre un punto  $x_0$  tale che

$$f(x_0)f''(x_0) > 0$$

Di solito si guarda agli estremi come papabili  $x_0$ : se poniamo  $x_0 = 0.5$  la condizione è soddisfatta e si ha convergenza col metodo di Newton.

## 4.4 Metodi iterativi in $\mathbb{R}^n$

### 4.4.1 Introduzione

#### 4.4.1.1 Funzione su più variabili

La teoria dei metodi iterativi, precedentemente esposta, può essere estesa alle funzioni non lineari del tipo

$$f(x) : \mathbb{R}^n \rightarrow \mathbb{R}^n$$

Vogliamo trovare le soluzioni dell'equazione  $f(x) = 0$ . Questo significa risolvere un sistema di  $n$  equazioni in  $n$  incognite:

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

#### 4.4.1.2 Schema iterativo

Lavorando su funzioni non lineari possiamo ricorrere solo a metodi iterativi. Introduciamo una matrice  $G(x) \in \mathbb{R}^{n \times n}$ , che deve avere  $\det(G(x)) \neq 0$  in un insieme di indeterminazione  $D$

$$D = \{x \in \mathbb{R}^n \mid a_i < x_i < b_i, i = 1, \dots, n\}$$

Premoltiplichiamo  $f(x)$  con  $G(x)$

$$G(x)f(x) = 0$$

Non abbiamo alterato le soluzioni, visto che  $\det G \neq 0$  nell'intervallo. Aggiungiamo  $x$  a primo e secondo membro

$$x + G(x)f(x) = x$$

Isoliamo la  $x$  a primo membro, segue

$$x = x - G(x)f(x)$$

Abbiamo ottenuto uno schema analogo a quello scalare, ma si lavora con matrici e vettori. Analogamente

$$x = \phi(x) \longrightarrow x^{(k+1)} = \phi(x^{(k)})$$

con  $k = 0, 1, 2, \dots$   $x \in \mathbb{R}^n$ ,  $\phi \in \mathbb{R}^n$ ,

### 4.4.2 Teorema di convergenza locale su $\mathbb{R}^n$

**Teorema.** Se  $\alpha$  è un punto fisso di  $\phi(x)$  condizione sufficiente per la convergenza ad  $\alpha$  del metodo iterativo

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

è che esistano due numeri positivi  $K$  e  $\rho$ , con  $K < 1$ , tali che si abbia

$$\|\phi'(x)\| \leq K, \forall x \in D_\rho = \{x : \|x - \alpha\| \leq \rho\}$$

purchè  $x^{(0)}$  sia scelto in  $D_\rho$ : in tal caso  $\alpha$  è l'unico punto fisso di  $\phi$  in  $D_\rho$ .  $\phi'(x)$  consiste nella matrice Jacobiana.

Non si ragiona più di un intervallo come visto in Analisi I, ma in termini di sfera (avente centro  $\alpha$  e raggio  $\rho$ ) come visto in Analisi II.

#### 4.4.2.1 Matrice jacobiana

Per quanto riguarda  $\phi'(x)$  ricorriamo alla matrice jacobiana, che possiamo definire se esistono continue le derivate parziali prime della funzione  $\phi$

$$\phi'(x) = \begin{pmatrix} \frac{\partial \phi_1}{\partial x_1} & \dots & \frac{\partial \phi_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial \phi_n}{\partial x_1} & \dots & \frac{\partial \phi_n}{\partial x_n} \end{pmatrix}$$

#### 4.4.3 Metodo di Newton-Raphson (Newton su più variabili)

##### 4.4.3.1 Introduzione

Introduciamo il *metodo di Newton* per funzioni  $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ . Se supponiamo che le funzioni  $f_i$  siano derivabili con continuità rispetto a ciascuna variabile possiamo scrivere la matrice Jacobiana.

$$J(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1} & \dots & \frac{\partial f_n}{\partial x_n} \end{pmatrix}$$

Ai fini dell'applicazione del metodo è necessario anche  $J$  sia non singolare nel dominio contenente al suo interno la soluzione  $\alpha$ : invece del reciproco della derivata prima calcoliamo l'inversa della matrice.

**Funzione  $\phi(x)$  e schema iterativo** Poniamo

$$\phi(x) = x - J^{-1}(x)f(x)$$

otteniamo il sistema  $x = x - J^{-1}(x)f(x)$ , che da origine al seguente schema iterativo

$$x^{(k+1)} = x^{(k)} - J^{-1}(x^{(k)})f(x^{(k)})$$

dove  $k = 0, 1, \dots$

##### 4.4.3.2 Costo computazionale

Quante funzioni dovremo valutare per svolgere l'iterazione?

- $n$  funzioni  $f_i(x)$
- $n^2$  derivate parziali per ottenere la matrice  $J$

In aggiunta abbiamo la necessità di calcolare l'inversa della matrice  $J$ .

##### 4.4.3.3 Variante: risoluzione senza il calcolo delle inverse

Abbiamo detto che il costo computazionale dipende soprattutto dalle operazioni attorno alla matrice  $J$ : calcolo delle derivate parziali e calcolo dell'inversa della matrice stessa.

- La matrice  $J$  deve essere calcolata per forza, ergo non possiamo non valutare le derivate parziali (a meno che non si cambi drasticamente il metodo).
- Il calcolo dell'inversa di  $J$  può essere evitato adottando una semplice variante del metodo.

Prendiamo la formula trovata prima

$$x^{(k+1)} = x^{(k)} - J^{-1}(x^{(k)})f(x^{(k)})$$

premultiplichiamo per  $J(x^{(k)})$  entrambi i membri

$$J(x^{(k)})x^{(k+1)} = J(x^{(k)})x^{(k)} - f(x^{(k)})$$

spostiamo il termine del secondo membro con  $x^{(k)}$  nel primo membro e raccogliamo

$$J(x^{(k)})(x^{(k+1)} - x^{(k)}) = -f(x^{(k)})$$

poniamo  $d^{(k)} = x^{(k+1)} - x^{(k)}$

$$\boxed{J(x^{(k)})d^{(k)} = -f(x^{(k)})}$$

in sostanza abbiamo trovato un sistema lineare da risolvere, cosa pesante ma migliore del calcolo dell'inversa. Trovato  $d^{(k)}$  calcoliamo  $x^{(k+1)}$  con una semplice somma

$$x^{(k+1)} = x^{(k)} + d^{(k)}$$

**Quindi** Per ogni iterazione dobbiamo risolvere un sistema lineare.

#### 4.4.3.4 Variante: metodo di Newton semplificato

Il metodo di Newton semplificato prevede che la matrice Jacobiana venga calcolata una sola volta, alla prima iterazione

$$\boxed{J(x^{(0)})d^{(k)} = -f(x^{(k)})}$$

con  $k = 0, 1, \dots$ . Chiaramente  $x^{(0)}$  deve essere una "buona" approssimazione, e chiaramente non aggiornare la matrice peggiora la velocità del metodo (si ha convergenza lineare).

**Osservazione** Abbiamo *in tempi diversi* sistemi lineari aventi matrice dei coefficienti comune

$$\begin{aligned} J(x^{(0)})d^{(0)} &= -f(x^{(0)}) \\ J(x^{(0)})d^{(1)} &= -f(x^{(1)}) \\ &\vdots \\ J(x^{(0)})d^{(n)} &= -f(x^{(n)}) \end{aligned}$$

Con "in tempi diversi" si intende il fatto che non otterremo il sistema  $k$ -esimo fino a quando non avremo l'approssimazione  $x^{(k-1)}$ . Si applica Gauss una volta soltanto: per trovare da  $J(x^{(0)})$  una corrispondente matrice triangolare superiore.

**Vie di mezzo su Matlab** L'apposita procedura su Matlab prevede, tra i parametri di ingresso, la possibilità di scegliere il ritmo di aggiornamento della matrice Jacobiana:

- sempre (Newton-Raphson);
- mai (variante Newton semplificato);
- ogni tot passi (via di mezzo).

#### 4.4.4 Metodo non lineare di Jacobi-Newton

##### 4.4.4.1 Spiegazione

Vogliamo evitare a tutti i costi il calcolo della matrice  $J(x)$ . Facciamolo calcolando l' $i$ -esima equazione del sistema  $f(x) = 0$  come una equazione nella sola incognita  $x_i$ , ed applicare a ciascuna equazione del sistema il metodo di Newton "classico". La differenza sta nella notazione delle derivate (qua si parla di derivate parziali)

$$x_i^{(k+1)} = x_i^{(k)} - \frac{f_i(x_1^{(k)}, \dots, x_n^{(k)})}{\left(\frac{\partial f_i(x_1^{(k)}, \dots, x_n^{(k)})}{\partial x_i}\right)}$$

con  $i = 1, \dots, n$  e  $k = 0, 1, \dots$

**Osservazione** Ad ogni passo  $(k+1)$ -esimo facciamo uso dell'intero vettore  $x^{(k)}$ . Si ripensi al metodo di Jacobi, detto *metodo delle sostituzioni simultanee*.

##### 4.4.4.2 Costo computazionale

Il costo computazionale è sicuramente migliore: col calcolo di  $J$  dovevamo calcolare  $n^2$  derivate parziali. Adesso valutiamo due funzioni per iterazione, ergo abbiamo  $2n$  funzioni totali da considerare.

##### 4.4.4.3 Variante: metodo non lineare di Gauss-Seidel

Una variante del metodo precedente, ispirata dal *metodo di Gauss-Seidel*, prevede il calcolo delle funzioni non più sul vettore del passo precedente  $x^{(k)}$

$$x_i^{(k+1)} = x_i^{(k)} - \frac{f_i(x_1^{(k+1)}, \dots, x_{i-1}^{(k+1)}, x_i^{(k)}, \dots, x_n^{(k)})}{\left(\frac{\partial f_i(x_1^{(k+1)}, \dots, x_{i-1}^{(k+1)}, x_i^{(k)}, \dots, x_n^{(k)})}{\partial x_i}\right)}$$

con  $i = 1, \dots, n$  e  $k = 0, 1, \dots$ . Si calcolano le due funzioni su un vettore avente per componenti:

- le componenti già aggiornate per il passo  $(k+1)$ -esimo  $(x_1^{(k+1)}, \dots, x_{i-1}^{(k+1)})$
- le componenti calcolate al passo  $k$ -esimo, ancora da aggiornare  $(x_i^{(k)}, \dots, x_n^{(k)})$

L'ordine di convergenza è chiaramente inferiore al metodo di Newton-Raphson, ma il costo della singola iterazione è decisamente inferiore.

## 4.5 Zeri di polinomi

### 4.5.1 Equazioni oggetto di studio

Studiamo il caso particolare in cui l'equazione da risolvere sia algebrica e di grado  $m \geq 2$

$$P(x) = a_m x^m + a_{m-1} x^{m-1} + \cdots + a_1 x + a_0 = 0$$

ovviamente  $a_i \in \mathbb{R}, i = 0, 1, \dots, m$  e  $a_m \neq 0$ . A queste equazioni algebriche possiamo applicare tutti i metodi visti nei capitoli precedenti, ma queste equazioni presentano particolari proprietà (e quindi ulteriori strumenti utilizzabili).

### 4.5.2 Successione di Sturm

**Definizione.** Data l'equazione da risolvere

$$P(x) = 0$$

possiamo definire il seguente insieme di polinomi come *successione di Sturm*

$$P_0(x) = P(x) \quad P_1(x) \quad P_2(x) \quad \dots \quad P_k(x) \quad k \leq m$$

se sono verificate le seguenti proprietà (ci muoviamo su un intervallo  $[a, b]$ ):

1.  $P_k(x) \neq 0, \forall x \in [a, b]$  (l'ultimo polinomio non si deve annullare);
2.  $P_i(\alpha) = 0 \implies P_{i-1}(\alpha)P_{i+1}(\alpha) < 0$ , con  $i = 1, 2, \dots, k - 1$  (il che equivale a dire che non è possibile avere due polinomi consecutivi che si annullano, inoltre deve esservi segno discordo);
3.  $P_0(\alpha) = 0 \implies P_1(\alpha)P'_0(\alpha) > 0$  (la soluzione  $\alpha$  deve essere di molteplicità 1).

### 4.5.3 Funzione variazione

**Definizione.** Data una successione di Sturm definiamo  $V(x_0)$  *funzione variazione* la somma delle variazioni di segno nei seguenti polinomi (non si considerano i polinomi  $j$ -esimi dove  $P_j(x_0) = 0$ ), calcolati tutti nel punto  $x_0$

$$P_0(x_0) \quad P_1(x_0) \quad P_2(x_0) \quad \dots \quad P_k(x_0)$$

Si prenda ad esempio il seguente caso, dove calcoliamo nel punto  $x_0$  tutti i polinomi e consideriamo esclusivamente il segno del risultato

$$\begin{aligned} & P_0(x_0) + \\ & P_1(x_0) + \\ & P_2(x_0) - \\ & P_3(x_0) - \\ & P_4(x_0) - \\ & P_5(x_0) + \\ & P_6(x_0) - \end{aligned}$$

Abbiamo tre variazioni di segno, quindi  $V(x_0) = 3$ .

#### 4.5.4 Teorema di Sturm

**Teorema.** Data una successione di Sturm relativa al polinomio  $P(x)$ , il numero degli zeri del polinomio  $P(x)$  nell'intervallo  $a < x \leq b$  è dato da

$$V(a) - V(b)$$

dove  $V(a)$  e  $V(b)$  sono *funzioni variazione*.

##### 4.5.4.1 Successione di Sturm completa, corollario sulle radici

**Corollario.** Se i polinomi della successione di Sturm sono  $m + 1$ , e quindi  $m = k$ , la successione è detta *completa*. In aggiunta, se tutti i polinomi hanno coefficienti dei termini di grado massimo dello stesso segno, allora l'equazione

$$P(x) = 0$$

ha  $m$  radici reali e distinte.

La cosa si dimostra agilmente prendendo gli  $m + 1$  polinomi: per ciascuno di essi si osserva il segno ponendo  $x \rightarrow +\infty$  e  $x \rightarrow -\infty$ .

$-\infty$	$+\infty$
-	+
+	+
-	+
+	+
:	:
-	+
+	+
$m$	0

Si osservi che con  $x \rightarrow +\infty$  non abbiamo variazioni di segno, mentre con  $x \rightarrow -\infty$  abbiamo  $m$  variazioni di segno. Segue

$$V(-\infty) = m \quad V(+\infty) = 0 \quad V(-\infty) - V(+\infty) = m - 0 = m$$

#### 4.5.5 Costruzione della successione di Sturm con Euclide

La successione di Sturm si costruisce ricorrendo all'*algoritmo di Euclide*.

**Divisione di polinomi.** Dati due polinomi  $P(x)$  ed  $S(x)$ , col grado di  $P$  maggiore o uguale del grado di  $S$ . Possiamo dire che esistono  $Q(x)$  ed  $R(x)$ , col grado di  $R$  minore del grado di  $S$ , tali che

$$P(x) = Q(x)S(x) + R(x)$$

**Premessa con l'Algoritmo di Euclide** L'algoritmo di Euclide prevede una successione di divisioni. Si parte con l'obiettivo di trovare

$$\text{M.C.D.}\{P(x), S(x)\}$$

Si divide  $P(x)$  per  $S(x)$  individuando così  $Q(x)$  ed  $R(x)$ . Possiamo affermare che

$$\text{M.C.D}\{P(x), S(x)\} = \text{M.C.D}\{S(x), R(x)\}$$

Costruiamo la successione di divisioni con i passi descritti e ci fermiamo quando  $R(x) = 0$ . A quel punto affermiamo che il MCD consiste nell'ultimo  $R(x) \neq 0$ , inoltre:

- se  $R(x)$  finale è costante allora i polinomi sono primi tra loro (non hanno fattori comuni);
- se  $R(x)$  finale è un polinomio allora i polinomi non sono primi tra loro (e troviamo i valori comuni che annullano  $P(x)$  ed  $S(x)$ )

**Algoritmo per la costruzione di successioni di Sturm** Per costruire una successione di Sturm ricorriamo a una versione modificata dell'algoritmo.

- Polinomi iniziali
  - $P_0$  è il polinomio  $P(x)$  (di grado  $m$ )
  - $P_1$  è la derivata prima del polinomio  $P(x)$   
(di grado  $m - 1$ , derivata di un polinomio di grado  $m$ )
- Per quanto riguarda il polinomio  $P_{r-1}(x)$  lo dividiamo per  $P_r(x)$ .
  - Si prenda il primo rapporto  $P_0/P_1$

$$P_0(x) = Q_1(x)P_1(x) - P_2(x)$$

troviamo  $P_2(x)$  come il resto della divisione  $P_0/P_1$ , cambiato di segno. Il resto trovato sarà di grado al più  $m - 2$ . Dalla formula si trae  $P_2(x)$

$$P_2(x) = Q_1(x)P_1(x) - P_0(x)$$

- Si prenda il rapporto  $P_1/P_2$

$$P_1(x) = Q_2(x)P_2(x) - P_3(x)$$

otteniamo  $P_3(x)$  come il resto di questa divisione, cambiato di segno. Il resto trovato sarà di grado al più  $m - 3$  (è chiaro che ogni volta che si divide il grado diminuisce almeno di 1). Dalla formula si trae  $P_3(x)$

$$P_3(x) = Q_2(x)P_2(x) - P_1(x)$$

- Arriveremo al più a  $P_k(x)$ , dove  $k \leq m$ . Ci fermeremo a un valore  $k < m$  se otteniamo un resto costante. In quel caso avremo

$$P_{k-1}(x) = P_k(x)Q_k(x)$$

Generalizziamo

$$P_{r-1}(x) = P_r(x)Q_r(x) - P_{r+1}(x) \longrightarrow P_{r+1}(x) = P_r(x)Q_r(x) - P_{r-1}(x)$$

con  $r = 1, 2, \dots$

#### 4.5.6 Riflessioni sulla molteplicità delle soluzioni

**Molteplicità delle soluzioni** L'Algoritmo di Euclide modificato restituisce la successione di Sturm, e fornisce il seguente MCD (visto che si parte da  $P_0$  e  $P_1$  dividendo  $P_0/P_1$ )

$$P_k(x) = \text{M.C.D.}\{P(x), P'(x)\}$$

- **Se una soluzione  $\alpha$  ha molteplicità  $> 1$  come sarà il MCD?**

Si ricordi che una soluzione  $\alpha$  ha molteplicità  $s$  in una funzione  $f \in C^s(D)$  se

$$f(\alpha) = 0 \quad f'(\alpha) = 0 \quad f^{(s-1)}(\alpha) = 0 \quad f^{(s)}(\alpha) \neq 0$$

Fortunatamente i polinomi sono  $\in C^\infty(D)$ .

$$P(\alpha) = 0 \quad P'(\alpha) = 0 \quad P^{(s-1)}(\alpha) = 0 \quad P^{(s)}(\alpha) \neq 0$$

Il polinomio si annulla in  $P(\alpha)$  e nelle derivate successive: sicuramente  $P(x)$  è divisibile per  $x - \alpha$  (Teorema di Ruffini), ma anche  $P'(x)$ . Segue che il MCD è un polinomio di almeno grado uno.

Non si ha una successione di Sturm, condizione (3) non soddisfatta

- **Cosa succede se  $P(x)$  e  $P'(x)$  non hanno zeri in comune?**

Se  $P(x)$  e  $P'(x)$  non hanno zeri reali in comune (e quindi non si ha un valore  $\alpha$  con molteplicità  $> 1$ ) allora il MCD sarà una costante.

Inoltre l'algoritmo di Euclide fornisce una successione di Sturm.

Se  $P(x)$  e  $P'(x)$  hanno zeri reali allora non avremo una successione di Sturm. Se poniamo la successione nel seguente modo

$$\frac{P_0(x)}{P_k(x)}, \frac{P_1(x)}{P_k(x)}, \dots, \frac{P_k(x)}{P_k(x)}$$

allora avremo una successione di Sturm con tanti zeri semplici quanti sono gli zeri distinti di  $P(x)$

**Rispetto delle condizioni alla base delle successioni** Consideriamo il caso in cui  $P(x)$  e  $P'(x)$  non hanno zeri in comune. Osserviamo che le proprietà della successione di Sturm sono soddisfatte.

- La proprietà (1) è rispettata perchè  $P_k(x)$  non ha zeri reali.
- La proprietà (2) è rispettata perchè ...

$$P_{r-1}(x) = P_r(x)Q_r(x) - P_{r+1}(x)$$

... se poniamo  $P_r(x)$  uguale a zero otteniamo

$$P_{r-1}(x) = -P_{r+1}(x)$$

- La proprietà (3) è rispettata poichè

$$P'_0(\alpha)P_1(\alpha) = [P'(\alpha)]^2 > 0$$

#### 4.5.7 Uso della successione di Sturm per l'individuazione degli zeri

Una successione di Sturm può essere usata per individuare un intervallo  $[a, b]$  contenente una sola radice reale  $\alpha$  di una equazione algebrica. Quello che si fa è:

1. calcolare la successione di Sturm;
2. applicare il teorema di Sturm per individuare il numero di radici presenti nell'intervallo  $[a, b]$ ;
3. applicare uno dei metodi iterativi noti per approssimare la soluzione  $\alpha$ .

Nel metodo di Newton, ad esempio, poniamo quanto segue

$$x_{n+1} = x_n - \frac{P(x_n)}{P'(x_n)}$$

# Capitolo 5

## Calcolo degli autovalori

### 5.1 Premessa: risoluzione dell'equazione caratteristica

Dalle definizioni introdotte a principio si potrebbe pensare che una buona via per individuare gli autovalori (o una loro approssimazione) sia ricorrere all'equazione caratteristica

$$\det(A - \lambda I) = 0$$

L'approssimazione si otterrebbe per mezzo dei metodi affrontati nei capitoli precedenti, risolvendo il sistema

$$(A - \lambda I)x = 0$$

In realtà questa cosa è fortemente sconsigliata: sia per il costo computazionale, sia per gli inevitabili errori che si introducono nel calcolo dei coefficienti (piccole variazioni potrebbero mutare profondamente gli autovalori, addirittura rendere complessi autovalori reali).

### 5.2 Metodo delle potenze

#### 5.2.1 Teorema del metodo delle potenze

**Teorema.** Sia  $A \in \mathbb{C}^{n \times n}$ , matrice diagonalizzabile e avente autovalore  $\lambda_1$  tale che

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \dots \geq |\lambda_m|$$

in sostanza  $\lambda_1$  deve avere modulo dominante rispetto al modulo degli altri autovalori (modulo massimo, unico ad averlo). Considero un vettore arbitrario  $z^{(0)} \in \mathbb{C}^n - \{0\}$  (non abbiamo particolari vincoli). Calcoliamo la seguente successione

$$\begin{cases} y^{(0)} = z^{(0)} \\ y^{(k)} = Ay^{(k-1)} \quad k = 1, 2, 3, \dots \end{cases}$$

Date tutte queste ipotesi possiamo verificare che...

- Il limite di un vettore, le cui componenti sono divise per una sua componente non nulla, è direttamente proporzionale all'autovettore  $x^{(1)}$ , associato a  $\lambda_1$ .

$$\lim_{k \rightarrow +\infty} \frac{y^{(k)}}{y_j^{(k)}} = x^{(1)} \alpha$$

- Il limite del quoziente di Rayleigh converge all'autovalore  $\lambda_1$

$$\lim_{k \rightarrow +\infty} \frac{y^{(k)^H} A y^{(k)}}{y^{(k)^H} y^{(k)}} = \lambda_1$$

Tanto per fare un inciso: gli algoritmi dei motori di ricerca adottano cose simili.

Prendiamo il vettore  $z^{(0)} = y^{(0)}$ . Se la matrice  $A$  è diagonalizzabile allora le molteplicità algebriche coincidono con quelle geometriche: questo significa avere  $n$  autovettori linearmente indipendenti tra loro. Questi  $n$  autovettori costituiscono una base dello spazio  $\mathbb{C}^n$ : segue che possiamo scrivere  $z^{(0)}$  come una combinazione lineare degli autovettori

$$z^{(0)} = y^{(0)} = c_1 x^{(1)} + c_2 x^{(2)} + \cdots + c_n x^{(n)} = \sum_{i=1}^n c_i x^{(i)}$$

Prendiamo  $y^{(k)}$ , possiamo dire

$$y^{(k)} = A y^{(k-1)} = A^2 y^{(k-2)} = A^3 y^{(k-3)} = \cdots = A^k y^{(0)}$$

La potenza  $k$ -esima! Ecco da dove viene il nome di *metodo delle potenze*. Se  $y^{(0)}$  è combinazione lineare allora

$$y^{(k)} = A^k \sum_{i=1}^n c_i x^{(i)} = \sum_{i=1}^n c_i A^k x^{(i)} = \dots$$

Sappiamo, dalla definizione di autovettore, che

$$A x^{(i)} = \lambda_i x^{(i)}$$

sappiamo inoltre che nel caso di potenze di  $A$  si mantengono gli stessi autovettori, mentre gli autovalori sono le potenze  $k$ -esime degli autovalori di  $A$ . Segue

$$\dots = \sum_{i=1}^n c_i \lambda_i^k x^{(i)}$$

Isoliamo il termine di modulo massimo  $|\lambda_1|$ .

$$y^{(k)} = \lambda_1^k \left( c_1 x^{(1)} + \sum_{i=2}^n c_i \frac{\lambda_i^k}{\lambda_1^k} x^{(i)} \right) = \lambda_1^k (c_1 x^{(1)} + \omega^{(k)})$$

La sommatoria  $\omega^{(k)}$  tende a zero con  $k \rightarrow +\infty$ . Definiamo la componente  $j$ -esima nel seguente modo

$$y_j^{(k)} = \lambda_1^k \left( c_1 x_j^{(1)} + \omega_j^{(k)} \right)$$

Facciamo il rapporto e calcoliamo il limite

$$\lim_{k \rightarrow +\infty} \frac{\lambda_1^k (c_1 x^{(1)} + \omega^{(k)})}{\lambda_1^k (c_1 x_j^{(1)} + \omega_j^{(k)})} = \frac{c_1 x^{(1)}}{c_1 x_j^{(1)}}$$

Se  $c_1 \neq 0$  concludiamo la dimostrazione del primo punto ( $\alpha = \frac{1}{x_j^{(1)}}$ )

$$\dots = \frac{1}{x_j^{(1)}} x^{(1)} = \alpha x^{(1)}$$

Concludiamo dimostrando il secondo punto

$$\lim_{k \rightarrow +\infty} \frac{y^{(k)H} A y^{(k)}}{y^{(k)H} y^{(k)}} = \lim_{k \rightarrow +\infty} \frac{\left( \frac{y^{(k)}}{y_j^{(k)}} \right)^H A \left( \frac{y^{(k)}}{y_j^{(k)}} \right)}{\left( \frac{y^{(k)}}{y_j^{(k)}} \right)^H \left( \frac{y^{(k)}}{y_j^{(k)}} \right)} = \lambda_1$$

Poniamo i rapporti dividendo entrambi i membri per le componenti: se quei rapporti tendono all'autovettore allora il quoziente di Rayleigh convergerà all'autovalore.

**Problema di fondo** Difficile verificare che  $A$  sia diagonalizzabile. Sicuramente le matrici normali sono matrici diagonalizzabili.

**Nota** La convergenza del metodo delle potenze all'autovalore di modulo massimo e all'autovettore associato si può dimostrare anche con  $A$  non diagonalizzabile, purchè valga la condizione detta sugli autovalori.

### 5.2.2 Criterio di arresto

Consideriamo il quoziente di Rayleigh, che come abbiamo già detto ci permette di ottenere un'approssimazione dell'autovalore

$$R(y^{(k)}) = \frac{y^{(k)H} A y^{(k)}}{y^{(k)H} y^{(k)}}$$

Si adotta come criterio di arresto il seguente (dove  $E \in \mathbb{R}^+$ )

$$|R(y^{(k)}) - R(y^{(k-1)})| < E$$

### 5.2.3 Metodo delle potenze normalizzato

La versione normalmente utilizzata è il cosiddetto *metodo normalizzato*: il metodo classico soffre di problemi di *underflow* (nei casi in cui si ottengono vettori  $\bar{x} \approx 0$ ) e *overflow*. Dato un vettore arbitrario  $z^{(0)} \in \mathbb{C} - \{0\}$

$$\begin{cases} y^{(k)} = Az^{(k-1)} & k = 1, 2, \dots \\ z^{(k)} = y^{(k)}/\alpha_k \end{cases}$$

Normalmente la scelta più gettonata per il coefficiente  $\alpha_k$  è la norma seguente

$$\alpha_k = \|y^{(k)}\|_\infty$$

**Perchè si risolve** Il modulo indicato consiste nel modulo massimo tra le componenti del vettore:

1. la componente di modulo massimo risulterà divisa per il modulo stesso, segue che questa avrà modulo 1 (si risolve l'*overflow*);
2. se la componente di modulo massimo ha modulo unitario allora tutti gli altri moduli saranno minori di 1, in ogni caso non avremo un vettore  $\bar{x} \approx 0$  (si risolve l'*underflow*).

Le conclusioni del teorema per il metodo delle potenze si applicano anche nel metodo normalizzato.

### 5.2.4 Estensione del teorema

Il teorema del metodo delle potenze si può estendere al caso più generale dove  $\lambda_1$  ha molteplicità  $r \geq 1$ , cioè

$$\lambda_1 = \lambda_2 = \cdots = \lambda_r \quad |\lambda_1| > |\lambda_{r+1}| \geq \cdots \geq |\lambda_n|$$

si osservi che non è un discorso di primi  $r$  moduli uguali, ma di primi  $r$  autovalori uguali!

**Osservazione** La successione non converge a un autovettore, ma converge a un vettore che fa parte dello spazio generato dagli  $r$  autovettori associati a  $\lambda_1$ . Segue che il vettore ottenuto cambia in base al vettore iniziale  $z^{(0)}$  scelto.

### 5.2.5 Applicazione del metodo alle matrici normali: deflazione

Applichiamo il metodo alle matrici normali: ricordarsi che una matrice è normale se

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

Si osservi la seguente proprietà

**Teorema.** Gli autovettori di una qualunque matrice normale  $A$  sono ortogonali tra loro. Vale a dire: in un contesto dove gli autovettori sono normalizzati (norma unitaria) avremo

$$x^{(i)^H} x^{(j)} = \delta_{ij}$$

dove  $1 \leq i, j \leq n$ . La  $\delta_{ij}$  è la cosiddetta *delta di Kronecker*.

$$\delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$

Quanto affermiamo a breve è valido con matrici normali dove

$$|\lambda_1| > |\lambda_2| > \dots > |\lambda_n|$$

cioè ad ogni passo dobbiamo avere un matrice con modulo dominante.

1. Si parte dalla matrice  $A_0 = A$ . Applichiamo ad essa il metodo delle potenze individuando l'autovalore  $\lambda_1$ .
2. Consideriamo l'autovettore  $x^{(1)}$  e il relativo autovalore  $\lambda_1$ . Calcoliamo la matrice  $A_1$

$$A_1 = A_0 - \lambda_1 x^{(1)} x^{(1)^H}$$

Il prodotto tra vettori è il prodotto tra un vettore riga e un vettore colonna, compatibile con  $A_0$ . Quello che osserviamo è che la matrice  $A_1$  ha:

- come autovettori gli stessi autovettori di  $A$ ;
- come autovalori  $0, \lambda_2, \dots, \lambda_n$ .

**Dimostriamolo.** Post-moltiplichiamo per  $x^{(1)}$ : si osservi che  $x^{(1)^H} x^{(1)} = I$ , segue

$$A_1 x^{(1)} = A_0 x^{(1)} - \lambda_1 x^{(1)} x^{(1)^H} x^{(1)} = A x^{(1)} - \lambda_1 x^{(1)} = 0$$

otteniamo 0 visto che  $\lambda_1$  e  $x^{(1)}$  sono rispettivamente autovalore e autovettore. Poichè

$$A_1 x^{(1)} = \lambda_1 x^{(1)}$$

otteniamo  $\lambda_1 x^{(1)} = 0 \rightarrow \lambda_1 = 0$ . Consideriamo un autovettore  $j \geq 2$

$$A_1 x^{(j)} = A_0 x^{(j)} - \lambda_1 x^{(1)} x^{(1)^H} x^{(j)}$$

Il secondo termine non c'è visto che  $\lambda_1 = 0$ . Otteniamo i rimanenti autovettori

$$A_1 x^{(j)} = A x^{(j)} = \lambda_j x^j$$

3. Calcolo la matrice  $A_2$

$$A_2 = A_1 - \lambda_2 x^{(2)} x^{(2)^H}$$

Possiamo immaginare che questa matrice avrà come autovalori  $0, 0, \lambda_3, \dots, \lambda_n$ . Si applica il metodo delle potenze e si individua l'autovalore  $\lambda_3$ .

4. In generale

$$A_i = A - \sum_{r=1}^i \lambda_r x^{(r)} x^{(r)^H}$$

dove  $i = 1, 2, \dots, n-1$

## 5.3 Metodo di Jacobi per matrici reali e simmetriche

### 5.3.1 Spiegazione

Il *metodo di Jacobi* permette di approssimare tutti gli autovalori di una matrice hermitiana. Per semplicità considereremo solo la matrici  $A$  reali e simmetriche.

**Trasformazioni per similitudine** Il metodo genera una successioni di matrici  $A$  simili tra loro: si attuano trasformazioni per similitudine per mezzo di matrici di rotazione  $G_{rt}$ . Ricordiamone la forma

$$G_{rt} = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & & \cos \phi & & -\sin \phi \\ & & & & 1 & \\ & & & & & \ddots \\ & & & & & & 1 \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{pmatrix}$$

Dove  $r$  e  $t$  sono indici di riga e colonna dove andremo a collocare gli elementi diversi da 0 e 1 (con  $1 \leq r < t \leq n$ ). Si consideri che la matrice di rotazione è una matrice ortogonale

$$G_{rt} G_{rt}^T = G_{rt}^T G_{rt} = I \longrightarrow G_{rt}^{-1} = G_{rt}^T$$

**Passo (metodo classico)** Il metodo di Jacobi è un metodo iterativo in cui al passo  $k$ -esimo si trasforma una matrice  $A_k$  mediante una matrice ortogonale  $G_{rt}^{(k)}$  (ad ogni passo abbiamo una matrice di rotazione con certi indici  $r$  e  $t$ )

$$\begin{aligned} A_1 &= A \\ A_{k+1} &= G_{rt}^{(k)^T} A_k G_{rt}^{(k)} \quad k = 1, 2, \dots \end{aligned}$$

Ad ogni passo lavoriamo su una delle due triangolari (visto che si parla di matrici simmetriche)

1. scelgo gli indici  $r$  e  $t$  individuando l'elemento non diagonale  $a_{rt}^{(k)}$  di modulo massimo;
2. scelgo l'angolo  $\phi$  in modo tale che nella matrice ottenuta dalla trasformazione si abbia

$$a_{rt}^{(k+1)} = 0$$

Siamo certi che ad ogni iterazione andremo a manipolare soltanto gli elementi lungo le righe e le colonne di indici  $r$  e  $t$ .

**Problema** Rendere nulla una componente durante una trasformazione non garantisce che questa rimanga nulla anche nei passi successivi. Questa cosa contribuisce a rallentare la convergenza.

**Formule** Nelle diapositive sono indicate delle formule che permettono di ottenere le singole componenti (nell'implementazione non si andrà mai a effettuare il prodotto matriciale vero e proprio, visto che la matrice è caratterizzata da tantissimi zeri) e il valore  $\phi$ <sup>1</sup>.

### 5.3.2 Teorema di Jacobi

**Teorema.** La successione  $\{A_k\}$  generata con la versione classica del metodo di Jacobi converge alla matrice diagonale

$$D = \text{diag}(\lambda_1, \lambda_2, \lambda_n)$$

dove  $\lambda_1, \dots, \lambda_n$  sono gli autovalori di  $A$ .

Questo teorema rende chiaro l'utilità nel calcolare questa successione.

### 5.3.3 Criterio di arresto

Il criterio di arresto solitamente adottato è il seguente

$$\max_{i>j} |a_{ij}^{(k+1)}| \leq E$$

dove  $E > 0$  è un valore prefissato. Ribadiamo anche qua (lo si vede da  $i > j$ ) che basta lavorare su una delle due triangolari, visto che la matrice è simmetrica.

### 5.3.4 Variante: metodo di Jacobi ciclico

L'elemento più costoso dal punto di vista computazionale è la ricerca della componente avente modulo massimo su tutta la matrice. Quello che facciamo è sopprimere questa ricerca!

1. Ci poniamo su una delle due triangolari (prendiamo quella superiore come esempio)
2. Annulliamo sistematicamente tutti gli elementi non nulli che si incontrano percorrendo per righe gli elementi

$$(12), (13), \dots, (1n), (23), (24), \dots, (2n), \dots, (n-1, n)$$

Ci si muove lungo la diagonale superiore e concludiamo con l'elemento avente indici  $n-1$  ed  $n$ .

Vale anche in questa variante il teorema di Jacobi, tuttavia permane il problema delle componenti nulle modificate in passi successivi.

---

<sup>1</sup>Non vanno sapute, non le ricordo nemmeno io (cit.).

## 5.4 Riduzione in forma tridiagonale e di Hessenberg

### 5.4.1 Introduzione

Se una matrice è tridiagonale, oltre che hermitiana, allora il metodo di Jacobi (ma anche altri metodi) possono essere applicati in modo più agile.

**Quindi** Effettuiamo una trasformazione per similitudine per ottenere, da una generica matrice hermitiana  $A$ , una matrice tridiagonale simile.

### 5.4.2 Premessa: matrice tridiagonale

Una matrice è detta *tridiagonale* se presenta la seguente struttura (prendiamo per comodità una matrice  $\in \mathbb{R}^{4 \times 4}$ )

$$\begin{pmatrix} x & x & 0 & 0 \\ x & x & x & 0 \\ 0 & x & x & x \\ 0 & 0 & x & x \end{pmatrix}$$

Si pongono per definizioni nulle tutte le componenti che non si trovano lungo la diagonale principale, lungo la codiagonale superiore (gli elementi immediatamente accanto alla diagonale principale, nella triangolare superiore) e lungo la codiagonale inferiore.

### 5.4.3 Metodo di Givens per la tridiagonalizzazione

Il *metodo di Givens* prevede ancora l'uso delle matrici di rotazione, ma a differenza di prima siamo certi che se un elemento è stato annullato questo rimarrà nullo in tutte le successive trasformazioni.

#### 5.4.3.1 Applicazione del metodo a matrici simmetriche

Per semplicità lavoreremo anche qua con matrici reali e simmetriche (quindi anche qua lavoriamo solo su uno dei due triangoli).

- **Cosa si fa?**

Si annullano ordinatamente i termini non nulli fra gli elementi  $a_{ij}$  tali che  $i - j \geq 2$  (le codiagonali), considerati per colonne

$$a_{31}, a_{41}, \dots, a_{n1}; a_{42}, a_{52}, \dots, a_{n2}; \dots; a_{n,n-2}$$

Si annullano gli elementi per mezzo delle seguenti matrici di rotazione

$$G_{23}, G_{24}, \dots, G_{2n}; G_{34}, G_{35}, \dots, G_{3n}; \dots; G_{n-1,n}$$

Come trovo gli indici  $< i, r >$  della matrice di rotazione? Dato l'elemento  $a_{ij} \neq 0$ :

- incremento  $j$ ;
- inverti di posizione i due indici

Quindi  $a_{ij} \Rightarrow G_{j+1,i}$ .

- **Perchè si risolve il problema degli elementi nulli?**

Perchè si svolgono combinazioni lineari di elementi nulli: nulli sono e nulli rimangono.

- **Numero di passaggi?**

Abbiamo un metodo che termina dopo un numero preciso di passi! Ad ogni passo si annulla un elemento (e il suo simmetrico visto che siamo in una matrice simmetrica).

- Nella prima colonna annullo  $n - 2$  elementi.
- Nella seconda annullo  $n - 3$  elementi.
- ... ci si muove fino alla  $n$ -esima colonna

Il numero di elementi da annullare è la somma dei primi  $n - 2$  interi. Segue il seguente numero di rotazioni:

$$\frac{(n-2)(n-1)}{2}$$

#### 5.4.3.2 Applicazione del metodo a matrici non simmetriche: Hessenberg

Supponiamo di applicare il metodo di Givens a una matrice non simmetrica, lavorando sulla triangolare inferiore. Il risultato, a seguito di  $(n-2)(n-1)/2$  trasformazioni, è la cosiddetta *matrice di Hessenberg superiore*.

$$H = \begin{pmatrix} h_{11} & h_{12} & \dots & h_{1n} \\ h_{21} & h_{22} & \dots & h_{2n} \\ \ddots & \ddots & \ddots & \vdots \\ & h_{n-1,n} & h_{nn} \end{pmatrix}$$

## 5.5 Metodo QR

### 5.5.1 Fattorizzazione QR

**Teorema.** Per ogni matrice  $A \in \mathbb{R}^{n \times n}$  esiste una fattorizzazione data dal prodotto di una matrice  $Q$  ortogonale per una matrice  $R$  triangolare superiore.

In generale l'approssimazione di tutti gli autovalori viene effettuata per mezza del metodo QR. Per comodità affrontiamo il solo caso con componenti reali (in quel caso  $Q$  matrice unitaria). Dimostriamo costruendo le due matrici.

- Supponiamo di avere una matrice  $A \in \mathbb{R}^{4 \times 4}$

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix}$$

- **Individuazione della matrice  $R$ .**

Partiamo dalla matrice  $R$ , che abbiamo detto essere una matrice triangolare superiore. Effettuiamo trasformazioni per similitudine dove andiamo ad eliminare gli elementi inferiori alla diagonale principale.

1. Prendiamo l'elemento  $a_{21}$  e consideriamo la matrice di rotazione  $G_{12}$  (solo scambio degli indici). Premoltiplichiamo  $A$  con la matrice detta in modo tale da avere  $a_{21} = 0$ , ponendo un opportuno angolo  $\phi$

$$G_{12}A$$

2. Prendiamo l'elemento  $a_{31}$  e consideriamo la matrice di rotazione  $G_{13}$ . Premoltiplichiamo  $G_{12}A$  con la matrice detta in modo tale da avere  $a_{31} = 0$ , ponendo un opportuno angolo  $\phi$

$$G_{13}G_{12}A$$

3. Prendiamo l'elemento  $a_{41}$  e consideriamo la matrice di rotazione  $G_{14}$ . Premoltiplichiamo  $G_{13}G_{12}A$  con la matrice detta in modo tale da avere  $a_{31} = 0$ , ponendo un opportuno angolo  $\phi$

$$G_{14}G_{13}G_{12}A$$

4. Prendiamo l'elemento  $a_{32}$  e consideriamo la matrice di rotazione  $G_{23}$ . Premoltiplichiamo  $G_{14}G_{13}G_{12}A$  con la matrice detta in modo tale da avere  $a_{32} = 0$ , ponendo un opportuno angolo  $\phi$

$$G_{23}G_{14}G_{13}G_{12}A$$

5. Prendiamo l'elemento  $a_{42}$  e consideriamo la matrice di rotazione  $G_{24}$ . Premoltiplichiamo  $G_{23}G_{14}G_{13}G_{12}A$  con la matrice detta in modo tale da avere  $a_{42} = 0$ , ponendo un opportuno angolo  $\phi$

$$G_{24}G_{23}G_{14}G_{13}G_{12}A$$

6. Concludiamo con l'elemento  $a_{43}$  e la matrice di rotazione  $G_{34}$

$$G_{34}G_{24}G_{23}G_{14}G_{13}G_{12}A$$

Svolgiamo  $n(n - 1)/2$  rotazioni (la somma delle trasformazioni fatte è la somma dei primi  $n - 1$  interi). Il risultato è la matrice triangolare superiore  $R$

$$R = G_{34}G_{24}G_{23}G_{14}G_{13}G_{12}A$$

• **Individuazione della matrice  $Q$ .**

Sappiamo che le matrici di rotazione sono matrici ortogonali, quindi l'inversa equivale alla trasposta. Per mezzo di una serie di premoltiplicazioni otteniamo

$$R = G_{34}G_{24}G_{23}G_{14}G_{13}G_{12}A \longrightarrow (G_{12}^T G_{13}^T \dots G_{24}^T G_{34}^T)R = A$$

La proprietà rilevante è che il prodotto di matrici ortogonali restituisce una matrice ortogonale! Abbiamo trovato  $Q$

$$Q = G_{12}^T G_{13}^T \dots G_{24}^T G_{34}^T$$

**Osservazione** Quanto è visto è solo un esempio, non esiste una fattorizzazione univoca!

### 5.5.2 Fattorizzazione come metodo diretto per risoluzione di sistemi

Alcuni libri classificano la fattorizzazione QR come un metodo diretto per la risoluzione di sistemi lineari. Si prenda  $Ax = b$ . Sostituiammo  $A$

$$QRx = b$$

Premoltiplico per l'inversa di  $Q$  (la trasposta, quindi facile da calcolare<sup>2</sup>).

$$Rx = Q^T b$$

---

<sup>2</sup>Semplificazione rispetto alla fattorizzazione LR.

### 5.5.3 Algoritmo del metodo QR

L'algoritmo del metodo QR per l'individuazione di autovalori di una matrice  $A \in \mathbb{R}^{n \times n}$  ricorrere alla fattorizzazione QR precedentemente introdotta. Costruiamo una successione di matrici

$$\begin{aligned} A_1 &= A \\ A_k &= Q_k R_k \\ A_{k+1} &= R_k Q_k \quad k = 1, 2, \dots \end{aligned}$$

Otteniamo la nuova matrice invertendo la posizione dei fattori.

**Perchè si fa questo?** Prendiamo  $A_2 = R_1 Q_1$  e poniamo la matrice identica nel secondo membro.

$$A_2 = I R_1 Q_1 = Q_1^T Q_1 R_1 Q_1$$

Il prodotto  $Q_1 R_1$  è uguale ad  $A_1$ , segue

$$A_2 = Q_1^T A_1 Q_1$$

si osservi la forma del prodotto matriciale: è una trasformazione per similitudine. Segue che le matrici  $A_1$  e  $A_2$  sono simili, quindi hanno gli stessi autovalori. Generalizziamo

$$\begin{aligned} A_k &= Q_k R_k \\ A_{k+1} &= R_k Q_k = Q_k^T Q_k R_k Q_k = Q_k^T A_k Q_k \end{aligned}$$

L'algoritmo genera una successione di matrici simili!

### 5.5.4 Teorema di Schur

**Teorema.** Per ogni matrice  $A \in \mathbb{R}^{n \times n}$  esiste una matrice reale ortogonale  $B$  tale che

$$S = B^{-1}AB = B^T AB$$

consiste in una matrice triangolare a blocchi con i blocchi diagonali di ordine uno o due (detta *matrice di Shur*)

$$S = \begin{pmatrix} S_{11} & S_{12} & \dots & S_{1r} \\ & S_{22} & \dots & S_{2r} \\ & & \ddots & \vdots \\ & & & S_{rr} \end{pmatrix}$$

I blocchi diagonali di ordine 1 sono autovalori reali di  $A$ , mentre i blocchi diagonali di ordine due hanno come autovalori una coppia di autovalori di  $A$  complessi coniugati.

### 5.5.5 Teorema del metodo QR

**Teorema.** Se gli autovalori  $A \in \mathbb{R}^{n \times n}$  sono reali e distinti in modulo e quindi verificano la condizione

$$|\lambda_1| > |\lambda_2| > \dots > |\lambda_n|$$

e se gli autovettori formano una matrice  $X$  tale che  $X^{-1}$  sia fattorizzabile LR, allora le matrici  $A_k$  (per  $k \rightarrow \infty$ ) tendono ad una matrice triangolare superiore e gli elementi diagonali  $a_{ii}^{(k)}$  di  $A_k$  tendono agli autovalori  $\lambda_i$  di  $A$  ordinati per modulo decrescente.

**E se mancasse l'ipotesi di fattorizzabilità?** Non è un dramma: avremo gli autovalori in ordine sparso.

**Se gli autovalori sono complessi?** Abbiamo convergenza a una matrice triangolare superiore **a blocchi** (*matrice di Schur*, teorema precedente). Ricordiamo che i blocchi diagonali di ordine 1 danno gli autovalori reali, mentre i blocchi diagonali di ordine 2 restituiscono autovalori come coppie di complessi coniugati.

**Autovalori distinti, ma con lo stesso modulo?** Si potrebbe applicare la traslazione dello spettro per distinguere gli autovalori.

### 5.5.6 Uso della matrice di Hessenberg

In generale prima di applicare il metodo QR si preferisce condurre la matrice  $A$  nella forma di Hessenberg superiore mediante trasformazioni per similitudine.

$$H = \begin{pmatrix} h_{11} & h_{12} & \dots & h_{1n} \\ h_{21} & h_{22} & \dots & h_{2n} \\ \ddots & \ddots & & \vdots \\ & h_{n-1,n} & h_{nn} \end{pmatrix}$$

L'aspetto rilevante è che tutte le matrici della successione generata col metodo QR risultano tutte matrici di Hessenberg superiori.

# Capitolo 6

## Interpolazione e approssimazione di funzioni

### 6.1 Introduzione

#### 6.1.1 Cosa vogliamo fare

Consideriamo  $k + 1$  punti  $x_0, x_1, x_2, \dots, x_k$ . Questi punti sono a due a due distinti, cioè  $x_i \neq x_j$  se  $i \neq j$ . Consideriamo i valori ottenuti da una presunta funzione  $f$

$$f(x_0), f(x_1), \dots, f(x_k)$$

Noi conosciamo i punti  $x_1, \dots, x_k$  e i valori  $f(x_1), \dots, f(x_k)$ , ma non conosciamo la funzione  $f$ ! Vogliamo individuare la cosiddetta *funzione interpolante*.

#### 6.1.2 Funzione interpolante

**Definizione.** Definiamo  $g(x)$  funzione interpolante di  $f(x)$  se possiamo dire

$$g(x_i) = f(x_i), \quad i = 0, 1, \dots, k$$

Tra le possibili funzioni ci occuperemo, in particolare di polinomi di interpolazione (che adottiamo per le loro caratteristiche, decisive nel calcolo di derivate e integrali di funzioni di interpolazione), cioè di *interpolazione parabolica*.

### 6.2 Interpolazione parabolica

#### 6.2.1 Polinomio di interpolazione

**Definizione.** Definiamo  $P(x)$  polinomio di interpolazione della funzione  $f(x)$  se

$$P(x_i) = f(x_i), \quad i = 0, 1, \dots, k$$

Dell'insieme dei possibili polinomi  $\pi(x)$  consideriamo un sottoinsieme  $\pi_k(x)$ , cioè un insieme dei polinomi fino al grado  $k$ !

$$P_k(x) = a_k x^k + a_{k-1} x^{k-1} + \dots + a_1 x + a_0$$

dove  $a_i \in \mathbb{R}$ , con  $i = 0, \dots, k$ .

## 6.2.2 Matrice di Vandermonde e risoluzione di un sistema

Le condizioni che dobbiamo imporre nell'individuazione di un polinomio di interpolazione sono le seguenti

$$\begin{cases} f(x_0) = a_k x_0^k + a_{k-1} x_0^{k-1} + \cdots + a_1 x_0 + a_0 \\ f(x_1) = a_k x_1^k + a_{k-1} x_1^{k-1} + \cdots + a_1 x_1 + a_0 \\ f(x_2) = a_k x_2^k + a_{k-1} x_2^{k-1} + \cdots + a_1 x_2 + a_0 \\ \vdots \\ f(x_k) = a_k x_k^k + a_{k-1} x_k^{k-1} + \cdots + a_1 x_k + a_0 \end{cases}$$

Un sistema lineare! Distinguiamo la matrice dei coefficienti dal vettore delle incognite e quello dei termini noti

$$\begin{pmatrix} x_0^k & x_0^{k-1} & \cdots & x_0 & 1 \\ x_1^k & x_1^{k-1} & \cdots & x_1 & 1 \\ \vdots & & & & \\ x_{k-1}^k & x_{k-1}^{k-1} & \cdots & x_{k-1} & 1 \\ x_k^k & x_k^{k-1} & \cdots & x_k & 1 \end{pmatrix} \begin{pmatrix} a_k \\ a_{k-1} \\ \vdots \\ a_1 \\ a_0 \end{pmatrix} = \begin{pmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_{k-1}) \\ f(x_k) \end{pmatrix}$$

La matrice trovata è detta *matrice di Vandermonde* (le colonne sono le potenze di un insieme di numeri).

**Problema** Sono facilmente malcondizionate. Segue che non risolveremo il problema di interpolazione per mezzo di questi calcoli

### 6.2.2.1 Grado massimo del polinomio di interpolazione

Il vettore delle incognite è il seguente

$$(a_k \ a_{k-1} \ \cdots \ a_1 \ a_0)^T$$

Si osservi che è possibile avere come soluzione  $a_k = 0$ , oppure  $a_k = 0, a_{k-1} = 0$ , e così via. Questo significa che il polinomio interpolante potrebbe avere grado inferiore al grado  $k$ .

**Conclusione** Con  $k$  punti il polinomio di interpolazione risulta di grado al più  $k - 1$ .

### 6.2.2.2 Unicità del polinomio di interpolazione

Si potrebbe dimostrare (noi non lo facciamo) che il determinante delle matrici di Vandermonde è uguale a

$$|\det(V)| = \prod_{0 \leq i < j \leq k} |x_i - x_j|$$

Segue che avremo matrice non singolare solo se avremo valori  $x_i, x_j$  distinti, a due a due. Se ci si pensa avere valori uguali non avrebbe senso

- se  $x_i = x_j$  allora  $f(x_i)f(x_j)$  (non ha senso cercare due volte lo stesso valore  $f(x_i)$ );
- non è possibile avere  $f(x_i) \neq f(x_k)$  con  $x_i = x_j$  (si rompe la definizione di funzione).

Se la matrice è non singolare allora il polinomio di interpolazione possibile è unico!

### 6.2.3 Interpolazione di Lagrange

#### 6.2.3.1 Polinomio fondamentale di interpolazione

Consideriamo la seguente funzione

$$l_r(x) = \frac{(x - x_0) \dots (x - x_{r-1})(x - x_{r+1}) \dots (x - x_k)}{(x_r - x_0) \dots (x_r - x_{r-1})(x_r - x_{r+1}) \dots (x_r - x_k)}$$

dove  $r = 0, 1, \dots, k$ . La funzione è di grado  $k$ :

- il denominatore è un numero da calcolare;
- il numeratore è un prodotto di polinomi di grado 1, ergo risulterà essere un polinomio di grado  $k$  (dove  $k$  è il numero di valori  $x$  in nostro possesso, ma anche il numero di polinomi di grado 1 al numeratore).

Adesso prendiamo, tra i valori in nostro possesso, un particolare valore  $x_s$

$$l_r(x_s) = \begin{cases} 1 & r = s \\ 0 & r \neq s \end{cases}$$

- Se  $r = s$  al numeratore la  $x$  sarà sostituita da  $x_s$ , quindi potremo semplificare tutti i fattori fino ad ottenere 1.
- Se  $r \neq s$  il numeratore si annullerà in  $(x - x_s) = (x_s - x_s)$ , quindi otterrete 0.

#### 6.2.3.2 Polinomio di interpolazione di Lagrange

Definiamo quello che è effettivamente noto come *polinomio di Lagrange*

$$L_k(x) = \sum_{i=0}^k l_i(x) f(x_i)$$

Si parla di polinomio di interpolazione in quanto

$$L_k(x_s) = f(x_s)$$

con  $s = 0, 1, \dots, k$ .

#### 6.2.3.3 Esempio

Si prenda la seguente tabella di valori

$x$	0	1	2	-1
$y$	1	1	3	3

Si vuole calcolare il polinomio di interpolazione di Lagrange  $L_3(x)$ .

**Risoluzione** Abbiamo da calcolare quattro polinomi fondamentali di interpolazione. Recuperiamo la formula generica

$$l_r(x) = \frac{(x - x_0) \dots (x - x_{r-1})(x - x_{r+1}) \dots (x - x_k)}{(x_r - x_0) \dots (x_r - x_{r-1})(x_r - x_{r+1}) \dots (x_r - x_k)}$$

e scriviamo i polinomi  $l_0(x)$ ,  $l_1(x)$ ,  $l_2(x)$  ed  $l_3(x)$

$$\begin{aligned} l_0(x) &= \frac{(x-1)(x-2)(x-(-1))}{(0-1)(0-2)(0-(-1))} = \frac{(x-1)(x-2)(x+1)}{(-1)(-2)(1)} \\ l_1(x) &= \frac{(x-0)(x-2)(x-(-1))}{(1-0)(1-2)(1-(-1))} = \frac{x(x-2)(x+1)}{(1)(-1)(2)} \\ l_2(x) &= \frac{(x-0)(x-1)(x-(-1))}{(2-0)(2-1)(2-(-1))} = \frac{x(x-1)(x+1)}{(2)(1)(3)} \\ l_3(x) &= \frac{(x-0)(x-1)(x-2)}{(-1-0)(-1-1)(-1-2)} = \frac{x(x-1)(x-2)}{(-1)(-2)(-3)} \end{aligned}$$

Segue

$$L_3(x) = l_0(x)f(x_0) + l_1(x)f(x_1) + l_2(x)f(x_2) + l_3(x)f(x_3) = l_0(x) \cdot 1 + l_1(x) \cdot 1 + l_2(x) \cdot 3 + l_3(x) \cdot 3$$

Svolgendo i calcoli si ottiene

$$L_3(x) = x^2 - x + 1$$

Ricordarsi (in eventuali esercizi d'esame) che il polinomio di interpolazione  $L_k(x)$  può essere di al più grado  $k-1$ .

## 6.2.4 Interpolazione di Newton

### 6.2.4.1 Premessa: differenze divise di ordine $k$

**Definizione.** Definiamo le *differenze divise* riferite a una funzione  $f(x) : D \subseteq \mathbb{R} \rightarrow \mathbb{R}$ . Siano  $x_0, x_1, \dots, x_{k-1} \in D$ , con  $x_i \neq x_j$  se  $i \neq j$

$$f[x_0, x_1, \dots, x_{k-1}, x] = \frac{f[x_0, x_1, \dots, x_{k-2}, x] - f[x_0, x_1, \dots, x_{k-1}]}{x - x_{k-1}}$$

dove, per  $k=1$  abbiamo  $f[x_0, x] = \frac{f(x) - f(x_0)}{x - x_0}$  è definita la *differenza divisa di ordine 1*. Questo  $\forall x \in D, x \neq x_i, i = 0, 1, \dots, j-1$ .

Per dare un'idea si osservi che il numero di variabili è  $k$ , che nel primo termine si sostituisce il penultimo termine con l'ultimo e nel secondo si sostituisce l'ultimo col penultimo

$$\begin{aligned} f[x] &= f(x) && \text{differenza divisa di ordine 0} \\ f[x_0, x] &= \frac{f(x) - f(x_0)}{x - x_0} && \text{differenza divisa di ordine 1} \\ f[x_0, x_1, x] &= \frac{f[x_0, x] - f[x_0, x_1]}{x - x_1} && \text{differenza divisa di ordine 2} \\ f[x_0, x_1, x_2, x] &= \frac{f[x_0, x_1, x] - f[x_0, x_1, x_2]}{x - x_2} && \text{differenza divisa di ordine 2} \end{aligned}$$

### Proprietà delle differenze divise

#### 1. Simmetria.

Consideriamo una permutazione  $i_0, i_1, \dots, i_k$  dei numeri  $0, 1, \dots, k$ . Possiamo dire che

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

## 2. Prolungamento per continuità.

Se  $f(x) \in C^1(D)$  allora la funzione  $f[x_0, x_1, \dots, x_{k-1}, x]$  è prolungabile per continuità su tutto  $D$ .

## 3. Derivabilità.

Se  $f(x) \in C^k(D)$  esiste almeno un valore  $\xi$  tale che

$$f[x_0, x_1, \dots, x_k] = \frac{f^{(k)}(\xi)}{k!}$$

Si osservi che  $\min_i x_i < \xi < \max_i x_i$ .

### 6.2.4.2 Teorema di espansione per le differenze divise

**Teorema.** Sia  $f(x) : D \subseteq \mathbb{R} \rightarrow \mathbb{R}$  e siano  $x_0, x_1, \dots, x_k \in D$  con  $x_i \neq x_j$  se  $i \neq j$ . Vale l'identità

$$\begin{aligned} f(x) = & f(x_0) + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2] + \\ & + \dots + (x - x_0)(x - x_1) \dots (x - x_{k-1})f[x_0, x_1, \dots, x_k] + \\ & + (x - x_0)(x - x_1) \dots (x - x_k)f[x_0, x_1, \dots, x_k, x] \end{aligned}$$

Si osservi che:

- per ogni termine avente differenza divisa di ordine  $k$  si moltiplica per  $k - 1$  fattori;
- tutti le differenze divise, tranne l'ultima, non sono funzioni (non dipendono da  $x$ );
- senza l'ultima differenza divisa si avrebbe  $f(x)$  polinomio.

### Dimostriamo per induzione

- Per  $k = 0$  abbiamo come somma primo e ultimo termine della sommatoria introdotta prima

$$\begin{aligned} f(x) &= f(x_0) + (x - x_0)f[x_0, x] = f(x_0) + (x - x_0) \frac{f(x) - f(x_0)}{x - x_0} \\ &= f(x_0) + f(x) - f(x_0) = f(x) \end{aligned}$$

nel secondo termine si semplifica  $x - x_0$ , le sottrazioni seguono in modo immediato.

- Il fatto che per  $k$  sia verificata deve implicare che sia valida anche per  $k + 1$ . La seguente espansione con  $k$  sappiamo essere valida

$$\begin{aligned} f(x) = & f(x_0) + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2] + \\ & + \dots + (x - x_0)(x - x_1) \dots (x - x_{k-1})f[x_0, x_1, \dots, x_k] + \\ & + (x - x_0)(x - x_1) \dots (x - x_k)f[x_0, x_1, \dots, x_k, x] \end{aligned}$$

Prendiamo la differenza divisa di ordine  $k + 1$

$$f[x_0, \dots, x_{k+1}, x] = \frac{f[x_0, \dots, x_k, x] - f[x_0, \dots, x_{k+1}]}{x - x_{k+1}}$$

$$(x - x_{k+1})f[x_0, \dots, x_{k+1}, x] = f[x_0, \dots, x_k, x] - f[x_0, \dots, x_{k+1}]$$

$$f[x_0, \dots, x_k, x] = f[x_0, \dots, x_{k+1}] + (x - x_{k+1})f[x_0, \dots, x_{k+1}, x]$$

Sostituiamo in  $f(x)$  la differenza divisa di ordine  $k$ . Il risultato è la dimostrazione dell'espansione (se è valida per  $k$  lo è anche per  $k+1$ )

$$\begin{aligned} f(x) &= f(x_0) + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2] + \\ &\quad + \dots + (x - x_0)(x - x_1) \dots (x - x_{k-1})f[x_0, x_1, \dots, x_k] + \\ &\quad + (x - x_0)(x - x_1) \dots (x - x_k)f[x_0, x_1, \dots, x_{k+1}] + \\ &\quad + (x - x_0)(x - x_1) \dots (x - x_{k+1})f[x_0, x_1, \dots, x_{k+1}, x] \end{aligned}$$

#### 6.2.4.3 Teorema di Newton (Polinomio di interpolazione di Newton)

**Teorema.** Data la funzione  $f : D \subseteq \mathbb{R} \rightarrow \mathbb{R}$  affermiamo che il seguente polinomio

$$\begin{aligned} P_k(x) &= f(x_0) + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2] + \\ &\quad \dots + (x - x_0) \dots (x - x_{k-1})f[x_0, \dots, x_k] \end{aligned}$$

è interpolante ed è noto come *polinomio di interpolazione di Newton*.

Si prenda l'espansione introdotta ieri: tutti i termini tranne l'ultimo permettono di ottenere un polinomio, visto che le funzioni divise non sono funzioni (non dipendono da  $x$ ). Segue che noi potremo dire

$$f(x) = P_k(x) + (x - x_0) \dots (x - x_k)f[x_0, \dots, x_k, x]$$

Se noi ragioniamo per valori  $x_i$ , dove  $i = 0, 1, \dots, k$  allora

$$P_k(x_i) = f(x_i)$$

visto che il secondo termine risulterà annullato per  $x_i - x_i = 0$ . Segue il<sup>1</sup> polinomio di interpolazione visto nel teorema.

#### Dimostrazione per induzione

- Per  $k = 0$  abbiamo

$$P_0(x_0) = f(x_0)$$

L'unica condizione presente con  $k = 0$  è rispettata.

- Per un valore  $k$  possiamo dire

$$P_k(x_i) = f(x_i)$$

con  $i = 0, 1, \dots, k$ . Se il polinomio è interpolante sui primi  $k$  punti allora deve esserlo anche con  $k+1$  punti

$$P_{k+1}(x_i) = f(x_i)$$

con  $i = 0, 1, \dots, k+1$ .

- Si osservi che

$$P_{k+1}(x) = P_k(x) + (x - x_0) \dots (x - x_k)f[x_0, \dots, x_{k+1}]$$

Calcoliamo in  $x_i$ , con  $i = 0, \dots, k$

$$P_{k+1}(x_i) = P_k(x_i) = f(x_i)$$

visto che l'ultimo termine si annulla (occhio ai valori assunti da  $i$ ). Il polinomio  $P_{k+1}$  interpola i primi  $k+1$  punti esattamente come  $P_k$

---

<sup>1</sup>Ricordarsi l'unicità

- Ci manca da verificare

$$P_{k+1}(x_{k+1}) = f(x_{k+1})$$

sviluppiamo

$$P_{k+1}(x_{k+1}) = P_k(x_{k+1}) + (x_{k+1} - x_0) \dots (x_{k+1} - x_k) f[x_0, \dots, x_{k+1}]$$

calcoliamo  $f(x_{k+1})$  (si riprenda  $f(x)$  posto subito dopo il teorema)

$$f(x_{k+1}) = P_k(x_{k+1}) + (x_{k+1} - x_0) \dots (x_{k+1} - x_k) f[x_0, \dots, x_k, x_{k+1}]$$

Abbiamo ottenuto

$$f(x_{k+1}) = P_{k+1}(x_{k+1})$$

#### 6.2.4.4 Errore nell'interpolazione con le differenze divise

Possiamo scrivere, per mezzo delle differenze divise, una formula per calcolare l'errore del processo di interpolazione. La cosa vale in generale, tenendo a mente che il polinomio di interpolazione è unico.

$$E(x) = f(x) - P_k(x) = (x - x_0) \dots (x - x_k) f[x_0, \dots, x_k, x]$$

Nel caso in cui  $f \in C^{k+1}(D)$  allora

$$E(x) = f(x) - P_k(x) = (x - x_0) \dots (x - x_k) \frac{f^{(k+1)}(\xi_x)}{(k+1)!}$$

dove  $\min\{x_0, \dots, x_k, x\} < \xi_x < \max\{x_0, x_1, \dots, x_k, x\}$ .

#### 6.2.4.5 Quadro di Newton

Per il calcolo del polinomio di Newton è necessario andare a calcolarsi le differenze divise. Ci serve un algoritmo per tale fine. L'idea è di costruire la seguente tabella. Supponiamo di voler costruire un polinomio  $P_4$

$x$	$f(x)$	DD1	DD2	DD3
$x_0$	$f(x_0)$	-	-	-
$x_1$	$f(x_1)$	$f[x_0, x_1]$	-	-
$x_2$	$f(x_2)$	$f[x_0, x_2]$	$f[x_0, x_1, x_2]$	-
$x_3$	$f(x_3)$	$f[x_0, x_3]$	$f[x_0, x_1, x_3]$	$f[x_0, x_1, x_2, x_3]$

Ricordando la definizione di differenze divise possiamo scrivere

$x$	$f(x)$	DD1	DD2	DD3
$x_0$	$f(x_0)$	-	-	-
$x_1$	$f(x_1)$	$\frac{f(x_1) - f(x_0)}{x_1 - x_0}$	-	-
$x_2$	$f(x_2)$	$\frac{f(x_2) - f(x_0)}{x_2 - x_0}$	$\frac{f[x_0, x_2] - f[x_0, x_1]}{x_2 - x_1}$	-
$x_3$	$f(x_3)$	$\frac{f(x_3) - f(x_0)}{x_3 - x_0}$	$\frac{f[x_0, x_3] - f[x_0, x_1]}{x_3 - x_1}$	$\frac{f[x_0, x_1, x_3] - f[x_0, x_1, x_2]}{x_3 - x_2}$

L'algoritmo adottabile è il seguente:

- costruisco la tabella con  $x$  e  $f(x)$
- inizio a estenderla con le colonne necessarie per trovare le differenze divise;
- in DD1 ignoro la prima riga, nelle colonne successive ignoro un elemento in più rispetto alla precedente;
- calcolo le differenze divise di una colonna dall'alto verso il basso;
- ogni differenza divisa si ottiene sottraendo al termine che si trova nella stessa riga e nella colonna immediatamente a sinistra il primo termine di questa colonna, e dividendo per la differenza dei corrispondenti valori  $x_i$ .

#### 6.2.4.6 Primo esempio

Prendiamo la seguente tabella

$x$	0	-1	2	-2	3
$f(x)$	5	3	3	-9	11

Si vuole calcolare il polinomio di interpolazione di Newton  $P_4(x)$ .

**Risoluzione** Calcoliamo il polinomio di interpolazione ricorrendo al quadro di Newton visto precedentemente. I primi elementi nelle colonne DD1, DD2, DD3 ci restituiranno le differenze divise necessarie per i nostri calcoli.

$x$	$f(x)$	DD1	DD2	DD3
0	5	-	-	-
-1	3	2	-	-
2	3	-1	-1	-
-2	-9	7	-5	1
3	11	2	0	1

Applichiamo la formula

$$P_k(x) = f(x_0) + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, 2] + \dots + (x - x_0)\dots(x - x_{k-1})f[x_0, \dots, x_k]$$

e concludiamo

$$P_4(x) = 5 + (x - 0)*2 + (x - 0)(x + 1)*(-1) + (x - 0)(x + 1)(x - 2)*1 = x^3 - 2x^2 - x + 5$$

#### 6.2.4.7 Secondo esempio

Data una matrice  $A \in \mathbb{C}^{3 \times 3}$  e i seguenti valori

$$\det(A - I) = 1 \quad \det(A - 2I) = -3 \quad \det(A + I) = 3 \quad \det(A) = -1$$

Vogliamo calcolare il polinomio caratteristico  $P(\lambda) = \det(A - \lambda I)$ .

**Risoluzione** Abbiamo a tutti gli effetti delle coppie  $(x, f(x))$ . Costruiamo il quadro di Newton e calcoliamo il polinomio di interpolazione.

$\lambda$	$\det(A - \lambda I)$	DD1	DD2	DD3
1	-1	-	-	-
2	-3	-2	-	-
-1	3	-2	0	-
0	-1	0	-1	-1

Concludiamo

$$\det(A - \lambda I) = -1 - 2(\lambda - 1) + 0 * (\lambda - 1)(\lambda - 2) - 1(\lambda - 1)(\lambda - 2)(\lambda - 3) = -\lambda^3 + 2\lambda^2 - \lambda - 1$$

#### 6.2.4.8 Terzo esempio

Prendiamo la seguente tabella

$x$	0	-1	1	2	$\alpha$
$f(x)$	-1	$\beta$	-1	1	5

dove  $\alpha, \beta \in \mathbb{R}$ . Si vogliono calcolare i valori reali  $\alpha$  e  $\beta$  in modo tale che il polinomio di interpolazione risulti di grado minimo.

**Risoluzione** Anche in questo caso costruiamo il quadro di Newton. Suggerimento appassionato del professore: riformulare la tabella mettendo in fondo alla tabella le righe con i parametri  $\alpha$  e  $\beta$ .

$x$	0	1	2	-1	$\alpha$
$f(x)$	-1	-1	1	$\beta$	5

Costruiamo un polinomio di interpolazione con i primi tre punti, quelli senza parametri:  $P_2(x)$ . Dopo aver calcolato questo polinomio dovremo pensare a valori  $\alpha, \beta$  che non innalzino il valore del polinomio di interpolazione precedentemente ottenuto. Andremo a imporre

$$\beta = P_2(-1) \quad P_2(\alpha) = 5$$

Costruiamo il quadro

$x$	$f(x)$	DD1	DD2
0	-1	-	-
1	-1	0	-
2	1	1	1

Otteniamo il seguente polinomio

$$P_2(x) = -1 + 0 * (x - 0) + 1 * (x - 0)(x - 1) = x^2 - x - 1$$

Calcoliamo  $\beta$

$$\beta = P_2(-1) = 1$$

Calcoliamo  $\alpha$ , otteniamo un'equazione di secondo grado con cui possiamo concludere solo in presenza di soluzioni

$$\alpha^2 - \alpha - 1 = 5 \longrightarrow \alpha^2 - \alpha - 6 = 0$$

Troviamo  $\alpha_1 = -2$  e  $\alpha_2 = 3$ .

#### 6.2.4.9 Quarto esempio

Prendiamo il seguente polinomio

$$P(x) = x^4 - 3x^3 + x^2 + x + 1$$

Possiamo dire che questo è il polinomio di interpolazione di Newton ottenuto dalla seguente tabella?

$x$	0	1	-1	2
$f(x)$	1	1	5	-1

#### Risoluzione

- L'idea base che si potrebbe avere è di sostituire i vari valori  $x$  in  $P(x)$  e verificare che si ottenga  $f(x) = P(x)$  per ogni  $x$  trattato.
- Se facciamo quanto detto otteniamo che le corrispondenze sono valide. E' un polinomio di interpolazione? NO! Ha grado 4, il polinomio di interpolazione ha al più grado 3.

## 6.3 Interpolazione osculatoria di Hermite

### 6.3.1 Cosa abbiamo

Siano assegnati  $k + 1$  punti reali  $x_0, x_1, \dots, x_k \in D$  due a due distinti. Conosciamo in corrispondenza  $2k + 2$  valori reali: quelli della funzione e quelli della derivata prima

$$\begin{array}{cccc} f(x_0) & f(x_1) & \dots & f(x_k) \\ f'(x_0) & f'(x_1) & \dots & f'(x_k) \end{array}$$

### 6.3.2 Definizione di interpolazione osculatoria di Hermite

**Definizione.** L'interpolazione osculatorie di Hermite consiste nel determinare un polinomio  $H_{2k+1}(x)$  di grado al più  $2k + 1$  tale che

$$H_{2k+1}(x_r) = f(x_r) \quad H'_{2k+1}(x_r) = f'(x_r) \quad r = 0, 1, \dots, k$$

### 6.3.3 Polinomio di interpolazione di Hermite

**Definizione.** Il polinomio di interpolazione di Hermite presenta la seguente struttura

$$H_{2k+1}(x) = \sum_{r=0}^k h_{0r}(x)f(x_r) + \sum_{r=0}^k h_{1r}(x)f'(x_r)$$

Si osservi che derivando otteniamo

$$H'_{2k+1}(x) = \sum_{r=0}^k h'_{0r}(x)f(x_r) + \sum_{r=0}^k h'_{1r}(x)f'(x_r)$$

Quello che noi vogliamo dire, in generale, è che

$$H_{2k+1}(x_i) = f(x_i) \quad H'_{2k+1}(x_i) = f'(x_i)$$

Dobbiamo lavorare sulle funzioni  $h_{0r}$  e  $h_{1r}$ . In ogni caso avremo una sommatoria annullata nella sua interezza, e l'altra sommatoria con termini tutti nulli tranne uno (quello con  $f(x_i)$  o  $f'(x_i)$ ).

- **Funzione.**

Per quanto riguarda la funzione  $f$  dobbiamo porre la funzione  $h_{0r}$  e la derivata prima  $h'_{0r}$  in modo tale che

$$\begin{cases} h_{0r}(x_s) = \delta_{rs} & r, s = 0, 1, \dots, k \\ h'_{0r}(x_s) = 0 & r, s = 0, 1, \dots, k \end{cases}$$

- **Derivata prima.**

Per quanto riguarda le derivate prima la cosa si rovescia

$$\begin{cases} h_{1r}(x_s) = 0 & r, s = 0, 1, \dots, k \\ h'_{1r}(x_s) = \delta_{rs} & r, s = 0, 1, \dots, k \end{cases}$$

### 6.3.3.1 Polinomi $h_{0r}(x)$

Ricerchiamo i polinomi  $h_{0r}(x)$  nella forma<sup>2</sup>

$$h_{0r}(x) = (Ax + B)l_r^2(x)$$

dove  $l_r(x)$  è il polinomio fondamentale della interpolazione di Lagrange (visto nell'interpolazione parabolica di Lagrange). Riscriviamolo, allegando proprietà importante

$$l_r(x) = \frac{(x - x_0) \dots (x - x_{r-1})(x - x_{r+1}) \dots (x - x_k)}{(x_r - x_0) \dots (x_r - x_{r-1})(x_r - x_{r+1}) \dots (x_r - x_k)} \quad l_r(x_s) = \delta_{rs}$$

Dalla forma possiamo pensare che  $h_{0r}(x)$  saranno polinomi di grado  $2k+1$ , visto la presenza di un polinomio di grado 1 e  $l_r(x)$  di grado  $k$  (che diventa di grado  $2k$  visto l'elevazione a quadrato).

- **Funzione.**

Imponiamo come condizione  $h_{0r}(x_s) = \delta_{rs}$ . Ma sappiamo che  $\delta_{rs} = 0$  se  $r \neq s$ . Segue per costruzione

$$h_{0r}(x_s) = 0$$

Poniamo  $h_{0r}(x_r) = 1$

$$h_{0r}(x_r) = Ax_r + B = 1$$

Questa condizione non è per costruzione.

- **Derivata della funzione.**

Deriviamo  $h_{0r}$

$$h'_{0r}(x) = Al_r^2(x) + (Ax + B)2l_r(x)l'_r(x)$$

Si osservi che per costruzione abbiamo  $h'_{0r}(x_s) = 0$ , inoltre vogliamo che si verifichi  $h'_{0r}(x_r) = 0$

$$h'_{0r}(x_r) = A + (Ax_r + B)2l'_r(x_r) = 0$$

- **Unione delle condizioni trovate.**

Troviamo  $A$  e  $B$  in grado di soddisfare le condizioni poste in box. Otteniamo

$$\begin{aligned} Ax_r + B &= 1 \longrightarrow B = 1 - 2x_r l'_r(x_r) \\ A + 2l'_r(x_r) &= 0 \longrightarrow A = -2l'_r(x_r) \end{aligned}$$

In conclusione (Non ve lo chiederò mai a memoria - cit.)

$$h_{0r}(x) = (1 - 2l'_r(x_r)(x - x_r))l_r^2(x)$$

### 6.3.3.2 Polinomi $h_{1r}(x)$

Ricerchiamo i polinomi  $h_{1r}(x)$  in una forma analoga a quella già vista per  $h_{0r}$ <sup>3</sup>

$$h_{1r}(x) = (Cx + D)l_r^2(x)$$

dove  $l_r(x)$  è il polinomio fondamentale della interpolazione di Lagrange (visto nell'interpolazione parabolica di Lagrange). Vogliamo imporre

$$h_{1r}(x_s) = 0 \quad h'_{1r}(x_s) = \delta_{rs}$$

con  $s = 0, 1, \dots, k$ .

---

<sup>2</sup>Precisazione forse non troppo scontata: il professore ha usato  $A$  e  $B$  maiuscoli, ma non sono matrici.

<sup>3</sup>Precisazione forse non troppo scontata: il professore ha usato  $C$  e  $D$  maiuscoli, ma non sono matrici.

- **Funzione.**

Prendiamo  $x_s$

$$(Cx_s + D)l_r^2(x_s) = 0$$

il fattore è sicuramente nullo per costruzione, visto che  $l_r(x_s) = 0$ . Vogliamo che sia nullo anche in  $x_r$ , segue la prima condizione

$$Cx_r + D = 0$$

- **Derivata della funzione.**

Prendiamo la derivata di  $h_{1r}$

$$h'_{1r}(x) = Cl_r^2(x) + (Cx + D)2l_r(x)l'_r(x)$$

poniamo  $x_s$

$$Cl_r^2(x_s) + (Cx_s + D)2l_r(x_s)l'_r(x_s) = \delta_{rs}$$

Se  $r \neq s$  abbiamo  $\delta_{rs} = 0$  per costruzione. Consideriamo invece il caso dove  $r = s$ , ecco la seguente condizione

$$C + (Cx_r + D)2l'_r(x_r) = 1$$

- **Unione delle condizioni trovate.**

Se vale la prima condizione allora la seconda condizione sarà

$$Cl_r^2(x_r) = 1 \implies C = 1$$

inoltre

$$D = -x_r$$

Possiamo concludere!

$$h_{1r}(x) = (x - x_r)l_r^2(x)$$

con  $r = 0, 1, \dots, k$

#### 6.3.4 Errore nell'interpolazione

**Teorema.** Se  $f(x) \in C^{2k+2}(D)$  si ha

$$E(x) = f(x) - H(x) = (x - x_0)^2 \dots (x - x_k)^2 \frac{f^{(2k+2)}(\xi_x)}{(2k+2)!}$$

dove  $\min\{x_0, x_1, \dots, x_k, x\} < \xi_x < \max\{x_0, x_1, \dots, x_k, x\}$

## 6.4 Interpolazione con funzioni spline

### 6.4.1 Perchè ne parliamo

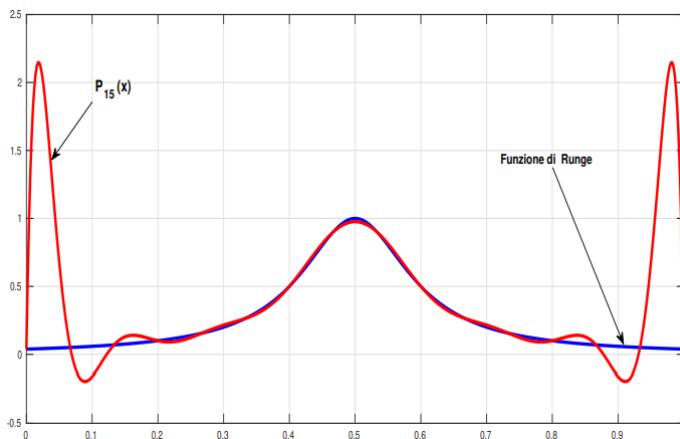
I precedenti metodi di interpolazione presentano un problema: se si aumenta il grado  $k$  allora aumenta il grado del polinomio di interpolazione. Il risultato è un polinomio che:

- è poco maneggevole;
- può discostarsi sensibilmente dai valori della funzione  $f$

Un esempio è la cosiddetta *funzione di Runge*, che rappresentiamo in  $[0, 1]$

$$f(x) = \frac{1}{100x^2 - 100x + 26}$$

Nell'intervallo possiamo dire  $f(x) \in C^\infty([0, 1])$ . Prendiamo il polinomio di Hermite di grado 15, e scegliamo punti equidistanti all'interno dell'intervallo: il risultato è una forte divergenza rispetto ad  $f(x)$  in prossimità degli estremi dell'intervallo.



Addirittura

$$\lim_{k \rightarrow \infty} \max_{0 \leq x \leq 1} |E_k(x)| = +\infty$$

### 6.4.2 Definizione di funzione spline

**Definizione.** Dicesi *funzione spline di grado  $m$*  relativa a un insieme di  $k+1$  punti reali

$$x_0, x_1, \dots, x_k$$

una funzione  $S_m : [x_0, x_k] \rightarrow \mathbb{R}$  tale che:

- $S_m(x)$  è un polinomio di grado non superiore ad  $m$  in ogni possibile intervallo  $[x_{i-1}, x_i]$ , con  $i = 1, 2, \dots, k$ ;
- $S_m(x_i) = f(x_i)$ ,  $i = 0, 1, \dots, k$
- $S_m \in C^{m-1}([x_0, x_k])$

In sostanza quello che facciamo è considerare un polinomio  $p_i(x)$  per ogni intervallo  $[x_{i-1}, x_i]$ . Avremo tanti polinomi quanti gli intervalli presenti. In sostanza definiamo una funzione polinomiale a tratti  $S_m(x)$ , dove  $m$  è il grado massimo assumibile dai polinomi dei vari intervalli!

### 6.4.3 Esempio: spline cubiche

Consideriamo il caso  $m = 3$ , le *spline cubiche*. Supponiamo di avere un intervallo  $[x_0, x_4]$  con i punti

$$x_0 < x_1 < x_2 < x_3 < x_4$$

Abbiamo quattro intervalli definiti, e per ciascuno di esso un polinomio. I polinomi devono rispettare le seguenti condizioni, relative agli estremi

- $p_1(x_0) = f(x_0)$  e  $p_1(x_1) = f(x_1)$
- $p_2(x_1) = f(x_1)$  e  $p_2(x_2) = f(x_2)$
- $p_3(x_2) = f(x_2)$  e  $p_3(x_3) = f(x_3)$
- $p_4(x_3) = f(x_3)$  e  $p_4(x_4) = f(x_4)$

8 condizioni imposte ( $2k$ ). Abbiamo quattro polinomi di al più grado 3: se ogni polinomio ha quattro coefficienti allora avremo 16 coefficienti da determinare ( $4k$ ). Se queste condizioni sono rispettate non avremo solo l'interpolabilità, ma anche la continuità  $C^0$  (la continuità in ciascun intervallo è per costruzione, visto che lavoriamo coi polinomi)! Si tenga a mente che è richiesto avere una funzione  $C^2$  nell'esempio affrontato

1. Si ha una funzione  $C^1$  introducendo le seguenti 3 condizioni ( $k - 1$ )

$$p'_1(x_1) = p'_2(x_1) \quad p'_2(x_2) = p'_3(x_2) \quad p'_3(x_3) = p'_4(x_3)$$

2. Si ha una funzione  $C^2$  ponendo le condizioni necessarie per  $C^1$  e le seguenti 3 ( $k - 1$ )

$$p''_1(x_1) = p''_2(x_1) \quad p''_2(x_2) = p''_3(x_2) \quad p''_3(x_3) = p''_4(x_3)$$

Complessivamente abbiamo trovato  $8 + 3 + 3 = 14$  condizioni, ma ce ne mancano due visto che i coefficienti da trovare sono 16. Poniamo una tra le seguenti:

- **Spline naturale.**  $p''_1(x_0) = p''_k(x_k) = 0$
- **Spline periodica.**  $p'_1(x_0) = p'_k(x_k)$  e  $p''_1(x_0) = p''_k(x_k)$
- **Spline vincolata.**  $p'_1(x_0) = y'_0$  e  $p'_k(x_k) = y'_k$  dove  $y'_0 = f'(x_0)$  e  $y'_k = f'(x_k)$

### 6.4.4 Teorema sull'unicità della funzione spline

**Teorema.** Se i punti  $x_0, \dots, x_k$  sono in progressione aritmetica, cioè

$$x_i - x_{i-1} = h$$

con  $i = 1, 2, \dots, k$  allora esiste una unica funzione spline cubica naturale  $S_3(x)$

- Introdotti  $k + 1$  valori arbitrari  $m_i$ , si costruiscono i polinomi  $p_i$  ricorrendo al polinomio di interpolazione di Hermite: abbiamo una sommatoria dove alcuni termini sono conosciuti, mentre in altri abbiamo delle incognite (i termini relativi alle derivate).
- Se applichiamo tutte le condizioni introdotte precedentemente otteniamo un sistema lineare con matrice tridiagonale. Nella matrice dei coefficienti si ha predominanza diagonale forte: segue matrice non singolare ai sensi del primo teorema di Gershgorin, e quindi unicità.

*Io questi calcoli non ho voglia di farli* (cit.).

## 6.5 Metodo dei minimi quadrati nel discreto

### 6.5.1 Spiegazione

Il metodo trova ampio uso per l'interpolazione a partire da moli di dati considerevoli. I primi metodi hanno i problemi detti, mentre la funzione Spline è una funzione a tratti. Vogliamo risolvere i problemi detti riguardo i primi metodi evitando funzioni a tratti.

- Abbiamo  $k + 1$  punti  $x_j \in D, j = 0, 1, \dots, k$ . Per ogni  $x_j$  ho il valore  $f(x_j)$ . Contrariamente ai metodi precedenti non consideriamo derivate.
- La funzione interpolante  $\phi_x(x)$  si individua con la combinazione lineare di  $m + 1$  funzioni

$$\phi_0(x), \phi_1(x), \dots, \phi_m(x)$$

dove  $m \leq k$ , cioè

$$\phi(x) = c_0\phi_0(x) + c_1\phi_1(x) + \dots + c_m\phi_m(x)$$

Si consideri che i valori  $x$  non sono incognite, ma termini noti del problema che stiamo affrontando. Segue che ciò che dovremo determinare sono i coefficienti  $c_0, c_1, \dots, c_m$  della combinazione lineare.

- **Sommatoria delle differenze (degli scarti quadratici).**

Se parliamo di  $\phi(x)$  ed  $f(x)$  allora metteremo a confronto i valori  $f(x_i)$  e  $\phi(x_i)$ . L'idea base è quella di valutare le differenze tra i vari valori  $i$ -esimi (le differenze sono dette *scarti quadratici*) per mezzo della seguente funzione

$$\psi(c_0, c_1, \dots, c_m) = \sum_{j=0}^k [\phi(x_j) - f(x_j)]^2$$

Siamo arrivati a questa funzione dopo due proposte non ottimizzanti:

$$-\psi(c_0, c_1, \dots, c_m) = \sum_{j=0}^k [\phi(x_j) - f(x_j)]$$

Non ci piace perchè è possibile avere un risultato di sommatoria negativo, o un risultato nullo pur avendo differenze considerevoli in alcuni punti

$$-\psi(c_0, c_1, \dots, c_m) = \sum_{j=0}^k |\phi(x_j) - f(x_j)|$$

Ci piace di più poichè tutti i termini sono minimo uguali a zero, ma non siamo ancora soddisfatti per problemi di differenziabilità (ricordarsi le proprietà della funzione valore assoluto viste ad Analisi I, a proposito della derivabilità).

La proposta definitiva non ha sicuramente valori negativi ed è differenziabile su tutto  $\mathbb{R}^{m+1}$ .

- **Perchè metodo dei minimi quadrati?**

Il metodo dei minimi quadrati consiste nello scegliere i coefficienti  $c_i$  per i quali la funzione  $\psi(c)$  risulta minimizzata

$$c = (c_0, c_1, \dots, c_m)^T \quad \psi(c) = \sum_{j=0}^k \left[ \sum_{i=0}^m c_i \phi_i(x_j) - f(x_j) \right]^2$$

L'individuazione del minimo passa dal gradiente  $\nabla\psi(c) = 0$ , consideriamo come esempio una generica derivata parziale rispetto a  $c_s$

$$\frac{d\psi}{dc_s} = 2 \sum_{j=0}^k \left[ \sum_{i=0}^m c_i \phi_i(x_j) - f(x_j) \right] \phi_s(x_j) = 0 \implies \sum_{j=0}^k \left[ \sum_{i=0}^m c_i \phi_i(x_j) - f(x_j) \right] \phi_s(x_j) = 0$$

con  $s = 0, 1, \dots, m$ .

### 6.5.2 Sistema delle equazioni normali

Per porre  $\nabla\phi(c) = 0$  è necessario porre uguale a zero ogni derivata parziale. Rapresentiamo il tutto in modo compatto per mezzo del *sistema delle equazioni normali*. Poniamo

$$c = (c_0 \ c_1 \ \dots \ c_m)^T \quad b = (f(x_0) \ f(x_1) \ \dots \ f(x_k))^T$$

$$A = \begin{pmatrix} \phi_0(x_0) & \phi_1(x_0) & \dots & \phi_m(x_0) \\ \phi_0(x_1) & \phi_1(x_1) & \dots & \phi_m(x_1) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(x_k) & \phi_1(x_k) & \dots & \phi_m(x_k) \end{pmatrix} \in \mathbb{R}^{(k+1) \times (m+1)}$$

Quello che noi affermiamo è che

$$A^T A c = A^T b$$

dove  $A^T A \in \mathbb{R}^{(m+1) \times (m+1)}$ . Si osservi che

$$Ac = (\phi(x_0) \ \phi(x_1) \ \dots \ \phi(x_k))^T$$

ergo  $Ac - b$  ci restituisce una matrice contenente gli scarti (non elevati al quadrato). In modo compatto otteniamo la somma dei quadrati degli scarti:

$$\phi(c) = (Ac - b)^T (Ac - b) = \|Ac - b\|_2^2$$

- **Esistenza della soluzione.**

Un sistema di questo tipo ha sicuramente soluzioni! Si potrebbe dimostrare

$$r(A^T A) = r(A^T A | A^T b)$$

- **Unicità della soluzione.**

Riprendiamo il teorema di Binet esteso (matrice quadrata ottenuta a partire da due matrici non quadrate).

- $m \leq k$  altrimenti avrei  $\det(A^T A)$  nullo per il numero di righe superiore al numero di colonne.
- Se  $m = k$  abbiamo una matrice quadrata allora

$$\det(A^T A) = \det(A^T) \det(A) = \det(A)^2$$

dipende tutto dalla matrice  $A$ .

- Se  $m < k$  si sommano i prodotti dei minori di ordine massimo estraibili dalle due matrici. Se il rango della matrice  $A$  è massimo (quindi estraggo un minore di ordine  $m + 1$ ) allora il determinante della matrice  $A^T A$  è diverso da zero.

Le funzioni  $\phi_i$  si prendono linearmente indipendenti! La cosa funziona bene nel continuo, un po' meno nel discreto (esistono casi di funzioni linearmente indipendenti nel continuo, che non lo sono nel discreto)

- **Matrice definita positiva.**

La matrice  $A^T A$  è definita positiva, lo si verifica col quoziente di Rayleigh

$$\frac{x^H B x}{x^H x} = \frac{x^T A^T A x}{x^T x} = \frac{(Ax)^T (Ax)}{x^T x} \geq 0$$

### 6.5.3 Sistemi lineari sovradeterminati e minimi quadrati

#### 6.5.3.1 Definizione di sistema lineare sovradeterminato

**Definizione.** Siano  $k, m \in \mathbb{N}$  con  $k > m$ . Siano  $A \in \mathbb{R}^{k \times m}$  e  $b \in \mathbb{R}^k$ . Un sistema lineare è detto sovradeterminato se il numero di equazioni è superiore al numero di incognite. Questo sistema

$$Ax = b$$

ha soluzione se e solo se  $r(A) = r(A|b)$  (nulla di nuovo, Rouchè-Capelli).

Se  $r(A) < r(A|b)$  allora non si ha soluzione. Possiamo calcolare il vettore dei residui  $r \in \mathbb{R}^k$ , che in questo contesto sarà  $\neq 0$

$$b - Ax = r$$

#### 6.5.3.2 Metodo dei minimi quadrati

Vogliamo risolvere il sistema lineare sovradeterminato nel senso dei minimi quadrati: questo significa trovare il vettore  $x$  che rende minima la norma al quadrato del vettore dei residui  $r$ . Si recuperi la definizione di prodotto scalare nell'insieme dei complessi:

$$\phi(x) = (b - Ax)^T (b - Ax) = r^T r = \sum_{i=1}^k r_i^2 = \|r\|_2^2$$

Se vogliamo annullare il prodotto guarderemo il vettore gradiente e le derivate parziali

$$\nabla \phi(x) = 0 \implies \frac{\partial}{\partial x_i} [(b - Ax)^T (b - Ax)] = 0, \quad i = 1, 2, \dots, m$$

Ma noi abbiamo già visto che il sistema delle derivate parziali uguali a zero può essere scritto così:

$$A^T Ax = A^T b$$

Se  $r(A) = m$  allora avremo sicuramente  $\det(A^T A) \neq 0$ : segue soluzione unica.

#### 6.5.3.3 Primo esempio

Si consideri il seguente sistema lineare sovradeterminato

$$\begin{cases} 2x_1 - x_2 = 1 \\ -x_1 + x_2 = 0 \\ x_1 + 2x_2 = 1 \end{cases}$$

e si calcoli la soluzione nel senso dei minimi quadrati.

## Svolgimento

$$A = \begin{pmatrix} 2 & -1 \\ -1 & 1 \\ 1 & 2 \end{pmatrix} \quad b = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \quad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

Per risolvere il sistema, e quindi individuare  $x$ , dobbiamo risolvere il sistema

$$A^T A x = A^T b$$

$$\begin{aligned} A^T A &= \begin{pmatrix} 2 & -1 \\ -1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 2 & -1 & 1 \\ -1 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 6 & -1 \\ -1 & 6 \end{pmatrix} \\ A^T b &= \begin{pmatrix} 2 & -1 & 1 \\ -1 & 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \end{aligned}$$

Seguono i seguenti calcoli

$$\begin{cases} 6x_1 - x_2 = 3 \\ -x_1 + 6x_2 = 1 \end{cases} \implies \begin{cases} x_2 = 6x_1 - 3 \\ -x_1 + 36x_1 - 18 = 1 \end{cases} \implies \begin{aligned} 35x_1 &= 19 \implies x_1 = \frac{19}{35} \\ x_2 &= \frac{6*19-3*35}{35} = \frac{9}{35} \end{aligned}$$

### 6.5.3.4 Secondo esempio

Si consideri il sistema lineare sovradeterminato

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & \alpha & 1 \\ 0 & 0 & \alpha \\ \beta & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}$$

dove  $\alpha, \beta \in \mathbb{R}$ . Determinare le coppie di valori  $\alpha, \beta$  per le quali il sistema non ha soluzione unica nel senso dei minimi quadrati.

**Risoluzione** Possiamo trovare  $\alpha, \beta$  imponendo  $\det(A^T A) = 0$ , ma fare questo significa richiedere che  $r(A) < m$ . Nell'esempio che stiamo affrontando significa affermare che la matrice abbia al più rango 2 (rango 4 non può averlo per costruzione, rango 3 dobbiamo fare in modo che non si manifesti).

Niente calcoli da fare, solo vedere i minori!!!! Ricordarselo in sede d'esame

Scriviamo i minori possibili e calcoliamoli per mezzo dello sviluppo di Laplace (spiegato in appendice in fondo alla dispensa)

$$\begin{aligned} \left| \begin{array}{ccc} 1 & 1 & 1 \\ 1 & \alpha & 1 \\ 0 & 0 & \alpha \end{array} \right| &= \alpha(\alpha - 1) & \left| \begin{array}{ccc} 1 & 1 & 1 \\ 1 & \alpha & 1 \\ \beta & 0 & 0 \end{array} \right| &= \beta(1 - \alpha) \\ \left| \begin{array}{ccc} 1 & 1 & 1 \\ 0 & 0 & \alpha \\ \beta & 0 & 0 \end{array} \right| &= \beta(\alpha) & \left| \begin{array}{ccc} 1 & \alpha & 1 \\ 0 & 0 & \alpha \\ \beta & 0 & 0 \end{array} \right| &= \beta(\alpha^2) \end{aligned}$$

La soluzione è  $\alpha = 0, \beta = 0$ : si annullano tutti i minori di ordine 3. Altrettanto valida è la soluzione  $\alpha = 1, \beta = 0$ .

### 6.5.3.5 Terzo esempio

Si consideri il sistema lineare sovradeterminato

$$\begin{pmatrix} 1 & \alpha \\ \alpha & \alpha \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

dove  $\alpha \in \mathbb{R}$ . Determinare i valori reali di  $\alpha$  per i quali il sistema ha un'unica soluzione nel senso dei minimi quadrati.

**Risoluzione** Questa volta si richiede un'unica soluzione. Vogliamo imporre  $\det(A^T A) \neq 0$ , ma questo significa affermare che il rango è massimo

$$r(A) = \min\{2, 3\} = 2$$

Consideriamo i seguenti minori di ordine 2, e imponiamo che almeno uno di questi sia diverso da zero.

$$\begin{vmatrix} 1 & \alpha \\ \alpha & \alpha \end{vmatrix} = \alpha - \alpha^2 \quad \begin{vmatrix} 1 & \alpha \\ 1 & 1 \end{vmatrix} = 1 - \alpha \quad \begin{vmatrix} \alpha & \alpha \\ 1 & 1 \end{vmatrix} = O$$

Individuiamo che avremo minori di ordini 2 diversi da zero imponendo  $\alpha \neq 1$ .

### 6.5.3.6 Quarto esempio

Si consideri il sistema lineare sovradeterminato  $Ax = b$  con matrice dei coefficienti.

$$A = \begin{pmatrix} \beta & 1 \\ \alpha & 1 \\ \beta & 1 \\ \alpha & -1 \end{pmatrix}$$

dove  $\alpha, \beta \in \mathbb{R}$ . Indicare i valori reali di  $\alpha$  e  $\beta$  per i quali il sistema ha un'unica soluzione nel senso dei minimi quadrati.

**Risoluzione** Vogliamo imporre che  $\det(A^T A) \neq 0$ , questo significa affermare che il rango è massimo

$$r(A) = \min\{4, 2\} = 2$$

Consideriamo i seguenti minori di ordine 2

$$\begin{vmatrix} \beta & 1 \\ \alpha & 1 \end{vmatrix} = \beta - \alpha \quad \begin{vmatrix} \beta & 1 \\ \alpha & -1 \end{vmatrix} = -\beta - \alpha \quad \begin{vmatrix} \alpha & 1 \\ \alpha & -1 \end{vmatrix} = -2\alpha$$

Individuiamo che  $|\alpha| \neq 0$  e  $|\alpha| + |\beta| \neq 0$ .

### 6.5.3.7 Quinto esempio

Si consideri il sistema lineare sovradeterminato  $Ax = b$  con matrice dei coefficienti.

$$A = \begin{pmatrix} \alpha & 1 \\ 2 & \alpha \\ 1 & \alpha \end{pmatrix}$$

dove  $\alpha \in \mathbb{R}$ . Indicare i valori reali di  $\alpha$  per i quali il sistema ha un'unica soluzione nel senso dei minimi quadrati.

**Risoluzione** Vogliamo imporre che  $\det(A^T A) \neq 0$ , questo significa affermare che il rango è massimo

$$r(A) = \min\{2, 3\} = 2$$

Consideriamo i seguenti minori di ordine 2

$$\begin{vmatrix} \alpha & 1 \\ 2 & \alpha \end{vmatrix} = \alpha^2 - 2 \quad \begin{vmatrix} \alpha & 1 \\ 1 & \alpha \end{vmatrix} = \alpha^2 - 1 \quad \begin{vmatrix} 2 & \alpha \\ 1 & \alpha \end{vmatrix} = \alpha$$

Individuiamo che si ha unicità  $\forall a \in \mathbb{R}$ .

#### 6.5.3.8 Sesto esempio

Si consideri il sistema lineare sovradeterminato  $Ax = b$  con matrice dei coefficienti.

$$A = \begin{pmatrix} 1 & -\alpha \\ \alpha & 1 \\ 1 & -\alpha \\ \alpha & 1 \end{pmatrix}$$

dove  $\alpha \in \mathbb{R}$ . Indicare i valori reali di  $\alpha$  per i quali il sistema ha un'unica soluzione nel senso dei minimi quadrati.

**Risoluzione** Vogliamo imporre che  $\det(A^T A) \neq 0$ , questo significa affermare che il rango è massimo

$$r(A) = \min\{4, 2\} = 2$$

L'unico minore di ordine 2 da considerare è il seguente (gli altri sono nulli oppure sono il seguente con righe invertite)

$$\begin{vmatrix} 1 & -\alpha \\ \alpha & 1 \end{vmatrix} = 1 + \alpha^2 \neq 0$$

Individuiamo che si ha unicità  $\forall a \in \mathbb{R}$ .

# Capitolo 7

## Integrazione numerica

### 7.1 Introduzione

#### 7.1.1 Promemoria: definizione di primitiva<sup>1</sup>

Data una funzione  $f : I \rightarrow \mathbb{R}$ , definita su un intervallo  $I \subset \mathbb{R}$  si definisce *primitiva* una funzione  $F : I \rightarrow \mathbb{R}$  tale che

$$F'(x) = f(x)$$

per ogni  $x \in I$ . Se  $F$  è una primitiva di  $f$  allora tutte e sole le primitive di  $f$  sono nella forma

$$F(x) + c$$

dove  $c \in \mathbb{R}$ . Cioè

$$\int f(x)dx = F(x) + c$$

#### 7.1.2 Perchè si parla di integrazione?

Vogliamo lavorare su integrali definiti, precisamente andremo ad approssimare integrali definiti. Perchè abbiamo bisogno di approssimare?

- **Non si conosce la funzione primitiva.**

Di molte funzioni integrabili non si conosce una primitiva esprimibile in modo elementare

$$f(x) = e^{\cos x}$$

Per capire meglio:  $\cos x$  non lo introduciamo mai esatto, ergo non potremo mai introdurre esatto  $e^{\cos x}$ : dobbiamo approssimare per forza.

- **Si conosce la primitiva, ma con forte rischio di errore.**

In molti casi funzioni semplici hanno funzioni primitive alquanto complicate, dove è facile introdurre perturbazioni. Il paradosso è che le formule approssimate sono più facili da calcolare e il rischio di errore è decisamente minore.

$$f(x) = \frac{x^2}{1+x^4} \quad F(x) = \frac{\sqrt{2}}{4} \left[ \frac{1}{2} \log \frac{x^2 - \sqrt{2}x + 1}{x^2 + \sqrt{2}x + 1} + \arctan \frac{\sqrt{2}x}{1-x^2} \right]$$

- **Insieme discreto.**

Spesso della funzione integranda si conosce solo una restrizione a un insieme discreto (in quel caso dobbiamo pensare, per forza, a un'approssimazione dell'integrale).

---

<sup>1</sup>Da Wikipedia

## 7.2 Grado di precisione ed errore

### 7.2.1 Definizione di funzione peso e di momenti

**Definizione.** Sia  $f(x)$  sufficientemente regolare sull'intervallo  $[a, b]$  dell'asse reale. Sia  $\rho(x)$  una funzione peso non negativa in  $[a, b]$  e tale che esistano finiti i momenti

$$m_k = I(x^k \rho) = \int_a^b x^k \rho(x) dx, k = 0, 1, \dots$$

### 7.2.2 Approssimazione per mezzo di formula di quadratura

Si pone il problema di approssimare

$$I(\rho f) = \int_a^b \rho(x) f(x) dx$$

con una formula di quadratura della forma

$$J_n(f) = \sum_{i=0}^n a_i f(x_i)$$

dove i numeri  $a_i \in \mathbb{R}, i = 0, 1, \dots, n$  sono detti *pesi* o *coefficienti*. Inoltre i punti  $x_i, i = 0, 1, \dots, n$  sono detti *nodi*

$$x_0 < x_1 < \dots < x_n$$

in generale appartengono all'intervallo  $[a, b]$  (non è vietato, ma se si va fuori dall'intervallo non sappiamo se la funzione  $f$  è definita).

### 7.2.3 Errore nell'approssimazione

**Definizione.** L'errore, per quanto detto prima, consiste nella seguente definizione

$$E_n(f) = I(\rho f) - J_n(f)$$

Ricordarsi che con  $n$  al pedice si fa riferimento a una formula avente  $n + 1$  nodi.

Si osservi che parliamo di errore algoritmico in quanto traduciamo l'integrale esatto in una formula di quadratura.

### 7.2.4 Grado di precisione

**Definizione.** Si consideri la base  $1, x, x^2, \dots, x^m, x^{m+1}$  dello spazio vettoriale dei polinomi algebrici di grado al più  $m + 1$ . Affermiamo che la formula di quadratura  $J_n(f)$  ha grado di precisione  $m \in \mathbb{N}$  se si verifica

$$E_n(1) = E_n(x) = \dots = E_n(x^m) = 0, E_n(x^{m+1}) \neq 0$$

cioè integra esattamente la funzione 1, la funzione  $x, \dots$ , fino alla funzione  $x^m$ , ma compie errore nell'approssimare l'integrale di  $x^{m+1}$ .

### 7.2.5 Primo esempio (determinare pesi in $J_1$ )

Prendiamo il seguente integrale

$$\int_{-1}^1 f(x)dx = I(f) \quad \rho(x) = 1$$

Vogliamo calcolare la seguente formula di quadratura

$$J_1(f) = a_0 f(-1) + a_1 f(1)$$

dove i nodi sono  $x_0 = -1$  e  $x_1 = 1$ . Calcolare la formula di quadratura significa determinare i pesi in modo tale che la formula stessa abbia il massimo grado di precisione possibile.

**Risoluzione** Tenendo a mente i seguenti calcoli (ricordarsi che tra parentesi c'è la funzione  $f$  dei nostri calcoli)

$$\begin{aligned} I(1) &= \int_{-1}^1 dx = 1 - (-1) = 2 & I(x) &= \int_{-1}^1 xdx = \frac{(-1)^2}{2} - \frac{1}{2} = 0 \\ J_1(1) &= a_0 + a_1 & J_1(x) &= a_1 - a_0 \end{aligned}$$

affermiamo che

$$\begin{aligned} E_1(f) &= I(f) - J_1(f) \\ E_1(1) &= 0 \implies I(1) - J_1(1) = 0 \implies 2 - a_0 - a_1 = 0 \\ E_1(x) &= 0 \implies I(x) - J_1(x) = 0 \implies 0 + a_0 - a_1 = 0 \end{aligned}$$

Abbiamo trovato il seguente sistema

$$\begin{cases} a_0 + a_1 = 2 \\ a_0 - a_1 = 0 \end{cases}$$

Il risultato è  $a_0 = a_1 = 1$ , quindi

$$J_1(f) = f(-1) + f(1)$$

Quale grado di precisione della formula abbiamo ottenuto?

$$E_1(x^2) = \frac{2}{3} - 2 \neq 0$$

abbiamo ottenuto grado di precisione 1.

#### 7.2.5.1 Conseguenza dell'esempio: formula trapezoidale

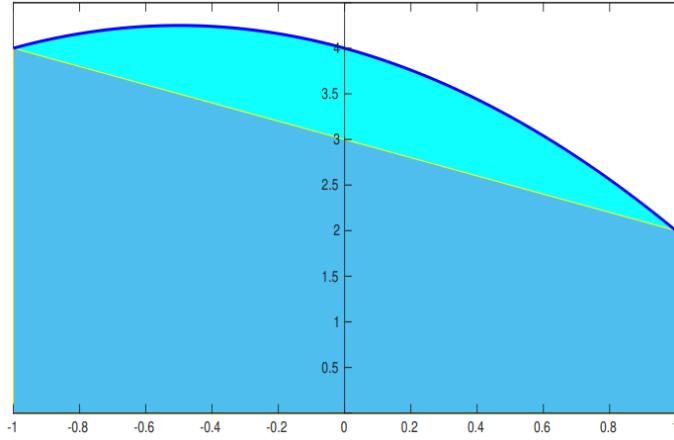
La formula di integrazione numerica ottenuta è nota come *formula trapezoidale*, visto l'interpretazione geometrica. Approssimiamo la seguente funzione

$$\int_a^b f(x)dx$$

con la seguente formula

$$J_1(f) = \frac{b-a}{2}(f(a) + f(b))$$

**Interpretazione grafica** Possiamo dare un'interpretazione grafica ai calcoli fatti, distinguendo l'area calcolata con l'integrale dall'area ottenuta per mezzo di approssimazione.



Abbiamo approssimato l'area della funzione  $f$  per mezzo di una trapezio.

### 7.2.6 Secondo esempio (determinare pesi in $J_2$ )

Prendiamo ancora l'esempio precedente, ma lavoriamo su  $J_2$ :

$$\int_{-1}^1 f(x) dx \quad J_2(f) = a_0 f(-1) + a_1 f(0) + a_2 f(1)$$

Vogliamo determinare i tre pesi in modo tale che la forma sia la migliore approssimazione possibile dell'integrale.

**Risoluzione** Facciamo i seguenti calcoli

$$\begin{aligned} f(x) = 1 &\rightarrow a_0 + a_1 + a_2 = 2 \\ f(x) = x &\rightarrow -a_0 + a_2 = 0 \\ f(x) = x^2 &\rightarrow a_0 + a_2 = \frac{2}{3} \end{aligned}$$

Abbiamo ricavato un sistema a tre equazioni e tre incognite. Il risultato sarà  $a_0 = a_2 = \frac{1}{3}$  e  $a_1 = \frac{4}{3}$ . Segue

$$J_2(f) = \frac{1}{3}f(-1) + \frac{4}{3}f(0) + \frac{1}{3}f(1)$$

Per verificare l'ordine di precisione consideriamo  $f(x) = x^3$

$$-\frac{1}{3} + \frac{1}{3} = 0$$

Quindi il grado di precisione è almeno 3, andiamo avanti con  $f(x) = x^4$

$$\frac{1}{3} + \frac{1}{3} \neq \frac{2}{5}$$

abbiamo trovato che il grado di precisione è proprio 3.

### 7.2.6.1 Conseguenza dell'esempio: formula di Simpson

La formula di integrazione numerica ottenuta è nota come *formula di Simpson*, visto l'interpretazione geometrica. Approssimiamo la seguente funzione

$$\int_a^b f(x)dx$$

con la seguente formula

$$J_2(f) = \frac{b-a}{6} \left( f(a) + 4f\left(\frac{b+a}{2}\right) + f(b) \right)$$

### 7.2.7 Terzo esempio (determinare pesi e nodi in $J_1$ )

Prendiamo il solito esempio

$$\int_{-1}^1 f(x)dx = I(f) \quad J_1(f) = a_0 f(x_0) + a_1 f(x_1)$$

Vogliamo determinare nodi  $x_0, x_1$  e pesi  $a_0, a_1$  in modo tale da avere la migliore approssimazione della funzione.

**Risoluzione** Chiaramente dobbiamo trovare un numero di equazioni sufficiente. Inoltre visto che stiamo ricercando anche nodi non andremo a risolvere un sistema lineare.

$$\begin{cases} a_0 + a_1 = 2 \\ a_0 x_0 + a_1 x_1 = 0 \\ a_0 x_0^2 + a_1 x_1^2 = \frac{2}{3} \\ a_0 x_0^3 + a_1 x_1^3 = 0 \end{cases}$$

- Partiamo dalla ultima equazione

$$a_0 x_0^3 + a_1 x_1^3 = 0$$

e moltiplichiamola per la seconda, ottenendo

$$a_0 x_0^3 + a_1 x_1 x_0^2 = 0$$

Che è valida con  $x_0 \neq 0$ . Prima di proseguire riflettiamo sul caso  $x_0 = 0$ , che verificherebbe al volo questa equazione.

Aggiorniamo il sistema ponendo  $x_0 = 0$  e vediamo se emergono risultati rilevanti

$$\begin{cases} a_0 + a_1 = 2 \\ a_1 x_1 = 0 \\ a_1 x_1^2 = \frac{2}{3} \end{cases}$$

In realtà troviamo che non ci sono risultati: se la seconda equazione è valida allora la terza non può esserlo.

**Quindi.** L'errore è nullo per  $f(x) = 1, f(x) = x$ , ma non per  $f(x) = x^2$ . Segue che avremo precisione  $m = 1$ .

- Adesso supponiamo che  $x_0 \neq 0$ , recuperiamo l'equazione ottenuta precedente

$$a_0x_0^3 + a_1x_1x_0^2 = 0$$

e facciamo sottrazione tra la quarta equazione e l'equazione precedente. Otteniamo

$$a_1x_1^3 - a_1x_1x_0^2 = 0$$

raccogliamo

$$a_1x_1(x_1^2 - x_0^2) = 0$$

le possibilità solo  $a_1 = 0, x_1 = 0$  oppure  $x_1 = x_0$  e  $x_1 = -x_0$ .

- $x_1 = 0$  ci restituisce un sistema identico a quello già visto. Otteniamo un sistema non risolubile e precisione  $m = 1$ .
- $a_1 = 0$ . Otteniamo il seguente sistema

$$\begin{cases} a_0 = 2 \\ a_0x_0 = 0 \end{cases}$$

Anche in questo caso abbiamo grado di precisione  $m = 1$ : unico modo per risolvere il sistema è imporre  $x_0 = 0$ .

- $x_1 = x_0$ . Otteniamo il seguente sistema

$$\begin{cases} a_0 + a_1 = 2 \\ x_0(a_0 + a_1) = 0 \end{cases}$$

L'unico modo per avere il sistema valido è imporre  $x_0 = 0$ .

- $x_1 = -x_0$ . Otteniamo il seguente sistema (la quarta è automaticamente verificata se è verificata la seconda equazione, quindi la escludiamo)

$$\begin{cases} a_0 + a_1 = 2 \\ x_0(a_0 - a_1) = 0 \\ x_0^2(a_0 + a_1) = \frac{2}{3} \end{cases}$$

Se  $x_0 = 0$  abbiamo grado di precisione  $m = 1$ . Altrimenti dobbiamo imporre  $a_0 = a_1$ : dalla prima equazione  $a_0 = a_1 = 1$ , dalla quarta  $x_0 = x_1 = \pm\frac{\sqrt{3}}{3}$ . Otteniamo in questo caso grado di precisione  $m = 3$ !

L'equazione trovata è una sola

$$J_1(f) = f\left(\frac{\sqrt{3}}{3}\right) + f\left(-\frac{\sqrt{3}}{3}\right)$$

### 7.2.8 Teorema di Peano per la rappresentazione dell'errore

**Teorema.** Vogliamo dare all'errore  $E_n(f)$  una rappresentazione generale, evitando formule specifiche. Ribadiamo le condizioni su cui abbiamo lavorato fino ad ora

$$\rho(x) = 1 \quad a = x_0 < x_1 < \dots < x_n = b$$

Sia  $f(x) \in \mathbb{C}^{m+1}([a, b])$  e  $J_n(f)$  avente grado di precisione  $m$ , allora possiamo dire

$$E_n(f) = \frac{1}{m!} \int_a^b f^{(m+1)}(t)G(t) dt$$

dove  $G(t) = E_n(s_m(x-t))$  (detta *nucleo di Peano*) e

$$s_m(x-t) = \begin{cases} (x-t)^m & t < x \\ 0 & t \geq x \end{cases}$$

L'osservazione principale è che l'errore dipende dalla derivata  $(m+1)$ -esima: come mai?

- Prendiamo una formula con grado di precisione  $m=3$ , quella vista prima. Se io uso quella formula allora potrò integrare in modo esatto i polinomi di grado al più 3.
- La derivata quarta di un polinomio di al più grado 3 è identicamente nulla (nulla  $\forall x$ ).

#### 7.2.8.1 Semplificazione con nucleo di Peano costante in segno

Se il nucleo di Peano non cambia segno in  $[a, b]$  allora possiamo dire

$$E_n(f) = \frac{1}{m!} f^{(m+1)}(\theta) \int_a^b G(t) dt$$

dove  $\theta$  è un valore prefissato  $\in ]a, b[$ . Supponiamo di voler calcolare l'errore con  $f(x) = x^{m+1}$ , otteniamo (si tenga a mente che la derivata  $(m+1)$ -esima è una costante)

$$E_n(x^{m+1}) = \frac{1}{m!} (m+1)! \int_a^b G(t) dt = (m+1) \int_a^b G(t) dt$$

L'unica cosa che non conosco in  $E_n(f)$  è l'integrale del nucleo di Peano. Quindi poniamo

$$\int_a^b G(t) dt = \frac{1}{m+1} E_n(x^{m+1})$$

in conclusione (i calcoli vanno bene visto che  $G(t)$  non dipende dalla funzione  $f$ )

$$E_n(f) = \frac{f^{(m+1)}(\theta)}{(m+1)!} E_n(x^{m+1})$$

#### 7.2.8.2 Errore nella formula trapezoidale

Avevo visto nell'esempio 1 che

$$\int_{-1}^1 f(x) dx = f(-1) + f(1) + E_1(f)$$

se voglio porre l'uguale è chiaro che dovrò rappresentare anche l'errore, lo facciamo con le nozioni appena viste. E' dimostrabile che il nucleo di Peano è costante in segno, quindi possiamo applicare la formula precedentemente dimostrata. Sapendo che  $m=1$  e

$$E_1(x^2) = \frac{2}{3} - 2 = -\frac{4}{3}$$

otteniamo

$$E_1(f) = -\frac{4}{3} \frac{f^{(2)}(\theta)}{2} = -\frac{2}{3} f^{(2)}(\theta)$$

generalizzando (intervallo generico  $[a, b]$ )

$$E_1(f) = -\frac{(b-a)^3}{12} f^{(2)}(\theta)$$

### 7.2.8.3 Errore nella formula di Simpson

Avevo visto nell'esempio 2 che

$$\int_{-1}^1 f(x)dx = \frac{1}{3}(f(-1) + 4f(0) + f(1)) + E_2(f)$$

se voglio porre l'uguale è chiaro che dovrò rappresentare anche l'errore, lo facciamo con le nozioni appena viste. È dimostrabile che il nucleo di Peano è costante in segno, quindi possiamo applicare la formula precedentemente dimostrata. Sapendo che  $m = 3$  e

$$E_2(x^4) = \frac{2}{5} - \frac{2}{3} = -\frac{4}{15}$$

otteniamo

$$E_2(f) = -\frac{4}{14} \frac{f^{(4)}(\theta)}{24} = -\frac{1}{90} f^{(4)}(\theta)$$

generalizzando (intervallo generico  $[a, b]$ )

$$E_2(f) = -\frac{(b-a)^5}{2880} f^{(4)}(\theta)$$

### 7.2.8.4 Esempio di esercizio

Supponiamo di voler approssimare il seguente integrale

$$I(f) = \int_{\sqrt{2}}^{\sqrt{8}} f(x)dx$$

la formula di quadratura proposta è la seguente

$$J_1(f) = \frac{\sqrt{2}}{2} \left( f(\sqrt{2}) + f(\sqrt{8}) \right)$$

supposto che il nucleo di Peano sia costante possiamo affermare che l'errore è posto nella seguente forma

$$E_1(f) = K f^{(s)}(\xi)$$

Vogliamo trovare due valori:  $k$  ed  $s$ .

#### Risoluzione

- Ricordiamo che  $s = m + 1$ , segue che ci basta trovare il grado di precisione della funzione  $f(x)$ .

$$E_1(1) = \sqrt{8} - \sqrt{2} - \left( \frac{\sqrt{2}}{2} 2 \right) = 2\sqrt{2} - \sqrt{2} - \sqrt{2} = 0$$

$$E_1(x) = 4 - 1 - \left( \frac{\sqrt{2}}{2} (\sqrt{2} + \sqrt{8}) \right) = 3 - 3 = 0$$

$$E_1(x^2) = \frac{1}{3} \left( 16\sqrt{2} - 2\sqrt{2} \right) - \frac{\sqrt{2}}{2} (2+8) = \frac{14\sqrt{2}}{3} - 5\sqrt{2} = -\frac{1}{3}\sqrt{2}$$

Abbiamo trovato che il grado di precisione è  $m = 1$ , segue  $s = 2$ .

- Abbiamo appena trovato il valore  $E_1(x^2)$ . Sappiamo anche che la derivata  $f^{(s)}(\xi)$  è una costante. Segue

$$E_1(x^2) = k2 \longrightarrow -\frac{1}{3}\sqrt{2} = 2k$$

concludiamo

$$k = -\frac{1}{6}\sqrt{2}$$

## 7.3 Formule di tipo interpolatorio

### 7.3.1 Dimostrazione di premessa alla definizione

Per approssimare il seguente integrale

$$I(\rho f) = \int_a^b \rho(x)f(x)dx$$

prendiamo in esame la formula di quadratura  $J_n(f)$ , dove abbiamo fissato i nodi  $x_0, x_1, \dots, x_n$  ( $x_i \neq x_j$  se  $i \neq j$ ). Vogliamo ricavare una *formula di tipo interpolatorio*.

**Svolgimento** Scriviamo  $f(x)$  per mezzo di un polinomio di interpolazione di Lagrange

$$f(x) = L_n(x) + E_n(x)$$

cioè

$$I(\rho f) = \int_a^b \rho(x)(L_n(x) + E_n(x))dx = \int_a^b (\rho(x)L_n(x) + \rho(x)E_n(x))dx =$$

scomponiamo l'integrale ricorrendo alla proprietà degli integrali di somme

$$= \int_a^b \rho(x)L_n(x)dx + \int_a^b \rho(x)E_n(x)dx$$

Ricordiamo che

$$L_n(x) = \sum_{i=0}^n l_i(x)f(x_i)$$

e sostituiamo

$$\begin{aligned} &= \int_a^b \rho(x) \left( \sum_{i=0}^n l_i(x)f(x_i) \right) dx + \int_a^b \rho(x)E_n(x)dx \\ &= \sum_{i=0}^n f(x_i) \int_a^b \rho(x)l_i(x)dx + \int_a^b \rho(x)E_n(x)dx \end{aligned}$$

l'integrale  $a_i = \int_a^b \rho(x)l_i(x)dx$  dipende solo dai nodi e non dalla funzione  $f$ . Poniamo

$$= \sum_{i=0}^n a_i f(x_i) + \int_a^b \rho(x)E_n(x)dx = J_n(f) + E_n(f)$$

Abbiamo trovato  $J_n(f)$ , ed è di tipo interpolatorio perchè si coinvolge il polinomio interpolatorio dell'equazione di Lagrange. La sostanza è che approssimiamo l'integrale esatto con l'integrale del polinomio di interpolazione di Lagrange!

### 7.3.2 Definizione di formule di quadratura di tipo interpolatorio

**Definizione.** Si dicono *formule di quadratura di tipo interpolatorio* le formule del tipo

$$J_n(f) = \sum_{i=0}^n a_i f(x_i)$$

dove, fissati i nodi  $x_i$ ,  $i = 0, 1, \dots, n$  abbiamo i pesi

$$a_i = \int_a^b \rho(x)l_i(x)dx$$

### 7.3.3 Unicità della formula di quadratura

Se i nodi della formula sono prefissati allora i pesi sono determinati dalle prime  $n + 1$  equazioni del sistema  $V\alpha = \mu$ , dove

$$V = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_0 & x_1 & \dots & x_n \\ \dots & \dots & \ddots & \dots \\ x_0^2 & x_1^n & \dots & x_n^n \end{pmatrix} \quad \alpha = \begin{pmatrix} a_0 \\ a_1 \\ \dots \\ a_n \end{pmatrix} \quad \mu = \begin{pmatrix} m_0 \\ M_1 \\ \dots \\ m_n \end{pmatrix}$$

Essendo una matrice Vandermonde, che ha determinante diverso da zero se i nodi sono distinti, avremo unicità della formula di quadratura.

### 7.3.4 Formule di Newton-Cotes

#### 7.3.4.1 Caratteristiche

Le formule di Newton-Cotes presentano le seguenti caratteristiche:

- la funzione peso è  $\rho(x) = 1$ ;
- i nodi sono fissati in progressione aritmetica, cioè sono scelti in modo tale che

$$x_0 = a \quad x_i = x_0 + ih \quad x_n = b$$

dove  $i = 0, 1, \dots, n$ .  $h$  è detto *passo della formula*;

- il nucleo di Peano  $G(t)$  è costante in segno nell'intervallo  $[a, b]$  considerato, ergo possiamo utilizzare la formula dell'errore semplificata.

I nodi della formula si individuano con la tecnica interpolatoria introdotta:  $a_i = I(l_i(x))$ .

**Esempi** La formula trapezoidale e la formula di Simpson sono formule di Newton-Cotes.

- Nella formula trapezoidale il passo  $h$  è  $b - a$ .
- Nella formula di Simpson il passo è  $h = \frac{b-a}{2}$

**Problema** Nelle formule calcoliamo i pesi per  $n$  valori. Se calcoliamo fino a  $n = 7$  pesi questi sono positivi, ma per  $n > 7$  compaiono pesi negativi e le formule diventano numericamente instabili (capacità di amplificare l'errore aumenta).

#### 7.3.4.2 Formula dei trapezi (generalizzazione della formula trapezoidale)

Se per una formula a  $n + 1$  punti il passo di integrazione è troppo ampio

$$h = \frac{b-a}{n}$$

allora si divide  $[a, b]$  in  $L$  parti uguali, mediante i punti

$$a = x_0 < x_1 < \dots < x_L = b$$

L'integrale può essere scritto nella seguente forma

$$\int_a^b f(x)dx = \int_{x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \dots + \int_{x_{L-1}}^{x_L} f(x)dx = \sum_{j=1}^L \int_{x_{i-1}}^{x_i} f(x)dx$$

Supponiamo che i punti  $x_i$  siano presi in progressione aritmetica

$$x_j = a + j \frac{b-a}{L}, \quad j = 0, 1, \dots, L$$

e applichiamo a ciascun integrale la formula trapezoidale. Ricordiamoci le formule

$$J_1(f) = \frac{b-a}{2} (f(a) + f(b)) \quad E_1(f) = -\frac{(b-a)^3}{12} f''(\theta)$$

Se andiamo a sommare tutti i contributi (sommiamo  $L$  volte) otteniamo

$$J_1(f) = \frac{b-a}{2L} \left( f(x_0) + 2 \sum_{j=1}^{L-1} f(x_j) + f(x_L) \right)$$

Facciamo la stessa cosa con l'errore, otteniamo

$$E_1(f) = -\frac{1}{12} \left( \frac{b-a}{L} \right)^3 L f''(\theta) = -\frac{1}{12} \frac{(b-a)^3}{L^2} f''(\theta)$$

dove sostituiamo  $L$  valutazioni della derivata seconda

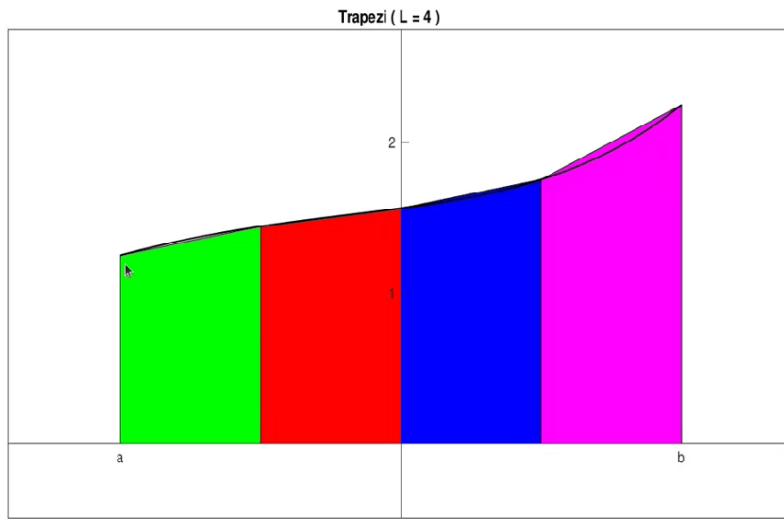
$$f''(\theta_1) + f''(\theta_2) + \dots + f''(\theta_L)$$

con un'unica valutazione  $L \cdot f''(\theta)$ .

**Grado di precisione** Il grado di precisione della formula del trapezio è lo stesso della formula trapezoidale. Allora perchè facciamo tutti questi calcoli? Perchè otteniamo  $L$  al denominatore, ergo possiamo ridurre la derivata seconda. Inoltre, l'errore tende a zero con  $L$  tendente a  $\infty$ .

**Numero di punti** Il numero di valutazioni considerate nella formula è  $N_T = L + 1$ .

**Interpretazione grafica** L'interpretazione grafica è simile a quella vista per la forma trapezoidale, ma quanto detto sulla formula trapezoidale si applica ai singoli intervalli.



#### 7.3.4.3 Formula di Cavalieri-Simpson (generalizzaz. della formula di Simpson)

Se applichiamo lo stesso approccio, ma applicando la formula di Simpson in ogni intervallo, otteniamo

$$I(f) = \frac{b-a}{6L} \left[ f(x_0) + 4 \sum_{i=0}^{L-1} f\left(\frac{x_i + x_{i+1}}{2}\right) + 2 \sum_{i=1}^{L-1} f(x_i) + f(x_L) \right] - \frac{(b-a)^5}{2880L^4} f^{(4)}(\theta)$$

**Numero di punti** Il numero di valutazioni considerate nella formula è  $N_{CS} = 2L + 1$ . Si raddoppia  $L$  perchè si considerano anche i punti medi.

#### 7.3.4.4 Esempio con confronto tra Trapezi e Cavalieri-Simpson

Si vuole approssimare il seguente integrale

$$\int_0^1 \frac{1}{1+x} dx$$

con un massimo errore assoluto  $E$  con  $|E| \leq 10^{-2}$ .

**Risoluzione** Se risolviamo in modo classico l'integrale ci accorgiamo come noi stiamo approssimando un numero irrazionale

$$\int_0^1 \frac{1}{1+x} dx = \log(2)$$

Per prima cosa andiamo a fare le derivate

$$\begin{aligned} f(x) &= (1+x)^{-1} \\ f'(x) &= -(1+x)^{-2} \\ f''(x) &= 2(1+x)^{-3} \\ f'''(x) &= -6(1+x)^{-4} \\ f^{(IV)}(x) &= 24(1+x)^{-5} \end{aligned}$$

- **Formula dei trapezi.**

Calcoliamo l'errore prendendo la formula relativa a quella dei trapezi. Il fatto è che noi non conosciamo il valore  $|f''(\theta)|$ , ma possiamo fare una stima

$$|E_1(f)| = \frac{b-a}{12L^2} |f''(\theta)|$$

con  $M_2 \geq \sup_{x \in [a,b]} |f''(x)|$ . In questo caso otteniamo  $M_2 = 2$ . Segue

$$\frac{b-a}{12L^2} 2 < \frac{1}{2} 10^{-2}$$

si osservi che abbiamo posto  $\frac{1}{2}$ , ricordando che stiamo trattando l'errore algoritmico e che l'errore complessivo si distingue in errore algoritmico ed errore dovuto alla trasmissione dei dati. Con pochi passaggi otteniamo

$$L \geq 6$$

Nel caso della formula dei trapezi otterremo  $N_T = 7$ .

- **Formula di Cavalieri-Simpson.**

Facciamo la stessa cosa con la formula di Cavalieri Simpson, dove  $M_4 \geq \sup_{x \in [a,b]} |f^{(iv)}(x)|$

$$|E_2(f)| = \frac{(b-a)^5}{2880L^4} |f^{(IV)}(\theta)|$$

Anche qua dividiamo per  $1/2$

$$\frac{1}{2880L^4} 24 < \frac{1}{2} 10^{-2}$$

il risultato è  $L \geq 2$ . Segue  $N_{CS} = 5$ .

Se facciamo un confronto è più conveniente la formula di Cavalieri-Simpson, visto il numero inferiore di valutazioni.

L'integrale proposto può essere calcolato esattamente e si ha

$$\log 2 = \int_0^1 \frac{1}{1+x} dx \simeq 0.69314718\dots$$

mentre risultano

$$J_1^{(G)}(f) = 0.69487734\dots \quad J_2^{(G)}(f) = 0.69325396\dots$$

#### 7.3.4.5 Tecnica di estrapolazione con la formula dei Trapezi

Si può dimostrare che se  $f(x)$  è sufficientemente regolare in  $[a, b]$  allora l'errore della formula dei trapezi ammette uno sviluppo in serie di potenze pari di  $h$ . Poniamo il primo passo

$$J_0^{(1)} = \frac{h}{2} \left[ f(x_0) + 2 \sum_{i=1}^{L-1} f(x_i) + f(x_L) \right]$$

Se  $f(x) \in C^{2r+2}([a, b])$  allora possiamo dire

$$I(f) = J_0^{(1)} + \alpha_1^{(1)} h^2 + \alpha_2^{(1)} h^4 + \alpha_3^{(1)} h^6 + \dots + \alpha_r^{(1)} h^{2r} + O(h^{2r+2})$$

dove con l'ultima notazione rappresentiamo termini che si comportano come  $h^{2r+2}$ . Si osservi che i coefficienti  $\alpha_1, \alpha_2, \dots$  non dipendono da  $h$ .

- $J_0^{(1)}$  è l'approssimazione che ho ottenuto prendendo un passo  $h$ .
- Calcoliamo con la formula dei trapezi  $J_1^{(1)}$ , scegliendo un passo  $\frac{h}{q}$ , dove  $q \in \mathbb{N}$ . Lo sviluppo detto precedentemente vale anche qua

$$I(f) = J_0^{(1)} + \alpha_1^{(1)} \left( \frac{h}{q} \right)^2 + \alpha_2^{(1)} \left( \frac{h}{q} \right)^4 + \alpha_3^{(1)} \left( \frac{h}{q} \right)^6 + \dots + \alpha_r^{(1)} \left( \frac{h}{q} \right)^{2r} + O \left( \left( \frac{h}{q} \right)^{2r+2} \right)$$

- Ricaviamoci  $\alpha_1^{(1)} h^2$  dallo sviluppo iniziale (quello con passo  $h$ ).

$$\alpha_1^{(1)} h^2 = I(f) - J_0^{(1)} - \alpha_2^{(1)} h^4 - \dots - \alpha_r^{(1)} h^{2r} + O(h^{2r+2})$$

sostituiamo nello sviluppo con passo  $\frac{h}{q}$

$$I(f) = J_1^{(1)} + \frac{1}{q^2} \left( I(f) - J_0^{(1)} - \alpha_2^{(1)} h^4 - \dots - \alpha_r^{(1)} h^{2r} \right) + \alpha_2^{(1)} \left( \frac{h}{q} \right)^4 + \alpha_3^{(1)} \left( \frac{h}{q} \right)^6 + \dots + \alpha_r^{(1)} \left( \frac{h}{q} \right)^{2r} + \dots$$

- Spostiamo il termine con l'integrale dal secondo al primo membro

$$I(f) - \frac{1}{q^2} I(f) = J_1^{(1)} - \frac{1}{q^2} J_0^{(1)} + \beta_2 h^4 + \beta_3 h^6 + \dots + \beta_r h^{2r} + O(h^{2r+1})$$

dove i coefficienti  $\beta_k$  si ottengono raccogliendo i termini e ricordando  $(h/q)^k = h^k/q^k$ .

- Concludiamo i calcoli

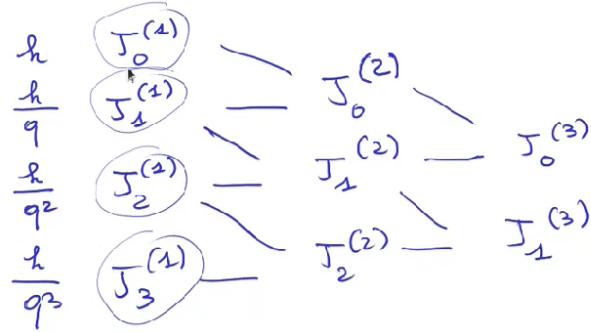
$$\frac{q^2 - 1}{q^2} I(f) = \frac{q^2 J_1^{(1)} - J_0^{(1)}}{q^2} + \beta_2 h^4 + \cdots + \beta_r h^{2r} + O(h^{2r+1})$$

$$I(f) = \frac{q^2 J_1^{(1)} - J_0^{(1)}}{q^2 - 1} + \alpha_2^{(2)} h^4 + \cdots + \alpha_r^{(2)} h^{2r} + O(h^{2r+1})$$

**Perchè tutti questi calcoli?** Si osservi che inizialmente la differenza  $I(f) - J_0^{(1)}$  va a 0 con  $h^2$ , mentre nella formula finale va a zero con  $h^4$ : segue una migliore approssimazione dell'integrale. Il primo termine ottenuto (detto *formula di Ronberg*) è una nuova approssimazione dell'integrale, migliore

$$J_0^{(2)} = \frac{q^2 J_1^{(1)} - J_0^{(1)}}{q^2 - 1}$$

Possiamo utilizzare due approssimazioni per ottenere una terza, più precisa



# **Parte III**

# **Appendici**

# Appendice A

## Calcolo di determinanti

Sezione posta per comodità personale, non c'è stata spiegazione del prof. a riguardo.

### Determinanti $2 \times 2$

Per i determinanti di matrici  $2 \times 2$  si ricorre banalmente alla seguente formula.

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - cb$$

### Determinanti $3 \times 3$

Per i determini di matrici  $3 \times 3$  si può ricorrere alla *regola di Sarrus*.

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = aei + bfg + dhc - (ceg + fhd + bdi)$$

Per ricordarsi la formula ricordarsi che

- facciamo la differenza tra due sommatorie;
- il primo termine della prima sommatoria è il prodotto dei termini della diagonale principale;
- il primo termine della seconda sommatoria è il prodotto dei termini dell'anti-diagonale;
- i termini successivi delle sommatorie si ottengono a partire dal prodotto dei termini delle codiagonali, a cui si aggiunge (nel prodotto) il termine opposto alla codiagonale.

### Sviluppo di Laplace per determinanti di ordine superiore

In generale non andremo a calcolare determinanti di matrici di dimensioni stratosferiche, salvo matrici che per le loro caratteristiche permettono un calcolo veloce del determinante. Molto utile è il cosiddetto *sviluppo di Laplace*. Si consideri la seguente matrice

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$$

Si introducono due tipologie di sviluppo, dove si sceglie una riga  $i$ -esima o una colonna  $j$ -esima. La scelta della riga o della colonna è fatta in modo tale da minimizzare i calcoli necessari (si veda l'esempio).

## Sviluppo per righe

Fissata una riga  $i$ -esima della matrice  $A$ , il determinante della matrice stessa consiste nella seguente sommatoria

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

dove la matrice  $A_{ij}$  è una sottomatrice ottenuta escludendo la riga  $i$ -esima e la colonna  $j$ -esima.

## Sviluppo per colonne

Fissata una colonna  $j$ -esima della matrice  $A$ , il determinante della matrice stessa consiste nella seguente sommatoria

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

dove la matrice  $A_{ij}$  è una sottomatrice ottenuta escludendo la riga  $i$ -esima e la colonna  $j$ -esima.

## Esempio

Si consideri la seguente matrice

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

Proviamo due possibili risoluzioni

- **Prima riga**  $i = 1$ .

$$\begin{aligned} \det(A) &= 2(-1)^{1+1} \det \begin{pmatrix} 2 & 0 \\ 4 & 3 \end{pmatrix} + 1(-1)^{1+2} \det \begin{pmatrix} 2 & 0 \\ 4 & 3 \end{pmatrix} + 1(-1)^{1+3} \det \begin{pmatrix} 2 & 2 \\ 4 & 4 \end{pmatrix} = \\ &= 2[6 - 0] - 1[6 - 0] + 1[8 - 8] = 12 - 6 = 6 \end{aligned}$$

- **Terza colonna**  $j = 3$ .

$$\begin{aligned} \det(A) &= 1(-1)^{1+3} \det \begin{pmatrix} 2 & 2 \\ 4 & 4 \end{pmatrix} + 0 + 3(-1)^{1+5} \det \begin{pmatrix} 2 & 1 \\ 2 & 2 \end{pmatrix} = \\ &= 1[8 - 8] + 3[4 - 2] = 6 \end{aligned}$$

Abbiamo ottenuto lo stesso risultato, ma prendendo la terza colonna ci siamo semplificati i calcoli (un elemento è nullo, quindi un determinante in meno da calcolare)!

# Differenze con il programma degli anni precedenti (corso prof. Ghelardoni)

## Argomenti aggiuntivi

- Numeri sottonormalizzati (detti anche subnormalizzati o denormalizzati).
- Costo computazionale delle operazioni aritmetiche matriciali e vettoriali.
- Calcolo della fattorizzazione LU (aka LR) con pivoting parziale e totale (ovvero dei fattori L, U e della matrice/i di permutazione).
- Costo computazionale dell'algoritmo di sostituzione all'indietro per sistemi lineari triangolari e dell'algoritmo di eliminazione di Gauss per sistemi generali.
- Dimostrazione del teorema: matrice irriducibile se e solo se il grafo orientato è fortemente connesso.
- Dimostrazione dei teoremi di convergenza del metodo di Gauss Seidel per matrici a predominanza diagonale.
- Fattorizzazione QR e come calcolarla mediante trasformazioni di Householder.
- Metodo basato sulla fattorizzazione QR per risolvere problemi lineari ai minimi quadrati.

## Argomenti non trattati

- Teorema di Sturm e sua applicazione alla soluzione di equazioni algebriche.
- Metodo di Jacobi per il calcolo degli autovalori di una matrice reale e simmetrica.

## Bibliografia

- [1] Metodi Numerici, R. Bevilacqua, D. Bini, M. Capovani, O. Menchi, 1992.
- [2] Metodi Numerici per l'Algebra Lineare, D. Bini, M. Capovani, O. Menchi, 1988.

Entrambi i testi sono liberamente scaricabili dalla pagina e-learning del corso.