

Calcolo Numerico ed Elementi di Analisi

Luca Dede'
Andrea Manzoni

21 marzo 2023

Indice

1	Introduzione	1
1.1	La Modellistica Matematica e il Calcolo Scientifico	1
1.1.1	Perché i Modelli e non solo i Dati?	1
1.1.2	A cosa serve il Calcolo Scientifico?	2
1.2	Tre Classi Fondamentali di Problemi	3
1.3	Soluzione Esatta e Approssimazione Numerica	6
1.4	A Cosa Servono i Metodi Numerici?	7
1.5	Rappresentazione dei Numeri e Operazioni al Calcolatore	9
1.5.1	Numeri Floating point	9
1.5.2	Aritmetica floating point	13
1.6	Dal Problema Matematico al Problema Numerico	14
1.6.1	Il problema matematico	14
1.6.2	Il problema numerico	16
1.6.3	Scelta di un metodo numerico	19
2	Sistemi Lineari	21
2.1	Motivazioni, Esempi e Classificazione dei Metodi	21
2.2	Metodi Diretti	22
2.2.1	Sistemi lineari “semplici”	22
2.2.2	Metodo di fattorizzazione LU	25
2.2.3	Algoritmo di Thomas	34
2.2.4	Metodo della fattorizzazione di Cholesky	35
2.2.5	Accuratezza della soluzione numerica ottenuta mediante metodi diretti	36
2.2.6	Sistemi sovradeterminati	41
2.2.7	Il comando \ di Matlab	43
2.3	Metodi Iterativi	43
2.3.1	Lo schema generale	43
2.3.2	Metodi di decomposizione additiva (metodi di splitting)	44
2.3.3	Metodi di Jacobi e Gauss–Seidel	45
2.3.4	Metodi di Richardson precondizionati	47
2.3.5	Metodo del gradiente	51
2.3.6	Metodo del gradiente coniugato	54
2.3.7	Metodi iterativi ed errore computazionale	56
2.3.8	Criteri d’arresto per metodi iterativi	58
2.4	Un (breve) Confronto tra Metodi Diretti e Iterativi	59
3	Autovalori e Autovettori	61
3.1	Definizioni ed Esempi	61
3.2	Metodo delle Potenze	63
3.3	Metodo delle Potenze Inverse	65
3.4	Metodi delle Potenze Inverse con Shift	66
3.5	Localizzazione Geometrica degli Autovalori: Criteri di Gershgorin	67

3.6	Metodo delle Iterazioni QR	69
3.7	Decomposizione ai Valori Singolari di una Matrice	70
4	Equazioni Non Lineari	75
4.1	Metodo di Bisezione	76
4.1.1	Costruzione del metodo di bisezione	76
4.1.2	Algoritmo e proprietà del metodo	77
4.1.3	Criterio d'arresto	78
4.2	Metodi di Newton	79
4.2.1	Metodo di Newton	79
4.2.2	Metodo di Newton modificato	81
4.2.3	Criteri d'arresto per i metodi di Newton	82
4.2.4	Metodi di quasi–Newton e inesatti	83
4.2.5	Metodo di Newton per sistemi di equazioni non lineari	84
4.3	Iterazioni di Punto Fisso	85
4.3.1	Equazioni non lineari, zeri, punti fissi e funzioni di iterazione	85
4.3.2	Algoritmo delle iterazioni di punto fisso	86
4.3.3	Proprietà di convergenza del metodo delle iterazioni di punto fisso	87
4.3.4	Criterio d'arresto per iterazioni di punto fisso	90
4.3.5	Il metodo di Newton come metodo delle iterazioni di punto fisso	91
4.3.6	Il metodo delle corde come metodo delle iterazioni di punto fisso	92
4.3.7	Iterazioni di punto fisso per funzioni vettoriali	93
5	Approssimazione di Funzioni e Dati	95
5.1	Motivazioni ed Esempi	95
5.1.1	Approssimazione di funzioni tramite polinomi di Taylor	96
5.2	Interpolazione	96
5.2.1	Interpolazione polinomiale di Lagrange	97
5.2.2	Interpolazione trigonometrica	103
5.2.3	Interpolazione polinomiale a tratti	105
5.2.4	Splines	106
5.3	Metodo dei Minimi Quadrati	108
5.3.1	La retta di regressione	109
6	Integrazione Numerica	111
6.1	Scopo e Classificazione delle Formule di Quadratura	111
6.2	Formule di Quadratura del Punto Medio	112
6.3	Formule di Quadratura del Trapezio	114
6.4	Formule di Quadratura di Simpson	116
6.5	Formule di Quadratura Interpolatorie	117
6.5.1	Formule di quadratura Gaussiane	119
6.5.2	Formule di quadratura di Gauss–Legendre	119
6.5.3	Formule di quadratura di Gauss–Legendre–Lobatto	121
6.6	Integrazione Numerica in Dimensione $d > 1$	122
7	Equazioni Differenziali Ordinarie	123
7.1	Esempi Notevoli di EDO	123
7.2	Il Problema di Cauchy: Buona Posizione	124
7.2.1	Il problema di Cauchy (o ai valori iniziali)	125
7.2.2	Esistenza e unicità (globale) per il problema di Cauchy	128
7.2.3	Stabilità secondo Liapunov del problema di Cauchy	130
7.3	Derivazione Numerica	132
7.3.1	Schemi alle differenze finite per la derivata prima	133
7.3.2	Schema alle differenze finite per la derivata seconda	136

7.4	Approssimazione Numerica di EDO del Primo Ordine	136
7.4.1	Metodo di Eulero in avanti	137
7.4.2	Metodo di Eulero all'indietro	137
7.4.3	Metodo di Crank-Nicolson	139
7.4.4	Metodo di Heun	140
7.4.5	Analisi dell'errore dei metodi	140
7.4.6	Stabilità dei metodi numerici: zero-stabilità e stabilità assoluta	143
7.4.7	Metodi Runge-Kutta	148
7.4.8	Metodi multipasso	150
7.5	Approssimazione Numerica di Sistemi di EDO del Primo Ordine	151
7.5.1	Il problema di Cauchy nel caso vettoriale	151
7.5.2	Metodi numerici per sistemi di EDO del primo ordine	152
7.5.3	θ -metodo	152
7.6	Approssimazione Numerica di EDO del Secondo Ordine	154
7.6.1	Riscrittura dell'EDO del secondo ordine come sistema di EDO del primo ordine	155
7.6.2	Metodo Leap Frog (non svolto a lezione)	156
7.6.3	Metodo di Newmark (non svolto a lezione)	157
7.7	Approssimazione Numerica di EDO di Ordine Superiore a Due	158
8	Problemi ai Limiti e ai Valori Iniziali	159
8.1	Definizioni ed Esempi	159
8.1.1	Da dove viene un modello? Leggi fisiche e leggi costitutive	161
8.2	Differenze Finite per Problemi ai Limiti 1D	162
8.2.1	Differenze finite per il problema di Poisson con condizioni di Dirichlet	164
8.2.2	Differenze finite per problemi di diffusione-trasporto-reazione con condizioni di Dirichlet	167
8.2.3	Differenze finite per problemi di diffusione-trasporto e schema Upwind	169
8.2.4	Differenze finite per il problema di Poisson con condizioni miste di Dirichlet/Neumann	176
8.3	Differenze Finite per il Problema di Poisson 2D	178
8.4	Differenze Finite per l'Equazione del Calore (di Diffusione) 1D	181
8.4.1	Semi-discretizzazione in spazio	182
8.4.2	Discretizzazione in tempo	182
8.4.3	Stabilità (asintotica)	184

Capitolo 1

Introduzione

Illustriamo in questa introduzione il significato di modello matematico e di modelli costruiti a partire da leggi fisiche; in particolare, ci focalizziamo sul ruolo del moderno Calcolo Scientifico in relazione alla modellistica matematica. Dopo aver introdotto tre problemi modello, motiviamo l'esigenza di introdurre opportune tecniche di approssimazione numerica per la loro risoluzione, e alcuni concetti fondamentali riguardanti i metodi numerici, che verranno in futuro declinati per ciascun argomento. Dal momento che la propagazione degli errori di arrotondamento e troncamento impatta sul risultato di ciascuna operazione svolta al calcolatore, presentiamo anche una sintetica introduzione alla rappresentazione dei numeri macchina.

1.1 La Modellistica Matematica e il Calcolo Scientifico

La *modellistica matematica* gioca un ruolo cruciale al giorno d'oggi nella descrizione di numerosi fenomeni delle Scienze applicate e dell'Ingegneria. Con **modello matematico** intendiamo un insieme di equazioni (algebriche o differenziali) capaci di catturare le caratteristiche essenziali di un sistema complesso allo scopo di descrivere, prevedere e controllare il suo comportamento o la sua evoluzione.

In ambito industriale (basti pensare all'esigenza di progettare le componenti di un velivolo al fine di migliorare la loro efficienza aerodinamica) la modellistica matematica fornisce oggi tecniche e procedure estremamente diffuse grazie alle capacità computazionali ormai disponibili, e imprescindibili, insieme alle attività sperimentali, per lo sviluppo e il miglioramento tecnologico.

I modelli matematici basati sulla fisica sono alla base del moderno **Calcolo Scientifico** – si può parlare, a tal proposito, di *Modeling Based Scientific Computing* (MBSC), o anche di *Physics Based Modeling*. Tali modelli sono derivati da principi primi (come le leggi di conservazione della massa, del momento, dell'energia, della carica, ...) che codificano le leggi della natura e conducono alla scrittura di equazioni (molto spesso differenziali) la cui soluzione non è esprimibile, solitamente, in forma esplicita.

L'approssimazione di tali equazioni richiede algoritmi numerici accurati ed efficienti, che trasformano il problema di partenza (infinito-dimensionale) in un problema discreto (finito-dimensionale, algebrico), la cui dimensione può essere molto grande. Occorre infine validare la soluzione calcolata con un tale metodo numerico con i risultati sperimentali.

1.1.1 Perché i Modelli e non solo i Dati?

I modelli matematici sono tradizionalmente usati insieme alla teoria e agli esperimenti. Tuttavia, essi sono stati utilizzati anche nei casi in cui la teoria matematica non è ancora disponibile (ad esempio per simulare problemi di multi-fisica, ad esempio nella medicina computazionale) o quando i dati sperimentali sono pericolosi o impossibili da raggiungere (come per i test nucleari, il rientro di veicoli spaziali dall'atmosfera superiore, simulazione di eventi estremi come terremoti, ecc.).

Volendo suddividere la storia della Scienza in macro fasi di sviluppo, possiamo pensare che a una prima fase basata sulle osservazioni empiriche e a una seconda incentrata sulla scienza teorica e su appro-

fondimenti basati sulla matematica, ne sia seguita una terza (a partire dagli anni '50-'60 del '900) in cui protagoniste sono state le scienze computazionali e la simulazione numerica, e dunque i modelli derivati dalla fisica. Siamo da poco entrati in una quarta fase, caratterizzata dalla *Data Science*.

Al giorno d'oggi quantità praticamente illimitate di *dati* sono generate da più fonti, come Internet e le reti digitali, ampie reti di sensori, esperimenti o misure su larga scala (dalla microscala, come nella fisica delle alte energie, alla macroscala, come nel caso di dati acquisiti da satelliti per l'osservazione della Terra). Questa mole di dati, opportunamente combinata con modelli statistici, può fornire nuovi strumenti di indagine in quelle aree o situazioni in cui i modelli basati sulla fisica non sono applicabili perché i principi primi o sono inappropriate per modellare e simulare processi complessi, o addirittura non esistono.

Alternativamente, tali dati possono essere combinati con i modelli derivati dalla fisica allo scopo di (*i*) definire opportunamente un modello, che richiede dati in input (quali, ad esempio, forme geometriche, condizioni iniziali, condizioni al contorno, parametri fisici), oppure (*ii*) quantificare le incertezze nei risultati forniti dai modelli derivati dalla fisica mediante misurazioni e osservazioni. Si parla in quest'ultimo caso di *validazione* di un modello: tale operazione risulta ancora più rilevante quando i modelli derivati dalla fisica vengono applicati in nuove aree tra cui le scienze sociali, le discipline umanistiche, le imprese, la finanza, dove l'incertezza, il rischio e la casualità svolgono un ruolo importante.

I modelli basati sui dati, invece, si fondano su un diverso paradigma secondo il quale fenomeni complessi possono essere analizzati e previsti a partire appunto dai dati, mediante tecniche statistiche di apprendimento (*statistical learning*). Se con il tradizionale paradigma basato sulla fisica, in altri termini, i dati sono accessori rispetto al modello, con il paradigma basato sui dati il sogno è consentire agli algoritmi statistici (a patto di disporre di un'enorme potenza computazionale) di svelare le leggi e i modelli che governano i sistemi di dati complessi quando i primi principi non sono in grado di farlo.

Ad esempio, la rivoluzione del *deep learning*, una moderna reincarnazione delle reti neurali artificiali che, secondo alcuni, mira a trovare rappresentazioni comuni in tutti i domini sostituendo la simulazione numerica con l'apprendimento basato sui dati: si tratta di una potente classe di modelli di apprendimento automatico, sotto forma di funzioni matematiche semplici e addestrabili, che sono compatibili con molte varianti dell'apprendimento automatico. Il riconoscimento vocale e delle immagini, il riconoscimento e il rilevamento di oggetti, la traduzione automatica, la modellizzazione del linguaggio, sono domini in cui il deep learning sta mostrando risultati fantastici. Non ci sono (ancora) prove, tuttavia, che l'"apprendimento profondo" possa avere lo stesso successo nel simulare processi basati sulla fisica come, per esempio, complessi campi di flusso con interazioni su più scale (e, naturalmente, molti altri). Tuttavia, alcuni recenti lavori mostrano come alcune reti neurali, opportunamente allenate, riescano a ricostruire la soluzione di problemi differenziali, senza tuttavia poterne controllare (in modo semplice e diretto) l'accuratezza, come invece succede per qualsiasi solutore numerico. Questo ambito, e più in generale il fatto che i modelli basati sulla fisica possono estrarre da grandi insiemi di dati scientifici preziose intuizioni che possono andare ben oltre ciò che può essere recuperato dalla modellazione statistica black-box, mostra come le scienze computazionali basate sui modelli e la Data Science siano, nella pratica, fortemente interconnesse nel modo moderno di progettare processi di simulazione numerica.

I modelli basati sui principi primi della fisica risultano dunque componenti essenziali dei sistemi che estraggono preziose intuizioni da enormi quantità di dati, intuizioni che tendono ad andare molto al di là di ciò che può essere recuperato solo dalla modellizzazione statistica. Inoltre, alla base di algoritmi impiegati nello statistical learning, si trovano molto spesso algoritmi sviluppati nell'ambito del Calcolo Scientifico, quali ad esempio i metodi per l'ottimizzazione numerica, la soluzione di sistemi lineari di grandi dimensioni, la decomposizione di una matrice in autovalori e autovettori. Da tutte queste considerazioni, emerge ancora più forte la centralità dei metodi numerici nell'odierno panorama scientifico.

1.1.2 A cosa serve il Calcolo Scientifico?

I modelli derivati dalla fisica sono dunque un pilastro fondamentale nella comprensione e nella previsione di fenomeni più disparati. L'esigenza di effettuare simulazioni numeriche emerge infatti ogni qualvolta siamo interessati a comprendere o ricostruire scenari noti (come in Medicina e nelle Scienze della Vita), ottimizzare tali scenari (ad esempio nella produzione industriale o nel design manifacturing) o ancora

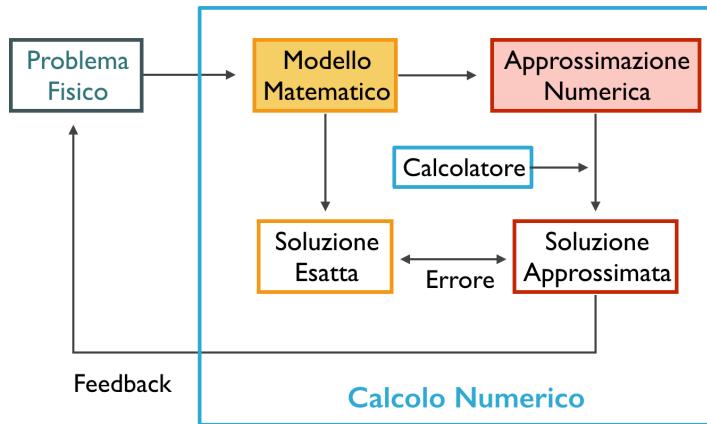


Figura 1.1: Modello matematico, approssimazione numerica e Calcolo Numerico.

predire scenari incogniti (come accade, ad esempio, nell’elaborazione di previsioni meteorologiche o nelle Scienze dei Materiali).

Tali modelli, tuttavia, conducono a problemi matematici che solo in rarissimi casi possono essere risolti *analiticamente*, o in maniera esatta: salvo poche eccezioni, non è infatti possibile scrivere in forma esplicita la loro soluzione. Per questo motivo, occorre introdurre opportune tecniche di approssimazione numerica. È dunque importante comprendere come sfruttare un calcolatore nella soluzione di problemi matematici fondamentali, introducendo gli strumenti chiave del Calcolo Scientifico. Quest’ultimo fornisce la base matematica alla simulazione numerica, e risulta per natura un ambito interdisciplinare, coinvolgendo matematica, scienze computazionali, ingegneria e scienze applicate. Uno schema generale che mostra l’interazione tra problemi fisici, modelli matematici e approssimazione numerica, è mostrato in Fig. 1.1.

1.2 Tre Classi Fondamentali di Problemi

Il Calcolo Scientifico si prefigge dunque di costruire, analizzare e applicare metodi computazionali per la soluzione di problemi derivati da modelli matematici. Un modello matematico si basa su due ingredienti principali: **leggi generali** e **relazioni costitutive**. Con leggi costitutive intendiamo tutte quelle leggi generali provenienti dalla meccanica dei continui, sotto forma di leggi di conservazione o di equilibrio (ad es. di massa, energia, carica elettrica, ecc.). Le relazioni costitutive sono di natura sperimentale e dipendono strettamente dalle caratteristiche dei fenomeni in esame. Esempi di relazioni costitutive sono la legge di Fourier di conduzione del calore, o la legge di Fick per la diffusione di una sostanza. Il risultato della combinazione dei due ingredienti è di solito un’equazione a derivate parziali o un sistema di equazioni di questo tipo.

Durante questo corso non affronteremo direttamente lo studio di EDP – forniremo solo alcuni dettagli su un particolare metodo per approssimare semplici EDP, il metodo delle *differenze finite*. Tuttavia, siccome alcuni argomenti basilari del Calcolo Scientifico (la risoluzione di sistemi lineari, l’approssimazione di funzioni, derivate e integrali) sono alla base dei più comuni metodi di approssimazione numerica per problemi differenziali come le EDP, in questo capitolo introduttivo consideriamo anche alcuni esempi di problemi a derivate parziali.

Un’equazione differenziale è un’equazione che coinvolge una o più derivate di una funzione incognita. In un’**equazione differenziale ordinaria** (EDO), tutte le derivate sono prese rispetto a una singola variabile indipendente; se sono invece presenti derivate parziali, si parla di **equazione a derivate parziali** (EDP).

Le equazioni differenziali ordinarie e a derivate parziali in generale ammettono un numero infinito di soluzioni. Per trovarne una, dobbiamo impostare ulteriori condizioni che prescrivano, ad esempio, il valore

assunto da tale soluzione in uno o più punti dell'intervallo di integrazione.

A seconda di quante (e quali) condizioni imponiamo, si originano problemi differenziali di diverso tipo. Presentiamo nel seguito tre *problemi modello* di tipo differenziale che vengono utilizzati per descrivere molte applicazioni.

Problema di Cauchy (o ai valori iniziali) per una EDO

Si tratta di un problema differenziale in cui viene assegnata un'unica condizione sulla soluzione, in un punto (solitamente, l'estremo sinistro dell'intervallo di integrazione). La sua forma risulta essere la seguente: trovare $u : I \subset \mathbb{R} \rightarrow \mathbb{R}$ tale che

$$\begin{cases} u'(t) = f(t, u(t)) & t \in I, \\ u(t_0) = u_0, \end{cases} \quad (1.1)$$

dove $I \subset \mathbb{R}$ designa un intervallo contenente il punto t_0 e $f : I \times \mathbb{R} \rightarrow \mathbb{R}$. Di solito, t indica la variabile temporale; l'equazione descrive l'evoluzione di una quantità scalare, u , funzione del tempo, ma che non è distribuita in spazio; qui $u'(t) = \frac{du}{dt}(t)$.

Ad esempio, nel caso di un circuito RC, il potenziale $v(t)$ attraverso un condensatore di capacità C si ottiene risolvendo il seguente problema:

$$\begin{cases} v'(t) + \frac{1}{RC}v(t) = 0 & t \in I, \\ v(t_0) = v_0, \end{cases}$$

per cui la soluzione è data da $v(t) = v_0 e^{-t/RC}$. Si noti che la condizione iniziale è obbligatoria per selezionare una soluzione tra le infinite soluzioni possibili dell'equazione differenziale.

Nel caso vettoriale, lo stato del sistema è descritto da un vettore $\mathbf{u} = \mathbf{u}(t)$, con $\mathbf{u} = (u_1, \dots, u_m) \in \mathbb{R}^m$ e il problema diventa: trovare $\mathbf{u} : I \subset \mathbb{R} \rightarrow \mathbb{R}^m$ tale che

$$\begin{cases} \mathbf{u}'(t) = \mathbf{F}(t, \mathbf{u}(t)) & t \in I, \\ \mathbf{u}(t_0) = \mathbf{u}_0, \end{cases}$$

dove $\mathbf{F} : I \times \mathbb{R}^m \rightarrow \mathbb{R}^m$.

Problema ai valori al contorno per una EDP stazionaria

Si tratta di un problema differenziale definito su un intervallo $(a, b) \subset \mathbb{R}$ o in una regione multidimensionale aperta $\Omega \subset \mathbb{R}^d$ ($d = 2, 3$) per cui il valore della soluzione sconosciuta (o della sua derivata) viene prescritto nei punti finali a e b dell'intervallo, o sul bordo $\partial\Omega$. Nel caso multidimensionale, l'equazione differenziale coinvolgerà derivate parziali della soluzione rispetto alle coordinate spaziali.

Nel caso $d = 1$, in cui si ha una sola variabile indipendente x e compaiano nell'equazione solo derivate dell'incognita rispetto alla sola variabile indipendente x , si considera tale problema nella classe delle EDP per la natura delle condizioni imposte sulla sua soluzione (che risultano di natura *globale*, riguardando non più soltanto un singolo punto del dominio, e non locale).

Dato l'intervallo $\Omega = (a, b) \subset \mathbb{R}$ introduciamo l'equazione di ordine 2 (detta *equazione di Poisson*) corredata di due condizioni al contorno che assegnano il valore di u nel punti a e b :

$$\begin{cases} -u''(x) = f(x) & x \in (a, b), \\ u(a) = u(b) = 0; \end{cases} \quad (1.2)$$

qui $u''(x) = \frac{d^2u}{dx^2}(x)$. Questa equazione modella un fenomeno stazionario (non compare infatti la variabile temporale) e rappresenta il più semplice modello di diffusione, come la diffusione di un inquinante lungo un canale monodimensionale (a, b) o lo spostamento verticale di un filo elastico (detto anche *linea elastica*) fissato ai suoi estremi. Nel primo caso $f = f(x)$ indica la sorgente dell'inquinante lungo il flusso, mentre nel secondo caso f è la forza trasversale che agisce sul filo elastico, nell'ipotesi di massa trascurabile e piccoli spostamenti.

Problema ai valori iniziali e al contorno per una EDP

Si tratta di problemi che riguardano equazioni che dipendono dallo spazio, ma anche dal tempo (indicato con t), come l'equazione del calore e l'equazione delle onde. In tal caso, devono essere prescritte anche le condizioni iniziali a $t = 0$, oltre che le condizioni al contorno agli estremi dell'intervallo (o sul bordo del dominio). Nel caso in cui $d = 1$ e $u = u(x, t)$ dipenda anche dal tempo t , possiamo considerare il seguente problema:

$$\begin{cases} \frac{\partial u}{\partial t} - D \frac{\partial^2 u}{\partial x^2} = f & x \in (a, b), t > 0, \\ u(a, t) = u(b, t) = 0 & t > 0, \\ u(x, 0) = u_0(x) & x \in (a, b). \end{cases} \quad (1.3)$$

L'equazione (7.5) è detta *equazione del calore (o equazione di diffusione)*: la quantità $u(x, t)$ descrive la temperatura nel punto x e all'istante t di una sbarra di metallo monodimensionale che occupa l'intervallo $\Omega = (a, b)$. Il coefficiente di diffusione D rappresenta la risposta termica del materiale, ovvero $D = \frac{\kappa}{\rho c_p}$, dove $\kappa > 0$ è la conducibilità termica, ρ è la densità e c_p la capacità termica per unità di massa.

Le condizioni al contorno esprimono il fatto che le estremità della sbarra sono mantenute ad una temperatura di riferimento (zero gradi in questo caso), mentre all'istante $t = 0$, viene assegnata la temperatura in ciascun punto $x \in [a, b]$ dalla condizione iniziale $u_0(x)$. Infine, la sbarra risulta soggetta a una fonte di calore di densità lineare $f(x, t)$.

Problemi in regioni multidimensionali $\Omega \subset \mathbb{R}^d$, $d = 2, 3$

L'estensione del problema (1.2) in $d > 1$ dimensioni è rappresentata dall'equazione del potenziale (o di Poisson) con condizioni al contorno di Dirichlet omogenee:

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = 0 & \text{su } \partial\Omega, \end{cases} \quad (1.4)$$

dove

$$\Delta u = \sum_{i=1}^d \frac{\partial^2 u}{\partial x_i^2}$$

è l'*operatore di Laplace (o Laplaciano)* e $\Omega \subset \mathbb{R}^d$ è un dominio il cui contorno è indicato con $\partial\Omega$. Tale equazione permette di modellare, ad esempio, la diffusione di particelle in un fluido (in assenza di termini convettivi), oppure lo spostamento verticale di una membrana elastica; risulta infine un problema di assoluto rilievo anche in elettrostatica.

La versione multidimensionale ($d = 2, 3$) dell'equazione del calore (o equazione di diffusione) è

$$\frac{\partial u}{\partial t} - \Delta u = f \quad \text{in } \Omega, t > 0, \quad (1.5)$$

corredato di condizioni iniziali e al contorno adeguate. L'equazione del calore descrive un fenomeno evolutivo, mentre l'equazione di Laplace descrive i corrispondenti stati stazionari, quando la soluzione non dipende dal tempo.

Un'altra importante equazione di ordine 2 è l'*equazione delle onde*:

$$\frac{\partial^2 u}{\partial t^2} - c^2 \Delta u = 0 \quad \text{in } \Omega \subset \mathbb{R}^d, t > 0, \quad (1.6)$$

anch'essa corredata da opportune condizioni iniziali e al contorno. Tale modello descrive la propagazione delle onde trasversali di piccola ampiezza in una corda (ad esempio in un violino) se $d = 1$ o in una membrana elastica se $d = 2$, mentre se $d = 3$ descrive le onde sonore o onde elettromagnetiche nel vuoto. L'incognita $u = u(x, t)$ è collegata all'ampiezza delle vibrazioni e c è la velocità di propagazione.

1.3 Soluzione Esatta e Approssimazione Numerica

Tutte le equazioni presentate nella sezione precedente ammettono una soluzione analitica (ovvero, tale da poter essere espressa in forma esplicita) solo in rari casi; a titolo dimostrativo, consideriamo alcuni esempi.

- Nel caso del problema di Cauchy (1.1) con $n = 1$, a patto che f sia continua rispetto alla prima variabile, il problema si può trasformare in un problema integrale,

$$\int_{t_0}^t \frac{dy}{d\tau}(\tau)d\tau = \int_{t_0}^t f(\tau, y(\tau))d\tau \quad \forall \tau \in I,$$

da cui, se richiediamo che $y \in C^1(I)$,

$$y(t) = y(t_0) + \int_{t_0}^t f(\tau, y(\tau))d\tau = y_0 + \int_{t_0}^t f(\tau, y(\tau))d\tau.$$

In questo caso risulta possibile determinare una soluzione del problema di Cauchy solo quando f assume una forma particolare. Ad esempio, se $f(t, y) = a(t)y + b(t)$, si ha:

$$y(t) = e^{\int_{t_0}^t a(s)ds} \left(y_0 + \int_{t_0}^t b(s) e^{-\int_{t_0}^s a(z)dz} ds \right).$$

In questo caso è necessario conoscere le primitive di $a(t)$, e saper integrare la funzione $b(s) e^{-\int_{t_0}^s a(z)dz}$, per poter risalire a una soluzione del problema di Cauchy. Più in generale, solo poche classi di equazioni differenziali ordinarie si possono risolvere analiticamente (equazioni di ordine 1 o 2, lineari, omogenee o non omogenee, a coefficienti costanti; equazioni a variabili separabili, etc.).

Nel caso di un sistema di EDO lineare, omogeneo, a coefficienti costanti,

$$\begin{cases} \mathbf{y}'(t) &= A\mathbf{y}(t) \quad t \in I, \\ \mathbf{y}(t_0) &= \mathbf{y}_0, \end{cases} \quad (1.7)$$

a patto che $A \in \mathbb{R}^{n \times n}$ sia diagonalizzabile e $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n) \in \mathbb{R}^{n \times n}$ contenga i suoi autovalori, la soluzione si può esprimere in forma analitica, e risulta pari a

$$\mathbf{y}(t) = e^{A(t-t_0)} \mathbf{y}_0,$$

dove $e^{At} = Pe^{\Lambda t}P^{-1}$ è la matrice esponenziale e P è la matrice degli autovettori di A . Tuttavia, in questo caso occorre saper calcolare gli autovalori di A . Anche nel caso di una matrice reale e simmetrica, tale problema equivale a trovare gli zeri di un polinomio di grado n , problema non risolvibile analiticamente in generale (nel caso n sia grande). L'approssimazione degli autovalori di una matrice è un problema affrontabile con gli strumenti del Calcolo Scientifico.

- Nel caso del problema ai valori al contorno (1.2) per l'equazione di Poisson, sappiamo trovare una soluzione in esplicita nel caso in cui ad esempio $f(x) = c \in \mathbb{R}$ è una costante. In questo caso, è semplice verificare che

$$u(x) = \frac{c}{2}(x - a)(b - x) \quad (1.8)$$

soddisfa sia l'equazione differenziale che le condizioni al contorno.

Più in generale, se consideriamo per semplicità il caso in cui $(a, b) = (0, 1)$, si ha che a patto di considerare $f \in C^0([0, 1])$, esiste un'unica soluzione $u \in C^2([0, 1])$ del problema (1.2) ed essa risulta data da

$$u(x) = x \int_0^1 (1-s)f(s)ds - \int_0^x (x-s)f(s)ds. \quad (1.9)$$

Anche in questo caso, è possibile trovare la soluzione esplicitamente a patto di saper determinare le primitive delle funzioni che compaiono sotto il segno di integrale.

- Nel caso del problema ai valori iniziali e al contorno per l'equazione del calore (7.5), prendendo per semplicità $(a, b) = (0, 1)$ e a patto di considerare $f = 0$, ovvero

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0 \quad \text{in } \Omega = (0, 1) \text{ per } t > 0,$$

con $u(0, t) = u(1, t) = 0$, $t > 0$, è ad esempio possibile esprimere la soluzione sotto forma di serie infinita di termini

$$u(x, t) = \sum_{j=1}^{\infty} u_{0,j} e^{-(j\pi)^2 t} \sin(j\pi x),$$

dove $u_{0,j}$ indica il j -esimo coefficiente di Fourier,

$$u_{0,j} = 2 \int_0^1 u_0(x) \sin(j\pi x) dx, \quad j = 1, 2, \dots$$

Oltre a richiedere il calcolo di integrali, la soluzione esatta così ottenuta va necessariamente approssimata, qualora il dato iniziale ammetta uno sviluppo in serie di Fourier con infiniti termini. Ancora una volta, occorre dunque ricorrere a opportuni metodi numerici per poter operare in tal senso.

In generale, non possiamo risolvere analiticamente una EDP a causa della forma del dominio e della necessità di imporre condizioni al contorno. Questo fatto fornisce l'importanza di disporre di **metodi numerici** che consentono di costruire un'approssimazione u_h della soluzione esatta u e la valutazione (in una norma appropriata) dell'errore corrispondente $u_h - u$. Qui, h designa un parametro reale (di discretizzazione) che caratterizza l'approssimazione numerica e nelle applicazioni sarà piccolo e destinato ad avvicinarsi a zero per migliorare l'accuratezza della soluzione numerica. Possiamo rappresentare schematicamente questo approccio come segue:

$\mathcal{P}(u; g) = 0$	EDP esatta (problema matematico)
↓	<i>metodo numerico</i>
$\mathcal{P}_h(u_h; g_h) = 0$	EDP approssimata (problema numerico)

dove g_h è un'approssimazione dei dati g , mentre \mathcal{P}_h è la nuova funzione che caratterizza il problema approssimato. Uno strumento molto flessibile e potente per approssimare la soluzione delle EDP è rappresentato dal metodo degli elementi finiti (*Finite Element Method*, FEM). Faremo accenno a tale metodo alla fine del corso.

1.4 A Cosa Servono i Metodi Numerici?

Per avere una prima idea di quali metodi numerici risultino di fondamentale importanza in Ingegneria, consideriamo ancora una volta il problema di determinare la configurazione di equilibrio di una struttura, ovvero di trovare lo spostamento u di una struttura dato un carico f . Per semplicità consideriamo il caso monodimensionale che ci ha condotto al modello (1.2): supponiamo di voler determinare lo spostamento verticale del filo u e di voler poi calcolare il lavoro del carico esterno, fornito da

$$W = \int_0^1 f u \, dx. \tag{1.10}$$

Abbiamo già notato come, nel caso in cui f sia uguale a una costante $c > 0$, il problema (1.2) possa essere risolto analiticamente, dando

$$u(x) = \frac{c}{2}x(1-x);$$

analogamente, anche la valutazione dell'integrale (1.10) è banale, risultando in questo caso

$$W = \int_0^1 \frac{c}{2}x(1-x)dx = \frac{c}{12}.$$

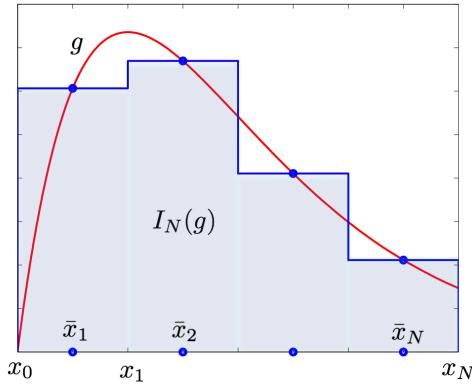


Figura 1.2: Formula del punto medio composita per l'approssimazione di $I(g)$.

La valutazione degli integrali che appaiono in (1.10), così come in (8.13), potrebbe non essere semplice (o addirittura impossibile) per particolari scelte di f . Per questo motivo, si può approssimare, ad esempio, l'integrale (1.10) numericamente.

Consideriamo per semplicità il caso di una generica funzione $g(x)$ da integrare su un intervallo $[a, b]$,

$$I(g) = \int_a^b g(x) dx.$$

Introduciamo sull'intervallo di integrazione $[a, b]$ una partizione in N sottointervalli $I_j = [x_{j-1}, x_j]$ per ogni $j = 1, \dots, N$, con $x_j = a + jh$ dei nodi, supponendo che tutti gli intervalli abbiano la stessa lunghezza $h = (b - a)/N$, e usiamo la cosiddetta formula di quadratura del punto medio composita (si veda la Figura 1.2) per approssimare $I(g)$ con

$$I_N(g) = \frac{b-a}{N} \sum_{j=1}^N g(\bar{x}_j), \quad \text{dove } \bar{x}_j = \frac{x_{j-1} + x_j}{2} \quad \text{per } j = 1, \dots, N.$$

Per N sufficientemente grande, si ha che $I_N(g) \rightarrow I(g)$, e come vedremo l'errore $|I(g) - I_N(g)|$ si comporta come $1/N$, ovvero proporzionale ad h .

D'altra parte, per approssimare la soluzione del problema (1.2), un'opzione possibile è quella di supporre che l'equazione differenziale debba essere soddisfatta in qualsiasi nodo x_j interno a $(0, 1)$ di una partizione di $[0, 1]$ in $N+1$ nodi $x_j = j h$ con $h = \frac{1}{N+1}$ e $j = 0, \dots, N+1$, ovvero

$$-u''(x_j) = f(x_j) \quad \forall j = 1, \dots, N.$$

Possiamo approssimare questo insieme di N equazioni sostituendo la seconda derivata con la quantità

$$\delta^2 u(\bar{x}) = \frac{u(\bar{x} + h) - 2u(\bar{x}) + u(\bar{x} - h)}{h^2},$$

chiamata *differenza finita*; come vedremo, se $u : [0, 1] \rightarrow \mathbb{R}$ è una funzione sufficientemente regolare in un intorno di un punto generico $\bar{x} \in (0, 1)$, allora $\delta^2 u(\bar{x})$ fornisce un'approssimazione di $u''(\bar{x})$ di ordine 2 rispetto a h , ovvero l'errore $|u''(\bar{x}) - \delta^2 u(\bar{x})|$ si comporta come h^2 .

Possiamo dunque approssimare il problema (1.2) come segue: al posto di cercare una funzione $u(x)$ che soddisfi il problema ai valori ai limiti, con il metodo delle differenze finite cerchiamo un insieme di N valori reali $\{u_j\}_{j=1}^N$ tali che

$$\begin{cases} \frac{-u_{j+1} - 2u_j + u_{j-1}}{h^2} = f(x_j) & \forall j = 1, \dots, N, \\ u_0 = u_{N+1} = 0. \end{cases} \quad (1.11)$$

Ovviamente, u_j costituirà un'approssimazione di $u(x_j)$. Le equazioni (8.15) riconducono a un *sistema lineare*

$$A\mathbf{u} = \mathbf{f}, \quad (1.12)$$

dove $\mathbf{f} = (f(x_1), f(x_2), \dots, f(x_{N-1}), f(x_N))^T$, $\mathbf{u} = (u_1, \dots, u_N)^T$ è il vettore di incognite e $A \in \mathbb{R}^{N \times N}$ è la matrice tridiagonale

$$A = \frac{1}{h^2} \text{tridiag}(-1, 2, 1) = \frac{1}{h^2} \begin{bmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & \ddots & & \vdots \\ 0 & \ddots & \ddots & -1 & 0 \\ \vdots & & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 2 \end{bmatrix}. \quad (1.13)$$

Come vedremo, questo sistema ammette una soluzione unica poiché A è simmetrica e definita positiva. Tale sistema può essere risolto dall'algoritmo di Thomas o, più in generale, con il metodo di eliminazione di Gauss, vale a dire, riducendo la matrice in forma a scala come visto in algebra lineare.

Si noti che se $f \in C^2([0, 1])$, la seguente stima vale per l'errore tra la soluzione esatta e la soluzione approssimata ottenuta risolvendo il sistema lineare derivante dall'approssimazione con differenze finite,

$$\max_{j=0, \dots, N+1} |u(x_j) - u_j| \leq \frac{h^2}{96} \max_{x \in [0, 1]} |f''(x)|;$$

osserviamo pertanto che (i) l'errore decade con un tasso proporzionale a h^2 (minore è $h < 1$, minore è l'errore) e (ii) i dati di problema matematico entra in gioco in una stima dell'errore anche per l'errore di discretizzazione. Quest'ultimo fatto verrà osservato più volte nel seguito.

Questo semplice esempio mostra come alcune esigenze comuni nella risoluzione di un problema differenziale possano essere affrontate mediante opportune tecniche di approssimazione numerica. Altre possibili esigenze, sempre considerando l'esempio in esame, potrebbero riguardare l'analisi dei modi di risonanza del filo elastico, che condurrebbe al calcolo degli autovalori di A , oppure la stima di proprietà del materiale (qualora incognite) a partire da misure sperimentali dello spostamento, che richiederebbe la soluzione di un problema di approssimazione nel senso dei minimi quadrati.

Anche nel caso del metodo degli elementi finiti per la soluzione di problemi a derivate parziali¹ si giunge a ottenere a un sistema lineare analogo a (8.16), la cui soluzione sarà resa possibile dagli strumenti del Calcolo Numerico.

1.5 Rappresentazione dei Numeri e Operazioni al Calcolatore

Un calcolatore può gestire solo un numero finito di numeri ed effettuare un numero finito di operazioni: l'insieme dei numeri reali \mathbb{R} è dunque rappresentato da un insieme *finito* di numeri macchina $\mathbb{F} = \{-\tilde{a}_{min}, \dots, \tilde{a}_{max}\}$, detti **numeri floating point** (la posizione della virgola non è infatti fissata).

1.5.1 Numeri Floating point

Definizione 1.5.1. L'insieme dei numeri floating point \mathbb{F} è il sottoinsieme di numeri reali che possono essere rappresentati al calcolatore, ovvero $\mathbb{F} \subset \mathbb{R}$ con $\text{card}(\mathbb{F}) < +\infty$. In generale, $\mathbb{F} = \mathbb{F}_0 \cup \{0\}$, dove \mathbb{F}_0 indica i numeri floating point escluso lo zero.

¹Il metodo degli elementi finiti, come si vedrà in futuro, è una tecnica assolutamente generale e facilmente estendibile anche al caso di problemi definiti in due o tre dimensioni, su domini arbitrariamente complessi. L'introduzione di tale tecnica richiede opportuni metodi analitici per poter formulare e poi analizzare una formulazione alternativa (detta *formulazione debole*) del problema differenziale.

Esempio 1.5.1. $x = \frac{1}{3} = 0.\overline{3} = 0.\underbrace{333\cdots 3}_{\infty \text{ cifre}}, \text{ con } fl(x) = 0.\underbrace{333\cdots 3}_N, \text{ con } N < +\infty.$

Per comodità di notazione, indicheremo d'ora in avanti con \mathbb{F} i numeri floating point escluso lo zero. L'insieme $\mathbb{F} = \mathbb{F}(\beta, t, L, U)$ è caratterizzato da quattro parametri β, t, L e U tali per cui ogni numero reale $x \in \mathbb{F}$ può essere scritto come:

$$x = (-1)^s m \beta^{e-t} = (-1)^s (a_1 a_2 \cdots a_t)_\beta \beta^{e-t},$$

dove:

- $\beta \geq 2$ è la *base*, un numero intero che determina il sistema numerico;
- $m = (a_1 a_2 \cdots a_t)_\beta$ è la *mantissa*, $(0 < m \leq \beta^t - 1)$ essendo t il *numero di cifre* (dette anche cifre significative) tali per cui $0 < a_1 \leq \beta - 1$ e $0 \leq a_i \leq \beta - 1$ per $i = 2, \dots, t$;
- $e \in \mathbb{Z}$ è l'*esponente* tale che $L \leq e \leq U$, con $L < 0$ e $U > 0$;
- $s = \{0, 1\}$ è il suo *segno*.

Il numero di cifre dell'esponente definisce il range dei numeri macchina; il numero di cifre della mantissa la sua precisione. Una volta che l'insieme $\mathbb{F}(\beta, t, L, U)$ è caratterizzato, il numero $x \in \mathbb{F}$ è rappresentato dai valori assunti dai parametri s, m ed e .

Poiché $0.a_1 a_2 \dots a_t = a_1 a_2 \dots a_t \cdot \beta^{-t}$, una rappresentazione equivalente dei numeri floating point è data da

$$x = (-1)^s \beta^e \sum_{i=1}^t a_i \beta^{-i}.$$

Per rendere unica la rappresentazione di un numero macchina $fl(x)$ (osserviamo per esempio che $1.2345 \cdot 10^3 = 0.0012345 \cdot 10^6$) richiediamo le seguenti condizioni (numeri normalizzati): $a_1 \neq 0$ e $m \geq \beta^{t-1}$, peraltro già soddisfatte dai requisiti precedentemente indicati per la mantissa m .

Esempio 1.5.2. Consideriamo i seguenti due casi:

- $-3.154 \cdot 10^5 = -0.3154 \cdot 10^6 = (-1)^1 \cdot 10^6 \left(\frac{3}{10^1} + \frac{1}{10^2} + \frac{5}{10^3} + \frac{4}{10^4} \right)$, ovvero $s = 1, \beta = 10, m = 3154, e = 6, t = 4$;
- $(4.25)_{10} = (100.01)_2 = 1.0001 \cdot 2^2 = 0.10001 \cdot 2^3$; ovvero $s = 0, \beta = 2, m = 10001, e = 3_{10} = 11_2, t = 5$.

Nel caso dell'insieme dei numeri floating point $\mathbb{F}(\beta, t, L, U)$ abbiamo:

- il valore detto **epsilon macchina**

$$\varepsilon_M = \beta^{1-t},$$

che rappresenta la distanza tra 1 e il più piccolo numero floating point maggiore di 1, ovvero il più piccolo numero reale maggiore di zero tale per cui $fl(1 + \varepsilon_M) > 1$.

- l'**errore di arrotondamento** (roundoff) è l'errore relativo che si commette sostituendo $x \in \mathbb{R} \setminus \{0\}$ con il suo rappresentante $fl(x) \in \mathbb{F}$ è limitato da

$$\boxed{\frac{|x - fl(x)|}{|x|} \leq \frac{1}{2} \varepsilon_M \quad x \neq 0,}$$

essendo $\frac{1}{2} \varepsilon_M$ l'**unità di arrotondamento** (ovvero il massimo errore relativo che la macchina può commettere nella rappresentazione di un numero). Si noti che anche se l'errore di arrotondamento $\frac{1}{2} \varepsilon_M$ è "piccolo", cioè l'errore relativo è "piccolo", l'errore assoluto $|x - fl(x)|$ potrebbe essere molto "grande", specialmente se $|x|$ è "grande";

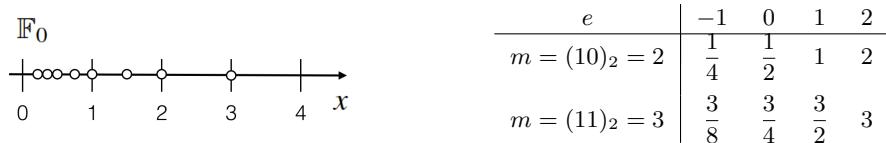
- poiché L ed U sono finiti, non è possibile rappresentare numeri in valore assoluto arbitrariamente piccoli o grandi. Il più piccolo e il più grande numero positivo sono

$$x_{min} = \beta^{L-1}, \quad x_{max} = \beta^U (1 - \beta^{-t});$$

- non è quindi possibile rappresentare alcun numero (a parte lo zero) minore in valore assoluto di x_{min} , o maggiore di x_{max} ;
 - la cardinalità di $\mathbb{F}(\beta, t, L, U)$ (considerando solo i positivi), escluso lo zero, è data da

$$\text{card}(\mathbb{F}) = 2(\beta - 1)\beta^{t-1}(U - L + 1).$$

Esempio 1.5.3. Si consideri l'insieme dei numeri floating point $\mathbb{F}(2, 2, -1, 2)$, cioè tali per cui $\beta = 2$ (sistema numerico in base 2), $t = 2$ (numero di cifre della mantissa), $L = -1$ e $U = 2$. Di conseguenza, si hanno $\epsilon_M = \beta^{1-t} = \frac{1}{2}$, $x_{min} = \beta^{L-1} = \frac{1}{4}$ e $x_{max} = \beta^U (1 - \beta^{-t}) = 3$; la cardinalità dell'insieme \mathbb{F} è uguale a $2(\beta - 1)\beta^{t-1}(U - L + 1) = 16$. I valori che l'esponente e può assumere sono $-1, 0, 1$ e 2 . La mantissa è $m = (a_1 a_2)_\beta$ essendo $t = 2$; ne consegue che, essendo $\beta = 2$, abbiamo $a_1 = 1$, mentre a_2 è uguale a 0 oppure a 1 . I valori ammessi per m sono perciò $m = (10)_2 = 2$ oppure $(11)_2 = 3$. Scegliendo il segno $s = 0$, i numeri reali positivi in \mathbb{F}_0 sono pertanto rappresentabili come $x = m\beta^{e-t} = m2^{e-2}$ e sono riportati nella tabella seguente.



Osserviamo come, tanto più grandi sono i valori $|f(x)|$, tanto meno densi sono i numeri in \mathbb{R} .

Lo standard floating point IEEE *doppia precisione* usa una stringa di $N = 64$ bits per rappresentare i numeri macchina e corrisponde a $\mathbb{F} = \mathbb{F}(2, 52, -1022, 1023)$:



Esempio 1.5.4. Il numero reale $x = \frac{1}{10}$ non può essere rappresentato esattamente in base binaria $\beta = 2$ in quanto richiede una serie infinita $x = \frac{1}{2^4} + \frac{1}{2^5} + \frac{1}{2^4} + \frac{0}{2^5} + \frac{0}{2^6} + \frac{1}{2^7} + \dots$. La sua rappresentazione in base binaria a doppia precisione è $fl(x) = \frac{1}{2^4} + \frac{1}{2^5} + \frac{1}{2^4} + \frac{0}{2^5} + \frac{0}{2^6} + \frac{1}{2^7} + \dots + \frac{0}{2^{51}} + \frac{1}{2^{52}}$, ovvero un'approssimazione di x .

Osservazione 1.5.1. Poiché in base $\beta = 2$ necessariamente $a_1 = 1$, in questo caso non serve un bit per a_1 , dato che tale cifra è nota a priori; occorrerà semplicemente memorizzare nella mantissa la frazione $m < 1$ del numero $1 + \tilde{m}$. In base $\beta = 2$ si utilizza dunque la rappresentazione seguente²:

$$x = (-1)^s 2^e (1 + \tilde{m}), \quad \tilde{m} = a_2 2^{-1} + a_3 2^{-2} + \dots + a_{t+1} 2^{-t}$$

dove a_2 può essere indifferentemente 1 o 0; la normalizzazione del numero avviene dunque sulla cifra delle unità, essendo

$$1 + \tilde{m} = \sum_{i=1}^t a_i 2^{-i} + a_{t+1} 2^{-t}.$$

Si ha quindi, nel caso in cui si usino 52 bit per \tilde{m} e 11 bit per e (doppia precisione),

$$1 = (-1)^0 2^0 (1 + 00000\ldots000) \quad \text{ovvero} \quad s = 0, e = 0, a_2 = \dots = a_{53} = 0.$$

^2Si ha $1.m = 1 + \tilde{m}$, ovvero $\tilde{m} = m \cdot 2^-$

Il più piccolo dei numeri macchina maggiori di 1 è invece dato da

$$(-1)^0 2^0 (1 + 00000...001) \quad \text{ovvero} \quad s = 0, e = 0, a_2 = \dots = a_{52} = 0, a_{53} = 1$$

e non si tratta del più piccolo numero rappresentabile! Il più piccolo numero rappresentabile si ottiene invece per

$$(-1)^0 2^{-(11\dots1)_2} (1 + 0) \text{ ovvero } s = 0, e = -(2^{10} - 1), a_2 = \dots = a_{53} = 0.$$

Per non memorizzare il segno dell'esponente, si memorizza in c l'esponente e maggiorato di 1023: in questo modo $0 < c < 2047 = 2^{11} - 1$. Si ha dunque

$$x = (-1)^s \cdot 2^{c-1023} \cdot (1.m)_2.$$

Esempio 1.5.5. 0 100 0000 0001 1010 0000 0000 0000 coincide con $+1 \cdot 2^{1025-1023} \cdot (1 + \frac{1}{2^1} + \frac{1}{2^3}) = 6.5$.

In Matlab®, per CPUs a 64-bit (doppia precisione), i numeri sono rappresentati in base $\beta = 2$. Sulla base delle considerazioni precedenti (Osservazione 1.5.1), si ha dunque $t = 52$. Tuttavia, il numero di cifre t usate per la mantissa m è di fatto $52 + 1 = 53$ (dal momento che la prima cifra della mantissa a_1 è sempre 1). Tali cifre in base 2 corrispondono alle 15 cifre significative in base 10 nel format long di Matlab®. Ne consegue che in Matlab® :

- $\epsilon_M = 2^{-52} \approx 2.22 \cdot 10^{-16}$;
- usando i comandi `realmin` e `realmax`, si hanno:

$$\begin{aligned} x_{\min} &= (1.000\dots000)_2 \cdot 2^{-1022} = 2.225073858507201 \cdot 10^{-308}, \\ x_{\max} &= (1.111\dots111)_2 \cdot 2^{1023} = 2^{1023}(2 - 2^{-52}) = 1.797693134862316 \cdot 10^{+308}; \end{aligned}$$

- un numero positivo maggiore di x_{\max} produce una segnalazione di *overflow* e viene memorizzato nella variabile `Inf`. `Inf` è rappresentato prendendo $c = 2047$ e $m = 0$; soddisfa le relazioni $1/\text{Inf} = 0$ e $\text{Inf} + \text{Inf} = \text{Inf}$;
- analogamente, dà overflow un numero negativo minore di $-x_{\max}$, memorizzato come `-Inf`;
- un numero positivo minore di x_{\min} (o maggiore di $-x_{\min}$) produce una segnalazione di *underflow* e viene trattato come 0;
- se un'operazione cerca di produrre un valore che non è definito nemmeno in \mathbb{R} (come ad esempio $0/0$ e $\text{Inf} - \text{Inf}$) il risultato è Not-a-Number, `NaN`, rappresentato prendendo $c = 2047$ e $m \neq 0$.

Esempio 1.5.6. I numeri di \mathbb{F} sono molto addensati vicino a x_{\min} diventando sempre più radi all'avvicinarsi a x_{\max} . Infatti, in Matlab®, il numero di \mathbb{F} immediatamente precedente x_{\max} e immediatamente successivo a x_{\min} sono rispettivamente

$$x_{\max}^- = 1.797693134862315 \cdot 10^{+308} \quad \text{e} \quad x_{\min}^+ = 2.225073858507202 \cdot 10^{-308},$$

da cui $x_{\min}^+ - x_{\min} \approx 10^{-323}$, $x_{\max} - x_{\max}^- \approx 10^{292}$ (la distanza relativa è comunque piccola).

1.5.2 Aritmetica floating point

Le operazioni algebriche che coinvolgono numeri floating point in \mathbb{F} non beneficiano delle stesse proprietà dei numeri reali in \mathbb{R} . Infatti, *errori di arrotondamento* (round-off) possono propagarsi ed eventualmente crescere a seconda del numero e tipo di operazioni algebriche coinvolte nel calcolo. Il termine *flops* viene usato per indicare il numero di operazioni algebriche in aritmetica floating point.

Esempio 1.5.7. Alcuni esempi di effetti di errori di arrotondamento in Matlab[®].

```
>> a = 1 - 3 * ( 4 / 3 - 1 )
a = 2.2204e-16

>> b = sqrt(1e-16 + 1) - 1
b = 0

>> c = 1e-16 - 1e-16 + 1
>> d = 1e-16 + 1 - 1e-16
>> f = c - d
f = 1.1102e-16
```

Siccome \mathbb{F} è un sottoinsieme proprio di (ovvero strettamente contenuto in) \mathbb{R} , operazioni algebriche elementari sui numeri floating-point non soddisfano tutte le proprietà di operazioni analoghe in \mathbb{R} : ad esempio, la proprietà commutativa continua a valere per addizione e moltiplicazione, ma altre proprietà come quella associativa o distributiva possono essere violate (come si è visto nell'esempio precedente). Inoltre, lo zero non è più unico: esiste infatti almeno un numero $b \neq 0$ tale che $a + b = a$ ($a + b$ è sempre uguale ad a se $b < \epsilon_M(b)$).

La proprietà associativa viene violata ogniqualvolta si verifica una situazione di overflow o underflow. Prendendo ad esempio $a = 1.0 \cdot 10^{+308}$, $b = 1.1 \cdot 10^{+308}$ e $c = -1.001 \cdot 10^{+308}$, troveremmo

$$a + (b + c) = 1.0990 \cdot 10^{+308}, \quad (a + b) + c = \text{Inf}.$$

Come abbiamo visto in precedenza, questo è un caso particolare di ciò che accade quando si sommano due numeri con segno opposto ma valore assoluto simile (ricordiamo che il numero di condizionamento dell'operazione di sottrazione in questo caso diviene molto grande): si parla a tal proposito di perdita (o cancellazione) di cifre significative.

Esempio 1.5.8. Per ogni numero reale escluso lo zero $x \in \mathbb{R} \setminus \{0\}$ si ha $\frac{(1+x)-1}{x} \equiv 1$. Tuttavia, in aritmetica floating point $\frac{fl(1+fl(x))-1}{fl(x)} = y$, dove y è un numero reale in generale diverso da 1. Qualora provassimo a verificare la prima identità in Matlab[®], otterremmo un numero $y \neq 1$ con i seguenti errori a seconda del valore del numero reale x scelto.

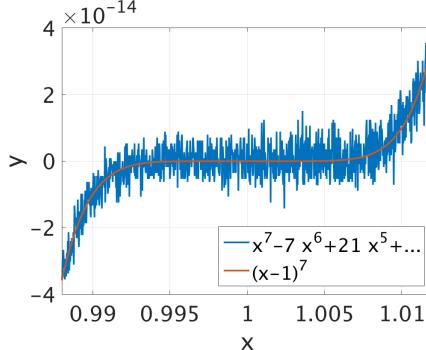
	10^{-10}	10^{-14}	10^{-15}	10^{-16}
errore relativo	$8 \cdot 10^{-6}\%$	$8 \cdot 10^{-2}\%$	11%	100%

Esempio 1.5.9. Si consideri il seguente sistema lineare:

$$\begin{cases} 17x_1 + 5x_2 &= 22 \\ 1.7x_1 + 0.5x_2 &= 2.2 \end{cases}$$

dotato di una matrice $A = \begin{bmatrix} 17 & 5 \\ 1.7 & 0.5 \end{bmatrix}$ singolare e di infinite soluzioni. Risolvendo in Matlab[®] tale sistema si ottiene invece la soluzione $\tilde{x}_1 = -1.8431$, $\tilde{x}_2 = 10.6667$, che Matlab[®] considera come unica. Infatti, la rappresentazione floating point della matrice è non singolare; per esempio si hanno $fl(1.7) \neq 1.7$ e $\det(fl(A)) = 9.4369 \cdot 10^{-16} \neq 0$.

Esempio 1.5.10. Si consideri la funzione $f(x) = (x - 1)^7$ che può essere equivalentemente scritta (in aritmetica esatta) come $f(x) = x^7 - 7x^6 + 21x^5 - 35x^4 + 35x^3 - 21x^2 + 7x - 1$. La valutazione di tale funzione $f(x)$ in Matlab® dà risultati differenti a seconda dell'espressione utilizzata a causa di errori di arrotondamento e cancellazione. Per esempio, per $x \in [0.988, 1.012]$ si ottiene il risultato riportato nell'immagine seguente.



Gli errori di arrotondamento sono generalmente molto “piccoli”; tuttavia, se ripetuti all’interno di algoritmi “lunghi” e complessi, possono avere effetti vistosi o addirittura catastrofici se non opportunamente controllati.

Esempio 1.5.11. Esempi di effetti catastrofici legati alla propagazione di errori di arrotondamento sono stati l’esplosione del missile Ariane (4 Giugno 1996), dovuta all’insorgenza di overflow nel computer di bordo, e quello di un missile Patriot caduto, durante la prima guerra del Golfo del 1991, su una caserma americana in seguito a un errore di arrotondamento nel calcolo della traiettoria.

1.6 Dal Problema Matematico al Problema Numerico

In questa sezione introduciamo un ambiente comune per l’analisi di molti metodi numerici considerati durante il corso, e concetti chiave come la buona posizione di un problema, la consistenza, la stabilità, e la convergenza di un metodo numerico, nonché l’idea di numero di condizionamento.

1.6.1 Il problema matematico

Iniziamo considerando un *problema fisico* (PF) dotato di una *soluzione fisica*, simbolicamente indicata come x_{ph} , e dipendente da dati genericamente indicati con d . Il *problema matematico* (PM) è rappresentato dalla formulazione matematica del PF ed è dotato di *soluzione matematica* x . Indichiamo il PM come:

$$F(x; d) = 0, \quad (1.14)$$

dove $x \in \mathcal{X}$ e $d \in \mathcal{D}$, essendo \mathcal{X} e \mathcal{D} due spazi adeguati. L’errore tra le soluzioni fisica e matematica è chiamato *errore di modello* e viene indicato come $e_m := x_{ph} - x$. Tale sorgente di errore tiene conto di tutte quelle caratteristiche del problema fisico che non vengono rappresentate o catturate dal modello matematico. Non ci occuperemo, nel corso, di quantificare tale sorgente di errore, ma è bene tenere a mente che ogni modello matematico rappresenta una opportuna sintesi di un fenomeno fisico.

Invitiamo il lettore a prestare attenzione al cambio di notazione rispetto a pagina 7 (dove PM era indicato con \mathcal{P} , la soluzione matematica con u , il dato con g , etc.).

Esempio 1.6.1. Consideriamo come PF l’equilibrio del filo elastico fissato alle estremità già visto in precedenza. Nel caso in cui il peso del filo sia trascurabile rispetto al peso supportato, la forza esercitata sia uniforme rispetto alla distanza orizzontale, il problema matematico di trovare lo spostamento del filo u (la soluzione) può essere scritto come

$$F(u, d) = u - x \int_0^1 (1-s)f(s)ds + \int_0^x (x-s)f(s)ds = 0,$$

i cui dati sono $d = ((0, 1), 0, 0, f)$. Qui $\mathcal{D} = \mathbb{R}^2 \times \mathbb{R} \times \mathbb{R} \times C^0([0, 1])$, mentre $\mathcal{X} = C^2([0, 1])$.

Prima di risolvere un problema matematico, è necessario assicurarsi che questo sia *ben posto*.

Definizione 1.6.1. Il problema matematico $F(x; d) = 0$ è ben posto (*stabile*) se e solo se esiste un'unica soluzione $x \in \mathcal{X}$ che dipende con continuità dai dati $d \in \mathcal{D}$.

Sia \mathcal{D} l'insieme dei dati ammissibili, ovvero l'insieme dei valori di d in corrispondenza del quale il problema (1.14) ammette una soluzione unica. *Dipendenza continua dai dati* significa che varazioni continue delle perturbazioni sui dati $d \in \mathcal{D}$ comportano variazioni continue sulla soluzione $x \in \mathcal{X}$.

Esempio 1.6.2. Trovare il numero di radici reali di un polinomio non è un problema ben posto. Ad esempio, considerando $F(x, d) = x^4 - x^2(2d - 1) + d(d - 1)$, si ha una variazione discontinua del numero di radici reali poiché, se d varia continuamente in \mathbb{R} , abbiamo 4 radici reali se $d \geq 1$, 2 se $d \in (0, 1)$, mentre non esistono radici reali se $d < 0$.

PM che sono formalmente ben posti possono esibire "grandi" variazioni della soluzione x anche per "piccole" variazioni dei valori dei dati d . Una misura di tale sensitività è data dal numero di condizionamento del PM.

Definizione 1.6.2. Il numero di condizionamento relativo del PM $F(x; d) = 0$ per i dati $d \in \mathcal{D}$ è definito come

$$K(d) := \sup_{\substack{\delta d : (d + \delta d) \in \mathcal{D} \\ \epsilon \|\delta d\| \neq 0}} \left\{ \frac{\|\delta x\| / \|x\|}{\|\delta d\| / \|d\|} \right\}.$$

Il numero di condizionamento relativo di un PM è tale per cui $K(d) \geq 1$ per definizione. Se $K(d)$ è "piccolo", il PM è *ben condizionato*; se $K(d)$ è "grande", il PM si dice *mal condizionato*.

In un problema ben condizionato significa che la soluzione ottenuta con dati leggermente perturbati non differisce molto dalla soluzione del problema con i dati originali: nei problemi mal condizionati la soluzione è invece molto sensibile a piccole perturbazioni dei dati. La proprietà di un PM di essere ben condizionato è indipendente dal metodo numerico usato per risolverlo. Tuttavia, anche per problemi matematici estremamente semplici la nozione di condizionamento può riservare sorprese, e spiegare come la propagazione di piccole perturbazioni possa comportare errori piuttosto grandi nel risultato.

Esempio 1.6.3. Consideriamo il problema di moltiplicare due numeri reali, $F(x, d) = x - d_1 d_2 = 0$, essendo $d = (d_1, d_2)^T$, e perturbiamone i dati ottenendo $\tilde{d} = (\tilde{d}_1, \tilde{d}_2)^T = d_1(1 + \epsilon), d_2(1 + \epsilon))^T$; in questo caso $\|\delta d\| = \epsilon \|d\|$ e

$$\frac{|\delta x|}{|x|} = \frac{|\tilde{d}_1 \tilde{d}_2 - d_1 d_2|}{|d_1 d_2|} = \frac{|d_1 d_2 (1 + \epsilon)^2 - d_1 d_2|}{|d_1 d_2|} = (1 + \epsilon)^2 - 1 = \epsilon^2 + 2\epsilon;$$

cosicché se le perturbazioni sono piccole, $\epsilon \ll 1$, possiamo trascurare il fattore ϵ^2 rispetto a 2ϵ e ottenere $K(d) = 2$. Il problema di moltiplicare due numeri reali risulta cioè sempre ben condizionato.

In modo simile, possiamo ricavare che il numero di condizionamento del problema di sottrarre due numeri reali è $K(d) = 1$ se $\text{sign}(d_1) = -\text{sign}(d_2)$, ovvero, l'addizione di due numeri reali è sempre un problema ben condizionato. Se invece i due numeri hanno lo stesso segno e $d_1 \approx d_2$,

$$K(d) = \frac{|d_1| + |d_2|}{|d_1 - d_2|}$$

diventa molto grande e la sottrazione risulta mal condizionata in questo caso. Per convincercene, basta prendere ad esempio $d_1 = 1/51$ e $d_2 = 1/52$; otteniamo in questo caso $K(d) = 103$, e $\tilde{d}_1 = 0.196 \cdot 10^{-1}, \tilde{d}_2 = 0.192 \cdot 10^{-1}$, da cui $\tilde{d}_1 - \tilde{d}_2 = 0.400 \cdot 10^{-3}$, questo risultato è piuttosto diverso da quello esatto, $d_1 - d_2 = 0.377 \cdot 10^{-3}$; il grande numero di condizionamento riflette il fatto che la soluzione di tale problema può essere soggetta a grandi errori dovuti alla cancellazione di cifre significative.

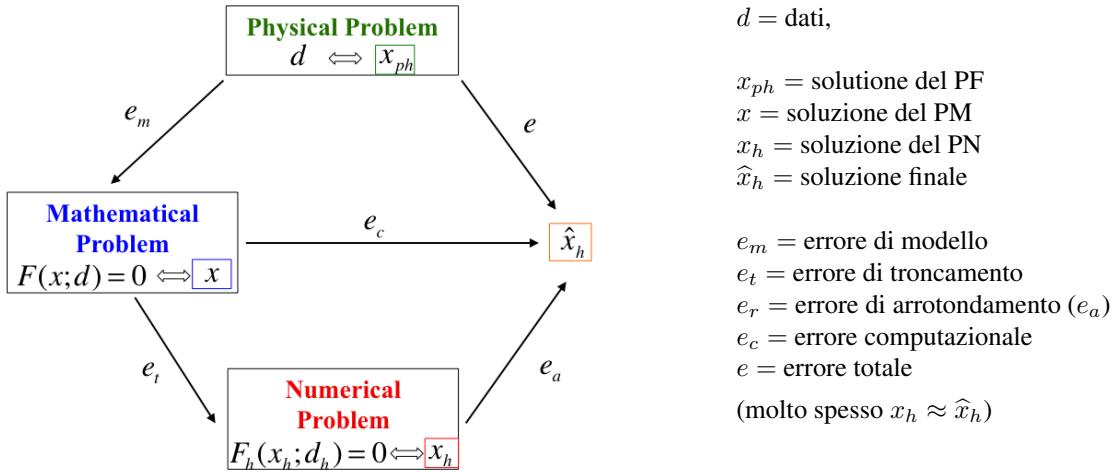


Figura 1.3: I problemi fisico, matematico e numerico e le sorgenti di errore.

1.6.2 Il problema numerico

Il *problema numerico* (PN) è un'approssimazione del PM (1.14); indichiamo la sua *soluzione numerica* come x_h , con h un adeguato parametro di *discretizzazione* (in altri casi si usa la notazione n per indicare il numero di iterazioni per un metodo iterativo). L'errore tra le soluzioni matematica e numerica è chiamato *errore di truncamento* $e_t := x - x_h$ (si veda lo schema di Fig. 1.3). Indichiamo il PN come:

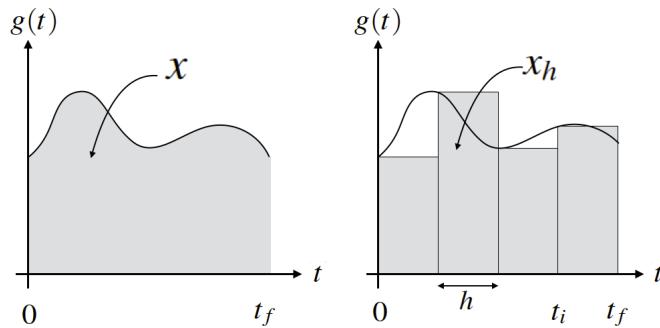
$$F_h(x_h; d_h) = 0, \quad (1.15)$$

dove $x_h \in \mathcal{X}_h$ e $d_h \in \mathcal{D}_h$, essendo \mathcal{X}_h e \mathcal{D}_h spazi appropriati.

La soluzione finale \hat{x}_h è in generale affetta da errori di arrotondamento $e_r := x_h - \hat{x}_h$. Tali errori sono legati alla soluzione del PN al *calcolatore*; essi dipendono dalla rappresentazione macchina al calcolatore e dall'aritmetica floating point.

Gli errori di truncamento e di arrotondamento concorrono a determinare l'*errore computazionale* $e_c := x - \hat{x}_h = e_t + e_r$. Osserviamo che per molti PN (anche se non sempre) $|e_r| \ll |e_t|$, per cui l'errore e_t è spesso identificato con e_c .

Esempio 1.6.4. Per il PM $F(x; d) = x - \int_0^{t_f} g(t) dt = 0$ con i dati $d = \{t_f, g(t)\}$, consideriamo per esempio il PN $F_h(x_h; d_h) = x_h - h \sum_{i=0}^{n-1} g(t_i) = 0$, dove $t_i = i h$ per $i = 0, \dots, n$, con $h = \frac{t_f}{n}$. Abbiamo già incontrato questo metodo per approssimare il calcolo di un integrale definito nella Sezione 1.4.



Come per un problema matematico, dobbiamo assicurarcene che anche il problema numerico sia ben posto; ugualmente, possiamo definire il numero di condizionamento di un problema numerico.

Definizione 1.6.3. Il problema $F_h(x_h; d_h) = 0$ di Eq. (1.15) è ben posto (*stabile*) se e solo se esiste un'unica soluzione $x_h \in \mathcal{X}_h$ che dipende con continuità dai dati $d_h \in \mathcal{D}_h$.

Definizione 1.6.4. Il numero di condizionamento relativo del PN $F_h(x_h; d_h) = 0$ per i dati $d_h \in \mathcal{D}_h$ è definito come:

$$K_h(d_h) := \sup_{\substack{\delta d_h : (d_h + \delta d_h) \in \mathcal{D}_h \\ e \|\delta d_h\| \neq 0}} \left\{ \frac{\|\delta x_h\|/\|x_h\|}{\|\delta d_h\|/\|d_h\|} \right\}.$$

Se $K_h(d_h)$ è piccolo, il problema numerico (1.15) è ben condizionato. Al contrario, se $K_h(d_h)$ è grande, il problema è mal condizionato.

Perché siamo così interessati a valutare la propagazione degli errori dovuti a piccole perturbazioni?

Vale il *principio di Wilkinson*, secondo il quale il risultato di un'operazione numerica al calcolatore (o in aritmetica floating point) equivale al risultato della medesima operazione in aritmetica esatta effettuata su dati affetti da una (piccola) perturbazione. Tale principio fornisce dunque uno strumento per quantificare l'effetto della propagazione degli errori di arrotondamento nel processo computazionale.

Siamo ovviamente interessati a sviluppare metodi numerici che permettano di ottenere errori computazionali che tendano a zero al migliorare della procedura di discretizzazione numerica. Tale concetto è espresso mediante la seguente

Definizione 1.6.5. Diciamo che un problema numerico è convergente quando l'errore computazionale tende a zero, ovvero:

$$\lim_{\substack{h \rightarrow 0 \\ (o n \rightarrow +\infty)}} e_c = 0.$$

Osserviamo che per alcuni PN il parametro di discretizzazione h è sostituito da un intero n , spesso con l'idea che $h \sim \frac{1}{n}$.

Un aspetto importante collegato alla convergenza del PN è *ordine di convergenza*; a questo fine, ridefiniamo l'errore computazionale e_c come $e_c = |x - \hat{x}_h|$.

Definizione 1.6.6. Se l'errore computazionale $e_c \leq C h^p$, con C una costante positiva indipendente da h e p , allora il PN è convergente con ordine p .

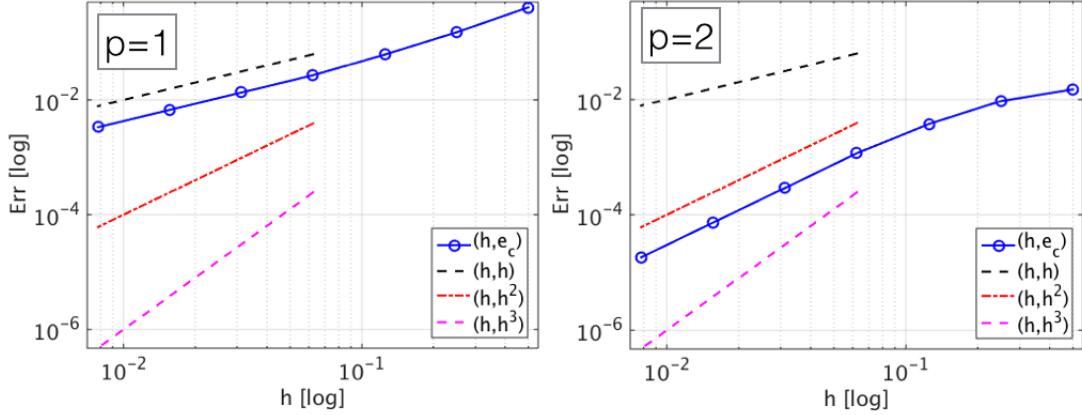
Se esiste una costante positiva $\tilde{C} \leq C$ indipendente da h e p tale per cui $\tilde{C} h^p \leq e_c \leq C h^p$, allora possiamo scrivere come $e_c \simeq C h^p$. Se possiamo scrivere $e_c \simeq C h^p$, allora l'ordine di convergenza p può essere stimato in due modi considerando un PM per cui l'esatta soluzione matematica x è nota:

- *Algebricamente.* Per prima cosa si calcolano gli errori e_{c1} e e_{c2} associati rispettivamente a due valori del parametro di discretizzazione h_1 e h_2 (che siano sufficientemente “piccoli”); in seguito, scrivendo $e_{c1} \simeq C h_1^p$, $e_{c2} \simeq C h_2^p$ e osservando che $\frac{e_{c1}}{e_{c2}} = \left(\frac{h_1}{h_2}\right)^p$, l'ordine p è stimato come:

$$p = \frac{\log(e_{c1}/e_{c2})}{\log(h_1/h_2)}.$$

- *Graficamente.* Gli errori e_c calcolati per valori differenti di h sono rappresentati graficamente vs. h in scala logaritmica (*log-log*) su entrambi gli assi. Dato che $\log e_c = \log(C h^p) = \log C + p \log h$,

abbiamo $p = \text{atan}(\theta)$, dove θ è la pendenza della curva (h, e_c) che è una linea retta nelle scale log–log. Invece di calcolare l'angolo θ di tale linea retta, si può verificare che le curve (h, e_c) e (h, h^p) siano *parallele* in scale log–log.



Affinché un problema numerico sia convergente, è sufficiente che sia ben posto (stabile)? No.

Occorre infatti che il problema numerico sia anche *consistente* con il problema matematico, ovvero ne risulti una *copia sufficientemente fedele*.

Definizione 1.6.7. Il PN (1.15) è consistente se e solo se $\lim_{h \rightarrow 0} F_h(x; d) = F(x; d) = 0$, con $d \in \mathcal{D}_h$.

Il PN (1.15) è fortemente consistente se e solo se $F_h(x; d) \equiv F(x; d) = 0$ per ogni $h > 0$, con $d \in \mathcal{D}_h$.

Esempio 1.6.5. Consideriamo due diversi PN associati al PM $F(x; d) = x - d = 0$, con $d = \sqrt{2}$, per cui la soluzione è $x = \sqrt{2}$.

- Definiamo il PN $F_n(x_n; d) = x_{n+1} - \frac{3}{4}x_n - \frac{1}{2x_n} = 0$ per $n \geq 0$, con $x_0 = 1$; in questo caso n assume il ruolo di parametro di discretizzazione e indica il numero di iterate. Dal momento che $F_n(x; d) = \sqrt{2} - \frac{3}{4}\sqrt{2} - \frac{1}{2\sqrt{2}} = 0$ per ogni $n \geq 0$, il PN è fortemente consistente.
- Consideriamo ora il seguente PN $F_n(x_n; d) = x_{n+1} - \frac{3}{4}x_n - \frac{1}{2x_n} + \frac{1}{(1+n)^5} = 0$ per $n \geq 0$, con $x_0 = 1$. Osserviamo che $F_n(x; d) = \frac{1}{(1+n)^5} \neq 0$ per $n \geq 0$, da cui si deduce che il PN non è fortemente consistente. Tuttavia, il PN è consistente essendo $\lim_{n \rightarrow +\infty} F_n(x; d) = 0$.

In sintesi, un PN deve essere: i) ben posto (ben condizionato); ii) consistente e; iii) convergente. La buona posizione è una condizione sufficiente per la convergenza a patto che il problema numerico sia consistente.

Teorema 1.6.1 (Lax–Richtmeyer, equivalenza). Se il PN $F_h(x_h; d_h) = 0$, per $x_h \in \mathcal{X}_h$ e $d_h \in \mathcal{D}_h$, è consistente, allora è ben posto se e solo se è anche convergente (cioè $x_h \rightarrow x$).

Come conseguenza del teorema di equivalenza, se il PN è consistente e ben posto, allora questo è anche convergente. Se invece, in maniera del tutto simile, il PN è consistente e convergente, allora esso è anche ben posto. Il precedente teorema di equivalenza è molto utile in quanto permette di verificare solo due delle proprietà di un PN per ottenere la terza; in generale, si osserva che è “facile” mostrare la consistenza di un PN mentre potrebbe essere più “difficile” mostrare la sua buona posizione e/o convergenza.

1.6.3 Scelta di un metodo numerico

La scelta di un metodo numerico (PN) per approssimare la soluzione x di un PM deve tenere in considerazione delle:

- proprietà matematiche del PM;
- l'efficienza computazionale in termini di: ordine di convergenza atteso dell'errore, i flops coinvolti nel calcolo, le prestazioni della CPU installata sul calcolatore, le modalità di accesso e la disponibilità della memoria di calcolo.

Supponiamo di indicare con m la dimensione del PN. I flops coinvolti nel calcolo della soluzione x_h del PN può dipendere dalla dimensione m del PN in modi differenti, come indicato nella seguente tabella.

	$O(1)$	$O(m)$	$O(m^\gamma)$	$O(\gamma^m)$	$O(m!)$
dipendenza flops	independente	lineare	polinomiale	esponenziale	fattoriale

Esempio 1.6.6. Da un punto di vista teorico, se la matrice $A \in \mathbb{R}^{n \times n}$ è non singolare, la soluzione $\mathbf{x} \in \mathbb{R}^n$ del sistema lineare $A\mathbf{x} = \mathbf{b}$, un PM molto semplice, si può ottenere applicando la *regola di Cramer*:

$$x_i = \frac{\det(B_i)}{\det(A)} \quad i = 1, \dots, n,$$

dove $B_i \in \mathbb{R}^{n \times n}$ è la matrice ottenuta sostituendo la i -esima colonna di A con il vettore $\mathbf{b} \in \mathbb{R}^n$:

$$B_i = \begin{bmatrix} a_{11} & \dots & b_1 & \dots & a_{1n} \\ a_{21} & \dots & b_2 & \dots & a_{2n} \\ \vdots & & \vdots & & \vdots \\ a_{n1} & \dots & b_n & \dots & a_{nn} \end{bmatrix}.$$

\uparrow
 i

Sfortunatamente, la soluzione di un sistema lineare con questo metodo richiede $O(e(n+1)!)$ operazioni. Se $n = 100$, sono richieste $100!e \approx 2.56 \cdot 10^{160}$ operazioni aritmetiche. Una macchina in grado di eseguire 10^{12} floating point operations (=flops) al secondo, cioè una potenza di calcolo di 1 TeraFlop, porterebbe a

$$\frac{2.56 \cdot 10^{160}}{10^{12}} = 2.56 \cdot 10^{148} \text{ secondi} \approx 8.11 \cdot 10^{140} \text{ anni}$$

In accordo con la teoria più accreditata, l'universo iniziò con il Big Bang circa $12.5(\pm) \cdot 10^9$ anni fa. Occorrono dunque metodi più efficienti. La tecnica più celebre è il *metodo di eliminazione di Gauss*³ che, come vedremo, richiede

$$O\left(\frac{2}{3}n^3\right) \text{ operazioni}$$

per risolvere un sistema lineare di dimensione n . Un sistema di dimensione $n = 100$ viene dunque risolto in 10^{-6} secondi avendo a disposizione una potenza di calcolo di 1 TeraFlop, e in meno di un secondo su un qualsiasi laptop.

³Carl Friedrich Gauss, 1777-1855; *Theoria motus corporum coelestium in sectionibus conicis solem ambientium* (1809)

Capitolo 2

Sistemi Lineari

Consideriamo la soluzione numerica di *sistemi lineari* tramite *metodi diretti* e *metodi iterativi*. Per semplicità, consideriamo il caso di sistemi lineari a valori reali, a meno che sia indicato diversamente; molte delle considerazioni e metodi presentati in questo capitolo possono essere applicati direttamente a sistemi lineari che coinvolgono numeri complessi.

2.1 Motivazioni, Esempi e Classificazione dei Metodi

Consideriamo la matrice quadrata $A \in \mathbb{R}^{n \times n}$ con $n \geq 1$, il vettore $\mathbf{b} \in \mathbb{R}^n$ e il vettore soluzione $\mathbf{x} \in \mathbb{R}^n$ del seguente sistema lineare:

$$A\mathbf{x} = \mathbf{b}. \quad (2.1)$$

L'obiettivo consiste nell'*approssimare numericamente* la soluzione $\mathbf{x} \in \mathbb{R}^n$ di tale sistema lineare. Ricordiamo le seguenti.

Definizione 2.1.1. La matrice $A \in \mathbb{R}^{n \times n}$ con $n \geq 1$ è non-singolare se e solo se $\det(A) \neq 0$.

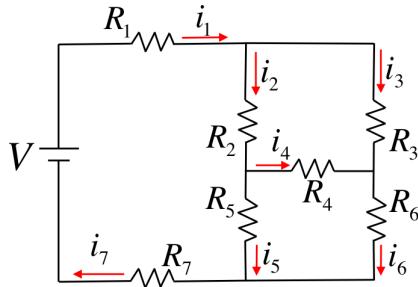
Proposizione 2.1.1. Se $A \in \mathbb{R}^{n \times n}$ è non-singolare, allora esiste un'unica soluzione $\mathbf{x} \in \mathbb{R}^n$ del sistema lineare (2.1).

In riferimento al sistema lineare (2.1), utilizziamo la notazione seguente per indicare gli elementi della matrice $A \in \mathbb{R}^{n \times n}$, ovvero $(A)_{ij} = a_{ij}$ per $i, j = 1, \dots, n$, e i vettori $\mathbf{x} \in \mathbb{R}^n$ e $\mathbf{b} \in \mathbb{R}^n$, ovvero rispettivamente $(\mathbf{x})_i = x_i$ e $(\mathbf{b})_i = b_i$ per $i = 1, \dots, n$. Inoltre, utilizzeremo la seguente notazione per indicare il sistema lineare in termini degli elementi di A , \mathbf{b} e \mathbf{x} :

$$\left\{ \begin{array}{lcl} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n & = & b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n & = & b_2 \\ \vdots & = & \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n & = & b_n \end{array} \right. \quad \text{o} \quad \left[\begin{array}{cccc} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{array} \right] \left[\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_n \end{array} \right] = \left[\begin{array}{c} b_1 \\ b_2 \\ \vdots \\ b_n \end{array} \right].$$

Osserviamo che i sistemi lineari possono essere direttamente interpretati come problemi matematici che modellizzano problemi fisici. Tuttavia, in molti casi, i sistemi lineari sono ottenuti come problemi numerici associati alla discretizzazione di problemi matematici; gli esempi mostrati nel capitolo precedente vanno in questa direzione. In tali casi, tanto maggiore è la dimensione n del sistema lineare, tanto maggiore è l'accuratezza dell'approssimazione del problema matematico che ha generato il sistema lineare; per tali problemi, è molto frequente risolvere sistemi lineari di dimensioni $O(n) = 10^5, 10^6$ oppure anche 10^7 .

Esempio 2.1.1. In questo esempio mostriamo come il problema matematico di determinare le correnti in un circuito elettrico corrisponda a un sistema lineare.



Il problema consiste nel trovare le correnti i_j per $j = 1, \dots, n$ distribuite nel circuito, dove $n = 7$, essendo specificate la tensione V e le resistenze R_j della parti che compongono il circuito. La chiusura del problema avviene per mezzo del bilancio delle tensioni ($V = V_1 + V_2 + V_5 + V_7$, $V_3 = V_2 + V_4$ e $V_5 = V_4 + V_6$), le leggi di Kirchhoff ($i_1 = i_2 + i_3$, $i_2 = i_4 + i_5$, $i_3 + i_4 = i_6$ e $i_5 + i_6 = i_7$) e le equazioni costitutive ($V_j = R_j i_j$ per ogni $j = 1, \dots, n$).

Otteniamo il seguente sistema lineare, la cui soluzione fornisce la distribuzione delle correnti $\{i_j\}_{j=0}^n$ nel circuito:

$$\begin{bmatrix} R_1 & R_2 & 0 & 0 & R_5 & 0 & R_7 \\ 0 & R_2 & -R_3 & R_4 & 0 & 0 & 0 \\ 0 & 0 & 0 & R_4 & -R_5 & R_6 & 0 \\ 1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & -1 \end{bmatrix} \begin{bmatrix} i_1 \\ i_2 \\ i_3 \\ i_4 \\ i_5 \\ i_6 \\ i_7 \end{bmatrix} = \begin{bmatrix} V \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Osserviamo che se la matrice $A \in \mathbb{R}^{n \times n}$ è *piena*, sarebbero necessarie, in principio, almeno n^2 operazioni per risolvere il sistema lineare. Anche se quest'assunzione è molto ottimistica, è opportuno considerare e sviluppare metodi per cui il numero di operazioni sia il più vicino possibile a questo numero ideale. Considerazioni differenti possono essere fatte per matrici *sparse*, ovvero matrici $A \in \mathbb{R}^{n \times n}$ per cui il numero di elementi non nulli è $O(n) \ll n^2$.

Osservazione 2.1.1. *La soluzione del sistema lineare (2.1) come $\mathbf{x} = A^{-1} \mathbf{b}$, ovvero calcolando e assemblando esplicitamente la matrice inversa di A , è una procedura computazionalmente inefficiente che dovrebbe essere evitata anche per matrici relativamente piccole.*

I metodi numerici per la soluzione di sistemi lineari possono essere classificati in due categorie: metodi *diretti* e metodi *iterativi*.

Definizione 2.1.2. *Con un metodo diretto la soluzione \mathbf{x} del sistema lineare (2.1) è ottenuta in un numero finito di passi. Al contrario, con un metodo iterativo la soluzione \mathbf{x} è ottenuta, in principio, in un numero infinito di passi.*

La scelta di un metodo diretto o iterativo per risolvere il sistema lineare (2.1) dipende da molteplici fattori, come la natura, la dimensione e la sparsità della matrice A , e anche dalle risorse computazionali disponibili (CPU e memoria).

2.2 Metodi Diretti

Consideriamo alcuni metodi diretti per la soluzione del sistema lineare $A \mathbf{x} = \mathbf{b}$ di Eq. (2.1) e analizziamo le loro proprietà. L'idea di fondo di questa famiglia di metodi consiste nel ricondurre la soluzione del generico sistema lineare $A \mathbf{x} = \mathbf{b}$ a quella di un *sistema lineare “più semplice”* attraverso un'opportuna manipolazione della matrice A .

2.2.1 Sistemi lineari “semplici”

Riportiamo di seguito alcuni esempi di sistemi lineari “semplici”, ovvero di “facile” risoluzione. Come anticipato, questa proprietà dipende dalla matrice $A \in \mathbb{R}^{n \times n}$ in considerazione.

Matrice diagonale

Consideriamo il caso di una *matrice diagonale* $D \in \mathbb{R}^{n \times n}$, cioè $(D)_{ii} = d_{ii}$ per $i = 1, \dots, n$ e $(D)_{ij} = 0$ per $i, j = 1, \dots, n$, ma $j \neq i$. In questo caso, la matrice diagonale D è il sistema lineare associato $D\mathbf{x} = \mathbf{b}$ sono dati rispettivamente da:

$$D = \begin{bmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_{nn} \end{bmatrix} \quad \text{e} \quad \begin{cases} d_{11}x_1 = b_1 \\ d_{22}x_2 = b_2 \\ \vdots = \vdots \\ d_{nn}x_n = b_n. \end{cases}$$

Assumiamo che in Eq. (2.1), sia $A = D$; allora la soluzione $\mathbf{x} \in \mathbb{R}^n$ del sistema diagonale $D\mathbf{x} = \mathbf{b}$ è:

$$x_i = \frac{b_i}{d_{ii}} \quad \text{per ogni } i = 1, \dots, n,$$

ed è ottenuta con n operazioni (divisioni).

Osservazione 2.2.1. Dal momento che D è una matrice diagonale, il suo determinante è calcolato come $\det(D) = \prod_{i=1}^n d_{ii}$. Ne consegue che $\det(D) \neq 0$ se e solo se $d_{ii} \neq 0$ per ogni $i = 1, \dots, n$.

Matrice triangolare inferiore: algoritmo delle sostituzioni in avanti

Definizione 2.2.1. $L \in \mathbb{R}^{n \times n}$ è una matrice triangolare inferiore se e solo se i suoi elementi sono tali che $(L)_{ij} = l_{ij} \in \mathbb{R}$ per $i = 1, \dots, n$, $j = 1, \dots, i$ e $(L)_{ij} = 0$ per $i = 1, \dots, n-1$, $j = i+1, \dots, n$; la matrice triangolare inferiore L è data da:

$$L = \begin{bmatrix} l_{11} & 0 & \cdots & & 0 \\ l_{21} & l_{22} & 0 & \cdots & 0 \\ l_{31} & l_{32} & l_{33} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \cdots & & l_{nn} \end{bmatrix}.$$

Data una matrice triangolare inferiore $L \in \mathbb{R}^{n \times n}$, consideriamo la soluzione del *sistema triangolare inferiore*:

$$L\mathbf{y} = \mathbf{b}, \quad \text{cioè} \quad \begin{cases} l_{11}y_1 & = b_1 \\ l_{21}y_1 + l_{22}y_2 & = b_2 \\ l_{31}y_1 + l_{32}y_2 + l_{33}y_3 & = b_3 \\ \vdots & \vdots \quad \vdots \quad \ddots \quad \vdots \\ l_{n1}y_1 + l_{n2}y_2 + l_{n3}y_3 + \cdots + l_{nn}y_n & = b_n, \end{cases}$$

dove, con riferimento a Eq. (2.1), poniamo $A = L$ e $\mathbf{y} = \mathbf{x}$. Il sistema triangolare inferiore $L\mathbf{y} = \mathbf{b}$ può essere risolto mediante l'*algoritmo delle sostituzioni in avanti*, ovvero:

$$\boxed{\begin{aligned} y_1 &= \frac{b_1}{l_{11}}, \\ y_i &= \frac{1}{l_{ii}} \left(b_i - \sum_{j=1}^{i-1} l_{ij}y_j \right) \quad \text{per } i = 2, \dots, n. \end{aligned}} \quad (2.2)$$

L'algoritmo delle sostituzioni in avanti risolve il sistema triangolare inferiore $L\mathbf{x} = \mathbf{b}$ in n^2 operazioni, dove n è la dimensione della matrice L ; infatti l'algoritmo esegue n divisioni, $\sum_{i=2}^n (i-1)$ sottrazioni e $\sum_{i=2}^n (i-1)$ moltiplicazioni, portando quindi il computo totale delle operazioni a $n + 2 \sum_{i=2}^n (i-1) = n^2$.

Osservazione 2.2.2. *Dal momento che L è una matrice triangolare inferiore, abbiamo $\det(L) = \prod_{i=1}^n l_{ii}$; pertanto, $\det(L) \neq 0$ se e solo se $l_{ii} \neq 0$ per ogni $i = 1, \dots, n$.*

Matrice triangolare superiore: algoritmo delle sostituzioni all'indietro

Definizione 2.2.2. *$U \in \mathbb{R}^{n \times n}$ è una matrice triangolare superiore se e solo se i suoi elementi sono tali che $(U)_{ij} = u_{ij} \in \mathbb{R}$ per $i = 1, \dots, n$, $j = i, \dots, n$ e $(U)_{ij} = 0$ per $i = 2, \dots, n$, $j = 1, \dots, i-1$; la matrice triangolare superiore U è data da:*

$$U = \begin{bmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ 0 & u_{22} & u_{23} & \cdots & u_{2n} \\ 0 & 0 & u_{33} & \cdots & u_{3n} \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & & \cdots & 0 & u_{nn} \end{bmatrix}. \quad (2.3)$$

Data una matrice triangolare superiore $U \in \mathbb{R}^{n \times n}$ consideriamo la soluzione del *sistema triangolare superiore*:

$$U\mathbf{x} = \mathbf{y}, \quad \text{cioè} \quad \left\{ \begin{array}{lcl} u_{11}x_1 + u_{12}x_2 + u_{13}x_3 + \cdots + u_{1n}x_n & = & y_1 \\ u_{22}x_2 + u_{23}x_3 + \cdots + u_{2n}x_n & = & y_2 \\ u_{33}x_3 + \cdots + u_{3n}x_n & = & y_3 \\ \ddots & \vdots & \vdots \\ u_{nn}x_n & = & y_n, \end{array} \right.$$

dove, con riferimento a Eq. (2.1), poniamo $A = U$ e $\mathbf{b} = \mathbf{y}$. Il sistema triangolare superiore $U\mathbf{x} = \mathbf{y}$ può essere risolto mediante l'*algoritmo delle sostituzioni all'indietro*, ovvero:

$$\begin{aligned} x_n &= \frac{y_n}{u_{nn}}, \\ x_i &= \frac{1}{u_{ii}} \left(y_i - \sum_{j=i+1}^n u_{ij}x_j \right) \quad \text{per } i = n-1, \dots, 1. \end{aligned} \quad (2.4)$$

L'algoritmo delle sostituzioni all'indietro risolve il sistema triangolare superiore $U\mathbf{y} = \mathbf{x}$ in n^2 operazioni, in analogia con l'algoritmo delle sostituzioni in avanti per sistemi triangolari inferiori.

Osservazione 2.2.3. *Dal momento che U è una matrice triangolare superiore, $\det(U) = \prod_{i=1}^n u_{ii}$, per cui $\det(U) \neq 0$ se e solo se $u_{ii} \neq 0$ per ogni $i = 1, \dots, n$.*

2.2.2 Metodo di fattorizzazione LU

Consideriamo la matrice non-singolare $A \in \mathbb{R}^{n \times n}$. La *fattorizzazione LU* (o decomposizione LU), ammesso che esista, della matrice A consiste nel determinare una matrice triangolare inferiore $L \in \mathbb{R}^{n \times n}$ e una matrice triangolare superiore $U \in \mathbb{R}^{n \times n}$ tali che:

$$A = L U.$$

Se la fattorizzazione LU della matrice A esiste ($A = L U$), allora il sistema lineare $A \mathbf{x} = \mathbf{b}$ può essere ri-

$$\begin{array}{ccc} \text{[Grey square]} & = & \text{[Top-left triangle]} \\ A & & L \\ & & \text{[Bottom-right triangle]} \\ & & U \end{array}$$

soltanto come soluzione in sequenza dei seguenti sistemi, il primo triangolare inferiore e il secondo triangolare superiore:

$$L \mathbf{y} = \mathbf{b} \quad \text{e} \quad U \mathbf{x} = \mathbf{y};$$

infatti, essendo $A = L U$, abbiamo $L U \mathbf{x} = \mathbf{b}$ da cui, introducendo il vettore ausiliario $\mathbf{y} = U \mathbf{x} \in \mathbb{R}^n$, otteniamo il precedente risultato.

Definizione 2.2.3. Il metodo di fattorizzazione LU per risolvere il sistema lineare $A \mathbf{x} = \mathbf{b}$ consiste nel:

1. determinare, se esiste, la fattorizzazione LU della matrice A ($A = L U$);
2. risolvere il sistema triangolare inferiore $L \mathbf{y} = \mathbf{b}$ con l'algoritmo delle sostituzioni in avanti (2.2);
3. risolvere il sistema triangolare superiore $U \mathbf{x} = \mathbf{y}$ con l'algoritmo delle sostituzioni all'indietro (2.4).

Dal momento che il metodo di fattorizzazione LU è basato sulla fattorizzazione LU di A , occorre determinare le matrici L e U di A , se queste esistono.

Esempio 2.2.1. Illustriamo la fattorizzazione LU della matrice $A \in \mathbb{R}^{n \times n}$ con $n = 2$, che è data da:

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix} \quad \text{oppure} \quad \begin{cases} l_{11} u_{11} & = a_{11} \\ l_{11} u_{12} & = a_{12} \\ l_{21} u_{11} & = a_{21} \\ l_{21} u_{12} + l_{22} u_{22} & = a_{22}. \end{cases}$$

Osserviamo che le matrici L e U coinvolgono in totale 6 incognite $l_{11}, l_{21}, l_{22}, u_{11}, u_{12}$ e u_{22} . D'altra parte, per la loro determinazione sono disponibili solo 4 vincoli.

Osservazione 2.2.4. Generalizzando l'esempio precedente, per la fattorizzazione LU di una generica matrice $A \in \mathbb{R}^{n \times n}$ ci sono $n^2 + n$ incognite delle matrici L e U , ma soltanto n^2 vincoli da impostare; infatti, abbiamo $a_{ij} = \sum_{r=1}^{\min\{i,j\}} l_{ir} u_{rj}$ per $i, j = 1, \dots, n$. Per ovviare a questo problema, per convenzione, gli elementi diagonali della matrice triangolare inferiore L ottenuta mediante fattorizzazione LU della matrice $A \in \mathbb{R}^{n \times n}$ sono posti pari a 1; cioè $l_{ii} = 1$ per ogni $i = 1, \dots, n$:

$$L = \begin{bmatrix} 1 & 0 & \cdots & & 0 \\ l_{21} & 1 & 0 & \cdots & 0 \\ l_{31} & l_{32} & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \cdots & l_{n,n-1} & 1 \end{bmatrix}. \quad (2.5)$$

Metodo di eliminazione di Gauss (MEG)

Il *metodo di eliminazione di Gauss* (MEG) è utilizzato per determinare la fattorizzazione LU di una matrice $A \in \mathbb{R}^{n \times n}$. Per illustrare l'algoritmo del MEG, introduciamo una notazione adeguata; in particolare, definiamo la matrice $\bar{A}^{(k)} \in \mathbb{R}^{n \times n}$ per un certo $k = 1, \dots, n$ come:

$$\bar{A}^{(k)} := \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2n}^{(2)} \\ 0 & 0 & \ddots & & \vdots \\ 0 & \cdots & 0 & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ 0 & \cdots & 0 & a_{k+1,k}^{(k)} & \cdots & a_{k+1,n}^{(k)} \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & a_{n,k}^{(k)} & \cdots & a_{n,n}^{(k)} \end{bmatrix} \quad \text{per } k = 1, \dots, n, \quad (2.6)$$

o, equivalentemente:

$$\left(\bar{A}^{(k)}\right)_{ij} = \begin{cases} a_{ij}^{(i)} & \text{per } i = 1, \dots, k-1, j = i, \dots, n \\ a_{ij}^{(k)} & \text{per } i, j = k, \dots, n \\ 0 & \text{altrimenti.} \end{cases} \quad \text{per } k = 1, \dots, n. \quad (2.7)$$

Per convenzione poniamo $\bar{A}^{(1)} \equiv A$, cioè $a_{ij}^{(1)} = a_{ij}$ per ogni $i, j = 1, \dots, n$.

Definizione 2.2.4. Dato un indice k , con $1 \leq k \leq n-1$, con riferimento alla matrice corrispondente $\bar{A}^{(k)}$ di Eq. (2.6), il suo elemento $a_{kk}^{(k)}$ è chiamato elemento pivotale (o pivot).

Il seguente algoritmo del MEG è utilizzato per determinare gli elementi della matrice $L \in \mathbb{R}^{n \times n}$ di Eq. (2.5) e $U \in \mathbb{R}^{n \times n}$ di Eq. (2.3) determinando la fattorizzazione LU di $A \in \mathbb{R}^{n \times n}$; la matrice U coincide con $\bar{A}^{(n)}$ ottenuta alla fine del MEG ($U = \bar{A}^{(n)}$).

Algorithm 2.1: Metodo di eliminazione di Gauss (MEG)

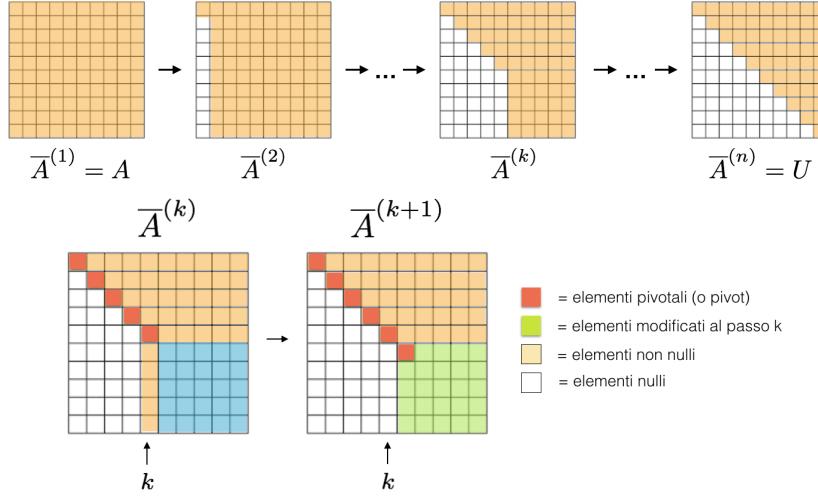
```

assegnare  $\bar{A}^{(1)} = A$ ;
for  $k = 1, \dots, n-1$  do
  for  $i = k+1, \dots, n$  do
     $l_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}};$ 
    for  $j = k+1, \dots, n$  do
       $| a_{ij}^{(k+1)} = a_{ij}^{(k)} - l_{ik} a_{kj}^{(k)};$ 
    end
  end
  assegnare  $\bar{A}^{(k+1)}$  come in Eq. (2.7);
end
assegnare  $L$  come in Eq. (2.5) e porre  $U = \bar{A}^{(n)}$ ;
```

Il numero di operazioni associate al MEG per la fattorizzazione LU di A è $O\left(\frac{2}{3}n^3\right)$.

Osservazione 2.2.5. Per effettuare la fattorizzazione LU della matrice A mediante il MEG, tutti gli elementi pivotali $a_{kk}^{(k)}$ associati alle matrici $\bar{A}^{(k)}$ devono essere non nulli, cioè $a_{kk}^{(k)} \neq 0$ per ogni $k = 1, \dots, n-1$.

Riportiamo di seguito una rappresentazione schematica dell'algoritmo MEG, mostrando in particolare come viene operata la trasformazione dalla matrice $\bar{A}^{(k)}$ nella matrice $\bar{A}^{(k+1)}$ al generico passo k .



Esempio 2.2.2. Formiamo la fattorizzazione LU della matrice $A = \begin{bmatrix} 3 & 1 & -1 \\ 1 & 4 & 2 \\ -1 & -1 & 4 \end{bmatrix}$ utilizzando il MEG. Iniziamo ponendo $\bar{A}^{(1)} = A$ e osservando che $n = 3$; quindi, otteniamo la fattorizzazione LU seguendo l'algoritmo del MEG 2.1.

$$\bullet \quad k = 1: \quad a_{11}^{(1)} = 3,$$

$$- \quad i = k + 1 = 2: \quad l_{21} = \frac{a_{21}^{(1)}}{a_{11}^{(1)}} = \frac{1}{3},$$

$$* \quad j = k + 1 = 2: \quad a_{22}^{(2)} = a_{22}^{(1)} - l_{21} a_{12}^{(1)} = \frac{11}{3},$$

$$* \quad j = n = 3: \quad a_{23}^{(2)} = a_{23}^{(1)} - l_{21} a_{13}^{(1)} = \frac{7}{3};$$

$$- \quad i = n = 3: \quad l_{31} = \frac{a_{31}^{(1)}}{a_{11}^{(1)}} = -\frac{1}{3},$$

$$* \quad j = k + 1 = 2: \quad a_{32}^{(2)} = a_{32}^{(1)} - l_{31} a_{12}^{(1)} = -\frac{2}{3},$$

$$* \quad j = n = 3: \quad a_{33}^{(2)} = a_{33}^{(1)} - l_{31} a_{13}^{(1)} = \frac{11}{3}.$$

$$L = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{3} & 1 & 0 \\ -\frac{1}{3} & ? & 1 \end{bmatrix}, \quad \bar{A}^{(2)} = \begin{bmatrix} 3 & 1 & -1 \\ 0 & \frac{11}{3} & \frac{7}{3} \\ 0 & -\frac{2}{3} & \frac{11}{3} \end{bmatrix}.$$

$$\bullet \quad k = 2: \quad a_{22}^{(2)} = \frac{11}{3},$$

$$- \quad i = k + 1 = n = 3: \quad l_{32} = \frac{a_{32}^{(2)}}{a_{22}^{(2)}} = -\frac{2}{11},$$

$$* \quad j = k + 1 = n = 3: \quad a_{33}^{(3)} = a_{33}^{(2)} - l_{32} a_{23}^{(2)} = \frac{45}{11}.$$

$$L = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{3} & 1 & 0 \\ -\frac{1}{3} & -\frac{2}{11} & 1 \end{bmatrix}, \quad U = \bar{A}^{(3)} = \begin{bmatrix} 3 & 1 & -1 \\ 0 & \frac{11}{3} & \frac{7}{3} \\ 0 & 0 & \frac{45}{11} \end{bmatrix}.$$

Osservazione 2.2.6. Una volta determinata la fattorizzazione LU della matrice A , le matrici L e U possono essere memorizzate in un'unica matrice, eventualmente sovrascritta alla matrice A . Infatti, al generico passo $k = 1, \dots, n$ della fattorizzazione LU, possiamo scrivere la matrice $\bar{A}^{(k)}$ di Eq. (2.6) e la matrice L , assemblata nelle sue prime k colonne, come:

$$\bar{A}^{(k)} := \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1n}^{(1)} \\ l_{21} & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2n}^{(2)} \\ l_{31} & l_{32} & \ddots & & \vdots \\ l_{k1} & \cdots & l_{k,k-1} & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ l_{k+1,1} & \cdots & l_{k+1,k-1} & a_{k+1,k}^{(k)} & \cdots & a_{k+1,n}^{(k)} \\ \vdots & & \vdots & \vdots & & \vdots \\ l_{n,1} & \cdots & l_{n,k-1} & a_{n,k}^{(k)} & \cdots & a_{n,n}^{(k)} \end{bmatrix}.$$

Qual è il ruolo dei moltiplicatori nella fattorizzazione LU?

Per meglio comprendere il motivo per cui la matrice L abbia per componenti della parte triangolare inferiore (diagonale esclusa) proprio i moltiplicatori, sfruttiamo il seguente risultato:

Proposizione 2.2.1. Per ogni $n \geq 2$, il prodotto di due matrici triangolari inferiori (o superiori) di ordine n risulta triangolare inferiore (o superiore). Inoltre, l'inversa di una matrice triangolare inferiore $L \in \mathbb{R}^{n \times n}$ è triangolare inferiore di ordine n , e se $l_{ii} = 1$ per ogni $i = 1, \dots, n$, allora anche $(L^{-1})_{ii} = 1$ per ogni $i = 1, \dots, n$.

Possiamo sfruttare questo risultato riformulando la fattorizzazione LU in termini di operazioni matriziali. Al primo passo otteniamo

$$\bar{A}^{(2)} = M_1 \bar{A}^{(1)} = M_1 A,$$

dove

$$M_1 = \begin{bmatrix} 1 & 0 & \cdots & & 0 \\ -l_{21} & 1 & 0 & \cdots & 0 \\ -l_{31} & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ -l_{n1} & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}, \quad l_{i1} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}}, \quad i = 2, \dots, n.$$

Procedendo, si ottiene

$$\bar{A}^{(3)} = M_2 \bar{A}^{(2)} = M_2 M_1 \bar{A}^{(1)} = M_2 M_1 A,$$

dove

$$M_2 = \begin{bmatrix} 0 & 0 & \cdots & & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & -l_{32} & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ 0 & -l_{n2} & 0 & \cdots & 0 & 1 \end{bmatrix}, \quad l_{i2} = \frac{a_{i2}^{(2)}}{a_{22}^{(2)}}, \quad i = 3, \dots, n.$$

Continuando in questo modo, otteniamo che

$$\bar{A}^{(n)} = M_{n-1}M_{n-2}\dots M_2M_1A = U,$$

da cui

$$A = (M_{n-1}M_{n-2}\dots M_2M_1)^{-1}U = M_1^{-1}M_2^{-1}\dots M_{n-2}^{-1}M_{n-1}^{-1}U,$$

ovvero

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

dove si può mostrare che

$$M_1^{-1} = \begin{bmatrix} 1 & 0 & \cdots & & 0 \\ l_{21} & 1 & 0 & \cdots & 0 \\ l_{31} & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ l_{n1} & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}, \quad M_2^{-1} = \begin{bmatrix} 0 & 0 & \cdots & & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & l_{32} & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ 0 & l_{n2} & 0 & \cdots & 0 & 1 \end{bmatrix}, \quad \dots$$

e che il prodotto $M_1^{-1}M_2^{-1}\dots M_{n-2}^{-1}M_{n-1}^{-1}$ fornisce una matrice le cui componenti risultano proprio i moltiplicatori.

Calcolo del determinante della matrice A tramite fattorizzazione LU

Se la matrice A ammette la fattorizzazione LU, allora si ha:

$$\det(A) = \det(LU) = \det(L)\det(U) = \det(U),$$

essendo $\det(L) = 1$. Perciò, la fattorizzazione LU può essere convenientemente utilizzata per calcolare il determinante della matrice A in un numero di operazioni $O(2n^3/3)$.

Proprietà della fattorizzazione LU e del metodo di fattorizzazione LU

Il MEG fornisce la fattorizzazione LU della matrice $A \in \mathbb{R}^{n \times n}$ richiesta per risolvere il sistema lineare $A\mathbf{x} = \mathbf{b}$ per mezzo del metodo di fattorizzazione LU di Definizione 2.2.3. Il numero di operazioni associate al metodo di fattorizzazione LU è $O(2n^3/3)$; infatti, $O(2n^3/3)$ operazioni sono richieste dal MEG, mentre n^2 per entrambi gli algoritmi delle sostituzioni in avanti e all'indietro.

Osservazione 2.2.7. La fattorizzazione LU della matrice $A \in \mathbb{R}^{n \times n}$ è indipendente dal vettore $\mathbf{b} \in \mathbb{R}^n$ associato al sistema lineare $A\mathbf{x} = \mathbf{b}$. Il metodo di fattorizzazione LU può dunque essere efficientemente utilizzato per risolvere il sistema lineare per diversi vettori \mathbf{b} dal momento che L e U possono essere assemblate una sola volta. I costi computazionali associati alla soluzione del sistema lineare per ogni nuovo vettore \mathbf{b} sono determinati solamente dalla soluzione dei sistemi triangolare inferiore $L\mathbf{y} = \mathbf{b}$ e triangolare superiore $U\mathbf{x} = \mathbf{y}$, per mezzo degli algoritmi delle sostituzioni in avanti e all'indietro.

Determiniamo i casi per cui la fattorizzazione LU di una matrice A non-singolare esiste ed è unica. A tal fine, richiamiamo la seguente definizione.

Definizione 2.2.5. La sottomatrice principale di $A \in \mathbb{R}^{n \times n}$ di ordine i , con $1 \leq i \leq n$, è la matrice $A_i \in \mathbb{R}^{i \times i}$ tale che $(A_i)_{lm} = (A)_{lm}$ per ogni $l, m = 1, \dots, i$.

La seguente proposizione esprime la condizione necessaria e sufficiente per l'esistenza e unicità della fattorizzazione LU di una matrice A non-singolare.

Proposizione 2.2.2 (Condizione necessaria e sufficiente per la fattorizzazione LU). *Data una matrice $A \in \mathbb{R}^{n \times n}$ non-singolare, la sua fattorizzazione LU esiste ed è unica se e solo se $\det(A_i) \neq 0$ per ogni $i = 1, \dots, n - 1$ (ovvero tutte le sottomatrici principali di A di ordine i , con $1 \leq i \leq n - 1$, sono non-singolari).*

Esempio 2.2.3. La fattorizzazione LU della matrice non-singolare $A = \begin{bmatrix} 1 & 1 & 4 \\ 2 & 2 & 3 \\ 4 & 6 & 7 \end{bmatrix}$, ottenuta utilizzando il MEG, non esiste. Infatti, dalla Proposizione 2.2.2, abbiamo $\det(A_1) = \det([1]) \neq 0$, ma $\det(A_2) = \det\left(\begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}\right) = 0$; nel caso specifico, si trova l'elemento pivotale $a_{22}^{(2)} = 0$ nell'applicazione dell'algoritmo MEG.

In alcuni casi, non è necessario verificare la condizione di Proposizione 2.2.2 per stabilire l'esistenza e unicità della fattorizzazione LU di A , ma basta verificare alcune condizioni solo sufficienti. A tal fine richiamiamo le seguenti definizioni.

Definizione 2.2.6. La matrice $A \in \mathbb{R}^{n \times n}$ è:

- simmetrica se e solo se $A^T \equiv A$;
- definita positiva se e solo se $\mathbf{z}^T A \mathbf{z} > 0$ per ogni $\mathbf{z} \in \mathbb{R}^n$ con $\mathbf{z} \neq \mathbf{0}$.

Definizione 2.2.7. La matrice $A \in \mathbb{R}^{n \times n}$ è:

- a dominanza diagonale per riga se e solo se $|a_{ii}| \geq \sum_{j=1, j \neq i}^n |a_{ij}|$ per ogni $i = 1, \dots, n$;
- a dominanza diagonale stretta per righe se e solo se $|a_{ii}| > \sum_{j=1, j \neq i}^n |a_{ij}|$ per ogni $i = 1, \dots, n$;
- a dominanza diagonale per colonna se e solo se $|a_{ii}| \geq \sum_{j=1, j \neq i}^n |a_{ji}|$ per ogni $i = 1, \dots, n$;
- a dominanza diagonale stretta per colonna se e solo se $|a_{ii}| > \sum_{j=1, j \neq i}^n |a_{ji}|$ per ogni $i = 1, \dots, n$.

La seguente proposizione esprime una serie di condizioni solo *sufficienti* per garantire l'esistenza e unicità della fattorizzazione LU di una matrice A non-singolare.

Proposizione 2.2.3 (Condizioni sufficienti per la fattorizzazione LU). *Data la matrice $A \in \mathbb{R}^{n \times n}$, se una delle seguenti condizioni è verificata:*

- A è simmetrica e definita positiva,
- A è a dominanza diagonale stretta per righe,
- oppure A è a dominanza diagonale stretta per colonne,

allora la fattorizzazione LU di A esiste ed è unica.

Esempio 2.2.4. La fattorizzazione LU della matrice $A = \begin{bmatrix} 4 & -2 & 1 \\ -2 & -5 & -1 \\ 1 & 3 & 9 \end{bmatrix}$ esiste ed è unica sulla base della

Proposizione 2.2.3 dato che A è a dominanza diagonale stretta per righe; infatti, $|4| > |-2| + |1|$, $|-5| > |-2| + |-1|$ e $|9| > |1| + |3|$.

Esempio 2.2.5. Consideriamo la matrice non-singolare $A = \begin{bmatrix} 1 & -2 & 8 \\ -2 & 5 & -1 \\ 1 & 1 & 0 \end{bmatrix}$. Nessuna delle condizioni sufficienti di Proposizione 2.2.3 è soddisfatta; pertanto, non è possibile trarre conclusioni sull'esistenza e unicità della fattorizzazione LU di A usando questa proposizione. In tal caso, l'esistenza e unicità della fattorizzazione LU di A deve essere verificata in termini delle condizioni necessarie e sufficienti di Proposizione 2.2.2, da cui deduciamo che esiste ed è unica essendo $\det(A_1) = \det([1]) \neq 0$ e $\det(A_2) = \det\left(\begin{bmatrix} 1 & -2 \\ -2 & 5 \end{bmatrix}\right) = 9 \neq 0$.

Tecnica del pivoting per righe (pivotazione per righe)

Se la condizione necessaria e sufficiente di Proposizione 2.2.2 non è soddisfatta, la fattorizzazione LU della matrice A non può essere determinata utilizzando l'algoritmo MEG 2.1. D'altra parte, necessitiamo ancora di risolvere il sistema lineare $A\mathbf{x} = \mathbf{b}$ per ogni matrice A che sia non-singolare per mezzo della fattorizzazione LU. A tale scopo, è possibile adottare la cosiddetta *tecnica di pivoting* (o *pivotazione*) in combinazione con il MEG per la fattorizzazione LU di A .

Definizione 2.2.8. La tecnica di pivoting per righe consiste nell'applicare opportune permutazioni delle righe della matrice non-singolare A in presenza di elementi pivotali $a_{kk}^{(k)}$ nulli incontrati durante l'applicazione dell'algoritmo MEG, per qualche indice $k = 1, \dots, n-1$.

Osservazione 2.2.8. L'applicazione del MEG con la tecnica del pivoting (per righe) assicura l'esistenza e l'unicità della fattorizzazione LU per ogni matrice $A \in \mathbb{R}^{n \times n}$ non-singolare.

Consideriamo il caso specifico di *tecnica di pivoting* con *permutozione per righe* della matrice A . Tale permutazione delle righe della matrice $A \in \mathbb{R}^{n \times n}$ consiste nel pre-moltiplicarla per una matrice di permutazione $P \in \mathbb{R}^{n \times n}$, ovvero PA . La matrice di permutazione P è ortogonale, cioè $P^T = P^{-1}$ ($P^T P = I$); se $P = I$, allora non sono applicate permutazioni alla matrice A . In generale, la matrice di permutazione P si ottiene simultaneamente all'assemblaggio delle matrici L e U durante l'uso del MEG con la tecnica del pivoting (per righe) applicato alla matrice non-singolare A .

Esempio 2.2.6. La matrice non-singolare A dell'Esempio 2.2.3 non ammette la fattorizzazione LU con il MEG standard (senza pivoting) dal momento che l'elemento pivotale (pivot) $a_{22}^{(2)} = 0$. Introducendo la matrice di permutazione

per righe $P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$ e applicandola ad A , otteniamo la matrice $\tilde{A} = PA = \begin{bmatrix} 1 & 1 & 4 \\ 4 & 6 & 7 \\ 2 & 2 & 3 \end{bmatrix}$ tale per cui

la seconda e terza riga sono state permutate. Applicando il MEG standard alla matrice permutata \tilde{A} , otteniamo la fattorizzazione LU con $L = \begin{bmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 2 & 0 & 1 \end{bmatrix}$ e $U = \begin{bmatrix} 1 & 1 & 4 \\ 0 & 2 & -9 \\ 0 & 0 & -5 \end{bmatrix}$, dove $\tilde{A} = PA = LU$; in tal caso i nuovi elementi pivotali sono $\tilde{a}_{11}^{(1)} = 1 \neq 0$ e $\tilde{a}_{22}^{(2)} = 2 \neq 0$.

In generale, la tecnica del pivoting è applicata contestualmente al MEG anche se gli elementi pivotali non sono necessariamente nulli. Infatti, la tecnica del pivoting può essere usata anche per ridurre e contenere la propagazione degli *errori di arrotondamento* associati all'applicazione del MEG al calcolatore. Ovvero, all'iterata generica $k = 1, \dots, n-1$ del MEG, la riga k è permutata con la riga l , dove

$$l = \arg \max_{i=k, \dots, n} |a_{ik}^{(k)}|,$$

essendo $\tilde{a}_{kk}^{(k)} = a_{lk}^{(k)}$ il nuovo elemento pivotale k -esimo:

Algorithm 2.2: Metodo di eliminazione di Gauss (MEG) con pivoting (per righe)

```

assegnare  $\bar{A}^{(1)} = A$  e  $P = I$  ;
for  $k = 1, \dots, n - 1$  do
    trovare  $\bar{r}$  tale che  $|a_{\bar{r}k}^{(k)}| = \max_{r=k, \dots, n} |a_{\bar{r}k}^{(k)}|$  e scambiare la riga  $k$  con la riga  $\bar{r}$  sia in  $\bar{A}^{(k)}$  che in  $P$ 
    ;
    for  $i = k + 1, \dots, n$  do
         $l_{ik} = a_{ik}^{(k)} / a_{kk}^{(k)}$  ;
        for  $j = k + 1, \dots, n$  do
             $a_{ij}^{(k+1)} = a_{ij}^{(k)} - l_{ik} a_{kj}^{(k)}$ ;
        end
    end
    assegnare  $\bar{A}^{(k+1)}$  come in Eq. (2.7) ;
end
assegnare  $L$  come in Eq. (2.5) e porre  $U = \bar{A}^{(n)}$ ;
```

Esempio 2.2.7. Si consideri la matrice $A = \begin{bmatrix} \epsilon & 1 \\ 1 & 1 \end{bmatrix}$ con $0 < \epsilon \ll 1$, ovvero molto “piccolo”. Applicando il MEG (senza pivoting) alla matrice A si hanno $L = \begin{bmatrix} 1 & 0 \\ 1/\epsilon & 1 \end{bmatrix}$ e $U = \begin{bmatrix} \epsilon & 1 \\ 0 & (1 - 1/\epsilon) \end{bmatrix}$, con elemento pivotale $a_{11}^{(1)} = \epsilon$. Il calcolo della fattorizzazione LU con MEG al calcolatore genera errori di arrotondamento rilevanti per ϵ molto “piccolo” dovuti al calcolo di $u_{22} = (1 - 1/\epsilon)$, essendo $1/\epsilon$ molto “grande”. Invece, applicando il MEG con pivotazione per righe di A , ovvero usando $P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, si ottengono: la matrice permutata $\tilde{A} = PA$, $\tilde{L} = \begin{bmatrix} 1 & 0 \\ \epsilon & 1 \end{bmatrix}$ e $\tilde{U} = \begin{bmatrix} 1 & 1 \\ 0 & (1 - \epsilon) \end{bmatrix}$, con elemento pivotale $\tilde{a}_{11}^{(1)} = 1$. In tal caso, il calcolo della fattorizzazione LU tramite MEG con pivoting per righe al calcolatore consente di contenere gli errori di arrotondamento; infatti, per ϵ molto “piccolo”, si ha $\tilde{u}_{22} = (1 - \epsilon)$.

Se la tecnica del pivoting per righe è applicata al fine della determinazione della fattorizzazione LU della matrice non-singolare A , appunto per mezzo della matrice di permutazione per righe P , allora le matrici L e U determinano la fattorizzazione LU della matrice permutata PA come:

$$PA = LU.$$

Ne consegue che il sistema lineare $A\mathbf{x} = \mathbf{b}$ può essere risolto per mezzo della soluzione in sequenza dei seguenti sistemi triangolari inferiori e superiori:

$$Ly = Pb \quad \text{e} \quad Ux = y;$$

infatti, si hanno $PA\mathbf{x} = Pb$ e $LU\mathbf{x} = Pb$ per cui, introducendo il vettore $y = U\mathbf{x}$, otteniamo il risultato precedente.

Definizione 2.2.9. Il metodo della fattorizzazione LU con pivoting per righe *per la soluzione del sistema lineare $A\mathbf{x} = \mathbf{b}$, cioè basato sulla permutazione per righe con la matrice P* , consiste in:

1. determinare la fattorizzazione LU della matrice PA ($PA = LU$);
2. risolvere il sistema lineare triangolare inferiore $Ly = Pb$ con l’algoritmo delle sostituzioni in avanti (2.2);
3. risolvere il sistema triangolare superiore $Ux = y$ con l’algoritmo delle sostituzioni all’indietro (2.4).

Tecnica del pivoting totale (pivotazione totale)

La tecnica di pivotazione può essere estesa alla ricerca dell'elemento pivotale "migliore" in tutta la sottomatrice $A^{(k)}$ di $\bar{A}^{(k)}$ di Eqs. (2.6) e (2.7) al generico passo $k = 1, \dots, n - 1$ del MEG, dove gli elementi di $A^{(k)}$ sono $(A^{(k)})_{i-k+1,j-k+1} = (\bar{A}^{(k)})_{ij} = a_{ij}^{(k)}$ per $i, j = k, \dots, n$. In tal caso, si parla di *pivotazione totale*, in quanto le permutazioni della matrice A coinvolgono non solo le righe, ma anche le colonne. Nello specifico, all'iterata generica $k = 1, \dots, n - 1$ del MEG, la riga k di $\bar{A}^{(k)}$ è permutata con la riga l e simultaneamente la colonna k è permutata con la colonna m , dove $(l, m) = \arg \max_{i, j=k, \dots, n} |a_{ij}^{(k)}|$, essendo infine $\tilde{a}_{kk}^{(k)} = a_{lm}^{(k)}$ il nuovo elemento pivotale della matrice permutata. Osserviamo che la permutazione totale comporta costi computazionali aggiuntivi rispetto alla pivotazione per righe; infatti, per ogni $k = 1, \dots, n - 1$, la ricerca dell'elemento pivotale "migliore" avviene su in insieme di $(n + 1 - k)^2$ elementi – della sottomatrice $A^{(k)} \in \mathbb{R}^{(n+1-k) \times (n+1-k)}$ – che è molto più grande di quello corrispondente alla riga k della sottomatrice $A^{(k)}$, ovvero gli $(n + 1 - k)$ elementi $\{(A^{(k)})_{ik}\}_{i=k}^n$.

La *pivotazione totale* applicata alla matrice $A \in \mathbb{R}^{n \times n}$ consiste pertanto nell'introdurre, oltre alla matrice di permutazione per righe $P \in \mathbb{R}^{n \times n}$ da pre-moltiplicare ad A , una matrice di *permutazione per colonne* $Q \in \mathbb{R}^{n \times n}$ da post-moltiplicare ad A come AQ . La matrice di permutazione per colonne Q è ortogonale, cioè $Q^T = Q^{-1}$ (per cui $Q^T Q = I$); se $Q = I$, allora non vi sono permutazioni di colonne della matrice A .

Applichiamo la tecnica di pivotazione totale per la determinazione della fattorizzazione LU della matrice non-singolare A ; allora, le matrici L e U determinano la fattorizzazione LU della matrice permutata, sia per righe che per colonne, PAQ come:

$$PAQ = LU.$$

Osserviamo che $QQ^{-1} = I$, per cui il sistema lineare $A\mathbf{x} = \mathbf{b}$ può essere riscritto come $PAQQ^{-1}\mathbf{x} = P\mathbf{b}$, ovvero $LUQ^{-1}\mathbf{x} = P\mathbf{b}$. Introducendo i vettori $\mathbf{x}^* = Q^{-1}\mathbf{x}$ e $\mathbf{y}^* = U\mathbf{x}^*$, si ottengono i seguenti sistemi lineari:

$$Ly^* = P\mathbf{b}, \quad U\mathbf{x}^* = \mathbf{y}^* \quad \text{e} \quad \mathbf{x} = Q\mathbf{x}^*.$$

Definizione 2.2.10. Il metodo della fattorizzazione LU con pivoting totale *per la soluzione del sistema lineare $A\mathbf{x} = \mathbf{b}$, cioè basato sulla permutazione per righe e colonne tramite le matrici P e Q , consiste in:*

1. determinare la fattorizzazione LU della matrice PAQ ($PAQ = LU$);
2. risolvere il sistema lineare triangolare inferiore $Ly^* = P\mathbf{b}$ con l'algoritmo delle sostituzioni in avanti (2.2);
3. risolvere il sistema triangolare superiore $U\mathbf{x}^* = \mathbf{y}^*$ con l'algoritmo delle sostituzioni all'indietro (2.4).
4. determinare $\mathbf{x} = Q\mathbf{x}^*$.

Osservazione 2.2.9. Se alla matrice simmetrica $A \in \mathbb{R}^{n \times n}$ è applicata la permutazione totale di righe e colonne tale per cui la matrice permutata è $\tilde{A} = PAQ$, allora $\tilde{A} \in \mathbb{R}^{n \times n}$ è simmetrica solo se $Q \equiv P$.

Osservazione 2.2.10. La tecnica di pivotazione totale può essere convenientemente applicata per prevenire e/o contenere il fenomeno del cosiddetto fill-in per cui una matrice A sparsa può dare luogo a una fattorizzazione LU con matrici L e U relativamente piene di elementi non nulli.

2.2.3 Algoritmo di Thomas

Consideriamo la matrice non-singolare $A \in \mathbb{R}^{n \times n}$, con $n \geq 2$, e *tridiagonale*, ovvero espressa come:

$$A = \begin{bmatrix} a_1 & c_1 & 0 & \cdots & & 0 \\ e_2 & a_2 & c_2 & 0 & \cdots & 0 \\ 0 & e_3 & a_3 & c_3 & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ 0 & \cdots & 0 & e_{n-2} & a_{n-2} & c_{n-2} & 0 \\ 0 & \cdots & 0 & e_{n-1} & a_{n-1} & c_{n-1} & 0 \\ 0 & \cdots & 0 & e_n & a_n & & \end{bmatrix},$$

con elementi a valori reali $\{a_i\}_{i=1}^n$, $\{c_i\}_{i=1}^{n-1}$ e $\{e_i\}_{i=2}^n$. Assumiamo che tale matrice A ammetta l'esistenza e unicità della fattorizzazione LU senza pivoting; nel caso di una matrice tridiagonale A , la fattorizzazione LU genera le seguenti matrici bidiagonali L e U :

$$L = \begin{bmatrix} 1 & 0 & \cdots & & 0 \\ \beta_2 & 1 & 0 & \cdots & 0 \\ 0 & \beta_3 & 1 & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \\ 0 & \cdots & 0 & \beta_{n-1} & 1 & 0 \\ 0 & \cdots & 0 & 0 & \beta_n & 1 \end{bmatrix} \quad \text{e} \quad U = \begin{bmatrix} \alpha_1 & c_1 & 0 & \cdots & 0 \\ 0 & \alpha_2 & c_2 & 0 & \cdots \\ 0 & 0 & \alpha_3 & c_3 & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \\ 0 & \cdots & 0 & 0 & \alpha_{n-1} & c_{n-1} \\ 0 & \cdots & 0 & 0 & 0 & \alpha_n \end{bmatrix},$$

con elementi a valori reali $\{\alpha_i\}_{i=1}^n$ e $\{\beta_i\}_{i=2}^n$ determinati come:

$$\alpha_1 = a_1,$$

$$\beta_i = \frac{e_i}{\alpha_{i-1}} \quad \text{e} \quad \alpha_i = a_i - \beta_i c_{i-1} \quad \text{per } i = 2, \dots, n.$$

(2.8)

Consideriamo ora il sistema lineare $A \mathbf{x} = \mathbf{b}$, con $A \in \mathbb{R}^{n \times n}$ la matrice tridiagonale sopra descritta, che risolviamo con il metodo della fattorizzazione LU. La fattorizzazione LU di A si effettua usando l'Eq. (2.8); in tal modo, il sistema lineare $L \mathbf{y} = \mathbf{b}$ viene risolto per mezzo del seguente algoritmo delle sostituzioni in avanti adattato alla matrice bidiagonale inferiore L :

$$y_1 = b_1,$$

$$y_i = b_i - \beta_i y_{i-1} \quad \text{per } i = 2, \dots, n.$$

(2.9)

Infine, il sistema lineare $U \mathbf{x} = \mathbf{y}$ si risolve per mezzo del seguente algoritmo delle sostituzioni all'indietro adattato alla matrice bidiagonale superiore U :

$$x_n = \frac{y_n}{\alpha_n},$$

$$x_i = \frac{y_i - c_i x_{i+1}}{\alpha_i} \quad \text{per } i = n-1, \dots, 1.$$

(2.10)

Definizione 2.2.11. L'algoritmo di Thomas per la soluzione del sistema lineare $A \mathbf{x} = \mathbf{b}$, con A una matrice non-singolare e tridiagonale che ammette un'unica fattorizzazione LU senza pivoting, consiste nel:

1. determinare la fattorizzazione LU della matrice A ($A = L U$) usando l'algoritmo di Eq. (2.8);
2. risolvere il sistema bidiagonale inferiore $L \mathbf{y} = \mathbf{b}$ con l'algoritmo di Eq. (2.9);
3. risolvere il sistema bidiagonale superiore $U \mathbf{x} = \mathbf{y}$ con l'algoritmo di Eq. (2.10).

L'algoritmo di Thomas richiede solo $O(8n)$ operazioni (precisamente $8n - 7$) per risolvere il sistema lineare associato alla matrice tridiagonale $A \in \mathbb{R}^{n \times n}$. Osserviamo che l'applicazione del metodo della fattorizzazione LU completa della matrice A avrebbe richiesto invece $O(2n^3/3)$ operazioni.

2.2.4 Metodo della fattorizzazione di Cholesky

Ricordiamo innanzitutto la seguente

Definizione 2.2.12. La matrice $A \in \mathbb{R}^{n \times n}$ è:

- simmetrica se e solo se $A^T \equiv A$;
- definita positiva se e solo se $\mathbf{z}^T A \mathbf{z} > 0$ per ogni $\mathbf{z} \in \mathbb{R}^n$ con $\mathbf{z} \neq \mathbf{0}$.

Se la matrice A è *simmetrica e definita positiva*, è possibile utilizzare la più computazionalmente conveniente *fattorizzazione di Cholesky* al posto della fattorizzazione LU.

Definizione 2.2.13. Sia $A \in \mathbb{R}^{n \times n}$ simmetrica e definita positiva. Allora, la sua fattorizzazione di Cholesky consiste nel determinare una matrice triangolare superiore $R \in \mathbb{R}^{n \times n}$ tale per cui:

$$A = R^T R.$$

L'espressione generale della matrice triangolare superiore $R \in \mathbb{R}^{n \times n}$ è:

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} & \cdots & r_{1n} \\ 0 & r_{22} & r_{23} & \cdots & r_{2n} \\ \vdots & \ddots & \ddots & \cdots & \vdots \\ 0 & \cdots & 0 & r_{nn} \end{bmatrix},$$

la quale, per $A \in \mathbb{R}^{n \times n}$ simmetrica e definita positiva, è determinata tramite l'*algoritmo di Cholesky*.

Algorithm 2.3: Fattorizzazione di Cholesky

```

 $r_{11} = \sqrt{a_{11}};$ 
for  $i = 2, \dots, n$  do
  for  $j = 1, \dots, i-1$  do
     $r_{ji} = \frac{1}{r_{jj}} \left( a_{ij} - \sum_{k=1}^{j-1} r_{ki} r_{kj} \right);$ 
  end
   $r_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} r_{ki}^2};$ 
end

```

L'algoritmo di Cholesky richiede $O(n^3/3)$ operazioni per determinare la matrice triangolare superiore R , un numero di flops circa la metà di quello associato alla fattorizzazione LU; inoltre, anche la memoria utilizzata dal calcolatore è inferiore.

Se A è simmetrica e definita positiva, la fattorizzazione di Cholesky esiste ($A = R^T R$) e la soluzione del sistema lineare $A \mathbf{x} = \mathbf{b}$ può essere ottenuta sequenzialmente come soluzione dei seguenti sistemi triangolari inferiore e superiore:

$$R^T \mathbf{y} = \mathbf{b} \quad \text{e} \quad R \mathbf{x} = \mathbf{y},$$

essendo R^T una matrice triangolare inferiore; infatti, dato che $A = R^T R$, abbiamo $R^T R \mathbf{x} = \mathbf{b}$ da cui il precedente risultato segue introducendo il vettore $\mathbf{y} = R \mathbf{x} \in \mathbb{R}^n$.

Definizione 2.2.14. Il metodo della fattorizzazione di Cholesky per la soluzione del sistema lineare $A \mathbf{x} = \mathbf{b}$, con A simmetrica e definita positiva, consiste nel:

1. determinare la fattorizzazione di Cholesky della matrice A ($A = R^T R$);
2. risolvere il sistema triangolare inferiore $R^T \mathbf{y} = \mathbf{b}$ con l'algoritmo delle sostituzioni in avanti (2.2);
3. risolvere il sistema triangolare superiore $R \mathbf{x} = \mathbf{y}$ con l'algoritmo delle sostituzioni all'indietro (2.4).

2.2.5 Accuratezza della soluzione numerica ottenuta mediante metodi diretti

In questa sezione ci occupiamo dell'accuratezza della soluzione del sistema lineare $A \mathbf{x} = \mathbf{b}$ ottenuta per mezzo di metodi diretti. Infatti, quando si utilizza un calcolatore, la soluzione numerica può essere influenzata dalla propagazione di errori di arrotondamento (*round-off*).

Nozioni preliminari e definizioni

Richiamiamo alcune definizioni di algebra lineare.

Definizione 2.2.15. Dato un vettore $\mathbf{v} \in \mathbb{R}^n$, la sua norma p è definita come

$$\|\mathbf{v}\|_p := \left(\sum_{i=1}^n |v_i|^p \right)^{1/p} \quad \text{per } 1 \leq p \leq +\infty.$$

In particolare, dato un vettore $\mathbf{v} \in \mathbb{R}^n$, abbiamo:

$$\|\mathbf{v}\|_2 = \sqrt{\mathbf{v} \cdot \mathbf{v}} = \sqrt{\sum_{i=1}^n |v_i|^2}, \quad \|\mathbf{v}\|_1 = \sum_{i=1}^n |v_i|, \quad \text{e} \quad \|\mathbf{v}\|_\infty = \max_{i=1,\dots,n} |v_i|.$$

Tipicamente, la norma 2 di un vettore \mathbf{v} è indicata semplicemente con $\|\mathbf{v}\| \equiv \|\mathbf{v}\|_2$.

Osservazione 2.2.11. Per ogni norma p , $1 \leq p \leq +\infty$, vale la seguente disegualanza triangolare:

$$\|\mathbf{u} + \mathbf{v}\|_p \leq \|\mathbf{u}\|_p + \|\mathbf{v}\|_p \quad \forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^n.$$

Osservazione 2.2.12. Osserviamo che $\|\mathbf{v}\|_\infty = \max_{i=1,\dots,n} |v_i|$ risulta essere $\|\mathbf{v}\|_\infty = \lim_{p \rightarrow \infty} \|\mathbf{v}\|_p$. Infatti, per ogni $\mathbf{v} \in \mathbb{R}^n$, $\mathbf{v} \neq \mathbf{0}$, si ha che $\tilde{\mathbf{v}} = \mathbf{v}/\|\mathbf{v}\|_\infty$ è tale che $1 \leq \|\tilde{\mathbf{v}}\|_p \leq n^{1/p}$, dal momento che le componenti di \mathbf{v} possono essere al più tutte pari a 1, e almeno una di esse deve risultare tale. Di conseguenza, passando al limite per $p \rightarrow \infty$ e sfruttando il teorema del confronto per i limiti, si ha che

$$\lim_{p \rightarrow \infty} \|\tilde{\mathbf{v}}\|_p = 1 \quad \text{da cui} \quad \|\mathbf{v}\|_\infty = \lim_{p \rightarrow \infty} \|\mathbf{v}\|_p.$$

Definizione 2.2.16. Data una matrice $A \in \mathbb{C}^{n \times n}$, i suoi autovalori $\{\lambda_i(A)\}_{i=1}^n \in \mathbb{C}$ e i corrispondenti autovettori $\{\mathbf{v}_i\}_{i=1}^n \in \mathbb{C}^n$ sono tali che $A \mathbf{v}_i = \lambda_i \mathbf{v}_i$ per ogni $i = 1, \dots, n$. Gli autovalori $\{\lambda_i(A)\}_{i=1}^n$ corrispondono agli zeri del polinomio caratteristico della matrice A , cioè $p_A(\lambda) := \det(A - \lambda I)$.

Definizione 2.2.17. Il raggio spettrale della matrice $A \in \mathbb{C}^{n \times n}$, di autovalori $\{\lambda_i(A)\}_{i=1}^n \in \mathbb{C}$, è definito come:

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

Data una matrice $A \in \mathbb{C}^{n \times n}$, osserviamo che:

- $\det(A) = \prod_{i=1}^n \lambda_i(A)$;
- $\lambda_i(A^{-1}) = 1/\lambda_{n+1-i}(A)$ per $i = 1, \dots, n$, se esiste l'inversa A^{-1} di A ;
- $\rho(A) \geq 0$.

Nel seguito ci concentriamo su un matrice a valori reali $A \in \mathbb{R}^{n \times n}$.

Proposizione 2.2.4. Se la matrice $A \in \mathbb{R}^{n \times n}$ è simmetrica, allora i suoi autovalori sono reali, cioè $\lambda_i(A) \in \mathbb{R}$ per ogni $i = 1, \dots, n$. Ne consegue che, se $A \in \mathbb{R}^{n \times n}$ è simmetrica, allora è anche definita positiva se e solo se i suoi autovalori sono tutti strettamente positivi, cioè $\lambda_i(A) > 0$ per ogni $i = 1, \dots, n$.

Definizione 2.2.18. Data la matrice $A \in \mathbb{R}^{n \times n}$, la sua norma p è definita come:

$$\|A\|_p := \sup_{\substack{\mathbf{v} \in \mathbb{R}^n, \\ \mathbf{v} \neq \mathbf{0}}} \frac{\|A\mathbf{v}\|_p}{\|\mathbf{v}\|_p} \quad \text{per } 1 \leq p \leq +\infty. \quad (2.11)$$

Data $A \in \mathbb{R}^{n \times n}$, abbiamo:

- $\|A\|_1 = \max_{j=1, \dots, n} \left(\sum_{i=1}^n |a_{ij}| \right)$ e $\|A\|_\infty = \max_{i=1, \dots, n} \left(\sum_{j=1}^n |a_{ij}| \right)$;
- $\|A\|_2 = \sup_{\mathbf{v} \in \mathbb{R}^n, \mathbf{v} \neq \mathbf{0}} \frac{\|A\mathbf{v}\|}{\|\mathbf{v}\|} = \sqrt{\lambda_{\max}(A^T A)} = \sqrt{\rho(A^T A)}$;
- se A è simmetrica e definita positiva¹, allora $\|A\|_2 = \lambda_{\max}(A)$ dal momento che $\lambda_{\max}(A^T A) = (\lambda_{\max}(A))^2$.

Osservazione 2.2.13. La norma matriciale definita in Eq. (2.11) si dice norma indotta dalla norma p vettoriale. In particolare, per una norma indotta, valgono le due seguenti proprietà:

1. compatibilità:

$$\|A\mathbf{v}\|_p \leq \|A\|_p \|\mathbf{v}\|_p \quad \forall \mathbf{v} \in \mathbb{R}^n;$$

2. submoltiplicatività:

$$\|AB\|_p \leq \|A\|_p \|B\|_p \quad \forall A, B \in \mathbb{R}^{n \times n}.$$

Definizione 2.2.19. Il numero di condizionamento in norma p di una matrice non-singolare $A \in \mathbb{R}^{n \times n}$ è definito come:

$$K_p(A) := \|A\|_p \|A^{-1}\|_p \quad \text{per } 1 \leq p \leq +\infty.$$

Per convenzione, se A è singolare, $K_p(A) = +\infty$.

¹Si noti che, se $A \in \mathbb{R}^{n \times n}$ è una matrice qualsiasi, $A^T A$ è simmetrica, dato che $(A^T A)^T = A^T (A^T)^T = A^T A$, e (semi)-definita positiva, essendo

$$\mathbf{x}^T (A^T A) \mathbf{x} = (\mathbf{x}^T A^T) A \mathbf{x} = (A \mathbf{x})^T A \mathbf{x} = (A \mathbf{x}) \cdot (A \mathbf{x}) = \|A \mathbf{x}\|^2 \geq 0;$$

se inoltre A è non singolare, allora necessariamente $A \mathbf{x} = \mathbf{0}$ se e solo se $\mathbf{x} = \mathbf{0}$, dunque in questo caso $A^T A$ è simmetrica e definita positiva. Dunque, si ha che $\lambda_{\max}(A^T A) = \rho(A^T A)$. Inoltre, se A è simmetrica, allora $\|A\|_2 = \sqrt{\rho(A^T A)} = \sqrt{\rho^2(A)} = \rho(A) = |\lambda_{\max}(A)|$; se in più A è anche definita positiva, allora $\|A\|_2 = \lambda_{\max}(A)$.

Definizione 2.2.20. Il numero di condizionamento spettrale di una matrice non-singolare $A \in \mathbb{R}^{n \times n}$ è definito come:

$$K(A) := \rho(A) \rho(A^{-1}) = \frac{\max_{i=1,\dots,n} |\lambda_i(A)|}{\min_{i=1,\dots,n} |\lambda_i(A)|},$$

dove $\rho(A)$ e $\rho(A^{-1})$ sono i raggi spettrali delle matrici A e A^{-1} , rispettivamente.

Data una matrice non-singolare $A \in \mathbb{R}^{n \times n}$ abbiamo:

- $K_p(A) \geq 1$ per ogni $1 \leq p \leq +\infty$;
- $K_2(A) = \|A\|_2 \|A^{-1}\|_2 = \sqrt{\frac{\lambda_{\max}(A^T A)}{\lambda_{\min}(A^T A)}}$;
- Se gli autovalori di A sono reali e strettamente positivi, $K(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$, dove $\lambda_{\max}(A)$ e $\lambda_{\min}(A)$ sono rispettivamente il massimo e il minimo autovalore di A ;
- se A è simmetrica e definita positiva, allora $K_2(A) \equiv K(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$.

Il numero di condizionamento di una matrice A fornisce una misura della sensibilità della soluzione del sistema lineare $A \mathbf{x} = \mathbf{b}$ rispetto a perturbazioni dei dati, cioè \mathbf{b} e A stessa. Il sistema è detto essere *ben condizionato* se $K_p(A)$ è relativamente “piccolo”, e *mal condizionato* se $K_p(A)$ è “molto grande” (esempio $O(10^9)$ o più grande...).

Accuratezza della soluzione numerica

Risolvere *numericamente* il sistema lineare $A \mathbf{x} = \mathbf{b}$ mediante un metodo diretto è equivalente a risolvere in aritmetica esatta il seguente sistema lineare *perturbato*:

$$(A + \delta A) \hat{\mathbf{x}} = \mathbf{b} + \delta \mathbf{b}, \quad (2.12)$$

dove $\hat{\mathbf{x}} \in \mathbb{R}^n$ è la soluzione numerica, $\delta A \in \mathbb{R}^{n \times n}$ è la matrice di perturbazione di A , e $\delta \mathbf{b} \in \mathbb{R}^n$ è il vettore di perturbazione di \mathbf{b} . Per poter quantificare l’accuratezza della soluzione numerica $\hat{\mathbf{x}}$, diamo le seguenti definizioni e stime per l’errore:

Definizione 2.2.21. Dato il sistema lineare $A \mathbf{x} = \mathbf{b}$ si definisce:

- L’errore (assoluto) $\mathbf{e} := \mathbf{x} - \hat{\mathbf{x}}$, dove $\mathbf{e} \in \mathbb{R}^n$;
- L’errore relativo $e_{rel} := \frac{\|\mathbf{x} - \hat{\mathbf{x}}\|}{\|\mathbf{x}\|}$, dove $\mathbf{x} \neq 0$, e dove $e_{rel} \in \mathbb{R}$;
- Il residuo $\mathbf{r} := \mathbf{b} - A \hat{\mathbf{x}}$, dove $\mathbf{r} \in \mathbb{R}^n$;
- Il residuo relativo $r_{rel} := \frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}$, per $\mathbf{b} \neq 0$, dove $r_{rel} \in \mathbb{R}$.

Osservazione 2.2.14. Se $\delta A = 0$ nel sistema perturbato (2.12), allora si ha $A \hat{\mathbf{x}} = \mathbf{b} + \delta \mathbf{b}$, da cui $\delta \mathbf{b} = -\mathbf{r}$.

Osservazione 2.2.15. In generale, il residuo \mathbf{r} e il residuo relativo r_{rel} sono utilizzati come stimatori dell’errore associato alla soluzione numerica $\hat{\mathbf{x}}$; infatti, la soluzione esatta \mathbf{x} del sistema lineare $A \mathbf{x} = \mathbf{b}$ è generalmente sconosciuta.

Proposizione 2.2.5. Per il sistema lineare perturbato (2.12), se $K_2(A) \frac{\|\delta A\|_2}{\|A\|_2} < 1$, allora l'errore relativo associato alla soluzione numerica \hat{x} è stimato (maggiorato) come:

$$e_{rel} \leq \frac{K_2(A)}{1 - K_2(A)} \frac{\|\delta A\|_2}{\|A\|_2} \left(\frac{\|\delta A\|_2}{\|A\|_2} + \frac{\|\delta b\|}{\|b\|} \right).$$

Dimostrazione. A partire da Eq. (2.12) e dal fatto che $Ax = b$, sottraendo membro a membro si ottiene

$$b + \delta b - b = (A + \delta A) \hat{x} - Ax = \delta A \hat{x} + A(\hat{x} - x)$$

da cui, essendo la matrice A non-singolare,

$$\delta x = \hat{x} - x = A^{-1}(-\delta A \hat{x} + \delta b).$$

Prendendo la norma 2, si ha che

$$\begin{aligned} \|\delta x\| &\leq \|A^{-1}(-\delta A \hat{x} + \delta b)\| \leq \|A^{-1}\| \|-\delta A \hat{x} + \delta b\| \\ &\leq \|A^{-1}\| (\|-\delta A \hat{x}\| + \|\delta b\|) \leq \|A^{-1}\| (\|\delta A\| \|\hat{x}\| + \|\delta b\|) \end{aligned}$$

avendo sfruttato la disegualanza triangolare e la compatibilità della norma p matriciale. Poiché risulta inoltre che

$$\|\hat{x}\| = \|\hat{x} - x + x\| \leq \|\hat{x} - x\| + \|x\| = \|\delta x\| + \|x\|$$

e che

$$\|b\| = \|Ax\| \leq \|A\| \|x\|,$$

possiamo dedurre che

$$\begin{aligned} \|\delta x\| &\leq \|A^{-1}\| \left(\frac{\|\delta A\|}{\|A\|} \|A\| \|\hat{x}\| + \frac{\|\delta b\|}{\|b\|} \|b\| \right) \\ &\leq \|A^{-1}\| \left(\frac{\|\delta A\|}{\|A\|} \|A\| \|\hat{x}\| + \frac{\|\delta b\|}{\|b\|} \|A\| \|x\| \right) = K_2(A) \left(\frac{\|\delta A\|}{\|A\|} \|\hat{x}\| + \frac{\|\delta b\|}{\|b\|} \|x\| \right) \\ &\leq K_2(A) \left(\frac{\|\delta A\|}{\|A\|} (\|x\| + \|\delta x\|) + \frac{\|\delta b\|}{\|b\|} \|x\| \right) \end{aligned}$$

ovvero che

$$\|\delta x\| \left(1 - K_2(A) \frac{\|\delta A\|}{\|A\|} \right) \leq K_2(A) \left(\frac{\|\delta A\|_2}{\|A\|_2} + \frac{\|\delta b\|}{\|b\|} \right) \|x\|.$$

Infine, dividendo per $1 - K_2(A) \frac{\|\delta A\|}{\|A\|} > 0$ grazie all'ipotesi fatta, si ricava la tesi, ovvero che

$$e_{rel} = \frac{\|\delta \hat{x}\|}{\|x\|} \leq \frac{K_2(A)}{1 - K_2(A)} \frac{\|\delta A\|_2}{\|A\|_2} \left(\frac{\|\delta A\|_2}{\|A\|_2} + \frac{\|\delta b\|}{\|b\|} \right). \quad \square$$

Osservazione 2.2.16. Il risultato di Proposizione 2.2.5 vale in generale per ogni norma p .

Corollario 2.2.1. Sotto le stesse ipotesi di Proposizione 2.2.5 e se $\frac{\|\delta A\|_2}{\|A\|_2} = 0$, o circa zero (per esempio per perturbazioni $\delta A \simeq 0$), allora l'errore relativo associato alla soluzione numerica \hat{x} è stimato come:

$$e_{rel} \leq K_2(A) r_{rel} = K_2(A) \frac{\|r\|}{\|b\|}. \quad (2.13)$$

La stima dell'errore di Eq. (2.13) è una *stima a posteriori dell'errore* e può essere valutata una volta che la soluzione numerica \hat{x} è stata calcolata.

Osservazione 2.2.17. Osserviamo come la stima a posteriori dell'errore di Eq. (2.13) possa essere dimostrata anche indipendentemente dal risultato della Proposizione 2.2.5. Infatti, dalla definizione di residuo, discende che

$$\mathbf{r} := \mathbf{b} - A\hat{\mathbf{x}} = A\mathbf{x} - A\hat{\mathbf{x}} \quad \Rightarrow \quad \mathbf{x} - \hat{\mathbf{x}} = A^{-1}\mathbf{r}$$

da cui otteniamo che

$$\|\mathbf{x} - \hat{\mathbf{x}}\| = \|A^{-1}\mathbf{r}\| \leq \|A^{-1}\| \|\mathbf{r}\|;$$

inoltre, essendo $\|\mathbf{b}\| = \|A\mathbf{x}\| \leq \|A\| \|\mathbf{x}\|$, risulta che $\frac{1}{\|\mathbf{x}\|} \leq \frac{\|A\|}{\|\mathbf{b}\|}$ da cui, infine,

$$e_{rel} := \frac{\|\mathbf{x} - \hat{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \|A^{-1}\| \frac{\|\mathbf{r}\|}{\|\mathbf{x}\|} \leq \|A^{-1}\| \|\mathbf{r}\| \frac{\|A\|}{\|\mathbf{b}\|} = K_2(A) \frac{\|\mathbf{r}\|}{\|\mathbf{b}\|} = K_2(A) r_{rel}.$$

Sulla base del risultato in (2.13), il *residuo relativo* r_{rel} rappresenta un *criterio* soddisfacente per stimare l'*errore* associato alla soluzione numerica $\hat{\mathbf{x}}$ del sistema lineare ottenuta mediante l'uso di un metodo diretto al calcolatore soltanto se il numero di condizionamento è “piccolo”, cioè quando la matrice A è ben condizionata. Al contrario, se il numero di condizionamento della matrice A è “grande”, cioè se A è mal condizionata, allora l'errore associato a $\hat{\mathbf{x}}$ potrebbe essere molto “grande” anche se r_{rel} è “piccolo”, a causa della propagazione di errori di *round-off* durante l'applicazione del metodo diretto al calcolatore.

La stima a posteriori dell'errore di Eq. (2.13) si ottiene nel caso in cui la perturbazione $\delta A \simeq 0$. Tale assunzione è però giustificabile solo in casi specifici, per esempio, quando la fattorizzazione LU della matrice A è ottenuta tramite applicazione del MEG con pivoting per righe o totale. Un'interpretazione viene fornita dal seguente risultato.

Teorema 2.2.1 (Wilkinson 1961, Higham, 2002). *Data la matrice $A \in \mathbb{R}^{n \times n}$ e il sistema lineare $A\mathbf{x} = \mathbf{b}$, sia $\hat{\mathbf{x}}$ la soluzione ottenuta al calcolatore tramite il metodo della fattorizzazione LU, ovvero applicando il MEG. Allora, si hanno:*

$$(A + \delta A)\hat{\mathbf{x}} = \mathbf{b} \quad e \quad \|\delta A\|_\infty \leq C n^3 \rho_n \varepsilon_M \|A\|_\infty,$$

dove $\rho_n = \frac{\max_{i,j,k=1,\dots,n} |a_{ij}^{(k)}|}{\max_{i,j=1,\dots,n} |a_{ij}|}$ è il fattore di crescita, $C > 0$ una costante positiva e ε_M l'epsilon macchina; invece, $\rho_n = \frac{\max_{i,j,k=1,\dots,n} |\tilde{a}_{ij}^{(k)}|}{\max_{i,j=1,\dots,n} |a_{ij}|}$ se il MEG è applicato con una tecnica di pivoting per righe o totale.

In generale, $\rho_n \geq 1$, ma può essere arbitrariamente “grande” se il MEG viene applicato senza alcuna tecnica di pivoting.

Al contrario, $\rho_n \leq 2^{n-1}$ se si utilizza il MEG con pivoting per righe, mentre $\rho_n \leq n^{1/2} n^{(\log n)/4}$ per il MEG con pivoting totale. Tuttavia, tali stime sono pessimistiche quando una tecnica di pivoting viene applicata; in tali casi infatti, si ha generalmente $\rho_n \leq 50$. Si deduce che lo scopo delle tecniche di pivoting consiste nel contenere e limitare il fattore di crescita ρ_n ; pertanto, quando tecniche di pivoting sono applicate contestualmente al MEG per determinare la fattorizzazione LU di A , l'assunzione $\delta A \simeq 0$ è giustificabile.

Esempio 2.2.8. Si consideri la matrice A dell'Esempio 2.2.7. Applicando il MEG senza pivoting, si ottiene $\rho_n = \frac{|1 - 1/\epsilon|}{1} \simeq \frac{1}{\epsilon}$, da cui, usando il Teorema 2.2.1, la perturbazione $\|\delta A\|_\infty \leq 16 C \frac{\varepsilon_M}{\epsilon}$ può risultare non trascurabile, ovvero “grande”, quando ϵ è “piccolo” o comparabile al valore dell'epsilon macchina ε_M .

Al contrario, quando viene applicato il MEG con pivoting per riga, si ottiene che $\rho_n = 1$, da cui si deduce che la perturbazione $\|\delta A\|_\infty \leq 16 C \varepsilon_M$ è del tutto trascurabile.

2.2.6 Sistemi sovradeterminati

Definizione 2.2.22. Si consideri il sistema lineare $A\mathbf{x} = \mathbf{b}$ con $A \in \mathbb{R}^{m \times n}$, $\mathbf{x} \in \mathbb{R}^n$ e $\mathbf{b} \in \mathbb{R}^m$, dove $m, n \geq 1$. Se $m > n$ il sistema si dice sovradeterminato. Se $m < n$ il sistema si dice sottodeterminato.

Osservazione 2.2.18. In generale, un sistema sovradeterminato (con $m > n$) non ammette soluzione $\mathbf{x} \in \mathbb{R}^n$ in senso classico a meno che $\mathbf{b} \in \mathbb{R}^m$ non sia un elemento del range di A , ovvero $\mathbf{b} \in \text{range}(A)$, dove $\text{range}(A) = \{\mathbf{z} \in \mathbb{R}^m : \mathbf{z} = A\mathbf{y} \text{ per ogni } \mathbf{y} \in \mathbb{R}^n\}$.

Si assuma ora che $\mathbf{b} \in \mathbb{R}^m$ sia un vettore arbitrario, ovvero non necessariamente appartenente a $\text{range}(A)$.

Definizione 2.2.23. Sia $A\mathbf{x} = \mathbf{b}$ con $A \in \mathbb{R}^{m \times n}$, $\mathbf{x} \in \mathbb{R}^n$ e $\mathbf{b} \in \mathbb{R}^m$, per $m > n$. Si dice che $\mathbf{x}^* \in \mathbb{R}^n$ è soluzione di $A\mathbf{x} = \mathbf{b}$ nel senso dei minimi quadrati se \mathbf{x}^* minimizza la norma euclidea del residuo, ovvero per $\Phi(\mathbf{y}) = \|A\mathbf{y} - \mathbf{b}\|^2$, si ha $\Phi(\mathbf{x}^*) \leq \Phi(\mathbf{y})$ per ogni $\mathbf{y} \in \mathbb{R}^n$. Se esiste, la soluzione \mathbf{x}^* nel senso dei minimi quadrati è soluzione del sistema delle equazioni normali:

$$A^T A \mathbf{x}^* = A^T \mathbf{b}. \quad (2.14)$$

I seguenti risultati determinano le condizioni per cui il sistema (2.14) ammette un'unica soluzione \mathbf{x}^* nel senso dei minimi quadrati.

Proposizione 2.2.6. La matrice $(A^T A) \in \mathbb{R}^{n \times n}$ è non-singolare se $A \in \mathbb{R}^{m \times n}$ ha rango pieno, ovvero $\text{rank}(A) = \min\{m, n\}$, dove $\text{rank}(A)$ è il massimo numero di vettori riga o colonna linearmente indipendenti.

Corollario 2.2.2. Se la matrice $A \in \mathbb{R}^{m \times n}$ ha rango pieno, allora la matrice $(A^T A) \in \mathbb{R}^{n \times n}$ è simmetrica e definita positiva.

Corollario 2.2.3. Se la matrice $A \in \mathbb{R}^{m \times n}$ ha rango pieno, allora il sistema $A\mathbf{x} = \mathbf{b}$ ammette un'unica soluzione $\mathbf{x}^* \in \mathbb{R}^n$ nel senso dei minimi quadrati.

Si consideri il caso di sistemi sovradeterminati $m \geq n$. Ammettendo che \mathbf{x}^* esista, allora la soluzione nel senso dei minimi quadrati può essere determinata risolvendo il sistema lineare (2.14). Tuttavia, dal punto di vista computazionale, ciò comporta l'assemblaggio di una matrice $A^T A$ al costo di $O(2n^3)$ operazioni. Inoltre, dal punto di vista numerico, l'assemblaggio al calcolatore della matrice $A^T A$ potrebbe comportare una perdita di accuratezza sulla matrice finale, che potrebbe risultare infatti non più definita positiva; infatti, se A ha rango pieno, allora dal Corollario 2.2.2 la matrice $A^T A$ deve essere simmetrica e definita positiva, anche se tale proprietà non è garantita eseguendo operazioni in aritmetica floating-point al calcolatore. Pertanto, per risolvere sistemi sovradeterminati, si prediligono alternativamente i metodi basati sulla fattorizzazione QR di A o della decomposizione a valori singolari (SVD) di A .

Fattorizzazione QR

Proposizione 2.2.7. Sia $A \in \mathbb{R}^{m \times n}$, con $m \geq n$, una matrice a rango pieno. Allora, esiste un'unica fattorizzazione QR di A , ovvero

$$A = Q R,$$

dove $Q \in \mathbb{R}^{m \times m}$ è una matrice quadrata ortogonale, ovvero per cui $Q^T Q = I$ e $R \in \mathbb{R}^{m \times n}$ è una matrice rettangolare i cui elementi sotto la diagonale principale sono nulli.

In molti casi di interesse pratico, risulta più conveniente utilizzare la versione “*ridotta*” della *fattorizzazione QR*.

Proposizione 2.2.8. *Sia $A \in \mathbb{R}^{m \times n}$, con $m \geq n$, una matrice a rango pieno. Allora, esiste un'unica fattorizzazione QR ridotta di A , ovvero*

$$A = \tilde{Q} \tilde{R},$$

dove, in riferimento alle matrici Q e R di Proposizione 2.2.7:

- $\tilde{Q} \in \mathbb{R}^{m \times n}$ è la sottomatrice rettangolare di Q , ovvero $\tilde{Q} = Q(1 : m, 1 : n)$, le cui colonne sono vettori ortonormali;
- $\tilde{R} \in \mathbb{R}^{n \times n}$ è la sottomatrice quadrata di R tale per cui $\tilde{R} = R(1 : n, 1 : n)$ ed è triangolare superiore.

Corollario 2.2.4. Sotto le stesse ipotesi di Proposizione 2.2.8, la matrice \tilde{R} è la matrice triangolare superiore ottenuta dalla fattorizzazione di Cholesky della matrice simmetrica e definita positiva $(A^T A)$, ovvero $(A^T A) = \tilde{R}^T \tilde{R}$.

Il calcolo della *fattorizzazione QR ridotta* di una matrice $A \in \mathbb{R}^{m \times n}$ si effettua in due passi:

1. determinazione della matrice \tilde{Q} tramite l'*algoritmo di ortogonalizzazione di Gram–Schmidt* 2.4;
2. calcolo di \tilde{R} come $\tilde{R} = \tilde{Q}^T A$; infatti, essendo \tilde{Q} ortogonale, ovvero $\tilde{Q}^T \tilde{Q} = I$, si verifica che $\tilde{Q}^T A = \tilde{Q}^T (\tilde{Q} \tilde{R})$.

Il costo computazionale della fattorizzazione QR ridotta è $O(mn^2)$.

Algorithm 2.4: Ortogonalizzazione di Gram–Schmidt

riscrivere $A \in \mathbb{R}^{m \times n}$ come $A = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n]$, dove $\mathbf{a}_i \in \mathbb{R}^m$ per ogni $i = 1, 2, \dots, n$;

$$\tilde{\mathbf{q}}_1 = \frac{\mathbf{a}_1}{\|\mathbf{a}_1\|};$$

for $k = 1, \dots, n - 1$ **do**

$$\mathbf{q}_{k+1} = \mathbf{a}_{k+1} - \sum_{j=1}^k (\tilde{\mathbf{q}}_j \cdot \mathbf{a}_{k+1}) \tilde{\mathbf{q}}_j;$$

$$\tilde{\mathbf{q}}_{k+1} = \frac{\mathbf{q}_{k+1}}{\|\mathbf{q}_{k+1}\|};$$

end

assemblare \tilde{Q} come $\tilde{Q} = [\tilde{\mathbf{q}}_1, \tilde{\mathbf{q}}_2, \dots, \tilde{\mathbf{q}}_n]$;

Essendo $A^T A \mathbf{x}^* = A^T \mathbf{b}$, $(A^T A) = \tilde{R}^T \tilde{R}$ e $A = \tilde{Q} \tilde{R}$, si ha $\tilde{R}^T \tilde{R} \mathbf{x}^* = \tilde{R}^T \tilde{Q}^T \mathbf{b}$, da cui si ottiene che la soluzione nel senso dei minimi quadrati è soluzione del sistema triangolare superiore:

$$\tilde{R} \mathbf{x}^* = \tilde{Q}^T \mathbf{b},$$

ovvero che $\mathbf{x}^* = \tilde{R}^{-1} \tilde{Q}^T \mathbf{b}$.

Definizione 2.2.24. Il metodo della fattorizzazione QR per la soluzione del sistema lineare sovradeterminato ($m \geq n$) $A \mathbf{x} = \mathbf{b}$ a rango pieno, consiste nel:

1. determinare la fattorizzazione QR ridotta della matrice A ($A = \tilde{Q} \tilde{R}$);
2. risolvere il sistema triangolare superiore $\tilde{R} \mathbf{x}^* = \tilde{Q}^T \mathbf{b}$ con l'algoritmo delle sostituzioni all'indietro (2.4).

2.2.7 Il comando \ di Matlab

Il comando Matlab \ rappresenta un'implementazione estremamente efficiente di metodi diretti per la soluzione del sistema lineare $A \mathbf{x} = \mathbf{b}$:

```
>> x = A \ b;
```

Il comando Matlab \ basa la scelta del metodo diretto da impiegarsi per risolvere il sistema sulle proprietà della matrice A . Se $A \in \mathbb{R}^{n \times n}$ è sparsa e a banda, allora Matlab utilizza, a seconda del tipo di banda della matrice, un metodo basato su generalizzazioni dell'algoritmo di Thomas. Se A è triangolare inferiore o superiore allora vengono utilizzati rispettivamente i metodi delle sostituzioni in avanti e all'indietro. Se A è simmetrica e definita positiva, Matlab utilizza il metodo della fattorizzazione di Cholesky. Infine, per una matrice $A \in \mathbb{R}^{n \times n}$ generica, viene applicato il metodo della fattorizzazione LU con pivoting totale.

2.3 Metodi Iterativi

Consideriamo ora *metodi iterativi* per la soluzione del sistema lineare $A \mathbf{x} = \mathbf{b}$. L'obiettivo è quello di risolvere $A \mathbf{x} = \mathbf{b}$ in principio in un numero infinito di passi, come ovvero di ottenere la soluzione esatta come $\mathbf{x} = \lim_{k \rightarrow +\infty} \mathbf{x}^{(k)}$, dove le iterate $\{\mathbf{x}^{(k)}\}_{k=0}^{+\infty}$ rappresentano una *sequenza* di vettori soluzione, mentre $\mathbf{x}^{(0)}$ è il *vettore iniziale* (o soluzione iniziale).

2.3.1 Lo schema generale

Un metodo iterativo per la soluzione di $A \mathbf{x} = \mathbf{b}$, con $A \in \mathbb{R}^{n \times n}$ non-singolare e $\mathbf{x}, \mathbf{b} \in \mathbb{R}^n$, può essere scritto in generale nella forma:

$$\boxed{\begin{aligned} &\text{dato } \mathbf{x}^{(0)} \in \mathbb{R}^n, \\ &\mathbf{x}^{(k+1)} = B \mathbf{x}^{(k)} + \mathbf{g} \quad \text{per } k = 0, 1, \dots, \end{aligned}} \tag{2.15}$$

dove $B \in \mathbb{R}^{n \times n}$ è la *matrice di iterazione* e $\mathbf{g} \in \mathbb{R}^n$ è il *vettore di iterazione*; B e \mathbf{g} dipendono dalla matrice A , dal vettore \mathbf{b} , e dallo specifico metodo iterativo considerato. In ogni caso, il metodo iterativo deve soddisfare la *condizione di consistenza forte*, per cui, se \mathbf{x} è la soluzione di $A \mathbf{x} = \mathbf{b}$, deve valere $\mathbf{x} = B \mathbf{x} + \mathbf{g}$; pertanto, il vettore iterazione deve essere dato da $\mathbf{g} = (I - B) A^{-1} \mathbf{b}$ dal momento che $\mathbf{x} = A^{-1} \mathbf{b}$.

Definizione 2.3.1. Definiamo l'errore $\mathbf{e}^{(k)} \in \mathbb{R}^n$ associato all'iterata $\mathbf{x}^{(k)} \in \mathbb{R}^n$ del metodo iterativo (2.15) come:

$$\mathbf{e}^{(k)} := \mathbf{x} - \mathbf{x}^{(k)} \quad \text{per } k = 0, 1, \dots,$$

mentre il residuo $\mathbf{r}^{(k)} \in \mathbb{R}^n$ come:

$$\mathbf{r}^{(k)} := \mathbf{b} - A \mathbf{x}^{(k)} \quad \text{per } k = 0, 1, \dots$$

Dalla definizione di errore e di consistenza forte abbiamo $\mathbf{e}^{(k+1)} = \mathbf{x} - \mathbf{x}^{(k+1)} = (B \mathbf{x} + \mathbf{g}) - (B \mathbf{x}^{(k)} + \mathbf{g}) = B (\mathbf{x} - \mathbf{x}^{(k)}) = B \mathbf{e}^{(k)}$ per $k = 0, 1, \dots$; e quindi, per ricorsione, segue che

$$\mathbf{e}^{(k)} = B^k \mathbf{e}^{(0)} \quad \text{per } k = 0, 1, \dots \tag{2.16}$$

In generale, $\lim_{k \rightarrow +\infty} \mathbf{e}^{(k)} = \mathbf{0}$ se e solo se $\lim_{k \rightarrow +\infty} B^k = \mathbf{0}$, che è verificato se e solo se $\rho(B) < 1$.

Osservazione 2.3.1. Osserviamo che, dalla relazione (2.16), otteniamo che

$$\|\mathbf{e}^{(k)}\|_2 = \|B^k \mathbf{e}^{(0)}\|_2 \leq \|B^k\|_2 \|\mathbf{e}^{(0)}\|_2;$$

risulta pertanto, essendo $\|B^k\|_2 \leq \|B\|_2^k$ per la submoltiplicatività della norma indotta, che $\|\mathbf{e}^{(k)}\|_2 \leq \|B\|_2^k \|\mathbf{e}^{(0)}\|_2$. Nel caso in cui B sia simmetrica e definitiva positiva, siccome $\|B\|_2 = \lambda_{\max}(B) = \rho(B)$, è dunque immediato provare che $\|\mathbf{e}^{(k)}\|_2 \leq (\rho(B))^k \|\mathbf{e}^{(0)}\|_2$ per ogni $k = 0, 1, \dots$ e dunque che $\lim_{k \rightarrow +\infty} \mathbf{e}^{(k)} = \mathbf{0}$ se e solo se $\rho(B) < 1$. Tale risultato, tuttavia, vale per qualsiasi matrice di iterazione B , sebbene la dimostrazione sia leggermente più laboriosa.

Proposizione 2.3.1 (Condizione necessaria e sufficiente per la convergenza). *Il metodo iterativo (2.15) è convergente alla soluzione esatta $\mathbf{x} \in \mathbb{R}^n$ del sistema lineare $A \mathbf{x} = \mathbf{b}$ per ogni scelta del vettore iniziale $\mathbf{x}^{(0)} \in \mathbb{R}^n$ se e solo se il raggio spettrale della matrice B è strettamente minore di uno, cioè $\rho(B) < 1$. Inoltre, tanto più piccolo è $\rho(B)$, tanto più rapida sarà la convergenza.*

2.3.2 Metodi di decomposizione additiva (metodi di splitting)

I metodi di decomposizione additiva (o metodi di *splitting*) sono una famiglia di metodi iterativi per cui la matrice di iterazione B è ottenuta per mezzo di operazioni di decomposizione additiva della matrice A . A questo scopo si introduce una matrice non-singolare $P \in \mathbb{R}^{n \times n}$, detta *matrice di precondizionamento* (o precondizionatore). Osservando che $A = P - P + A$ e $A \mathbf{x} = \mathbf{b}$, abbiamo che

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

da cui

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

Dall'ultima uguaglianza, in virtù della condizione di consistenza forte, otteniamo rispettivamente la matrice e il vettore di iterazione:

$$B = I - P^{-1} A \quad (2.17)$$

e $\mathbf{g} = P^{-1} \mathbf{b}$. Di conseguenza, il metodo iterativo (2.15) può essere scritto come

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

da cui

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

Definizione 2.3.2. Il residuo precondizionato $\mathbf{z}^{(k)} \in \mathbb{R}^n$ è la soluzione del sistema lineare:

$$P \mathbf{z}^{(k)} = \mathbf{r}^{(k)} \quad \text{per } k = 0, 1, \dots,$$

con $P \in \mathbb{R}^{n \times n}$ la matrice di precondizionamento non-singolare.

Ne consegue che il metodo iterativo (2.15) può essere scritto come:

$$\begin{aligned} &\text{dato } \mathbf{x}^{(0)} \in \mathbb{R}^n, \\ &\text{risolvere } P \mathbf{z}^{(k)} = \mathbf{r}^{(k)} \text{ e porre } \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{z}^{(k)} \quad \text{per } k = 0, 1, \dots \end{aligned} \quad (2.18)$$

Osserviamo che $\mathbf{r}^{(k+1)} = \mathbf{b} - A \mathbf{x}^{(k+1)} = \mathbf{b} - A \mathbf{x}^{(k)} - A \mathbf{z}^{(k)} = \mathbf{r}^{(k)} - A \mathbf{z}^{(k)}$. Pertanto, dall'Eq. (2.18) determiniamo il seguente metodo iterativo precondizionato (Algoritmo 2.5).

Il metodo iterativo deve essere fermato per mezzo di opportuni *criteri di arresto*. In particolare, possiamo considerare criteri basati sul *residuo* e il *residuo relativo* (normalizzato) per cui le iterazioni sono fermate al primo $k \geq 0$ per cui $\|\mathbf{r}^{(k)}\| < tol$ oppure $\frac{\|\mathbf{r}^{(k)}\|}{\|\mathbf{r}^{(0)}\|} < tol$, per una tolleranza specificata tol ; inoltre, il numero di iterazioni massime dovrebbe essere limitato ad un intero k_{\max} “sufficientemente” grande.

Algorithm 2.5: Metodo iterativo precondizionato

```

dato  $\mathbf{x}^{(0)} \in \mathbb{R}^n$ , porre  $\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}$ ;
for  $k = 0, 1, \dots$ , fino a che un criterio d'arresto è soddisfatto do
    | risolvere il sistema lineare  $P\mathbf{z}^{(k)} = \mathbf{r}^{(k)}$ ;
    | porre  $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{z}^{(k)}$ ;
    | porre  $\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - A\mathbf{z}^{(k)}$ ;
end

```

Osservazione 2.3.2. Ad ogni passo del metodo iterativo (2.18) è necessario risolvere il sistema lineare $P\mathbf{z}^{(k)} = \mathbf{r}^{(k)}$. Perciò, la scelta della matrice di precondizionamento P deve garantire che $P\mathbf{z}^{(k)} = \mathbf{r}^{(k)}$ possa essere risolto in modo computazionalmente efficiente per mezzo di un metodo diretto (in “poche” operazioni), ovvero dovrebbe costituire un sistema lineare “semplice”.

D'altra parte, la scelta di P deve garantire che il metodo iterativo sia convergente, ovvero che la matrice di iterazione associata $B = I - P^{-1}A$ abbia raggio spettrale strettamente inferiore all'unità ($\rho(B) < 1$); in aggiunta, è desiderabile che $\rho(B) \ll 1$ per assicurare una convergenza rapida alla soluzione esatta \mathbf{x} .

Si ponga $P = I$ per cui il sistema lineare è il più “semplice” possibile da risolvere dato che $\mathbf{z}^{(k)} = \mathbf{r}^{(k)}$; in questo caso, $B = I - A$ e, dato che P non “possiede” nessuna informazione della matrice A , si ha molto spesso che $\rho(B) > 1$ oppure, se il metodo converge, questo succede in un numero elevato di iterazioni.

D'altra parte, per $P = A$, il sistema lineare $A\mathbf{z}^{(k)} = \mathbf{r}^{(k)}$ possiede la stessa complessità dell'originale $A\mathbf{x} = \mathbf{b}$, ma $B = 0$ e $\rho(B) = 0$ per cui si ha convergenza in una sola iterazione per ogni $\mathbf{x}^{(0)} \in \mathbb{R}^n$.

In tale contesto, risulta evidente come la scelta della matrice di precondizionamento P sia un compromesso tra la “semplicità” del sistema lineare $P\mathbf{z}^{(k)} = \mathbf{r}^{(k)}$ da risolvere ad ogni iterazione e la necessità di assicurare la (rapida) convergenza del metodo iterativo (cioè $\rho(B) < 1$ ed eventualmente $\rho(B) \ll 1$).

2.3.3 Metodi di Jacobi e Gauss–Seidel

Consideriamo i metodi iterativi di Jacobi e Gauss–Seidel per la soluzione di $A\mathbf{x} = \mathbf{b}$; questi metodi rientrano nella categoria dei metodi di decomposizione additiva; vedi Eq. (2.18).

Metodo di Jacobi

Il metodo di Jacobi può essere utilizzato per una matrice non–singolare $A \in \mathbb{R}^{n \times n}$ con elementi diagonali non nulli, cioè quando $a_{ii} \neq 0$ per ogni $i = 1, \dots, n$. Il metodo di Jacobi consiste nel scegliere come precondizionatore P nello schema generale (2.18) la matrice diagonale estratta da A .

Precisamente, indicando con P_J la matrice di precondizionamento P per il metodo di Jacobi, abbiamo:

$$P_J = D,$$

dove $D \in \mathbb{R}^{n \times n}$ è la matrice diagonale estratta da A , cioè D ha come unici elementi non nulli $(D)_{ii} = a_{ii}$ per ogni $i = 1, \dots, n$. Osserviamo che $\det(P_J) \neq 0$, in accordo con l'ipotesi che gli elementi diagonali di A siano non nulli. Il sistema lineare $P_J\mathbf{z}^{(k)} = \mathbf{r}^{(k)}$ di Eq. (2.18) è “semplice” da risolvere tramite un metodo diretto, dal momento che $P_J = D$ è una matrice diagonale.

La matrice di iterazione associata al metodo di Jacobi, ossia B_J , è data da:

$$B_J = I - P_J^{-1}A = I - D^{-1}A,$$

da cui segue che la convergenza del metodo di Jacobi a \mathbf{x} per ogni scelta del vettore iniziale $\mathbf{x}^{(0)}$ dipende dal raggio spettrale $\rho(B_J)$ in accordo con la Proposizione 2.3.1.

In forma matriciale, l'iterata k –esima del metodo di Jacobi può essere scritta come:

$$D\mathbf{x}^{(k+1)} = \mathbf{b} - (A - D)\mathbf{x}^{(k)} \quad \text{per } k = 0, 1, \dots,$$

dato $\mathbf{x}^{(0)} \in \mathbb{R}^n$. Questo porta all'*algoritmo di Jacobi*:

$$\boxed{\begin{aligned} &\text{dato } \mathbf{x}^{(0)} \in \mathbb{R}^n, \\ &x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1, j \neq i}^n a_{ij} x_j^{(k)} \right) \quad i = 1, \dots, n, \quad \text{per } k = 0, 1, \dots \end{aligned}}$$

L'algoritmo di Jacobi è chiamato anche algoritmo di *aggiornamento simultaneo*.

Metodo di Gauss–Seidel

Il *metodo di Gauss–Seidel* può essere utilizzato per una matrice non–singolare $A \in \mathbb{R}^{n \times n}$ con *elementi diagonali non nulli*, cioè quando $a_{ii} \neq 0$ per ogni $i = 1, \dots, n$. Il metodo considera come matrice di precondizionamento P dell'Eq. (2.18) la matrice *triangolare inferiore* estratta da A .

Per convenzione, indichiamo con D la matrice diagonale estratta da A e con $E \in \mathbb{R}^{n \times n}$ la matrice triangolare inferiore (escludendo la diagonale principale) con unici elementi non nulli $(E)_{ij} = -a_{ij}$ per $i = 2, \dots, n$ e $j = 1, \dots, i-1$; infine $F \in \mathbb{R}^{n \times n}$ è la matrice triangolare superiore (escludendo la diagonale principale), con unici elementi non nulli $(F)_{ij} = -a_{ij}$ per $i = 1, \dots, n-1$ e $j = i+1, \dots, n$. In questo modo, abbiamo $A = D - E - F$.

Indicando con P_{GS} la matrice di precondizionamento P per il metodo di Gauss–Seidel, abbiamo:

$$P_{GS} = D - E,$$

dove $\det(P_{GS}) \neq 0$ in accordo con l'ipotesi che gli elementi diagonali di A siano non nulli. Il sistema lineare $P_{GS} \mathbf{z}^{(k)} = \mathbf{r}^{(k)}$ di Eq. (2.18) è “semplice” da risolvere attraverso metodi diretti dal momento che $P_{GS} = D - E$ è una matrice triangolare inferiore. Pertanto, la matrice di iterazione associata al metodo di Gauss–Seidel, ossia B_{GS} , è:

$$B_{GS} = I - P_{GS}^{-1} A = I - (D - E)^{-1} A;$$

le proprietà di convergenza del metodo di Gauss–Seidel dipendono dal raggio spettrale $\rho(B_{GS})$, in accordo con la Proposizione 2.3.1.

In forma matriciale, la k –esima iterata del metodo di Gauss–Seidel è data da:

$$(D - E) \mathbf{x}^{(k+1)} = \mathbf{b} - (A + E - D) \mathbf{x}^{(k)} \quad \text{per } k = 0, 1, \dots,$$

dato $\mathbf{x}^{(0)} \in \mathbb{R}^n$, da cui deduciamo l'*algoritmo di Gauss–Seidel*:

$$\boxed{\begin{aligned} &\text{dato } \mathbf{x}^{(0)} \in \mathbb{R}^n, \\ &x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij} x_j^{(k)} \right) \quad i = 1, \dots, n, \quad \text{per } k = 0, 1, \dots \end{aligned}}$$

L'algoritmo di Gauss–Seidel è anche detto algoritmo dell'*aggiornamento sequenziale*.

Condizioni sufficienti per la convergenza dei metodi di Jacobi e Gauss–Seidel

La condizione *necessaria e sufficiente* per la convergenza dei metodi di Jacobi e Gauss–Seidel (per ogni $\mathbf{x}^{(0)} \in \mathbb{R}^n$) è che il raggio spettrale delle corrispondenti matrici di iterazioni sia strettamente inferiore ad uno; si veda la Proposizione 2.3.1. Tuttavia, in alcuni casi, è possibile determinare la convergenza dei metodi semplicemente osservando la matrice A del sistema lineare $A \mathbf{x} = \mathbf{b}$ invece di assemblare la matrice di iterazione B e calcolare $\rho(B)$. Di seguito riportiamo alcune condizioni *sufficienti*.

Proposizione 2.3.2. *Se A è non–singolare e a dominanza diagonale stretta per riga, allora i metodi di Jacobi e Gauss–Seidel convergono a \mathbf{x} per ogni iterata iniziale $\mathbf{x}^{(0)} \in \mathbb{R}^n$.*

Proposizione 2.3.3. Se A è simmetrica e definita positiva, allora il metodo di Gauss–Seidel converge a \mathbf{x} per ogni iterata iniziale $\mathbf{x}^{(0)} \in \mathbb{R}^n$.

Proposizione 2.3.4. Se A è non–singolare e tridiagonale con tutti gli elementi diagonali non–nulli, allora i metodi di Jacobi e Gauss–Seidel sono entrambi divergenti o convergenti a \mathbf{x} . Se convergenti, il metodo di Gauss–Seidel converge più velocemente del metodo di Jacobi dato che $\rho(B_{GS}) = (\rho(B_J))^2$.

Osservazione 2.3.3. Le condizioni precedenti sono solo sufficienti; ovvero, se queste non sono soddisfatte, allora è necessario verificare le condizioni necessarie e sufficienti di Proposizione 2.3.1 per determinare la convergenza del metodo iterativo.

Esempio 2.3.1. Si consideri la matrice $A = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}$, che è non–singolare e a dominanza diagonale stretta per righe. Dato che le condizioni sufficienti di Proposizione 2.3.2 sono soddisfatte, allora entrambi i metodi di Jacobi e Gauss–Seidel sono convergenti per ogni $\mathbf{x}^{(0)} \in \mathbb{R}^2$ alla soluzione $\mathbf{x} \in \mathbb{R}^2$ del sistema lineare $A\mathbf{x} = \mathbf{b}$, per qualche $\mathbf{b} \in \mathbb{R}^2$. Osserviamo che anche le ipotesi delle Proposizioni 2.3.3 e 2.3.4 sono soddisfatte. Verifichiamo il risultato per mezzo delle condizioni necessarie e sufficienti di Proposizione 2.3.1; ovvero, verifichiamo che i raggi spettrali delle matrici di iterazione associate ai metodi di Jacobi e Gauss–Seidel sono strettamente inferiori ad uno. Per il metodo di Jacobi, abbiamo $P_J = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}$ e $B_J = I - P_J^{-1} A = \begin{bmatrix} 0 & -\frac{1}{3} \\ -\frac{1}{2} & 0 \end{bmatrix}$, da cui $\rho(B_J) = \frac{1}{\sqrt{6}} < 1$. Per il metodo di Gauss–Seidel, abbiamo $P_{GS} = \begin{bmatrix} 3 & 0 \\ 1 & 2 \end{bmatrix}$ e $B_{GS} = I - P_{GS}^{-1} A = \begin{bmatrix} 0 & -\frac{1}{3} \\ 0 & \frac{1}{6} \end{bmatrix}$, da cui $\rho(B_{GS}) = \frac{1}{6} < 1$.

Esempio 2.3.2. Consideriamo la matrice non–singolare $A = \begin{bmatrix} 1 & 0 & -1 \\ 3 & 2 & 0 \\ -1 & -1 & 2 \end{bmatrix}$. In questo caso, le ipotesi delle Proposizioni 2.3.2, 2.3.3, e 2.3.4 non sono soddisfatte, per cui, per verificare la convergenza dei metodi di Jacobi e Gauss–Seidel per tutti i vettori iniziali, è necessario verificare le ipotesi della Proposizione 2.3.1. Per il metodo

di Jacobi, abbiamo $P_J = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$ e $B_J = I - P_J^{-1} A = \begin{bmatrix} 0 & 0 & 1 \\ -\frac{3}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$, da cui $\rho(B_J) = \frac{109}{100} > 1$;

deduciamo che il metodo di Jacobi non converge per ogni scelta di $\mathbf{x}^{(0)} \in \mathbb{R}^3$ alla soluzione $\mathbf{x} \in \mathbb{R}^3$ del sistema lineare $A\mathbf{x} = \mathbf{b}$, per qualche $\mathbf{b} \in \mathbb{R}^3$. D'altra parte, per il metodo di Gauss–Seidel, abbiamo $P_{GS} = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 2 & 0 \\ -1 & -1 & 2 \end{bmatrix}$ e

$B_{GS} = I - P_{GS}^{-1} A = \begin{bmatrix} 0 & 0 & \frac{1}{3} \\ 0 & 0 & -\frac{1}{2} \\ 0 & 0 & -\frac{1}{4} \end{bmatrix}$, da cui $\rho(B_{GS}) = \frac{1}{4} < 1$; pertanto, il metodo di Gauss–Seidel converge a \mathbf{x} per ogni scelta di $\mathbf{x}^{(0)} \in \mathbb{R}^3$.

2.3.4 Metodi di Richardson precondizionati

Consideriamo una successione di parametri reali $\{\alpha_k\}_{k=0}^{+\infty} \in \mathbb{R}$; il *metodo di Richardson precondizionato* rappresenta una generalizzazione del metodo iterativo (2.18), ed è dato da:

dato $\mathbf{x}^{(0)} \in \mathbb{R}^n$,

risolvere $P\mathbf{z}^{(k)} = \mathbf{r}^{(k)}$ e porre $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{z}^{(k)}$ per $k = 0, 1, \dots$,

(2.19)

per una matrice di precondizionamento non-singolare $P \in \mathbb{R}^{n \times n}$. Se $\alpha_k = \alpha \in \mathbb{R}$ per ogni $k = 0, 1, \dots$, il metodo iterativo (2.19) è detto metodo di Richardson *stazionario*, mentre se α_k non è costante nell'indice di iterazione $k = 0, 1, \dots$, è detto metodo di Richardson *dinamico*. Osserviamo che, se $\alpha_k = \alpha = 1$, otteniamo il metodo di Eq. (2.18). Inoltre, $\mathbf{r}^{(k+1)} = \mathbf{b} - A\mathbf{x}^{(k+1)} = \mathbf{b} - A\mathbf{x}^{(k)} - \alpha_k A\mathbf{z}^{(k)} = \mathbf{r}^{(k)} - \alpha_k A\mathbf{z}^{(k)}$. Quindi, da Eq. (2.18) otteniamo il seguente metodo di Richardson precondizionato.

Algorithm 2.6: Metodo di Richardson precondizionato dinamico

```
Dato  $\mathbf{x}^{(0)} \in \mathbb{R}^n$ , porre  $\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}$ ;
for  $k = 0, 1, \dots$ , finché un criterio d'arresto non è soddisfatto do
    risolvere il sistema lineare  $P\mathbf{z}^{(k)} = \mathbf{r}^{(k)}$ ;
    scegliere  $\alpha_k$ ;
    porre  $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{z}^{(k)}$ ;
    porre  $\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k A\mathbf{z}^{(k)}$ ;
end
```

Osservazione 2.3.4. Per un metodo di Richardson precondizionato dinamico, abbiamo $\mathbf{x}^{(k+1)} = B_k \mathbf{x}^{(k)} + \mathbf{g}_k$ per $k = 0, 1, \dots$, dove la matrice di iterazione dinamica è data da

$$B_k = I - \alpha_k P^{-1} A \quad \text{for } k = 0, 1, \dots$$

e il vettore di iterazione $\mathbf{g}_k = \alpha_k P^{-1} \mathbf{b}$ varia con l'indice di iterazione. Pertanto, le proprietà di convergenza del metodo di Richardson precondizionato dinamico cambiano con l'indice di iterazione k , dal momento che anche il paramentro α_k cambia con k .

Osservazione 2.3.5. Per un metodo di Richardson precondizionato stazionario, la matrice di iterazione è data da

$$B_\alpha = I - \alpha P^{-1} A;$$

le proprietà di convergenza dipendono dal raggio spettrale B_α , cioè $\rho(\alpha) = \rho(B_\alpha)$.

Consideriamo ora le condizioni di convergenza del metodo di Richardson alla soluzione $\mathbf{x} \in \mathbb{R}^n$ di un generico sistema lineare $A\mathbf{x} = \mathbf{b}$ per ogni scelta del vettore iniziale $\mathbf{x}^{(0)} \in \mathbb{R}^n$.

Proposizione 2.3.5. Se le matrici A e $P \in \mathbb{R}^{n \times n}$ sono non-singolari, allora il metodo di Richardson stazionario converge a $\mathbf{x} \in \mathbb{R}^n$ per ogni scelta di $\mathbf{x}^{(0)} \in \mathbb{R}^n$ se e solo se

$$\alpha |\lambda_i(P^{-1}A)|^2 < 2 \operatorname{Re}\{\lambda_i(P^{-1}A)\} \quad \text{per ogni } i = 1, \dots, n,$$

con $\alpha \neq 0$, dove $\{\lambda_i(P^{-1}A)\}_{i=1}^n$ sono gli autovalori di $P^{-1}A$.

Dimostrazione. Condizione necessaria e sufficiente per la convergenza del metodo di Richardson stazionario è che $\rho(B_\alpha) < 1$, essendo $B_\alpha = I - \alpha P^{-1} A$ la matrice di iterazione del metodo. Tutti gli autovalori di $I - \alpha P^{-1} A$ devono dunque essere minori di 1 in modulo. Indicando con $\lambda_i = \lambda_i(P^{-1}A)$ gli autovalori della matrice precondizionata $P^{-1}A$, si ha che² la matrice B_α ha per autovalori $1 - \alpha\lambda_i$, $i = 1, \dots, n$. Deve risultare dunque che

$$|1 - \alpha\lambda_i| < 1 \quad \text{per ogni } i = 1, \dots, n.$$

Poiché a priori tali autovalori devono essere complessi, risulta

$$(\operatorname{Re}\{1 - \alpha\lambda_i\})^2 + (\operatorname{Im}\{1 - \alpha\lambda_i\})^2 < 1^2,$$

²Infatti $(I - \alpha P^{-1} A)\mathbf{w}_i = \mathbf{w}_i - \alpha\lambda_i \mathbf{w}_i = (1 - \alpha\lambda_i)\mathbf{w}_i$, dove $(\lambda_i, \mathbf{w}_i)$, $i = 1, \dots, n$ è una coppia autovalore-autovettore di $P^{-1}A$. In particolare, $I - \alpha P^{-1} A$ ha gli stessi autovettori di $P^{-1}A$.

da cui

$$(1 - \alpha \operatorname{Re} \{\lambda_i\})^2 + \alpha^2 (\operatorname{Im} \{\lambda_i\})^2 < 1$$

ovvero

$$2\alpha \operatorname{Re} \{\lambda_i\} > \alpha^2 ((\operatorname{Re} \{\lambda_i\})^2 + (\operatorname{Im} \{\lambda_i\})^2) = \alpha^2 |\lambda_i|^2,$$

da cui, dividendo tutto per $\alpha^2 |\lambda_i|^2 > 0$, si ottiene la tesi, ovvero

$$\frac{2\operatorname{Re} \{\lambda_i\}}{\alpha |\lambda_i|^2} > 1 \quad \text{per ogni } i = 1, \dots, n.$$

□

Corollario 2.3.1. Se le matrici A e $P \in \mathbb{R}^{n \times n}$ sono non-singolari e sono dotate di autovalori $\{\lambda_i(P^{-1}A)\}_{i=1}^n$ tutti a valori reali (cioè $\lambda_i(P^{-1}A) \equiv \operatorname{Re} \{\lambda_i(P^{-1}A)\}$ per ogni $i = 1, \dots, n$), allora il metodo di Richardson stazionario converge a $\mathbf{x} \in \mathbb{R}^n$ per ogni scelta di $\mathbf{x}^{(0)} \in \mathbb{R}^n$ se e solo se

$$0 < \alpha \lambda_i(P^{-1}A) < 2 \quad \text{per ogni } i = 1, \dots, n.$$

□

Dimostrazione. Il risultato è diretta conseguenza della Proposizione 2.3.5. Infatti, dovendo essere $|1 - \alpha \lambda_i| < 1$ per ogni $i = 1, \dots, n$, con $\lambda_i = \lambda_i(P^{-1}A) \in \mathbb{R}$, deve risultare, più semplicemente, che

$$-1 < 1 - \alpha \lambda_i < 1 \quad \text{per ogni } i = 1, \dots, n$$

da cui discende che, contemporaneamente, $\alpha > 0$ e $\alpha < 2/\lambda_i$ per ogni $i = 1, \dots, n$, da cui la tesi. □

□

Definizione 2.3.3. La norma energia di un vettore $\mathbf{v} \in \mathbb{R}^n$ rispetto a una matrice simmetrica e definita positiva $A \in \mathbb{R}^{n \times n}$ è definito come:

$$\|\mathbf{v}\|_A = \sqrt{\mathbf{v}^T A \mathbf{v}}.$$

□

Proposizione 2.3.6. Se le matrici A e $P \in \mathbb{R}^{n \times n}$ sono simmetriche e definite positive, allora il metodo di Richardson stazionario converge a $\mathbf{x} \in \mathbb{R}^n$ per ogni scelta di $\mathbf{x}^{(0)} \in \mathbb{R}^n$ se e solo se

$$0 < \alpha < \frac{2}{\lambda_{\max}(P^{-1}A)},$$

dove $\lambda_{\max}(P^{-1}A)$ è il massimo autovalore di $P^{-1}A$. Inoltre, il raggio spettrale della matrice di iterazione B_α è minimo per $\alpha = \alpha_{opt}$, dove

$$\alpha_{opt} := \frac{2}{\lambda_{\min}(P^{-1}A) + \lambda_{\max}(P^{-1}A)},$$

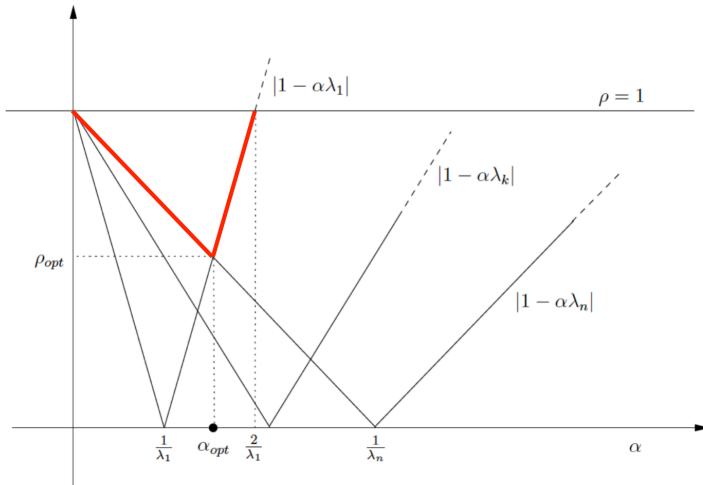
dove $\lambda_{\min}(P^{-1}A)$ è il minimo autovalore di $P^{-1}A$; in questo caso (per $\alpha = \alpha_{opt}$), abbiamo:

$$\|\mathbf{e}^{(k)}\|_A \leq d^k \|\mathbf{e}^{(0)}\|_A \quad \text{per } k = 0, 1, \dots, \tag{2.20}$$

con $d := \frac{K(P^{-1}A) - 1}{K(P^{-1}A) + 1}$, dove $K(P^{-1}A) = \frac{\lambda_{\max}(P^{-1}A)}{\lambda_{\min}(P^{-1}A)}$ è il numero di condizionamento spettrale di $P^{-1}A$.

□

Dimostrazione. Dimostriamo solo la convergenza del metodo e la validità dell'espressione di α_{opt} . Si tratta di verificare la condizione necessaria e sufficiente di convergenza per un metodo iterativo, ovvero che il raggio spettrale della matrice di iterazione $B_\alpha = I - \alpha P^{-1}A$ sia minore di 1.

Figura 2.1: Raggio spettrale di B_α in funzione di α

Osserviamo innanzitutto che, indicati con $\lambda_i = \lambda_i(P^{-1}A)$ gli autovalori³ di $P^{-1}A$, gli autovalori di B_α sono tali che

$$\lambda_i(B_\alpha) = 1 - \alpha\lambda_i.$$

Deve dunque risultare che $|1 - \alpha\lambda_i| < 1$ per ogni $i = 1, \dots, n$. Per convenzione, ordiniamo gli autovalori di $P^{-1}A$ in senso decrescente, con $\lambda_{max}(P^{-1}A) = \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n = \lambda_{min}(P^{-1}A) > 0$.

Il metodo risulta dunque convergente se e solo se $-1 < 1 - \alpha\lambda_i < 1$ per ogni $i = 1, \dots, n$. Poiché $\alpha > 0$, ciò equivale a richiedere che $-1 < 1 - \alpha\lambda_{max}$, da cui la condizione necessaria e sufficiente per la convergenza risulta $0 < \alpha < 2/\lambda_{max}$. Si può giungere alla medesima conclusione per via grafica, osservando che il raggio spettrale $\rho(B_\alpha)$ in funzione di α risulta minore di 1 a patto che $\alpha\lambda_1 - 1 < 1$, ovvero $\alpha < 2/\lambda_1 = 1/\lambda_{max}$, avendo considerato il ramo crescente della spezzata che costituisce il grafico di $\rho(B_\alpha)$ (evidenziato in rosso).

Inoltre, il raggio spettrale $\rho(B_\alpha)$ è minimo quando $1 - \alpha\lambda_n = \alpha\lambda_1 - 1$, ovvero per $\alpha_{opt} = 2/(\lambda_n + \lambda_1) = 2/(\lambda_{min}(P^{-1}A) + \lambda_{max}(P^{-1}A))$. Sostituendo il valore di α_{opt} nell'espressione del raggio spettrale, ne possiamo infine determinare il valore:

$$\rho_{opt} = \rho(B_{\alpha_{opt}}) = 1 - \alpha_{opt}\lambda_n = 1 - \lambda_{min} \frac{2}{\lambda_{min} + \lambda_{max}} = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{min} + \lambda_{max}} = \frac{\frac{\lambda_{max}}{\lambda_{min}} - 1}{\frac{\lambda_{max}}{\lambda_{min}} + 1},$$

da cui l'espressione di d in Eq. (2.20) ricordando la definizione di numero di condizionamento spettrale. \square

Osservazione 2.3.6. *Sotto le ipotesi della Proposizione 2.3.6, è disponibile una scelta ottimale per il parametro per un metodo di Richardson stazionario. D'altra parte, il risultato (2.20) indica anche che più la matrice di precondizionamento P è “vicina” alla matrice A , più il numero di condizionamento spettrale della matrice $P^{-1}A$ è vicino a uno, e più la convergenza del metodo è rapida; comunque, in questo caso, il sistema lineare $P\mathbf{z}^{(k)} = \mathbf{r}^{(k)}$ potrebbe essere relativamente complesso da risolvere. In particolare, per $P = A$, si ha $\alpha_{opt} = 1$ e $d = 0$, per cui la convergenza avviene in una sola iterazione. Al contrario, se $P = I$, si ha $\alpha_{opt} = \frac{2}{\lambda_{min}(A) + \lambda_{max}(A)}$ e $d = \frac{K(A) - 1}{K(A) + 1}$; in questo caso la convergenza del metodo iterativo può essere lenta se $K(A) \gg 1$, dal momento che $d \lesssim 1$. In generale, più $K(P^{-1}A)$ è vicino a uno, più la convergenza del metodo è rapida.*

³In base a un risultato teorico che non dimostriamo, se A e P sono simmetriche e definite positive, non è detto che anche $P^{-1}A$ lo sia; tuttavia, i suoi autovalori sono tutti reali e strettamente positivi.

Esempio 2.3.3. Consideriamo $A = \begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}$ e la matrice di precondizionamento $P = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$, che sono entrambe simmetriche e definite positive. Pertanto, per studiare le proprietà di convergenza del metodo di Richardson stazionario, è possibile utilizzare i risultati del Corollario 2.3.6. Tali proprietà di convergenza dipendono dalla matrice $P^{-1}A = \begin{bmatrix} 1 & 1/4 \\ 1/4 & 1/2 \end{bmatrix}$ e dai suoi autovalori $\lambda_{\min} = \lambda_{\min}(P^{-1}A) = \frac{3}{4} - \frac{\sqrt{2}}{4}$ e $\lambda_{\max} = \lambda_{\max}(P^{-1}A) = \frac{3}{4} + \frac{\sqrt{2}}{4}$. In particolare, il metodo di Richardson stazionario converge alla soluzione $\mathbf{x} \in \mathbb{R}^2$ di un sistema lineare associato ad A per ogni $\mathbf{x}^{(0)} \in \mathbb{R}^2$ se e solo se $0 < \alpha < \frac{2}{\lambda_{\max}} = \frac{8}{3+\sqrt{2}}$. Inoltre, il parametro ottimale $\alpha_{opt} = \frac{2}{\lambda_{\min} + \lambda_{\max}} = \frac{4}{3}$ porta al minimo raggio spettrale fra le matrici di iterazione B_α ; in particolare, $B_{\alpha_{opt}} = I - \alpha_{opt} P^{-1}A = \begin{bmatrix} -1/3 & -1/3 \\ -1/3 & 1/3 \end{bmatrix}$ e $\rho(B_{\alpha_{opt}}) = \frac{\sqrt{2}}{3} < 1$. Da Eq. (2.20), segue che $d = \frac{K(P^{-1}A) - 1}{K(P^{-1}A) + 1} = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} = \rho(B_{\alpha_{opt}}) = \sqrt{2}/3$; cioè l'errore in norma energia A si abbatte a ciascuna iterazione di un fattore $d \leq \sqrt{2}/3$.

In generale, per un metodo di Richardson precondizionato stazionario, la determinazione del parametro α_{opt} può essere molto costosa dal punto di vista computazionale, dal momento che è legata agli autovalori di $P^{-1}A$. Per evitare il loro calcolo esplicito per la determinazione del parametro α , è possibile utilizzare in modo opportuno un metodo di Richardson precondizionato dinamico.

2.3.5 Metodo del gradiente

Consideriamo il caso di una matrice $A \in \mathbb{R}^{n \times n}$ simmetrica e definita positiva. Il *metodo del gradiente precondizionato* è un metodo di Richardson precondizionato dinamico (2.19) per cui i parametri α_k sono scelti nel seguente modo:

$$\boxed{\alpha_k = \frac{(\mathbf{z}^{(k)})^T \mathbf{r}^{(k)}}{(\mathbf{z}^{(k)})^T A \mathbf{z}^{(k)}} \quad \text{per } k = 0, 1, \dots,}$$

dove P è una matrice simmetrica e definita positiva e $\mathbf{z}^{(k)}$ è il residuo precondizionato.

Similmente, il *metodo del gradiente* metodo di Richardson dinamico denza precondizionamento e con parametri α_k scelti nel seguente modo:

$$\boxed{\alpha_k = \frac{(\mathbf{r}^{(k)})^T \mathbf{r}^{(k)}}{(\mathbf{r}^{(k)})^T A \mathbf{r}^{(k)}} \quad \text{per } k = 0, 1, \dots} \quad (2.21)$$

È possibile ottenere il metodo del gradiente a partire dal metodo del gradiente precondizionato scegliendo $P = I$; in questo caso, $\mathbf{z}^{(k)} \equiv \mathbf{r}^{(k)}$ per ogni $k = 0, 1, \dots$. Per il metodo del gradiente, il vettore residuo $\mathbf{r}^{(k)}$ rappresenta la direzione di discesa per l'errore all'iterazione $k = 0, 1, \dots$, e, se A è simmetrica e definita positiva, la scelta di α_k data da (2.21) è quella che minimizza l'errore $\|\mathbf{e}^{(k+1)}\|_A$ lungo la direzione $\mathbf{r}^{(k)}$.

Forniamo la seguente interpretazione del metodo del gradiente:

Teorema 2.3.1. Sia $A \in \mathbb{R}^{n \times n}$ simmetrica e definita positiva, allora il vettore $\mathbf{x} \in \mathbb{R}^n$ è soluzione del sistema lineare $A\mathbf{x} = \mathbf{b}$ se e solo se la funzione energia del sistema $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}$, definita come $\Phi(\mathbf{y}) = \frac{1}{2}\mathbf{y}^T A \mathbf{y} - \mathbf{y}^T \mathbf{b}$, assume valore minimo per $\mathbf{y} = \mathbf{x}$.

Dimostrazione. Mostriamo innanzitutto che la soluzione \mathbf{x} del sistema lineare minimizza Φ . Scelto un qualsiasi vettore $\mathbf{v} \in \mathbb{R}^n$ si ha, sfruttando la simmetria di A , che $\Phi(\mathbf{x} + \mathbf{v}) = \frac{1}{2}(\mathbf{x} + \mathbf{v})^T A (\mathbf{x} + \mathbf{v}) -$

$(\mathbf{x} + \mathbf{v})^T \mathbf{b} = \Phi(\mathbf{x}) + \frac{1}{2} \mathbf{v}^T A \mathbf{v} + \mathbf{v}^T (A\mathbf{x} - \mathbf{b})$. Essendo \mathbf{x} soluzione del sistema lineare, si ottiene che $\Phi(\mathbf{x} + \mathbf{v}) = \Phi(\mathbf{x}) + \frac{1}{2} \mathbf{v}^T A \mathbf{v} \geq \Phi(\mathbf{x})$ per ogni $\mathbf{v} \in \mathbb{R}^n$, ovvero che $\Phi(\mathbf{x})$ è minimo.

Mostriamo ora che se Φ è minimo in \mathbf{x} , allora \mathbf{x} risolve il sistema lineare. Essendo Φ minimo in \mathbf{x} , si ottiene che $\nabla \Phi(\mathbf{x}) = \mathbf{0}$. Pertanto, essendo $\nabla \Phi(\mathbf{y}) = -(A\mathbf{y} - \mathbf{b})$, si ha che $\nabla \Phi(\mathbf{x}) = \mathbf{0}$ implica $A\mathbf{x} - \mathbf{b} = \mathbf{0}$, ovvero che \mathbf{x} risolve il sistema lineare. \square

Osservazione 2.3.7. Verifichiamo che effettivamente $\nabla \Phi(\mathbf{y}) = -(A\mathbf{y} - \mathbf{b})$. A tale scopo, riscriviamo la funzione energia $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}$ esplicitando i prodotti scalari che vi compaiono, come segue:

$$\Phi(\mathbf{y}) = \frac{1}{2} \mathbf{y}^T A \mathbf{y} - \mathbf{y}^T \mathbf{b} = \frac{1}{2} \sum_{i,j=1}^n y_i a_{ij} y_j - \sum_{i=1}^n b_i y_i.$$

Poiché A è simmetrica,

$$\frac{\partial \Phi}{\partial y_m} = \frac{1}{2} \left(\sum_{j=1}^n a_{mj} y_j + \sum_{i=1}^n y_i a_{im} \right) - b_m = \frac{1}{2} [(A\mathbf{y})_m + (A^T \mathbf{y})_m] - b_m = (A\mathbf{y})_m - b_m$$

ovvero, raggruppando tutte le componenti, $\nabla \Phi(\mathbf{y}) = \frac{1}{2} (A^T + A)\mathbf{y} - \mathbf{b} = A\mathbf{y} - \mathbf{b}$.

Assumiamo che la generica \mathbf{y} corrisponda ora all'iterata $\mathbf{x}^{(k)}$ del metodo, con $k = 0, 1, \dots$, ovvero $\Phi(\mathbf{x}^{(k)}) = \Phi(\mathbf{x}) + \frac{1}{2} \|\mathbf{e}^{(k)}\|_A^2$ e $\nabla \Phi(\mathbf{x}^{(k)}) = -\mathbf{r}^{(k)}$. L'obiettivo consiste nel determinare $\mathbf{x}^{(k+1)}$ tale che

$$\Phi(\mathbf{x}^{(k+1)}) \leq \Phi(\mathbf{x}^{(k)}),$$

per cui si può scegliere la direzione di discesa

$$-\nabla \Phi(\mathbf{x}^{(k)}) = \mathbf{r}^{(k)},$$

ovvero il *gradiente* di Φ , da cui il nome del metodo. Scriviamo pertanto

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha_k \nabla \Phi(\mathbf{x}^{(k)}) = \mathbf{x}^{(k)} + \alpha_k \mathbf{r}^{(k)},$$

con $\alpha_k \in \mathbb{R}$ da determinarsi. Una volta determinata la direzione di discesa $\mathbf{r}^{(k)}$, il parametro α_k esprime la distanza da percorrere lungo $\mathbf{r}^{(k)}$ per trovare $\mathbf{x}^{(k+1)}$. L'intersezione di $\Phi(\mathbf{y})$ con l'iperpiano passante per $(\mathbf{x}^{(k)}, \Phi(\mathbf{x}^{(k)}))$ e ortogonale a \mathbb{R}^n determina la funzione

$$\varphi(\alpha) = \Phi(\mathbf{x}^{(k)} + \alpha \mathbf{r}^{(k)}) = \frac{1}{2} (\mathbf{x}^{(k)} + \alpha \mathbf{r}^{(k)})^T A (\mathbf{x}^{(k)} + \alpha \mathbf{r}^{(k)}) - (\mathbf{x}^{(k)} + \alpha \mathbf{r}^{(k)})^T \mathbf{b}.$$

Vogliamo determinare il passo α_k in modo che, una volta presa la direzione $\mathbf{r}^{(k)}$, si ottenga il massimo decremento della funzione Φ ; in altri termini, vogliamo che per $\alpha = \alpha_k$ la funzione $\varphi(\alpha)$ presenti un punto di minimo. Imponiamo dunque che

$$\alpha_k : \left. \frac{d\varphi(\alpha)}{d\alpha} \right|_{\alpha=\alpha_k} = 0. \quad (2.22)$$

Poiché

$$\begin{aligned} \varphi'(\alpha) &= \frac{1}{2} (\mathbf{r}^{(k)})^T A \mathbf{x}^{(k)} + \frac{1}{2} (\mathbf{x}^{(k)})^T A \mathbf{r}^{(k)} + 2 \cdot \frac{1}{2} \alpha (\mathbf{r}^{(k)})^T A \mathbf{r}^{(k)} - (\mathbf{r}^{(k)})^T \mathbf{b} \\ &= (\mathbf{r}^{(k)})^T (A\mathbf{x}^{(k)} - \mathbf{b}) + \alpha (\mathbf{r}^{(k)})^T A \mathbf{r}^{(k)} = -(\mathbf{r}^{(k)})^T \mathbf{r}^{(k)} + \alpha (\mathbf{r}^{(k)})^T A \mathbf{r}^{(k)}, \end{aligned}$$

imponendo che sia soddisfatta (2.22), troviamo che α_k soddisfa l'equazione (2.21).

Osservazione 2.3.8. Analogamente, ricordando un risultato dell'analisi relativo alla derivazione di una funzione composta di più variabili, si ha che

$$\varphi'(\alpha) = \nabla \Phi(\mathbf{x}^{(k)} + \alpha \mathbf{r}^{(k)}) \cdot \mathbf{r}^{(k)} = (A(\mathbf{x}^{(k)} + \alpha \mathbf{r}^{(k)}) - \mathbf{b}) \cdot \mathbf{r}^{(k)}.$$

Imponendo che α_k sia tale da annullare la precedente derivata, ovvero $\varphi'(\alpha_k) = 0$ si ottiene

$$(A(\mathbf{x}^{(k)} + \alpha_k \mathbf{r}^{(k)}) - \mathbf{b}) \cdot \mathbf{r}^{(k)} = 0 \quad \Leftrightarrow \quad (-\mathbf{r}^{(k)} + \alpha_k A \mathbf{r}^{(k)}) \cdot \mathbf{r}^{(k)} = 0$$

da cui la precedente espressione di α_k .

Abbiamo riportato gli algoritmi del metodo del gradiente e del gradiente precondizionato, ovvero con matrice $P \neq I$ di precondizionamento. Riportiamo inoltre un risultato di convergenza del metodo del gradiente e del metodo del gradiente precondizionato.

Algorithm 2.7: Metodo del gradiente

```

dato  $\mathbf{x}^{(0)} \in \mathbb{R}^n$ , porre  $\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}$ ;
for  $k = 0, 1, \dots$ , finché un criterio d'arresto non è soddisfatto do
    porre  $\alpha_k = \frac{(\mathbf{r}^{(k)})^T \mathbf{r}^{(k)}}{(\mathbf{r}^{(k)})^T A \mathbf{r}^{(k)}}$ ;
    porre  $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{r}^{(k)}$ ;
    porre  $\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k A \mathbf{r}^{(k)}$ ;
end
```

Algorithm 2.8: Metodo del gradiente precondizionato

```

dato  $\mathbf{x}^{(0)} \in \mathbb{R}^n$ , porre  $\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}$ ;
for  $k = 0, 1, \dots$ , finché un criterio d'arresto non è soddisfatto do
    risolvere  $P \mathbf{z}^{(k)} = \mathbf{r}^{(k)}$ ;
    porre  $\alpha_k = \frac{(\mathbf{z}^{(k)})^T \mathbf{r}^{(k)}}{(\mathbf{z}^{(k)})^T A \mathbf{z}^{(k)}}$ ;
    porre  $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{z}^{(k)}$ ;
    porre  $\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k A \mathbf{z}^{(k)}$ ;
end
```

Proposizione 2.3.7. Se le matrici A e $P \in \mathbb{R}^{n \times n}$ sono simmetriche e definite positive, il metodo del gradiente precondizionato converge alla soluzione $\mathbf{x} \in \mathbb{R}^n$ per tutte le scelte di $\mathbf{x}^{(0)} \in \mathbb{R}^n$ e

$$\|\mathbf{e}^{(k)}\|_A \leq d^k \|\mathbf{e}^{(0)}\|_A \quad \text{per } k = 0, 1, \dots, \quad (2.23)$$

dove $d := \frac{K(P^{-1}A) - 1}{K(P^{-1}A) + 1}$; $K(P^{-1}A) = \frac{\lambda_{\max}(P^{-1}A)}{\lambda_{\min}(P^{-1}A)}$ è il numero di condizionamento spettrale di $P^{-1}A$.

Il precedente risultato può essere applicato al metodo del gradiente (non precondizionato), ponendo $P = I$. Inoltre, la stima di errore (2.23) può essere utilizzata per prevedere a priori il numero di iterazioni necessarie al metodo del gradiente precondizionato per convergere alla soluzione con la tolleranza desiderata.

Osserviamo che l'interpretazione geometrica data in precedenza del metodo del gradiente per A simmetrica e definita positiva, si può estendere al metodo del gradiente *precondizionato* con una matrice $P \in \mathbb{R}^{n \times n}$ anch'essa simmetrica e definita positiva. In tal caso, si può considerare il metodo del gradiente (non precondizionato) applicandolo però al sistema lineare $P^{-1}A\mathbf{x} = P^{-1}\mathbf{b}$ e dotato della funzione energia $\Phi(\mathbf{y}) = \frac{1}{2}\mathbf{y}^T (P^{-1}A)\mathbf{y} - \mathbf{y}^T (P^{-1}\mathbf{b})$.

Consideriamo nuovamente il metodo del gradiente (non precondizionato), applicato a una matrice A simmetrica e definita positiva. Osserviamo infine come, in virtù della scelta della direzione di discesa e del parametro α_k , risulti che le direzioni di discesa siano a due a due ortogonali, ovvero:

$$\mathbf{r}^{(k)} \cdot \mathbf{r}^{(k+1)} = (\mathbf{r}^{(k)})^T \mathbf{r}^{(k+1)} = 0 \quad \text{per ogni } k = 0, 1, \dots .$$

Infatti risulta

$$(\mathbf{r}^{(k)})^T \mathbf{r}^{(k+1)} = (\mathbf{r}^{(k)})^T (\mathbf{r}^{(k)} - \alpha_k A \mathbf{r}^{(k)}) = (\mathbf{r}^{(k)})^T \mathbf{r}^{(k)} - \frac{(\mathbf{r}^{(k)})^T \mathbf{r}^{(k)}}{(\mathbf{r}^{(k)})^T A \mathbf{r}^{(k)}} (\mathbf{r}^{(k)})^T A \mathbf{r}^{(k)} = 0.$$

Tale fatto indica che la nuova iterata $\mathbf{x}^{(k+1)}$ è ottimale rispetto alla direzione $\mathbf{r}^{(k)}$, ma non è tuttavia garantito che $\mathbf{x}^{(k+1)}$ sia ottima rispetto alle direzioni di discesa a tutti i passi precedenti.

In generale, un vettore $\mathbf{x}^{(k)}$ è ottimale rispetto a una generica direzione $\mathbf{p} \neq \mathbf{0}$ se

$$\Phi(\mathbf{x}^{(k)}) \leq \Phi(\mathbf{x}^{(k)} + \lambda \mathbf{p}) \quad \text{per ogni } \lambda \in \mathbb{R},$$

ovvero se Φ ammette minimo lungo la direzione \mathbf{p} per $\lambda = 0$. In modo equivalente, se

$$\left. \frac{d}{d\lambda} \Phi(\mathbf{x}^{(k)} + \lambda \mathbf{p}) \right|_{\lambda=0} = 0$$

Calcolando tale derivata, in modo analogo all'Osservazione 2.3.8, si ha

$$\frac{d}{d\lambda} \Phi(\mathbf{x}^{(k)} + \lambda \mathbf{p}) = \nabla \Phi(\mathbf{x}^{(k)} + \lambda \mathbf{p}) \cdot \mathbf{p} = (A(\mathbf{x}^{(k)} + \lambda \mathbf{p}) - \mathbf{b}) \cdot \mathbf{p} = -(\mathbf{b} - A\mathbf{x}^{(k)}) \cdot \mathbf{p} + \lambda \mathbf{p} \cdot (A\mathbf{p})$$

e, se $\lambda = 0$, si ottiene l'ottimalità $\left. \frac{d}{d\lambda} \Phi(\mathbf{x}^{(k)} + \lambda \mathbf{p}) \right|_{\lambda=0} = 0$ per $(\mathbf{r}^{(k)})^T \mathbf{p} = 0$. Di conseguenza, $\mathbf{x}^{(k)}$ è ottimale rispetto a una generica direzione \mathbf{p} se il residuo $\mathbf{r}^{(k)}$ associato a $\mathbf{x}^{(k)}$ è perpendicolare a \mathbf{p} . Nel caso del metodo del gradiente, $\mathbf{x}^{(k+1)}$ è ottimale rispetto alla direzione $\mathbf{p} = \mathbf{r}^{(k)}$ dal momento che, come abbiamo provato sopra, $(\mathbf{r}^{(k)})^T \mathbf{r}^{(k+1)} = 0$.

La ricerca di direzioni di discesa che preservino l'ottimalità di ciascuna iterazione (ovvero che risultino globalmente ottimali) è lo scopo del metodo del gradiente coniugato, introdotto nella sezione successiva.

2.3.6 Metodo del gradiente coniugato

Sempre considerando una matrice A simmetrica e definita positiva, possiamo dunque generare un nuovo metodo iterativo basato su direzioni di discesa richiedendo che

$$(\mathbf{p}^{(j)})^T \mathbf{r}^{(k+1)} = 0 \quad \text{per ogni } j = 0, 1, \dots, k$$

ovvero che la nuova iterata $\mathbf{x}^{(k+1)}$ sia ottimale non solo rispetto alla direzione $\mathbf{p}^{(k)}$, bensì rispetto a *tutte* le direzioni utilizzate fino a quel passo; vogliamo cioè identificare delle direzioni di discesa che preservino l'ottimalità a ciascuna iterazione, cioè globalmente. Cerchiamo di comprendere quale proprietà debbano soddisfare le direzioni di discesa per soddisfare questo requisito. Sia a tal proposito $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{q}$ e supponiamo che $\mathbf{x}^{(k)}$ sia ottimale rispetto a una certa direzione \mathbf{p} , ovvero che $(\mathbf{r}^{(k)})^T \mathbf{p} = 0$; imponiamo poi che $\mathbf{x}^{(k+1)}$ resti ottimale rispetto a \mathbf{p} , ovvero che $(\mathbf{r}^{(k+1)})^T \mathbf{p} = 0$: da quest'ultimo fatto discende che

$$0 = \mathbf{p} \cdot (\mathbf{b} - A\mathbf{x}^{(k+1)}) = \mathbf{p} \cdot (\mathbf{b} - A(\mathbf{x}^{(k)} + \mathbf{q})) = \mathbf{p} \cdot (\mathbf{r}^{(k)} - A\mathbf{q}) = \mathbf{p} \cdot \mathbf{r}^{(k)} - \mathbf{p} \cdot (A\mathbf{q}).$$

Poiché $\mathbf{p} \cdot \mathbf{r}^{(k)} = 0$ deve succedere che $\mathbf{p} \cdot (A\mathbf{q}) = 0$, ovvero che le direzioni \mathbf{p} e \mathbf{q} devono essere tra loro *A-coniugate* (o *A-ortogonali*).

Dunque, consideriamo una matrice simmetrica e definita positiva $A \in \mathbb{R}^{n \times n}$; il metodo del *gradiente coniugato* minimizza, a ogni iterazione $k = 0, 1, \dots$, l'errore $\|\mathbf{e}^{(k+1)}\|_A$ lungo la direzione di discesa $\mathbf{p}^{(k)} \in \mathbb{R}^n$ che è *A-coniugata* a tutte le direzioni di discesa precedentemente calcolate $\mathbf{p}^{(j)}$ per $j = 0, \dots, k-1$ (cioè $(\mathbf{p}^{(j)})^T A \mathbf{p}^{(k)} = 0$ per ogni $j = 0, \dots, k-1$, tali che $k \geq 1$). Il metodo del gradiente coniugato non rientra nella famiglia dei metodi di Richardson dinamici; infatti a ogni iterazione è necessario determinare il valore di due parametri α_k e β_k , quest'ultimo per calcolare la direzione di discesa $\mathbf{p}^{(k)}$.

Algorithm 2.9: Metodo del gradiente coniugato (non precondizionato)

```

dato  $\mathbf{x}^{(0)} \in \mathbb{R}^n$ , porre  $\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}$  e  $\mathbf{p}^{(0)} = \mathbf{r}^{(0)}$ ;
for  $k = 0, 1, \dots$ , finché un criterio d'arresto non è soddisfatto do
    porre  $\alpha_k = \frac{(\mathbf{p}^{(k)})^T \mathbf{r}^{(k)}}{(\mathbf{p}^{(k)})^T A \mathbf{p}^{(k)}}$ ;
    porre  $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{p}^{(k)}$ ;
    porre  $\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k A \mathbf{p}^{(k)}$ ;
    porre  $\beta_k = \frac{(\mathbf{p}^{(k)})^T A \mathbf{r}^{(k+1)}}{(\mathbf{p}^{(k)})^T A \mathbf{p}^{(k)}}$ ;
    porre  $\mathbf{p}^{(k+1)} = \mathbf{r}^{(k+1)} - \beta_k \mathbf{p}^{(k)}$ ;
end

```

Proposizione 2.3.8. Se $A \in \mathbb{R}^{n \times n}$ è simmetrica e definita positiva, il metodo del gradiente coniugato converge a $\mathbf{x} \in \mathbb{R}^n$ per ogni scelta di $\mathbf{x}^{(0)} \in \mathbb{R}^n$ al più n iterazioni (in aritmetica esatta) e

$$\|\mathbf{e}^{(k)}\|_A \leq \frac{2c^k}{1+c^{2k}} \|\mathbf{e}^{(0)}\|_A \quad \text{for } k = 0, 1, \dots, \quad (2.24)$$

dove $c := \frac{\sqrt{K(A)} - 1}{\sqrt{K(A)} + 1}$ e $K(A)$ è il numero di condizionamento spettrale di A .

Osservazione 2.3.9. Il metodo del gradiente coniugato può essere interpretato come un metodo diretto siccome la convergenza a $\mathbf{x} \in \mathbb{R}^n$ avviene in al più n iterazioni in aritmetica esatta. D'altra parte tipicamente l'algoritmo viene arrestato prima che le n iterazioni sono conclusive.

Osservazione 2.3.10. Per k sufficientemente grande, il termine $\frac{2c^k}{1+c^{2k}}$ nella stima dell'errore (2.24) decresce come $2c^k$. Pertanto, se A è simmetrica e definita positiva, il metodo del gradiente coniugato converge più rapidamente del metodo del gradiente, siccome $2c^k < d^k$ per k sufficientemente grande.

Data una matrice di precondizionamento $P \in \mathbb{R}^{n \times n}$ non-singolare, è possibile definire il metodo del gradiente coniugato precondizionato generalizzando il metodo del gradiente coniugato. Pur non riportando l'algoritmo, evidenziamo il seguente risultato.

Proposizione 2.3.9. Se A e $P \in \mathbb{R}^{n \times n}$ sono matrici simmetriche e definite positive, il metodo del gradiente precondizionato coniugato converge a $\mathbf{x} \in \mathbb{R}^n$ per ogni scelta di $\mathbf{x}^{(0)} \in \mathbb{R}^n$. L'errore $\|\mathbf{e}^{(k)}\|_A$ soddisfa la stima (2.24), essendo ora $c = \frac{\sqrt{K(P^{-1}A)} - 1}{\sqrt{K(P^{-1}A)} + 1}$.

Esempio 2.3.4. Riportiamo nella figura seguente le direzioni di discesa calcolate dal metodo del gradiente e del gradiente coniugato, entrambi non precondizionati, applicati al sistema lineare $A\mathbf{x} = \mathbf{b}$, con $A = \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}$ e $\mathbf{b} = (2, -8)^T$, la cui soluzione è $\mathbf{x} = (2, 2)^T$.

Come si può facilmente intuire, l'ortogonalità delle direzioni di discesa (prese a due a due) nel caso del metodo del gradiente comporta una convergenza più lenta del metodo; viceversa, il metodo del gradiente coniugato, sebbene consideri la stessa direzione di discesa alla prima iterazione, termina con la seconda iterazione trovando la soluzione esatta (a meno di errori dell'ordine dell'unità di round-off), come previsto dalla Proposizione 2.3.8.

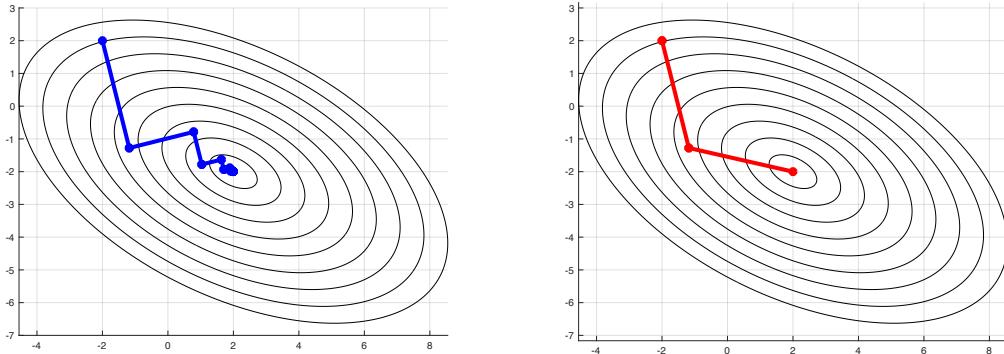


Figura 2.2: Iterate $\mathbf{x}^{(k)}$ del metodo del gradiente (a sinistra) e del metodo del gradiente coniugato (a destra) ottenute a partire dallo stesso vettore iniziale $\mathbf{x}^{(0)}$.

2.3.7 Metodi iterativi ed errore computazionale

I metodi iterativi approssimano la soluzione \mathbf{x} del sistema lineare $A\mathbf{x} = \mathbf{b}$ con una sequenza di iterate $\mathbf{x}^{(k)}$ tale per cui $\lim_{k \rightarrow +\infty} \mathbf{x}^{(k)} = \mathbf{x}$. Tuttavia, è necessario limitare la sequenza di iterate al passo k finito

tale per cui $\mathbf{x}^{(k)} \simeq \mathbf{x}$. Ciò introduce, in aritmetica esatta, l'errore di troncamento $e_t = \|\mathbf{x}^{(k)} - \mathbf{x}\| = \|\mathbf{e}^{(k)}\|$. L'uso di un calcolatore per l'applicazione di un metodo iterativo comporta inevitabilmente degli errori di arrotondamento, per cui si ottiene la soluzione approssimata $\hat{\mathbf{x}}^{(k)}$ affetta appunto da errori di arrotondamento. L'errore di arrotondamento è pertanto $e_r = \|\mathbf{x}^{(k)} - \hat{\mathbf{x}}^{(k)}\|$, mentre l'errore computazionale è $e_c = e_t + e_r$. Tipicamente, per un metodo iterativo si ha $e_t \gg e_r$, per cui l'errore computazionale e l'errore di troncamento sono dello stesso ordine di grandezza, ovvero $e_c \simeq e_t$.

Esempio 2.3.5. Confrontiamo su un semplice caso ($n = 2$) i quattro metodi iterativi introdotti in questo capitolo. A tal proposito, si consideri il sistema lineare

$$\begin{cases} 2x_1 + x_2 = 1 \\ x_1 + 3x_2 = 0 \end{cases} \quad (2.25)$$

la cui matrice $A = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$ è simmetrica e definita positiva. La soluzione esatta di questo sistema è $x_1 = 3/5 = 0.6$, $x_2 = -1/5 = -0.2$. Osserviamo che il metodo del gradiente e il metodo del gradiente coniugato convergono, indipendentemente dalla scelta del vettore iniziale $\mathbf{x}^{(0)}$. La convergenza del metodo di Gauss-Seidel è garantita dalla condizione sufficiente sulla matrice A simmetrica e definita positiva. Invece, la convergenza del metodo di Jacobi va verificata controllando che il raggio spettrale della corrispondente matrice di iterazione risulti inferiore a 1.

Vogliamo approssimare la soluzione del sistema a partire dal vettore iniziale

$$\mathbf{x}^{(0)} = \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \end{bmatrix} = \begin{bmatrix} 1 \\ 1/2 \end{bmatrix}$$

per cui

$$\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)} = \begin{bmatrix} -3/2 \\ -5/2 \end{bmatrix}$$

e

$$\|\mathbf{r}^{(0)}\| = \sqrt{(\mathbf{r}^{(0)})^T \mathbf{r}^{(0)}} = \frac{\sqrt{34}}{2} \approx 2.9155.$$

- Metodo di Jacobi:

$$\mathbf{x}^{(k+1)} = B_J \mathbf{x}^{(k)} + \mathbf{g}_J, \quad k \geq 0, \quad \text{dove } B_J = I - D^{-1}A \text{ e } \mathbf{g}_J = D^{-1}\mathbf{b}.$$

Abbiamo che

$$B_J = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 1/2 & 0 \\ 0 & 1/3 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 0 & -1/2 \\ -1/3 & 0 \end{bmatrix}, \quad \mathbf{g}_J = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/3 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1/2 \\ 0 \end{bmatrix}$$

e $\rho(B_J) = \max_{i=1,\dots,n} |\lambda_i(B_J)| = \max(\text{abs}(\text{eig}(B_J))) = 0.4082$. Dunque il metodo di Jacobi converge. Alla prima iterazione ($k = 0$) si ha:

$$\mathbf{x}^{(1)} = B_J \mathbf{x}^{(0)} + \mathbf{g}_J = \begin{bmatrix} 0 & -1/2 \\ -1/3 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1/2 \end{bmatrix} + \begin{bmatrix} 1/2 \\ 0 \end{bmatrix} = \begin{bmatrix} 1/4 \\ -1/3 \end{bmatrix} \approx \begin{bmatrix} 0.25 \\ -0.3333 \end{bmatrix}.$$

Notiamo che

$$\mathbf{r}^{(1)} = \mathbf{b} - A\mathbf{x}^{(1)} = \begin{bmatrix} 0.8333 \\ 0.75 \end{bmatrix} \quad \text{e} \quad \|\mathbf{r}^{(1)}\| = 1.1211.$$

- Metodo di Gauss-Seidel:

$$\mathbf{x}^{(k+1)} = B_{GS} \mathbf{x}^{(k)} + \mathbf{g}_{GS}, \quad k \geq 0, \quad \text{dove } B_{GS} = (D - E)^{-1}(D - E - A) \quad \text{e} \quad \mathbf{g}_{GS} = (D - E)^{-1}\mathbf{b}.$$

Abbiamo che

$$\begin{aligned} B_{GS} &= \begin{bmatrix} 2 & 0 \\ 1 & 3 \end{bmatrix}^{-1} \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1/2 & 0 \\ -1/6 & 1/3 \end{bmatrix} \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -1/2 \\ 0 & 1/6 \end{bmatrix} \\ \mathbf{g}_{GS} &= \begin{bmatrix} 1/2 & 0 \\ -1/6 & 1/3 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1/2 \\ -1/6 \end{bmatrix} \end{aligned}$$

In questo caso $\rho(B_{GS}) = \max_{i=1,\dots,n} |\lambda_i(B_{GS})| = \max(\text{abs}(\text{eig}(B_{GS}))) = 0.1667$, e dunque anche il metodo di Gauss-Seidel converge. Ci aspettiamo che la convergenza di questo secondo metodo sia più rapida di quella del metodo di Jacobi, dal momento che $\rho(B_{GS}) < \rho(B_J)$. Alla prima iterazione ($k = 0$) troviamo:

$$\mathbf{x}^{(1)} = B_{GS} \mathbf{x}^{(0)} + \mathbf{g}_{GS} = \begin{bmatrix} 0 & -1/2 \\ 0 & 1/6 \end{bmatrix} \begin{bmatrix} 1 \\ 1/2 \end{bmatrix} + \begin{bmatrix} 1/2 \\ -1/6 \end{bmatrix} = \begin{bmatrix} 1/4 \\ -1/12 \end{bmatrix} \approx \begin{bmatrix} 0.25 \\ -0.0833 \end{bmatrix}.$$

Si ha inoltre

$$\mathbf{r}^{(1)} = \mathbf{b} - A\mathbf{x}^{(1)} = \begin{bmatrix} 0.5833 \\ 0 \end{bmatrix} \quad \text{e} \quad \|\mathbf{r}^{(1)}\| = 0.5833.$$

- Metodo del gradiente precondizionato, con $P = D$: abbiamo che $\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 1/2 \end{bmatrix} = \begin{bmatrix} -3/2 \\ -5/2 \end{bmatrix}$. Per $k = 0$, troviamo dunque:

$$\begin{aligned} P\mathbf{z}^{(0)} &= \mathbf{r}^{(0)} \quad \Leftrightarrow \quad \mathbf{z}^{(0)} = P^{-1}\mathbf{r}^{(0)} = \begin{bmatrix} -3/4 \\ -5/6 \end{bmatrix} \\ \alpha_0 &= \frac{(\mathbf{z}^{(0)})^T \mathbf{r}^{(0)}}{(\mathbf{z}^{(0)})^T A \mathbf{z}^{(0)}} = 77/107 \\ \mathbf{x}^{(1)} &= \mathbf{x}^{(0)} + \alpha_0 \mathbf{z}^{(0)} = \begin{bmatrix} 0.4603 \\ -0.0997 \end{bmatrix} \\ \mathbf{r}^{(1)} &= \mathbf{r}^{(0)} - \alpha_0 A \mathbf{z}^{(0)} = \begin{bmatrix} 0.1791 \\ -0.1612 \end{bmatrix} \quad \text{e} \quad \|\mathbf{r}^{(1)}\| = 0.2410. \end{aligned}$$

- Metodo del gradiente coniugato, precondizionato con $P = D$: si ha che $\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}$, $\mathbf{z}^{(0)} = P^{-1}\mathbf{r}^{(0)}$ e $\mathbf{p}^{(0)} = \mathbf{z}^{(0)}$. Per $k = 0$, abbiamo dunque:

$$\begin{aligned} \alpha_0 &= \frac{(\mathbf{p}^{(0)})^T \mathbf{r}^{(0)}}{(\mathbf{p}^{(0)})^T A \mathbf{p}^{(0)}} = \frac{(\mathbf{z}^{(0)})^T \mathbf{r}^{(0)}}{(\mathbf{z}^{(0)})^T A \mathbf{z}^{(0)}} \\ \mathbf{x}^{(1)} &= \mathbf{x}^{(0)} + \alpha_0 \mathbf{p}^{(0)} = \mathbf{x}^{(0)} + \alpha_0 \mathbf{z}^{(0)} \\ \mathbf{r}^{(1)} &= \mathbf{r}^{(0)} - \alpha_0 A \mathbf{p}^{(0)} = \mathbf{r}^{(0)} - \alpha_0 A \mathbf{z}^{(0)}. \end{aligned}$$

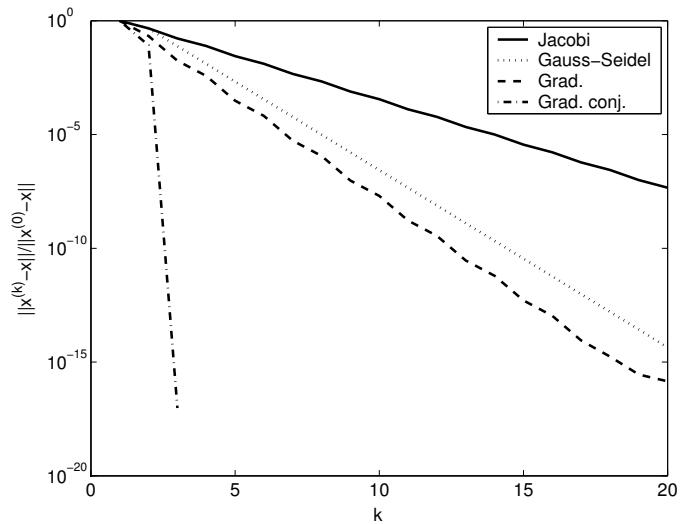
Notiamo che la prima iterata $\mathbf{x}^{(1)}$ coincide con quella trovata dal metodo del gradiente (precondizionato, con lo stesso precondizionatore P). Calcoliamo inoltre:

$$\begin{aligned} P\mathbf{z}^{(1)} &= \mathbf{r}^{(1)} \quad \Leftrightarrow \quad \mathbf{z}^{(1)} = P^{-1}\mathbf{r}^{(1)} = \begin{bmatrix} 0.0896 \\ -0.0537 \end{bmatrix} \\ \beta_0 &= \frac{(A\mathbf{p}^{(0)})^T \mathbf{z}^{(1)}}{(A\mathbf{p}^{(0)})^T A \mathbf{p}^{(0)}} = \frac{(A\mathbf{z}^{(0)})^T \mathbf{z}^{(1)}}{(A\mathbf{z}^{(0)})^T A \mathbf{z}^{(0)}} = -0.0077 \\ \mathbf{p}^{(1)} &= \mathbf{z}^{(1)} - \beta_0 \mathbf{p}^{(0)} = \mathbf{z}^{(1)} - \beta_0 \mathbf{z}^{(0)} = \begin{bmatrix} 0.0838 \\ -0.0602 \end{bmatrix}. \end{aligned}$$

Alla seconda iterata $\mathbf{x}^{(2)}$, per i quattro metodi considerati troviamo i risultati seguenti.

Metodo	$\mathbf{x}^{(2)}$	$\mathbf{r}^{(2)}$	$\ \mathbf{r}^{(2)}\ $
Jacobi	$\begin{bmatrix} 0.6667 \\ -0.0833 \end{bmatrix}$	$\begin{bmatrix} -0.2500 \\ -0.4167 \end{bmatrix}$	0.4859
Gauss-Seidel	$\begin{bmatrix} 0.5417 \\ -0.1806 \end{bmatrix}$	$\begin{bmatrix} 0.0972 \\ 0 \end{bmatrix}$	0.0972
Gradiente (precondizionato)	$\begin{bmatrix} 0.6070 \\ -0.1877 \end{bmatrix}$	$\begin{bmatrix} -0.0263 \\ -0.0438 \end{bmatrix}$	0.0511
Gradiente coniugato (precondizionato)	$\begin{bmatrix} 0.60000 \\ -0.2000 \end{bmatrix}$	$\begin{bmatrix} -0.2220 \\ -0.3886 \end{bmatrix} \cdot 10^{-15}$	$4.4755 \cdot 10^{-16}$

Notiamo come, a meno di errori di round-off, il metodo del gradiente coniugato termina dopo $n = 2$ iterazioni fornendo la soluzione esatta, in accordo con il risultato teorico di Proposizione 2.3.8. La storia di convergenza dei quattro metodi applicati al sistema (2.25) è riportata nella figura seguente.



2.3.8 Criteri d'arresto per metodi iterativi

Come anticipato, i metodi iterativi devono essere arrestati secondo un opportuno criterio d'arresto. Inoltre, il numero di iterazioni dell'algoritmo dovrebbe essere limitato da un qualche intero k_{max} "sufficientemente" grande. Il criterio d'arresto consiste nel terminare l'esecuzione dell'algoritmo all'iterata k tale per cui un adeguato *stimatore dell'errore* dell'errore vero, detto $\tilde{e}^{(k)}$, è più piccolo di una tolleranza prefissata tol , ovvero $\tilde{e}^{(k)} < tol$. I seguenti stimatori dell'errore e i corrispondenti criteri d'arresto sono tipicamente utilizzati:

- il *residuo* (assoluto), per cui $\tilde{e}^{(k)} = \|\mathbf{r}^{(k)}\|$;
- il *residuo normalizzato*, per cui $\tilde{e}^{(k)} = r_{rel}^{(k)} := \frac{\|\mathbf{r}^{(k)}\|}{\|\mathbf{b}\|}$ viene utilizzato per stimare l'errore relativo $e_{rel}^{(k)} := \frac{\|\mathbf{x} - \mathbf{x}^{(k)}\|}{\|\mathbf{x}\|}$, per $\mathbf{x} \neq \mathbf{0}$;
- la *differenza tra iterate successive*, per cui $\tilde{e}^{(k)} = \|\boldsymbol{\delta}^{(k-1)}\|$, dove $\boldsymbol{\delta}^{(k)} := \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}$ per $k \geq 0$.

Richiamiamo il risultato della Proposizione 2.2.5 e, specificamente, Eq. (2.13); ponendo $\hat{\mathbf{x}} = \mathbf{x}^{(k)}$ e $\mathbf{r} = \mathbf{r}^{(k)}$, si ha per un generico metodo iterativo (precondizionato):

$$e_{rel}^{(k)} \leq K_2(A) r_{rel}^{(k)}.$$

Dunque, il criterio d’arresto basato sul residuo (sia assoluto che normalizzato) è *soddisfacente* se il numero di condizionamento $K_2(A)$ della matrice A del sistema lineare da risolvere è relativamente “piccolo”, ovvero se A è *ben-condizionata*. In caso contrario, ovvero se la matrice A è *mal-condizionata*, il criterio d’arresto basato sul residuo è *insoddisfacente* dal momento che l’errore vero viene *sottostimato* dallo stimatore dell’errore (cioè dal residuo).

Si può inoltre mostrare che il criterio d’arresto basato sulla differenza tra iterate successive è *soddisfacente* se il raggio spettrale della matrice di iterazione B è molto piccolo, ovvero $\rho(B) \ll 1$. Al contrario, il criterio è *insoddisfacente* se $\rho(B) \gtrsim 1$ dal momento che l’errore vero viene *sottostimato* dallo stimatore dell’errore $\|\delta^{(k)}\|$.

2.4 Un (breve) Confronto tra Metodi Diretti e Iterativi

Lo standard per la soluzione di un sistema lineare $A\mathbf{x} = \mathbf{b}$ tramite un *metodo diretto* è rappresentato dal metodo della *fattorizzazione LU*; il metodo della fattorizzazione di Cholesky viene utilizzato per una matrice A simmetrica e definita positiva. L’algoritmo di Thomas viene convenientemente utilizzato per una matrice A tridiagonale; per matrici in forma diagonale (per esempio pentadiagonali, etc...), si possono usare algoritmi che generalizzano l’algoritmo di Thomas.

Per quanto riguarda i *metodi iterativi*, i metodi di Richardson precondizionati, sia stazionari che dinamici, che costituiscono una famiglia di metodi con un singolo parametro e matrice di precondizionamento P , vengono generalmente utilizzati per matrici A con proprietà ben definite e note a priori. I metodi del *gradiente precondizionato* e del *gradiente coniugato precondizionato* rappresentano lo stato dell’arte per metodi iterativi applicati alla soluzione di sistemi lineari con matrice A e matrice di precondizionamento P simmetriche e definite positive. Osserviamo che se la matrice A è non-singolare, la soluzione del sistema lineare $A\mathbf{x} = \mathbf{b}$ è equivalente alla soluzione di $A^T A\mathbf{x} = A^T \mathbf{b}$, per cui i metodi del gradiente e gradiente coniugato possono essere nominalmente applicati alla matrice $A^T A$ simmetrica e definita positiva. Tuttavia, il costo computazionale e la richiesta di memoria necessari all’assemblaggio della matrice $A^T A$ e del vettore $A^T \mathbf{b}$ divengono significativi quando la dimensione n del sistema è “grande”; inoltre, l’assemblaggio della matrice $A^T A$ al calcolatore, ovvero mediante operazioni in aritmetica floating-point, può comportare che la matrice $A^T A$ non sia effettivamente simmetrica. Per queste ragioni, l’assemblaggio diretto di una matrice $A^T A$ rappresenta un approccio sconsigliato. Per una generica matrice A non-singolare, lo stato dell’arte dei metodi iterativi è rappresentato dal metodo *GMRES* (Generalized Minimum RESidual) che è ampiamente utilizzato in diversi contesti applicativi.

Metodi diretti possono anche essere utilizzati per generare precondizionatori P da impiegarsi in metodi iterativi. Per esempio, una *fattorizzazione LU incompleta* (ILU) può essere applicata alla matrice A per generare la coppia di matrici triangolari inferiore L e superiore U da impiegarsi come opportuni precondizionatori P per metodi iterativi. In maniera analoga, una *fattorizzazione di Cholesky incompleta* (IC) può essere utilizzata per generare precondizionatori P di una matrice simmetrica e definita positiva A .

Come linea guida generale, la *scelta* di utilizzare un metodo *diretto* o *iterativo* per la soluzione del sistema lineare $A\mathbf{x} = \mathbf{b}$ dipende dalle proprietà della matrice A e dalle risorse computazionali a disposizione, sia in termini di potenza dell’unità di calcolo CPU che della memoria disponibile. A titolo di esempio, se la matrice A è di dimensione n molto “grande”, un’indicazione estremamente semplicistica e puramente indicativa consiste nello scegliere un metodo iterativo se la matrice A è piena, mentre un metodo diretto se A è sparsa e possiede una struttura a banda.

Capitolo 3

Autovalori e Autovettori

Consideriamo l'approssimazione numerica degli *autovalori* e *autovettori* di una matrice $A \in \mathbb{C}^{n \times n}$, ricordando che per tale matrice di dimensione n vi sono n autovalori e n autovettori corrispondenti. Ci concentreremo, in particolare, sull'approssimazione degli autovalori di modulo massimo e minimo di una matrice, tralasciando il problema (ancora più complicato) di approssimare l'intero spettro di una matrice, ovvero l'insieme di tutti i suoi autovalori.

Ricordiamo che se una matrice A è diagonalizzabile, allora $AV = VD$ dove $V = [\mathbf{x}_1 | \dots | \mathbf{x}_n] \in \mathbb{R}^{n \times n}$ è la matrice avente come colonne gli autovettori di A , e $D = \text{diag}(\lambda_1, \dots, \lambda_n) \in \mathbb{R}^{n \times n}$ è la matrice diagonale dei corrispondenti autovalori. Ogni matrice reale diagonalizzabile $A \in \mathbb{R}^{n \times n}$ ammette n autovettori linearmente indipendenti e n autovalori, a priori complessi, che si presentano a coppie di complessi coniugati. Solo nel caso in cui $A \in \mathbb{R}^{n \times n}$ sia simmetrica è possibile garantire che gli autovalori risultino reali.

Per questa ragione, nel seguito di queste note le operazioni indicate riguarderanno quantità a priori complesse, e dunque parleremo di vettore/matrice trasposto/a complesso/a coniugato/a qualora vettori $\mathbf{v} \in \mathbb{C}^n$ e matrici $A \in \mathbb{C}^{n \times n}$ siano complessi, anziché di vettore/matrice trasposto/a; l'operazione di trasposizione eseguita in Matlab® determina automaticamente se una quantità è complessa, e in tal caso restituisce il trasposto complesso coniugato.

3.1 Definizioni ed Esempi

Richiamiamo le seguenti definizioni dal momento che considereremo numeri complessi. Indichiamo con ι l'*unità immaginaria* tale per cui $\iota^2 = -1$.

Definizione 3.1.1. Consideriamo il vettore $\mathbf{v} \in \mathbb{C}^n$, con componenti $v_j = a_j + \iota b_j$, essendo a_j e $b_j \in \mathbb{R}$ per ogni $j = 1, \dots, n$. Il vettore $\bar{\mathbf{v}}$ indica il complesso coniugato di \mathbf{v} , che ha componenti $\bar{v}_j = a_j - \iota b_j$ per ogni $j = 1, \dots, n$. Il vettore $\mathbf{v}^H := (\bar{\mathbf{v}})^T$ è il vettore trasposto complesso coniugato di \mathbf{v} .

Definizione 3.1.2. La matrice $A \in \mathbb{C}^{n \times n}$ si dice Hermitiana se $A^H \equiv A$ (ovvero $(\bar{A})^T \equiv A$).

Ricordiamo cosa si intenda per *problema degli autovalori*.

Definizione 3.1.3. Data la matrice $A \in \mathbb{C}^{n \times n}$, il problema degli autovalori si scrive come: trovare $\lambda \in \mathbb{C}$ e $\mathbf{x} \in \mathbb{C}^n$ tali che $A\mathbf{x} = \lambda\mathbf{x}$, dove λ è un autovalore e \mathbf{x} il corrispondente autovettore. Il polinomio caratteristico della matrice A è $p_A(\lambda) = \det(A - \lambda I)$; gli n autovalori $\{\lambda_i(A)\}_{i=1}^n$ di A sono gli zeri di $p_A(\lambda)$.

La seguente definizione costituisce una generalizzazione del concetto di autovalore di una matrice.

Definizione 3.1.4. Data una matrice $A \in \mathbb{C}^{n \times n}$ e una matrice non-singolare $B \in \mathbb{C}^{n \times n}$, il problema degli autovalori generalizzato si scrive come: trovare $\lambda \in \mathbb{C}$ e $\mathbf{x} \in \mathbb{C}^n$ tali che $A\mathbf{x} = \lambda B\mathbf{x}$, dove λ è un autovalore generalizzato e \mathbf{x} il corrispondente autovettore. Il polinomio caratteristico della matrice A rispetto a B è $p_{A,B}(\lambda) = \det(A - \lambda B)$; gli n autovalori $\{\lambda_i(A; B)\}_{i=1}^n$ di A rispetto a B sono gli zeri di $p_{A,B}(\lambda)$.

Per una matrice $A \in \mathbb{C}^{n \times n}$ vi sono n autovalori e n autovettori corrispondenti.

Definizione 3.1.5. Consideriamo una matrice $A \in \mathbb{C}^{n \times n}$ con autovalori $\{\lambda_i\}_{i=1}^n \in \mathbb{C}$ e i corrispondenti autovettori $\{\mathbf{x}_i\}_{i=1}^n \in \mathbb{C}^n$; allora, il quoziente di Rayleigh riferito all'autovettore \mathbf{x}_i è:

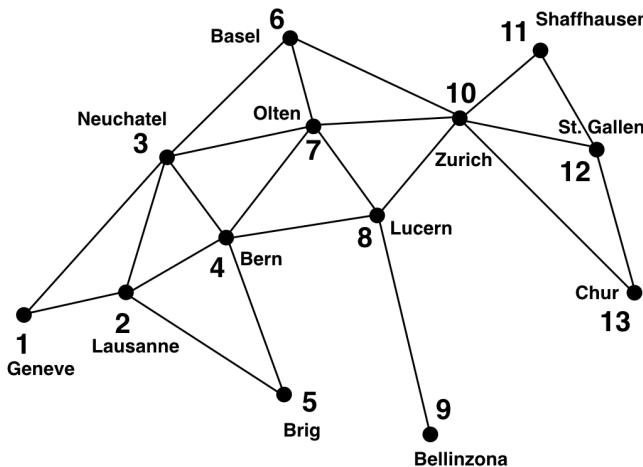
$$\lambda_i = \frac{\mathbf{x}_i^H A \mathbf{x}_i}{\mathbf{x}_i^H \mathbf{x}_i}, \quad (3.1)$$

per $\mathbf{x}_i \neq \mathbf{0}$.

La definizione del quoziente di Rayleigh (3.1) può essere usata per determinare gli autovalori della matrice A una volta che i corrispondenti autovettori sono noti. D'altra parte, se invece è l'autovalore λ_i ad essere noto, l'autovettore corrispondente \mathbf{x}_i può essere determinato risolvendo $(A - \lambda_i I) \mathbf{x}_i = \mathbf{0}$; osserviamo che tipicamente gli autovettori vengono normalizzati, ovvero $\|\mathbf{x}_i\| = 1$ per $i = 1, \dots, n$.

Esempio 3.1.1. Consideriamo la matrice $A = \begin{bmatrix} 3 & -1 \\ 0 & 1 \end{bmatrix}$. Il suo polinomio caratteristico è $p_A(\lambda) = (3 - \lambda)(1 - \lambda) = \lambda^2 - 4\lambda + 3$. Gli autovalori di A sono $\lambda_1 = 3$ e $\lambda_2 = 1$, che corrispondono agli zeri di $p_A(\lambda)$ (ovvero $p_A(\lambda_i) = 0$ per $i = 1$ e 2). Ora, calcoliamo gli autovettori corrispondenti \mathbf{x}_1 e \mathbf{x}_2 . Per $\lambda_1 = 3$ poniamo $(A - \lambda_1 I) \mathbf{x}_1 = \mathbf{0}$, da cui si ottiene $\begin{bmatrix} 0 & -1 \\ 0 & -2 \end{bmatrix} \mathbf{x}_1 = \mathbf{0}$ e quindi $\mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Per $\lambda_2 = 1$ poniamo $(A - \lambda_2 I) \mathbf{x}_2 = \mathbf{0}$, ovvero $\begin{bmatrix} 2 & -1 \\ 0 & 0 \end{bmatrix} \mathbf{x}_2 = \mathbf{0}$, ottenendo così $\mathbf{x}_2 = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$.

Esempio 3.1.2. Determiniamo la connettività di alcune città in Svizzera collegate dalla rete ferroviaria.



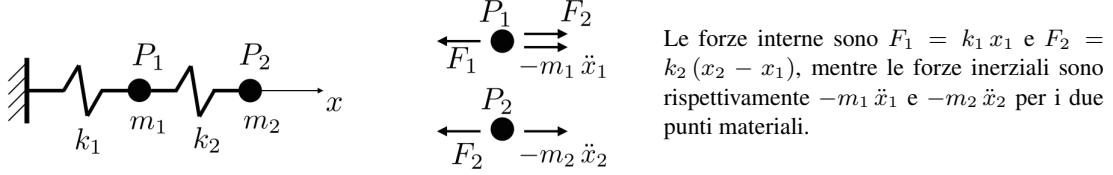
A tal fine, assembliamo una matrice $A \in \mathbb{R}^{n \times n}$, dove n è il numero di città collegate dalla rete ferroviaria; A rappresenta una matrice di connettività. Gli elementi della matrice A sono nulli, ad eccezione di quegli elementi $(A)_{ij} = 1$ per $i, j = 1, \dots, n$ tali per cui la città i è direttamente collegata alla città j ; $(A)_{ii} = 0$ per ogni $i = 1, \dots, n$. La matrice A corrisponde alla connettività della rete ferroviaria svizzera (pur semplificata) che è riportata di lato.

L'autovalore di A più grande in modulo, indicato come λ_1 , è indicativo della connettività complessiva della rete ferroviaria. Le componenti dell'autovettore corrispondente \mathbf{x}_1 forniscono una misura della connettività relativa delle città corrispondenti al resto della rete. Per la rete ferroviaria sopra riportata, si ottiene:

$$\mathbf{x}_1 \simeq (0.16, 0.28, 0.39, 0.39, 0.17, 0.29, 0.43, 0.31, 0.078, 0.36, 0.13, 0.15, 0.13)^T;$$

si deduce che Bellinzona ($\mathbf{x}_1)_9$ è la città meno connessa della rete ferroviaria, mentre Olten ($\mathbf{x}_1)_7$ risulta essere la città maggiormente interconnessa.

Esempio 3.1.3. Consideriamo la dinamica di due masse concentrate connesse da molle elastiche. Nel caso specifico, consideriamo due punti materiali P_1 e P_2 rispettivamente dotate di massa m_1 e m_2 e connesse attraverso molle elastiche con costanti k_1 e k_2 ; il punto materiale P_1 è ancorato come evidenziato nella figura seguente. Gli spostamenti dei punti materiali sono indicati come $x_1(t)$ e $x_2(t)$, dove t è la variabile indipendente che rappresenta il tempo.



L'imposizione dell'equilibrio tra le forze, ovvero $-m_1 \ddot{x}_1 = F_1 - F_2$ e $-m_2 \ddot{x}_2 = F_2$, conduce al seguente sistema lineare ad ogni istante di tempo t :

$$B \ddot{\mathbf{x}}(t) + A \mathbf{x}(t) = \mathbf{0},$$

dove $\mathbf{x}(t) = (x_1(t), x_2(t))^T$, $B = \begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix}$ e $A = \begin{bmatrix} k_1 + k_2 & -k_2 \\ -k_2 & k_2 \end{bmatrix}$. Assumendo che $x_i(t) = a_i \sin(\omega t + \phi)$, per qualche ω , ϕ , e a_i per $i = 1, 2$, il sistema dinamico può essere riscritto come:

$$(-\omega^2 B + A) \mathbf{a} = \mathbf{0},$$

dove $\mathbf{a} = (a_1, a_2)^T$. Quest'ultimo rappresenta un problema agli autovalori generalizzato per cui $\{\omega_i\}_{i=1}^2$ sono le frequenze naturali e $\{a_i\}_{i=1}^2$ i corrispondenti modi propri.

3.2 Metodo delle Potenze

Assumiamo che per una matrice $A \in \mathbb{C}^{n \times n}$ i suoi autovalori $\{\lambda_i\}_{i=1}^n$ siano ordinati come:

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

con i due autovalori più *grandi* in modulo *distinti*, ovvero non solo che $\lambda_1 \neq \lambda_2$, ma anche $|\lambda_1| \neq |\lambda_2|$. Inoltre, assumiamo che gli autovettori di A siano linearmente indipendenti, ovvero che $\det([\mathbf{x}_1, \dots, \mathbf{x}_n]) \neq 0$. Sotto queste ipotesi, il *metodo delle potenze* approssima l'*autovalore di modulo più grande* λ_1 della matrice A per mezzo del seguente algoritmo.

Algorithm 3.1: Metodo delle potenze

<p>Dato $\mathbf{x}^{(0)} \in \mathbb{C}^n$, con $\ \mathbf{x}^{(0)}\ \neq 0$;</p> <p>$\mathbf{y}^{(0)} = \frac{\mathbf{x}^{(0)}}{\ \mathbf{x}^{(0)}\ }$;</p> <p>$\lambda^{(0)} = (\mathbf{y}^{(0)})^H A \mathbf{y}^{(0)}$;</p> <p>for $k = 1, 2, \dots$, fino a che un criterio d'arresto è soddisfatto do</p> <p style="margin-left: 20px;">$\mathbf{x}^{(k)} = A \mathbf{y}^{(k-1)}$;</p> <p style="margin-left: 20px;">$\mathbf{y}^{(k)} = \frac{\mathbf{x}^{(k)}}{\ \mathbf{x}^{(k)}\ }$;</p> <p style="margin-left: 20px;">$\lambda^{(k)} = (\mathbf{y}^{(k)})^H A \mathbf{y}^{(k)}$;</p> <p>end</p>
--

Nell'algoritmo, $\lambda^{(k)}$ rappresenta un'approssimazione dell'autovalore λ_1 di A per mezzo del quoziente di Rayleigh (3.1), mentre $\mathbf{y}^{(k)}$ è un'approssimazione dell'autovettore corrispondente \mathbf{x}_1 . Sotto le ipotesi sopra indicate, il metodo delle potenze è tale che $\mathbf{y}^{(k)} = A^k \mathbf{y}^{(0)} / \|A^k \mathbf{y}^{(0)}\|$ approssima il primo autovettore di A , e che, per k sufficientemente grande, il quoziente di Rayleigh corrispondente approssima l'autovalore di A più grande in modulo, ovvero (per un'opportuna costante C)

$$(\mathbf{y}^{(k)})^H A \mathbf{y}^{(k)} = \frac{(\mathbf{x}^{(k)})^H A \mathbf{x}^{(k)}}{(\mathbf{x}^{(k)})^H \mathbf{x}^{(k)}} = \lambda_1 + C \left| \frac{\lambda_2}{\lambda_1} \right|^k.$$

Come *criterio d'arresto* per il metodo delle potenze in Algoritmo 3.1 si usa il seguente (tol è un'opportuna tolleranza) :

$$\frac{|\lambda^{(k)} - \lambda^{(k-1)}|}{|\lambda^{(k)}|} < tol \quad \text{per } k = 1, 2, \dots \quad (3.2)$$

Forniamo un'interpretazione dell'algoritmo delle potenze. Alla generica iterata $k \geq 2$ dell'Algoritmo 3.1, si ha

$$\mathbf{y}^{(k)} = \frac{\mathbf{x}^{(k)}}{\|\mathbf{x}^{(k)}\|} = \frac{1}{\|\mathbf{x}^{(k)}\|} A \mathbf{y}^{(k-1)} = \frac{1}{\|\mathbf{x}^{(k)}\| \|\mathbf{x}^{(k-1)}\|} A^2 \mathbf{y}^{(k-2)},$$

ne consegue che:

$$\mathbf{y}^{(k)} = \frac{1}{\prod_{j=1}^k \|\mathbf{x}^{(j)}\|} A^k \mathbf{y}^{(0)} = \frac{A^k \mathbf{y}^{(0)}}{\|A^k \mathbf{y}^{(0)}\|} \quad \text{per } k = 1, 2, \dots,$$

da cui il nome al metodo delle potenze. Poiché gli n autovettori $\{\mathbf{x}_i\}_{i=1}^n \in \mathbb{C}^n$ di A sono linearmente indipendenti, essi formano una base di \mathbb{C}^n e dunque, ogni vettore $\mathbf{v} \in \mathbb{C}^n$ si può scrivere come $\mathbf{v} = \sum_{i=1}^n \alpha_i \mathbf{x}_i$, per opportuni pesi $\{\alpha_i\}_{i=1}^n \in \mathbb{C}$. Esprimendo su tale base il vettore iniziale, si ha $\mathbf{x}^{(0)} = \sum_{i=1}^n \alpha_i \mathbf{x}_i$ e $\mathbf{y}^{(0)} = \frac{1}{\|\mathbf{x}^{(0)}\|} \sum_{i=1}^n \alpha_i \mathbf{x}_i$, da cui

$$\mathbf{x}^{(1)} = A \mathbf{y}^{(0)} = \frac{1}{\|\mathbf{x}^{(0)}\|} \sum_{i=1}^n \alpha_i A \mathbf{x}_i = \frac{1}{\|\mathbf{x}^{(0)}\|} \sum_{i=1}^n \alpha_i \lambda_i \mathbf{x}_i$$

e

$$\mathbf{y}^{(1)} = \frac{1}{\|\mathbf{x}^{(1)}\|} \mathbf{x}^{(1)} = \beta^{(1)} \sum_{i=1}^n \alpha_i \lambda_i \mathbf{x}_i,$$

con $\beta^{(1)} := \frac{1}{\|\mathbf{x}^{(1)}\| \|\mathbf{x}^{(0)}\|}$. Continuando in maniera analoga per $k = 2, 3, \dots$, si ha:

$$A^k \mathbf{y}^{(0)} = \alpha_1 \lambda_1^k \mathbf{x}_1 + \dots + \alpha_n \lambda_n^k \mathbf{x}_n = \alpha_1 \lambda_1^k \left[\mathbf{x}_1 + \sum_{i=2}^n \frac{\alpha_i}{\alpha_1} \left(\frac{\lambda_i}{\lambda_1} \right)^k \mathbf{x}_i \right] = \alpha_1 \lambda_1^k \left(\mathbf{x}_1 + \tilde{\mathbf{y}}^{(k)} \right),$$

avendo definito

$$\tilde{\mathbf{y}}^{(k)} = \mathbf{x}_1 + \sum_{i=2}^n \frac{\alpha_i}{\alpha_1} \left(\frac{\lambda_i}{\lambda_1} \right)^k \mathbf{x}_i.$$

Poiché per ipotesi $|\lambda_i/\lambda_1| < 1$, per $i = 2, \dots, n$, per k crescente il vettore $A^k \mathbf{y}^{(0)}$ (e dunque $\mathbf{y}^{(k)}$) tende ad assumere una componente sempre più dominante in direzione di \mathbf{x}_1 , mentre le altre sue componenti decrescono. In altri termini,

$$\mathbf{y}^{(k)} = \frac{\alpha_1 \lambda_1^k (\mathbf{x}_1 + \tilde{\mathbf{y}}^{(k)})}{\|\alpha_1 \lambda_1^k (\mathbf{x}_1 + \tilde{\mathbf{y}}^{(k)})\|} = \underbrace{\frac{\alpha_1 \lambda_1^k}{|\alpha_1 \lambda_1^k|}}_{= \pm 1} \frac{\mathbf{x}_1 + \tilde{\mathbf{y}}^{(k)}}{\|\mathbf{x}_1 + \tilde{\mathbf{y}}^{(k)}\|},$$

con $\lim_{k \rightarrow +\infty} \tilde{\mathbf{y}}^{(k)} = \mathbf{0}$, ovvero (essendo $\|\mathbf{x}_i\| = 1$, $i = 1, \dots, n$), otteniamo:

$$\lim_{k \rightarrow +\infty} \mathbf{y}^{(k)} = \mathbf{x}_1.$$

Una valutazione di quanto velocemente l'iterazione del metodo delle potenze riesca ad approssimare il primo autovalore è data dal seguente risultato: per una matrice $A \in \mathbb{C}^{n \times n}$, l'*errore assoluto* tra l'autovalore λ_1 e la sua approssimazione $\lambda^{(k)}$ ottenuta tramite il metodo delle potenze si può scrivere come:

$$e^{(k)} = |\lambda_1 - \lambda^{(k)}| \simeq C \left| \frac{\lambda_2}{\lambda_1} \right|^k \quad \text{per } k \text{ "sufficientemente" grande e } C > 0.$$

Se invece la matrice $A \in \mathbb{C}^{n \times n}$ è Hermitiana (ovvero se $A^H \equiv A$; vale a dire, simmetrica se $A \in \mathbb{R}^{n \times n}$), allora otteniamo:

$$e^{(k)} = |\lambda_1 - \lambda^{(k)}| \simeq C \left| \frac{\lambda_2}{\lambda_1} \right|^{2k} \quad \text{per } k \text{ "sufficientemente" grande e } C > 0.$$

Il metodo delle potenze consente l'approssimazione dell'autovalore λ_1 di A sotto le ipotesi per cui il metodo è applicabile. Se A è non-singolare, allora si ha la seguente relazione $\lambda_i(A^{-1}) = 1/\lambda_{n+1-i}(A)$ per $i = 1, \dots, n$. Pertanto, il metodo delle potenze può essere usato anche per approssimare l'autovalore di modulo più grande di A^{-1} ; a questo punto, l'autovalore di modulo più piccolo di A si può ottenere come $\lambda_n(A) = \frac{1}{\lambda_1(A^{-1})}$. Questo approccio tuttavia è del tutto sconsigliato in quanto computazionalmente costoso per via della necessità di assemblare la matrice inversa A^{-1} .

3.3 Metodo delle Potenze Inverse

Assumiamo che per la matrice *non-singolare* $A \in \mathbb{C}^{n \times n}$ i suoi autovalori $\{\lambda_i\}_{i=1}^n$ siano ordinati come:

$$|\lambda_1| \geq \dots \geq |\lambda_{n-1}| > |\lambda_n|,$$

con i due autovalori di modulo più piccolo *distinti*, ovvero $\lambda_n \neq \lambda_{n-1}$ e $|\lambda_n| \neq |\lambda_{n-1}|$. In aggiunta, assumiamo che gli autovettori di A siano linearmente indipendenti, ovvero che $\det([\mathbf{x}_1, \dots, \mathbf{x}_n]) \neq 0$. Sotto queste ipotesi, il *metodo delle potenze inverse* approssima l'autovalore di modulo più *piccolo* λ_n di A utilizzando il seguente algoritmo.

Algorithm 3.2: Metodo delle potenze inverse

```
Dato  $\mathbf{x}^{(0)} \in \mathbb{C}^n$ , con  $\|\mathbf{x}^{(0)}\| \neq 0$ ;
 $\mathbf{y}^{(0)} = \frac{\mathbf{x}^{(0)}}{\|\mathbf{x}^{(0)}\|}$ ;
 $\mu^{(0)} = (\mathbf{y}^{(0)})^H A^{-1} \mathbf{y}^{(0)}$ ;
for  $k = 1, 2, \dots$ , fino a che un criterio d'arresto è soddisfatto do
    | risolvere  $A \mathbf{x}^{(k)} = \mathbf{y}^{(k-1)}$ ;
    |  $\mathbf{y}^{(k)} = \frac{\mathbf{x}^{(k)}}{\|\mathbf{x}^{(k)}\|}$ ;
    |  $\mu^{(k)} = (\mathbf{y}^{(k)})^H A^{-1} \mathbf{y}^{(k)}$ ;
end
porre  $\lambda^{(k)} = \frac{1}{\mu^{(k)}}$ ;
```

Nell'algoritmo delle potenze inverse $\mu^{(k)}$ rappresenta un'approssimazione di $\frac{1}{\lambda_n}$, dove λ_n è l'autovalore di modulo più piccolo di A ; si ha infatti:

$$\lim_{k \rightarrow +\infty} \mu^{(k)} = \lim_{k \rightarrow +\infty} \frac{1}{\lambda^{(k)}} = \frac{1}{\lambda_n}.$$

Alternativamente, possiamo calcolare, a ogni passo del metodo delle potenze inverse¹,

$$\lambda^{(k)} = (\mathbf{y}^{(k)})^H A \mathbf{y}^{(k)},$$

per cui

¹La matrice inversa A^{-1} di A ha gli stessi autovettori di A : infatti, a patto che gli autovettori di A siano linearmente indipendenti, dalla relazione $AV = VD$ si ha che $V^{-1}A^{-1} = D^{-1}V^{-1}$, ovvero $A^{-1}V = VD^{-1}$ moltiplicando per V a destra e a sinistra la precedente relazione. Abbiamo sfruttato il risultato secondo cui, se $A, B \in \mathbb{R}^{n \times n}$ sono invertibili, anche AB lo è, ed ha per inversa il prodotto delle inverse, prese in ordine opposto ($(AB)^{-1} = B^{-1}A^{-1}$).

$$\lim_{k \rightarrow +\infty} \lambda^{(k)} = \lambda_n.$$

Come criterio d'arresto, possiamo considerare lo stesso già usato per il metodo delle potenze sostituendo $\lambda^{(k)}$ e $\lambda^{(k-1)}$ rispettivamente con $\mu^{(k)}$ e $\mu^{(k-1)}$ in Eq. (3.2).

Osservazione 3.3.1. *Dato che il metodo delle potenze inverse coinvolge la soluzione del sistema lineare $A\mathbf{x}^{(k)} = \mathbf{y}^{(k-1)}$ a ogni passo iterativo dell'Algoritmo 3.3, risulta computazionalmente conveniente utilizzare il metodo della fattorizzazione LU; infatti, la fattorizzazione LU della matrice A può essere effettuata una volta sola (contestualmente alla prima iterazione) e il sistema risolto con i metodi delle sostituzioni in avanti e indietro ad ogni passo iterativo $k \geq 2$.*

3.4 Metodi delle Potenze Inverse con Shift

Data $A \in \mathbb{C}^{n \times n}$, indichiamo con $\sigma(A) = \{\lambda_i\}_{i=1}^n$ lo spettro di A , ovvero l'insieme dei suoi autovalori.

Definizione 3.4.1. *Data la matrice $A \in \mathbb{C}^{n \times n}$, allora lo shift è un numero complesso $s \in \mathbb{C}$, $s \notin \sigma(A)$ tale per cui si ottiene la matrice (di shift) $A_s \in \mathbb{C}^{n \times n}$ come $A_s = A - sI$.*

Osservazione 3.4.1. *Siano $\{\lambda_i(A)\}_{i=1}^n$ gli autovalori di $A \in \mathbb{C}^{n \times n}$, allora gli autovalori della matrice $A_s \in \mathbb{C}^{n \times n}$ sono $\lambda_j(A_s) = \lambda_i(A) - s$ per qualche $i, j = 1, \dots, n$, essendo $s \in \mathbb{C}$ lo shift.*

Lo shift fornisce dunque una stima dell'autovalore che vogliamo calcolare. Per approssimare l'autovalore di A che in modulo risulta più vicino a s , applichiamo il metodo delle potenze inverse alla matrice shiftata A_s . Infatti, gli autovalori di A_s^{-1} sono $\zeta_i = \frac{1}{\lambda_i - s}$ per $i = 1, \dots, n$.

Se l'autovalore λ_m di A che è in modulo più vicino a s ha molteplicità 1, ovvero se

$$|\lambda_m - s| < |\lambda_i - s| \quad \forall i = 1, \dots, n, \text{ con } i \neq m,$$

allora ζ_m è l'autovalore di A_s^{-1} che ha modulo maggiore (o, il che è lo stesso, $\lambda_m - s$ è l'autovalore di A_s che ha modulo minimo). Se $s = 0$, ζ_m è l'autovalore di A con modulo minimo.

Selezionato lo shift $s \in \mathbb{C}$, possiamo dunque applicare il metodo delle potenze inverse a A_s , determinare $\lambda_n(A_s)$ determinare infine $\lambda_m = \lambda_m(A)$ come

$$\lambda_m(A) = \lambda_n(A_s) + s.$$

Algorithm 3.3: Metodo delle potenze inverse con shift

```

Dato  $s \in \mathbb{C}$ , dato  $\mathbf{x}^{(0)} \in \mathbb{C}^n$ , con  $\|\mathbf{x}^{(0)}\| \neq 0$ ;
 $\mathbf{y}^{(0)} = \frac{\mathbf{x}^{(0)}}{\|\mathbf{x}^{(0)}\|}$ ;
 $\mu^{(0)} = (\mathbf{y}^{(0)})^H (A - sI)^{-1} \mathbf{y}^{(0)}$ ;
for  $k = 1, 2, \dots$ , fino a che un criterio d'arresto è soddisfatto do
    | risolvere  $(A - sI) \mathbf{x}^{(k)} = \mathbf{y}^{(k-1)}$ ;
    |  $\mathbf{y}^{(k)} = \frac{\mathbf{x}^{(k)}}{\|\mathbf{x}^{(k)}\|}$ ;
    |  $\mu^{(k)} = (\mathbf{y}^{(k)})^H (A - sI)^{-1} \mathbf{y}^{(k)}$ ;
end
porre  $\lambda^{(k)} = \frac{1}{\mu^{(k)}} + s$ ;
```

In questo caso, abbiamo infatti:

$$\lim_{k \rightarrow +\infty} \mu^{(k)} = \frac{1}{\lambda_m(A) - s}.$$

Osserviamo tuttavia come la matrice shiftata $A_s = A - sI$ sia tale che, se A si può diagonalizzare come $A = VDV^{-1}$, allora

$$A_s = VDV^{-1} - sI = VDV^{-1} - sVV^{-1} = V(D - sI)V^{-1},$$

ovvero gli autovettori di A coincidono con quelli di A_s (e anche della sua inversa); dunque al posto del quoziente di Rayleigh che compare nell'algoritmo precedente si può direttamente valutare

$$\lambda^{(k)} = (\mathbf{y}^{(k)})^H A \mathbf{y}^{(k)}$$

e in questo caso si ha che

$$\lim_{k \rightarrow +\infty} \lambda^{(k)} = \lambda_m(A),$$

essendo $\lambda_m(A)$ l'autovalore che stiamo approssimando.

Esempio 3.4.1. Consideriamo la matrice $A \in \mathbb{R}^{3 \times 3}$ tale che i suoi autovalori sono $\lambda_1(A) = 5.3$, $\lambda_2(A) = 2.1$ e $\lambda_3(A) = 0.2$. Selezioniamo per esempio il valore di shift $s = 4.9$ per cui la matrice $A_s = A - sI$ possiede autovalori $\lambda_1(A_s) = -4.7$, $\lambda_2(A_s) = -2.8$ e $\lambda_3(A_s) = 0.4$. Applicando il metodo delle potenze inverse alla matrice A_s , allora viene approssimato l'autovalore $\lambda_3(A_s) = 0.4$, da cui si può rideterminare $\lambda_1(A)$ come $\lambda_1(A) = \lambda_3(A_s) + s$, per $s = 4.9$.

3.5 Localizzazione Geometrica degli Autovalori: Criteri di Gershgorin

La conoscenza a priori, seppur approssimativa, della dislocazione degli autovalori $\{\lambda_i\}_{i=1}^n$ della matrice $A \in \mathbb{C}^{n \times n}$ nel piano complesso è utile per la scelta del parametro di shift e, in generale, per inizializzare opportunamente i metodi iterativi per il calcolo dello spettro di A (anche solo parziale). I seguenti *criteri di Gershgorin* forniscono appunto una localizzazione geometrica degli autovalori.

Teorema 3.5.1 (dei cerchi per riga di Gershgorin). *Sia $A \in \mathbb{C}^{n \times n}$, allora si ha:*

$$\{\lambda_i\}_{i=1}^n \in \mathcal{S}_R := \bigcup_{k=1}^n \mathcal{R}_k,$$

dove

$$\mathcal{R}_k = \left\{ z \in \mathbb{C} : |z - a_{kk}| \leq \sum_{\substack{j=1 \\ j \neq k}}^n |a_{kj}| \right\} \quad \text{per } k = 1, \dots, n$$

sono i cerchi per riga di A .

Essendo $\{\lambda_i(A)\}_{i=1}^n \equiv \{\lambda_i(A^T)\}_{i=1}^n$, si ottiene il seguente risultato dal precedente.

Corollario 3.5.1 (dei cerchi per colonna di Gershgorin). *Sia $A \in \mathbb{C}^{n \times n}$, allora si ha:*

$$\{\lambda_i\}_{i=1}^n \in \mathcal{S}_C := \bigcup_{k=1}^n \mathcal{C}_k,$$

dove

$$\mathcal{C}_k = \left\{ z \in \mathbb{C} : |z - a_{kk}| \leq \sum_{\substack{j=1 \\ j \neq k}}^n |a_{jk}| \right\} \quad \text{per } k = 1, \dots, n$$

sono i cerchi per colonna di A .

Inoltre, come conseguenza dei due risultati precedenti, si ottiene il seguente.

Corollario 3.5.2 (Primo teorema di Gershgorin). Per $A \in \mathbb{C}^{n \times n}$ si ha:

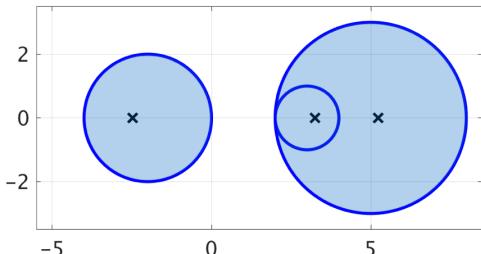
$$\lambda_i(A) \in \mathcal{S}_{\mathcal{R}} \cap \mathcal{S}_{\mathcal{C}} \quad \text{per ogni } i = 1, \dots, n.$$

Teorema 3.5.2 (Secondo teorema di Gershgorin). Per $A \in \mathbb{C}^{n \times n}$ siano:

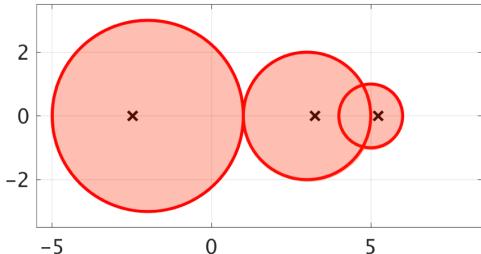
$$\mathcal{S}_1 = \bigcup_{k=1}^m \mathcal{R}_k \quad e \quad \mathcal{S}_2 = \bigcup_{k=m+1}^n \mathcal{R}_k$$

con $\mathcal{S}_1 \cap \mathcal{S}_2 = \emptyset$, allora \mathcal{S}_1 contiene esattamente m autovalori di A (ognuno contato per la loro molteplicità) e \mathcal{S}_2 contiene i restanti $n - m$.

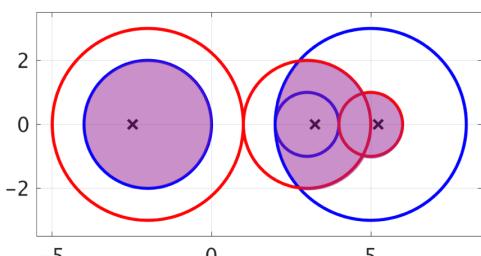
Esempio 3.5.1. Si consideri la matrice $A = \begin{bmatrix} 5 & -1 & 2 \\ 0 & 3 & 1 \\ 1 & 1 & -2 \end{bmatrix}$ con $n = 3$ i cui autovalori sono $\lambda_1 = 5.2288$, $\lambda_2 = 3.2459$ e $\lambda_3 = -2.4747$.



I cerchi per riga di A sono $\mathcal{R}_1 = \{z \in \mathbb{C} : |z - 5| \leq 3\}$, $\mathcal{R}_2 = \{z \in \mathbb{C} : |z - 3| \leq 1\}$ e $\mathcal{R}_3 = \{z \in \mathbb{C} : |z - 2| \leq 2\}$. Si ha pertanto $\mathcal{S}_{\mathcal{R}} = \mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3$ come evidenziato in color blu; inoltre, si osserva che gli autovalori si trovano in $\mathcal{S}_{\mathcal{R}}$. Dal secondo teorema di Gershgorin, si hanno $\mathcal{S}_1 = \mathcal{R}_1$ e $\mathcal{S}_2 = \mathcal{R}_2 \cup \mathcal{R}_3$ con $\mathcal{S}_1 \cap \mathcal{S}_2$, pertanto $m = 1$ autovalori sono contenuti in \mathcal{S}_1 e $n - m = 2$ in \mathcal{S}_2 .



I cerchi per colonna di A sono invece $\mathcal{C}_1 = \{z \in \mathbb{C} : |z - 5| \leq 1\}$, $\mathcal{C}_2 = \{z \in \mathbb{C} : |z - 3| \leq 2\}$ e $\mathcal{C}_3 = \{z \in \mathbb{C} : |z - 2| \leq 3\}$. Si ha $\mathcal{S}_{\mathcal{C}} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3$ come evidenziato in color rosso; si osserva che gli autovalori si trovano in $\mathcal{S}_{\mathcal{C}}$.



Utilizzando il primo teorema di Gershgorin, si ottiene il sottoinsieme del piano complesso $\mathcal{S}_{\mathcal{R}} \cap \mathcal{S}_{\mathcal{C}}$ evidenziato in color viola, che contiene gli autovalori della matrice A .

3.6 Metodo delle Iterazioni QR

I metodi per l'approssimazione simultanea di tutti gli autovalori della matrice A sono basati sulla trasformazione di A in una matrice simile T , ovvero con gli stessi autovalori di A , ma di tipo triangolare; in tal modo, gli autovalori di A sono gli elementi sulla diagonale principale di T .

Consideriamo il caso di una matrice reale $A \in \mathbb{R}^{n \times n}$ con *autovalori reali e distinti in modulo*, ovvero $\{\lambda_i\}_{i=1}^n \in \mathbb{R}$ e $|\lambda_1| > |\lambda_2| > \dots > |\lambda_n|$. Per il calcolo degli autovalori di A usiamo il *metodo delle iterazioni QR* per costruire una sequenza di matrici $A^{(k)}$ simili alla matrice A per $k = 1, 2, \dots$. Tale metodo, che sfrutta la fattorizzazione QR delle matrici simili $A^{(k)}$, eventualmente nella versione ridotta tramite l'algoritmo di ortogonalizzazione di Gram–Schmidt (si veda la Sez. 2.2.6), è illustrato schematicamente nell'Algoritmo 3.4.

Algorithm 3.4: Metodo delle iterazioni QR

```

porre  $A^{(0)} = A$ ;
for  $k = 0, 1, \dots$ , fino a che un criterio d'arresto è soddisfatto do
    determinare la fattorizzazione QR (ridotta) di  $A^{(k)}$ , ovvero le matrici quadrate  $Q^{(k+1)}$  e
     $R^{(k+1)}$  tali per cui  $Q^{(k+1)} R^{(k+1)} = A^{(k)}$ ;
    porre  $A^{(k+1)} = R^{(k+1)} Q^{(k+1)}$ ;
     $\lambda_i^{(k+1)} = (A^{(k+1)})_{ii}$  per  $i = 1, \dots, n$ ;
end
```

Osservazione 3.6.1. Ogni matrice $A^{(k)}$ è ortogonalmente simile ad A , infatti:

$$\begin{aligned} A^{(k)} &= R^{(k)} Q^{(k)} = \left(Q^{(k)}\right)^T Q^{(k)} R^{(k)} Q^{(k)} = \left(Q^{(k)}\right)^T A^{(k-1)} Q^{(k)} \\ &= \left(Q^{(0)} \cdots Q^{(k)}\right)^T A \left(Q^{(0)} \cdots Q^{(k)}\right) \quad \text{per } k = 0, 1, \dots, \end{aligned}$$

essendo le matrici $Q^{(k)}$ ortogonali in seguito alla fattorizzazione QR e $Q^{(0)} = I$.

Si osserva che, contrariamente ai metodi delle potenze, il metodo delle iterazioni QR non fornisce approssimazioni degli autovettori della matrice A .

Proposizione 3.6.1. Sia $A \in \mathbb{R}^{n \times n}$ con autovalori reali tali per cui $|\lambda_1| > |\lambda_2| > \dots > |\lambda_n|$. Allora, il metodo delle iterazioni QR (Algoritmo 3.4) converge a una matrice T triangolare superiore, ovvero:

$$\lim_{k \rightarrow \infty} A^{(k)} = T = \begin{bmatrix} \lambda_1 & t_{12} & t_{13} & \cdots & t_{1n} \\ 0 & \lambda_2 & t_{23} & \cdots & t_{2n} \\ 0 & 0 & \lambda_3 & \cdots & \vdots \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & 0 & \lambda_n & \end{bmatrix}.$$

Inoltre, si ha:

$$\left| a_{i,i-1}^{(k)} \right| = O\left(\left|\frac{\lambda_i}{\lambda_{i-1}}\right|^k\right) \quad \text{per } i = 2, \dots, n \quad \text{e } k \rightarrow \infty. \quad (3.3)$$

Se $A \in \mathbb{R}^{n \times n}$ è simmetrica, allora T è una matrice diagonale.

Il criterio d'arresto per il metodo delle iterazioni QR consiste nel terminare le iterazioni all'iterata k per cui un opportuno *stimatore dell'errore*, indicato come $\tilde{e}^{(k)}$, è inferiore a una tolleranza prescritta a priori tol , ovvero $\tilde{e}^{(k)} < tol$. Nel caso specifico, sfruttando il risultato di Proposizione 3.6.1, si sceglie:

$$\tilde{e}^{(k)} = \max_{\substack{i=2, \dots, n \\ j=1, \dots, i-1}} \left| a_{ij}^{(k)} \right|$$

dato che le matrici $A^{(k)} \rightarrow T$ per $k \rightarrow \infty$.

Osservazione 3.6.2. Dall'Eq. (3.3) si deduce che se gli autovalori di A sono distinti in modulo, ma non sono ben separati, allora la convergenza di $A^{(k)}$ alla matrice triangolare superiore T può essere molto lenta.

3.7 Decomposizione ai Valori Singolari di una Matrice

La *decomposizione ai valori singolari* di una matrice $A \in \mathbb{R}^{m \times n}$ riveste un ruolo particolarmente importante in algebra lineare numerica, per le sue connessioni con la compressione di dati e la possibilità di approssimare matrici con matrici di rango più basso. La decomposizione ai valori singolari è una tecnica di diagonalizzazione di una matrice generica (non necessariamente quadrata) che coinvolge il prodotto (sia sinistro che destro) per una matrice ortogonale. Più precisamente,

Definizione 3.7.1. Per ogni matrice $A \in \mathbb{R}^{m \times n}$, esistono due matrici ortogonali

$$U = [\mathbf{u}_1 | \dots | \mathbf{u}_m] \in \mathbb{R}^{m \times m}, \quad V = [\mathbf{v}_1 | \dots | \mathbf{v}_n] \in \mathbb{R}^{n \times n}$$

tali che

$$A = U \Sigma V^T, \quad \text{con } \Sigma := \text{diag}(\sigma_1, \dots, \sigma_p) \in \mathbb{R}^{m \times n} \quad (3.4)$$

e $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p \geq 0$, essendo $p = \min\{m, n\}$.

La fattorizzazione (3.4) è chiamata decomposizione ai valori singolari (o singular value decomposition, SVD) di A e i numeri $\sigma_i = \sigma_i(A)$ sono chiamati *valori singolari* di A ; $\mathbf{u}_1, \dots, \mathbf{u}_m$ sono chiamati *vettori singolari sinistri* di A , mentre $\mathbf{v}_1, \dots, \mathbf{v}_n$ *vettori singolari destri* di A , dal momento che

$$A \mathbf{v}_i = \sigma_i \mathbf{u}_i, \quad A^T \mathbf{u}_j = \sigma_j \mathbf{v}_j, \quad \text{per } i, j = 1, \dots, p = \min\{m, n\}.$$

Esempio 3.7.1. Illustriamo di seguito un esempio di decomposizione ai valori singolari di una matrice $A \in \mathbb{R}^{m \times n}$ con $m \geq n$.

$$\boxed{A} = \boxed{U} \boxed{\Sigma} \boxed{V^T}$$

Esempio 3.7.2. Illustriamo di seguito un esempio di decomposizione ai valori singolari di una matrice $A \in \mathbb{R}^{m \times n}$, questa volta con $m \leq n$.

$$\boxed{A} = \boxed{U} \boxed{\Sigma} \boxed{V^T}$$

Osserviamo che, essendo sia U che V matrici ortogonali, si ha che $U^T U = I \in \mathbb{R}^{m \times m}$, mentre $V^T V = I \in \mathbb{R}^{n \times n}$. Inoltre, la relazione (3.4) implica la seguente decomposizione spettrale di AA^T e di $A^T A$,

$$AA^T = U \Sigma \Sigma^T U^T \quad \text{e} \quad A^T A = V \Sigma^T \Sigma V^T$$

con

$$\Sigma \Sigma^T = \text{diag}(\sigma_1^2, \dots, \sigma_p^2, \underbrace{0, \dots, 0}_{m-p \text{ volte}}) \in \mathbb{R}^{m \times m} \quad \text{e} \quad \Sigma^T \Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_p^2, \underbrace{0, \dots, 0}_{n-p \text{ volte}}) \in \mathbb{R}^{n \times n},$$

rispettivamente. Poiché AA^T e $A^T A$ sono matrici simmetriche, i vettori singolari sinistri (rispettivamente, destri) di A risultano essere gli autovettori di AA^T (rispettivamente, di $A^T A$). Sussiste infatti una relazione molto stretta tra la decomposizione SVD di A e i problemi agli autovalori per le matrici $A^T A$ e AA^T , dal momento che

$$\sigma_i(A) = \sqrt{\lambda_i(A^T A)}, \quad \text{per } i = 1, \dots, p.$$

In particolare, il più grande e il più piccolo tra i valori singolari vengono indicati con

$$\sigma_{\max} = \max_{i=1, \dots, p} \sigma_i = \sigma_1, \quad \sigma_{\min} = \min_{i=1, \dots, p} \sigma_i = \sigma_p,$$

in considerazione dell'ordinamento dato nella definizione.

Osservazione 3.7.1. Se $A \in \mathbb{R}^{n \times n}$ è una matrice simmetrica, allora $\sigma_i(A) = |\lambda_i(A)|$, essendo $\lambda_1(A) \geq \lambda_2(A) \geq \dots \geq \lambda_n(A)$ gli autovalori di A .

Osservazione 3.7.2. A seconda che $m > n$ oppure $m < n$, può essere più vantaggioso calcolare gli autovalori di $A^T A \in \mathbb{R}^{n \times n}$ oppure di $AA^T \in \mathbb{R}^{m \times m}$. Tuttavia, il calcolo dei valori singolari di una matrice può essere effettuato evitando di assemblare una di queste matrici, sfruttando direttamente la fattorizzazione QR della matrice A . L'analisi dell'algoritmo per ottenere la fattorizzazione SVD di una matrice in base a tale metodo esula tuttavia dagli scopi di tale corso.

I valori singolari di una matrice sono direttamente collegati sia alla norma di una matrice che al suo numero di condizionamento. Valgono infatti i seguenti risultati: per ciascuna matrice $A \in \mathbb{R}^{m \times n}$,

$$\|A\|_2 = \sigma_{\max}, \quad \|A\|_F = \sqrt{\sum_{i=1}^p \sigma_i^2}, \tag{3.5}$$

dove la *norma di Frobenius* di una generica matrice $A \in \mathbb{R}^{m \times n}$ è definita come

$$\|A\|_F := \left(\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2 \right)^{1/2}. \tag{3.6}$$

Inoltre, se $A \in \mathbb{R}^{n \times n}$ è una matrice quadrata, non singolare, invertendo la relazione (3.4) otteniamo che

$$A^{-1} = V \Sigma^{-1} U^T,$$

dove $\Sigma^{-1} = \text{diag}(\sigma_1^{-1}, \dots, \sigma_n^{-1}) \in \mathbb{R}^{n \times n}$. Questo fatto mostra che σ_n^{-1} è il più grande valore singolare della matrice A^{-1} , da cui si deduce che

$$\|A^{-1}\|_2 = \frac{1}{\sigma_n}$$

e dunque che

$$K_2(A) = \|A\|_2 \|A^{-1}\|_2 = \frac{\sigma_1}{\sigma_n}.$$

La decomposizione ai valori singolari di una matrice riveste una fondamentale importanza per due ulteriori motivi. Il primo riguarda la possibilità di usare tale decomposizione per approssimare una data matrice con matrici di basso rango o di rango ridotto (*low-rank approximations*), risultato particolarmente utile nell'ambito dei cosiddetti modelli di ordine ridotto. Infatti, dal momento che $\text{rank}(A) = \text{rank}(\Sigma)$ e che il rango di una matrice diagonale è uguale al numero dei suoi elementi diagonali non nulli, se $A \in \mathbb{R}^{m \times n}$

ha r valori singolari strettamente positivi, allora² $\text{rank}(A) = r$. Se una matrice $A \in \mathbb{R}^{m \times n}$ ha rango r , essa può essere riscritta come somma di r matrici, ciascuna avente rango 1,

$$A = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^T.$$

Tale formula risulta molto utile per determinare approssimazioni a rango ridotto di una matrice poiché, grazie alle proprietà (3.5), la somma parziale di $k \leq r$ termini cattura la maggior quantità di *energia* della matrice A possibile, dove con *energia* ci riferiamo sia la norma 2 che la norma di Frobenius di una matrice. Più precisamente, tra tutte le matrici aventi rango k , la miglior approssimazione di rango k di una data matrice A è quella che si ottiene a partire dai vettori singolari destri e sinistri di A , pesando ciascun contributo di rango 1 per il valore singolare corrispondente, come stabilito dal seguente teorema.

Teorema 3.7.1 (Schmidt-Eckart-Young). *Per ogni matrice $A \in \mathbb{R}^{m \times n}$ di rango r , la matrice*

$$A_k = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^T, \quad 0 \leq k \leq r, \quad (3.7)$$

soddisfa la seguente proprietà di ottimalità:

$$\|A - A_k\|_F = \min_{\substack{\mathbb{B} \in \mathbb{R}^{m \times n} \\ \text{rank}(\mathbb{B}) \leq k}} \|A - \mathbb{B}\|_F = \sqrt{\sum_{i=k+1}^r \sigma_i^2}. \quad (3.8)$$

Se $k = r$ allora $A_k = A$ e la somma in (3.8) è pari a zero. Un risultato simile vale anche considerando la norma 2 al posto della norma di Frobenius: per ogni $0 \leq k \leq r$, la matrice A_k definita da (3.7) è tale che

$$\|A - A_k\|_2 = \min_{\substack{\mathbb{B} \in \mathbb{R}^{m \times n} \\ \text{rank}(\mathbb{B}) \leq k}} \|A - \mathbb{B}\|_2 = \sigma_{k+1}. \quad (3.9)$$

Inoltre, l'errore (in norma di Frobenius) commesso nell'approssimare una matrice A con la miglior matrice di rango k è pari alla somma dei valori singolari di A , presi al quadrato, corrispondenti ai modi trascurati nell'approssimazione – ovvero quelli da $k+1$ a r : risulta quindi evidente come, nel caso in cui il decadimento dei valori singolari sia molto rapido, l'errore commesso risulti piccolo. Per fare in modo che tale errore risulti al di sotto di una tolleranza desiderata ϵ , è sufficiente scegliere k come il più piccolo intero per il quale

$$I(k) = \frac{\sum_{i=1}^k \sigma_i^2}{\sum_{i=1}^r \sigma_i^2} \geq 1 - \epsilon^2, \quad (3.10)$$

ovvero tale per cui l'energia catturata dagli ultimi $r-k$ modi sia minore o uguale a ϵ^2 ; la quantità $I(k)$ rappresenta quindi la percentuale di energia delle colonne di A catturata dai primi k modi, ed è anche chiamata *contenuto di informazione relativo* della approssimazione a rango ridotto. In modo equivalente, il criterio (3.10) assicura che l'errore relativo tra A e la sua approssimazione di rango k A_k sia minore o uguale a ϵ , ovvero,

$$\frac{\|A - A_k\|_F}{\|A\|_F} \leq \epsilon.$$

Un'applicazione dell'approssimazione della matrice $A \in \mathbb{R}^{m \times n}$ tramite una matrice a rango ridotto $A_k \in \mathbb{R}^{m \times n}$ si trova nella compressione di immagini, appunto rappresentate dalla matrice A . Se la matrice

²Inoltre, è possibile definire una base ortonormale sia per il nucleo di A che per lo spazio delle colonne di A , come segue:

$$\ker(A) = \text{span}\{\mathbf{v}_{r+1}, \dots, \mathbf{v}_n\}, \quad \text{range}(A) = \text{span}\{\mathbf{u}_1, \dots, \mathbf{u}_r\}.$$

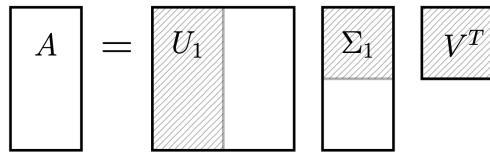


Figura 3.1: Decomposizione ai valori singolari in formato "thin" di una matrice

A contiene mn elementi da memorizzare, la matrice A_k può essere memorizzata complessivamente con $k(m+n+1)$ valori, ovvero mk elementi in $[\mathbf{u}_1 | \dots | \mathbf{u}_k] \in \mathbb{R}^{m \times k}$, nk elementi in $[\mathbf{v}_1 | \dots | \mathbf{v}_k] \in \mathbb{R}^{n \times k}$ e k valori singolari $\{\sigma_i\}_{i=1}^k$ sulla diagonale di Σ . Il rapporto di compressione è pertanto

$$\frac{mn}{k(m+n+1)}.$$

Memorizzare A_k tramite SVD diventa conveniente se k è sufficientemente piccolo, ovvero se il rapporto di compressione è molto maggiore di uno. Allo stesso tempo, dobbiamo verificare che l'energia di A catturata dall'approssimazione A_k sia "sufficiente", ovvero che l'errore (3.8) sia piccolo.

Osservazione 3.7.3. La decomposizione SVD di una matrice $A \in \mathbb{R}^{m \times n}$, per $m \geq n$, è spesso ottenuta mediante una matrice sinistra rettangolare $U_1 \in \mathbb{R}^{m \times n}$ anziché una matrice quadrata $U \in \mathbb{R}^{m \times m}$. In tal caso, invece di (3.4), otteniamo

$$A = U_1 \Sigma_1 V^T,$$

con $\Sigma_1 \in \mathbb{R}^{n \times n}$; tale decomposizione viene spesso indicata come decomposizione ai valori singolari in formato "thin" di una matrice. In questo caso,

$$U_1 = U(:, 1:n) = [\mathbf{u}_1 | \dots | \mathbf{u}_n] \in \mathbb{R}^{m \times n}$$

(si veda anche la Fig. 3.1) e

$$\Sigma_1 = \Sigma(1:n, 1:n) = \text{diag}(\sigma_1, \dots, \sigma_n) \in \mathbb{R}^{n \times n}.$$

Per comodità di notazione, nel seguito continueremo a usare la decomposizione SVD data da (3.4), sebbene le matrici U e Σ possano essere sostituite da U_1 e Σ_1 , rispettivamente, senza alcuna perdita di significato.

Il secondo motivo che giustifica l'introduzione della decomposizione ai valori singolari di una matrice riguarda la definizione di una matrice pseudo-inversa di ogni matrice rettangolare $A \in \mathbb{R}^{m \times n}$ e la conseguente possibilità di risolvere sistemi lineari di equazioni sia sovradeterminati che sottodeterminati. Nel primo caso, ricordiamo, una soluzione è possibile nel senso dei minimi quadrati, mediante il metodo della fattorizzazione QR (si veda la Sezione 2.2.6); nel secondo caso, non sono state ancora proposte strategie per la soluzione del problema. Vale innanzitutto la seguente definizione.

Definizione 3.7.2. Sia $A \in \mathbb{R}^{m \times n}$ una matrice di rango r . La matrice

$$A^\dagger = V \Sigma^\dagger U^T \in \mathbb{R}^{n \times m} \quad \text{con } \Sigma^\dagger := \text{diag}(\sigma_1^{-1}, \dots, \sigma_r^{-1}, 0, \dots, 0) \in \mathbb{R}^{n \times m}$$

è chiamata matrice pseudo-inversa (di Moore-Penrose) (o inversa generalizzata) della matrice A . Osserviamo che

$$A^\dagger = (A^T A)^{-1} A^T \quad \text{se } \text{rank}(A) = n < m$$

e che

$$A^\dagger = A^{-1} \quad \text{se } \text{rank}(A) = n = m.$$

Consideriamo innanzitutto il caso di un sistema sottodeterminato, in cui $A \in \mathbb{R}^{m \times n}$ con $n \leq m$, ovvero ci sono meno equazioni che incognite. In tal caso, la matrice A è solita avere rango pieno (per colonne),

ovvero $\text{rank}(A) = n$, e per ogni $\mathbf{b} \in \mathbb{R}^m$ esistono infinite soluzioni $\mathbf{x} \in \mathbb{R}^n$ – ovvero, non ci sono sufficienti valori in \mathbf{b} per determinare univocamente \mathbf{x} . Analogamente, nel caso sovradeterminato in cui $A \in \mathbb{R}^{m \times n}$ con $m \leq n$, tale matrice non può avere rango pieno per colonne, ovvero esistono vettori $\mathbf{b} \in \mathbb{R}^m$ tali che il sistema $A\mathbf{x} = \mathbf{b}$ non ammette soluzioni $\mathbf{x} \in \mathbb{R}^n$ – il sistema ammette infatti soluzione in senso classico solo a patto che $\mathbf{b} \in \text{range}(A)$.

Nel caso sovradeterminato, per un vettore $\mathbf{b} \in \mathbb{R}^m$ generico, abbiamo cercato (si veda la Sezione 2.2.6) una soluzione $\mathbf{x} \in \mathbb{R}^n$ nel senso dei minimi quadrati, ovvero tale da minimizzare la norma (al quadrato) del residuo, $\Phi(\mathbf{x}) = \|A\mathbf{x} - \mathbf{b}\|^2$; osserviamo come la soluzione nel senso dei minimi quadrati minimizzi anche $\|A\mathbf{x} - \mathbf{b}\|$. Nel caso sottodeterminato, possiamo cercare la soluzione di $A\mathbf{x} = \mathbf{b}$ di norma $\|\mathbf{x}\|$ minima. In entrambi i casi, la decomposizione SVD di A risulta di fondamentale importanza. Infatti, sfruttando la pseudo inversa A^\dagger di A , si ha che

$$A^\dagger = V\Sigma^\dagger U^T$$

e dunque

$$A^\dagger A = I \in \mathbb{R}^{n \times n}.$$

Da tale relazione è possibile determinare sia la soluzione nel senso dei minimi quadrati per un sistema $A\mathbf{x} = \mathbf{b}$ sovradeterminato, che la soluzione di norma minima per un sistema $A\mathbf{x} = \mathbf{b}$ sottodeterminato:

$$A^\dagger A \mathbf{x}^* = A^\dagger \mathbf{b}$$

da cui

$$\mathbf{x}^* = V\Sigma^\dagger U^T \mathbf{b}.$$

Infatti, risulta dalla definizione del sistema lineare, sostituendo l'espressione appena ottenuta per \mathbf{x}^* , che

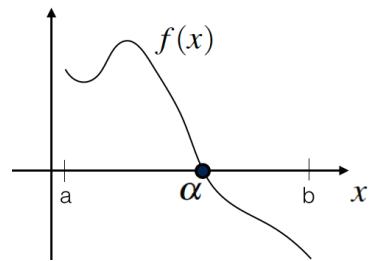
$$A\mathbf{x}^* = U\Sigma V^T V\Sigma^\dagger U^T \mathbf{b} = UU^T \mathbf{b}.$$

Osserviamo che UU^T non è necessariamente la matrice identità, ma rappresenta piuttosto la proiezione nel sottospazio generato dalle colonne di U , così che \mathbf{b} risulta una soluzione del sistema in senso classico a patto che $\mathbf{b} \in \text{range}(A)$.

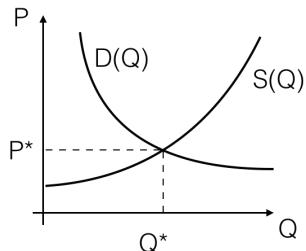
Capitolo 4

Equazioni Non Lineari

L'obiettivo consiste nell'*approssimare numericamente* lo zero $\alpha \in \mathbb{R}$ di una funzione $f(x)$ nell'intervallo $I = (a, b) \subseteq \mathbb{R}$. Il problema viene comunemente denominato come la soluzione numerica di un'*equazione non lineare*.



Esempio 4.0.1. *Domanda e offerta:* modelli microeconomici per la determinazione del prezzo di un bene in un mercato competitivo. Il prezzo unitario P di un bene varia fino a che viene raggiunto un equilibrio tra l'ammontare della domanda e dell'offerta di tale bene (Q).



Sia $P = S(Q)$ la funzione che rappresenta l'offerta sul mercato di tale bene, allora la quantità disponibile di tale bene cresce se il suo prezzo cresce. Sia $P = D(Q)$ la funzione domanda del bene, che cresce se il prezzo decresce. (Q^*, P^*) è il punto di equilibrio tale che $P^* = S(Q^*) = D(Q^*)$; posto $x = Q$, è necessario risolvere l'equazione non lineare $f(x) = S(x) - D(x) = 0$ per determinare l'equilibrio.

Esempio 4.0.2. Supponiamo di dover stabilire quale è la deformazione x di un provino costituito da materiale biologico, in risposta a un determinato livello di sforzo y , immaginando che una dimensione prevalga sulle altre. In presenza di un materiale elastico lineare, per il quale sforzi e deformazioni sono legati da una legge costitutiva lineare della forma $y = Ex$, essendo $E > 0$ il modulo di Young, la risposta a tale quesito sarebbe immediata, $x = y/E$. In presenza di un legame costitutivo non lineare, nella forma

$$y = E_0x + E_1e^{\gamma(x-x_0)} \quad \text{con } E_0, E_1, \gamma, x_0 > 0,$$

determinare x nota y presuppone la soluzione di un'equazione non lineare; in altri termini, occorre cercare lo zero della funzione $f(x) = E_0x + E_1e^{\gamma(x-x_0)} - y$.

Esempio 4.0.3. Un problema classico della dinamica delle popolazioni è determinare se una data popolazione, il cui numero di individui alla generazione $k \geq 0$ è indicato con $x^{(k)}$, raggiunge o meno uno stato di equilibrio. Immaginando che la dinamica della popolazione sia descritta da una legge della forma $x^{(k+1)} = \phi(x^{(k)})$, con $k \geq 0$, gli stati di equilibrio sono tutte e sole le soluzioni α dell'equazione $\alpha = \phi(\alpha)$. Un tale valore prende il nome di punto fisso della funzione ϕ . Semplici esempi di funzioni che esprimono la dinamica di una popolazione sono $\phi(x) = rx$, con $r > 0$ (tipico della situazione in cui una popolazione cresce indefinitamente, se $r > 1$, oppure si riduce progressivamente fino a estinguersi, se $r < 1$), oppure $\phi(x) = r/(1 + Kx)$, che descrive il caso in cui la crescita avviene in presenza di risorse limitate, per r e $K > 0$.

4.1 Metodo di Bisezione

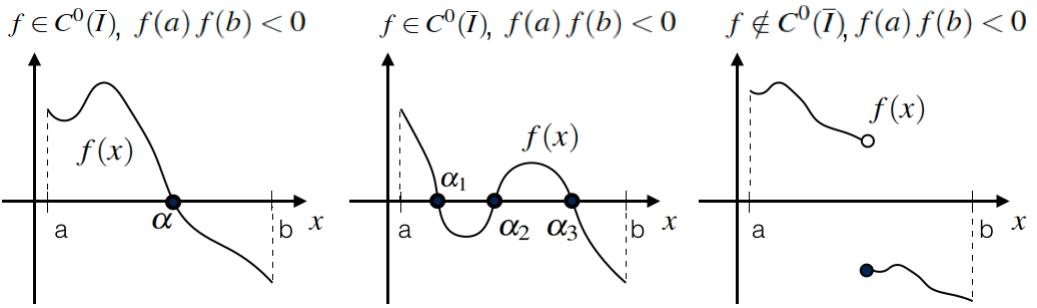
Consideriamo il *metodo di bisezione* per l'approssimazione dello zero $\alpha \in I$ di una funzione $f(x)$.

4.1.1 Costruzione del metodo di bisezione

Il metodo si basa sul seguente risultato teorico.

Teorema 4.1.1 (Zeri di una funzione continua). *Sia $f(x)$ una funzione continua nell'intervallo $I = (a, b)$, ovvero $f \in C^0(\bar{I}) \equiv C^0([a, b])$. Se $f(a) f(b) < 0$, allora esiste almeno uno zero $\alpha \in I$ di $f(x)$.*

Esempio 4.1.1. Illustriamo alcuni esempi di funzioni $f(x)$ tali che $f(a) f(b) < 0$.

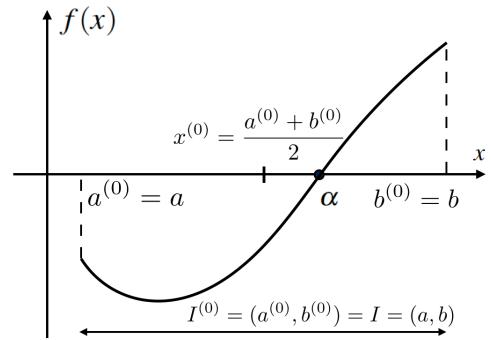


Assumiamo che esista un *unico* zero $\alpha \in I$ di una funzione $f \in C^0(\bar{I})$ tale che $f(a) f(b) < 0$. Allora, il metodo di bisezione cerca lo zero α approssimandolo con una sequenza di *punti medi* dei sottointervalli $I^{(k)}$ di I tale per cui la funzione $f(x)$ evidenzia un cambio di segno.

Esempio 4.1.2. Illustriamo il metodo di bisezione e il suo algoritmo attraverso i grafici seguenti.

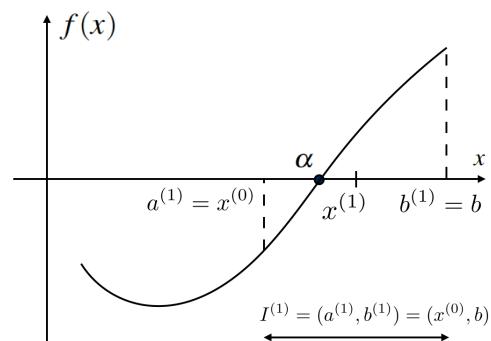
Passo 0.

$$\begin{aligned} I^{(0)} &= (a^{(0)}, b^{(0)}) = I = (a, b), \\ x^{(0)} &= \frac{a^{(0)} + b^{(0)}}{2} = \frac{a + b}{2} \end{aligned}$$



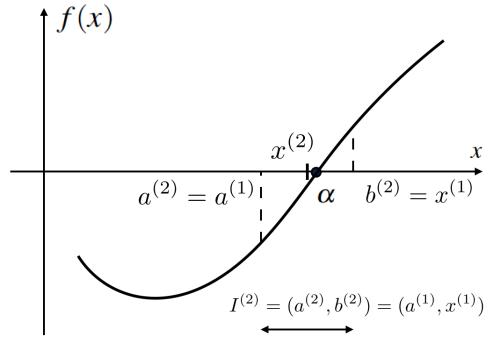
Passo 1.

$$\begin{aligned} \text{Dato che } f(x^{(0)}) f(b^{(0)}) &< 0: \\ a^{(1)} &= x^{(0)}, b^{(1)} = b, \\ I^{(1)} &= (a^{(1)}, b^{(1)}) = (x^{(0)}, b), \\ x^{(1)} &= \frac{a^{(1)} + b^{(1)}}{2} = \frac{x^{(0)} + b}{2}. \end{aligned}$$



Passo 2.

Dato che $f(x^{(1)}) f(a^{(1)}) < 0$:
 $a^{(2)} = a^{(1)}, b^{(2)} = x^{(1)}$,
 $I^{(2)} = (a^{(2)}, b^{(2)}) = (a^{(1)}, x^{(1)}),$
 $x^{(2)} = \frac{a^{(2)} + b^{(2)}}{2} = \frac{a^{(1)} + x^{(1)}}{2}$.



4.1.2 Algoritmo e proprietà del metodo

Riportiamo l'algoritmo e le proprietà del metodo di bisezione.

Algorithm 4.1: Metodo di bisezione

```

porre  $k = 0, a^{(0)} = a, b^{(0)} = b$  e  $x^{(0)} = \frac{a^{(0)} + b^{(0)}}{2}$ ;
for  $k = 1, 2, \dots$ , fino al soddisfacimento di un criterio d'arresto do
    if  $f(x^{(k-1)}) = 0$  then
        porre  $\alpha = x^{(k-1)}$  e terminare il ciclo;
    else
        if  $f(x^{(k-1)}) f(a^{(k-1)}) < 0$  then
            porre  $a^{(k)} = a^{(k-1)}$  e  $b^{(k)} = x^{(k-1)}$ ;
        end
        if  $f(x^{(k-1)}) f(b^{(k-1)}) < 0$  then
            porre  $a^{(k)} = x^{(k-1)}$  e  $b^{(k)} = b^{(k-1)}$ ;
        end
        porre  $x^{(k)} = \frac{a^{(k)} + b^{(k)}}{2}$ ;
    end
end
```

Per il sottointervallo $I^{(k)} = (a^{(k)}, b^{(k)})$ e il suo punto medio $x^{(k)} = \frac{a^{(k)} + b^{(k)}}{2}$, abbiamo che entrambi $x^{(k)}$ e $\alpha \in I^{(k)}$ per ogni $k \geq 0$. In aggiunta, essendo $|I^{(k)}| := b^{(k)} - a^{(k)} \equiv \frac{|I^{(k-1)}|}{2}$ per ogni $k \geq 1$, otteniamo:

$$|I^{(k)}| = \frac{|I^{(0)}|}{2^k} = \frac{b - a}{2^k} \quad \text{per ogni } k \geq 0.$$

Definiamo l'*errore* (di troncamento) associato al metodo di bisezione come $e^{(k)} := |x^{(k)} - \alpha|$. L'errore $e^{(k)} = |x^{(k)} - \alpha|$ può essere limitato dall'alto per mezzo della misura del sottointervallo $I^{(k+1)}$ per ogni $k \geq 0$; $|I^{(k+1)}|$ può essere usato per stimare l'errore e pertanto è noto come *stimatore dell'errore*. Si ottiene:

$$e^{(k)} \leq \tilde{e}^{(k)} := |I^{(k+1)}| = \frac{b - a}{2^{k+1}} \quad \text{per ogni } k \geq 0. \quad (4.1)$$

Questo implica che il metodo di bisezione è *convergente*; infatti, $\lim_{k \rightarrow +\infty} e^{(k)} = 0$ essendo $e^{(k)} \leq \tilde{e}^{(k)}$ per ogni $k \geq 0$ e $\lim_{k \rightarrow +\infty} \tilde{e}^{(k)} = \lim_{k \rightarrow +\infty} \frac{b - a}{2^{k+1}} = 0$.

Osservazione 4.1.1. Data la tolleranza $tol > 0$, è possibile calcolare il numero minimo di iterazioni del metodo di bisezione, indicato con k_{min} , tale che l'errore $e^{(k_{min})}$ risulti inferiore a tol , ovvero $e^{(k_{min})} < tol$. Infatti, da Eq. (4.1) si ottiene che k_{min} è il più piccolo intero tale per cui $\frac{b-a}{2^{k_{min}+1}} < tol$; ne consegue che

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

Sappiamo già che il metodo di bisezione è convergente. Tuttavia, vogliamo caratterizzare meglio tale convergenza. A questo scopo, richiamiamo la seguente definizione.

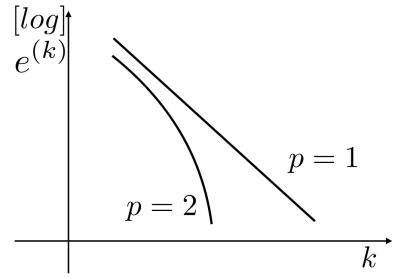
Definizione 4.1.1. Un metodo iterativo per l'approssimazione dello zero α della funzione $f(x)$ è convergente con ordine p se e solo se

$$\lim_{k \rightarrow +\infty} \frac{|x^{(k+1)} - \alpha|}{|x^{(k)} - \alpha|^p} = \mu, \quad (4.2)$$

dove $\mu > 0$ è un numero reale indipendente da k chiamato fattore di convergenza asintotico. Nel caso di convergenza lineare, ovvero per $p = 1$, è necessario che $0 < \mu < 1$.

Esempio 4.1.3. Illustriamo in un tipico grafico la sequenza di errori $e^{(k)}$ rispetto al numero di iterazioni k per ipotetici metodi iterativi con ordini di convergenza $p = 1$ e 2 .

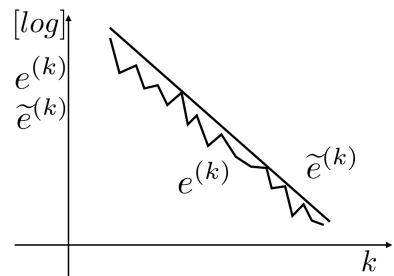
La scala logaritmica viene usata sull'asse dell'errore, mentre la scala lineare sull'asse del numero di iterazioni. Notiamo che la convergenza lineare ($p = 1$) è graficamente rappresentata da una linea retta la cui pendenza dipende dal fattore di convergenza asintotico μ . Una parabola si ottiene invece in corrispondenza della convergenza quadratica ($p = 2$).



Osservazione 4.1.2. L'errore per il metodo di bisezione è tende a zero, ma generalmente non in maniera monotona, ovvero $e^{(k+1)} \geq e^{(k)}$ per qualche $k \geq 0$; pertanto, anche se il metodo di bisezione è convergente, non è possibile stabilire un ordine di convergenza sulla base di Eq. (4.2). Anche il residuo assoluto $r^{(k)} := |f(x^{(k)})|$ non è in generale monotonicamente decrescente.

Esempio 4.1.4. Evidenziamo l'andamento tipico della sequenza di errori $e^{(k)}$ e degli stimatori degli errori $\tilde{e}^{(k)}$ ottenuti per il metodo di bisezione rispetto al numero di iterazioni k .

La scala logaritmica viene usata sull'asse dell'errore, mentre la scala lineare sull'asse del numero di iterazioni. Evidenziamo graficamente, sulla base dell'Osservazione 4.1.2, che un ordine di convergenza non può essere determinato per l'errore, mentre la convergenza è lineare per lo stimatore dell'errore.



4.1.3 Criterio d'arresto

Il criterio d'arresto dell'algoritmo di bisezione è basato sullo stimatore dell'errore $\tilde{e}^{(k)}$ di Eq. (4.1). Nell'Algoritmo 4.1, consideriamo il seguente criterio d'arresto in pseudocodice (al posto del ciclo *for*) indicando con tol la tolleranza selezionata e k_{max} il massimo numero di iterazioni consentito.

Algorithm 4.2: Metodo di bisezione. Criterio d'arresto

```

...;
while ( $\tilde{e}^{(k)} \geq tol \& k < k_{max}$ ) do
| ...
end

```

4.2 Metodi di Newton

Consideriamo i metodi di Newton e Newton modificato per approssimare lo zero α della funzione $f(x)$.

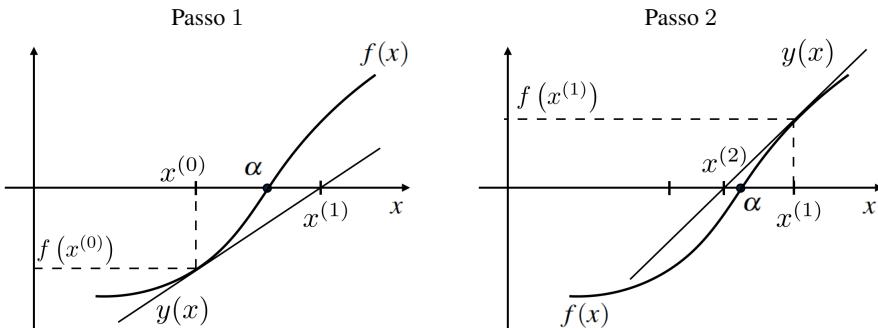
4.2.1 Metodo di Newton

Assumiamo che $f \in C^0(I)$ e che sia *differenziabile* nell'intervallo $I = (a, b) \subseteq \mathbb{R}$. Data una generica iterata $x^{(k)} \in I$, l'equazione della retta tangente alla curva $(x, f(x))$ nel punto $(x^{(k)}, f(x^{(k)}))$ è $y(x) = f(x^{(k)}) + f'(x^{(k)})(x - x^{(k)})$. Se assumiamo che $y(x^{(k+1)}) = 0$, allora calcoliamo l'iterata $x^{(k+1)}$ come:

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})} \quad \text{per ogni } k \geq 0, \quad (4.3)$$

data l'*iterata iniziale* $x^{(0)}$ e se $f'(x^{(k)}) \neq 0$ per ogni $k \geq 0$. L'Eq. (4.3) è detta *iterata di Newton*. Si ottiene lo zero α come limite delle sequenze di iterate $\{x^{(k+1)}\}_{k=0}^{+\infty}$ che risolvono l'equazione della tangente alla curva $(x, f(x))$ valutata in corrispondenza di ciascuna delle iterate $\{x^{(k)}\}_{k=0}^{+\infty}$.

Esempio 4.2.1. Illustriamo graficamente il metodo di Newton nelle figure seguenti in cui si evidenziano le prime due iterate del metodo.



Il *metodo di Newton* è applicabile a una funzione $f \in C^0(I)$ che sia differenziabile in I ; inoltre, dato $x^{(0)} \in I$, il metodo di Newton consiste nell'applicare in maniera sequenziale l'iterata di Newton (4.3), a condizione che $f'(x^{(k)}) \neq 0$ per ogni $k \geq 0$.

Algorithm 4.3: Metodo di Newton

```

porre  $k = 0$  e l'iterata iniziale  $x^{(0)}$ ;
while (il criterio d'arresto non è soddisfatto) do
| ...
|  $x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})};$ 
| porre  $k = k + 1$ ;
end

```

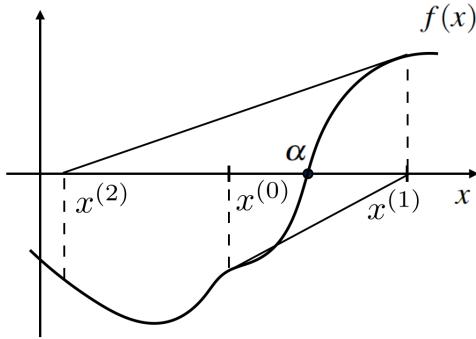
Osservazione 4.2.1. Assumiamo che $f \in C^2(I)$, allora l'espansione in serie di Taylor di $f(x)$ attorno a $x^{(k)}$ si scrive come

$$f(x^{(k+1)}) = f(x^{(k)}) + f'(x^{(k)}) \delta^{(k)} + O((\delta^{(k)})^2),$$

dove $\delta^{(k)} := x^{(k+1)} - x^{(k)}$ per $k \geq 0$ è la differenza tra iterate successive. Se $f(x^{(k+1)}) = 0$, allora il metodo di Newton rappresenta l'approssimazione di ordine uno dell'espansione in serie di Taylor di $f(x)$ attorno a $x^{(k)}$; osserviamo che tale assunzione è effettivamente soddisfatta se $\delta^{(k)}$ è "piccolo".

La scelta dell'iterata iniziale $x^{(0)}$ è cruciale per il successo del metodo di Newton. Infatti, è necessario scegliere $x^{(0)}$ "sufficientemente" vicino allo zero α . In pratica, la sequenza di iterate di Newton $\{x^{(k+1)}\}_{k=0}^{+\infty}$ potrebbe divergere, invece che convergere a α , qualora l'iterata iniziale $x^{(0)}$ non fosse "sufficientemente" vicino allo zero α . Dal momento che lo zero α non è noto, la scelta di $x^{(0)}$ potrebbe non essere banale; da questo punto di vista, il grafico della funzione o l'uso del metodo di *bisezione* sono estremamente utili per selezionare valori di $x^{(0)}$ "sufficientemente" vicino a α e dunque inizializzare il metodo di Newton.

Esempio 4.2.2. L'esempio seguente illustra come le iterate di Newton non convergano allo zero α ; ciò è dovuto al fatto che $x^{(0)}$ non è "sufficientemente" vicino a α .



Osservazione 4.2.2. Per funzioni di tipo affine (o lineari), ovvero nella forma $f(x) = cx + d$ con c e $d \in \mathbb{R}$, il metodo di Newton converge allo zero $\alpha = -\frac{d}{c}$ in una sola iterazione indipendentemente dalla scelta di $x^{(0)}$. Infatti, otteniamo da Eq. (4.3) che $x^{(1)} = x^{(0)} - \frac{f(x^{(0)})}{f'(x^{(0)})} = x^{(0)} - \frac{cx^{(0)} + d}{c} = -\frac{d}{c} = \alpha$, appunto per ogni $x^{(0)} \in \mathbb{R}$.

Caratterizziamo di seguito le proprietà di convergenza del metodo di Newton.

Proposizione 4.2.1 (Ordine di convergenza del metodo di Newton). Indichiamo con I_α un intorno di α . Se $f \in C^2(I_\alpha)$, $x^{(0)}$ è "sufficientemente" vicino ad α e $f'(\alpha) \neq 0$, allora il metodo di Newton è convergente con ordine 2 (quadraticamente) ad α , a condizione che $f'(x^{(k)}) \neq 0$ per ogni $k \geq 0$. In particolare, si ha:

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

sulla base di Eq. (4.2), $p = 2$ è l'ordine di convergenza e $\mu = \frac{1}{2} \frac{f''(\alpha)}{f'(\alpha)}$ il fattore di convergenza asintotico.

Dimostrazione. La dimostrazione è basata sulla interpretazione del metodo di Newton come metodo delle iterazioni di punto fisso; vedi Sez. 4.3.5. \square

Definizione 4.2.1. Sia $f \in C^m(I_\alpha)$, con $m \in \mathbb{N}$ tale che $m \geq 1$. Lo zero $\alpha \in I_\alpha$ è uno zero di molteplicità m se $f^{(i)}(\alpha) = 0$ per ogni $i = 0, \dots, m-1$ e $f^{(m)}(\alpha) \neq 0$. Se la condizione precedente è soddisfatta per $m = 1$, lo zero α è detto semplice, altrimenti è detto multiplo.

Proposizione 4.2.2 (Ordine di convergenza del metodo di Newton, zero multiplo). Se $f \in C^2(I_\alpha) \cap C^m(I_\alpha)$ e $x^{(0)}$ è “sufficientemente” vicino allo zero α di molteplicità $m > 1$, allora il metodo di Newton è convergente con ordine 1 (linearmente) a α , a condizione che $f'(x^{(k)}) \neq 0$ per ogni $k \geq 0$. In particolare, sulla base di Eq. (4.2), si ha:

$$\lim_{k \rightarrow +\infty} \frac{x^{(k+1)} - \alpha}{x^{(k)} - \alpha} = \mu,$$

con $p = 1$ l’ordine di convergenza e $\mu \in (0, 1)$ il fattore di convergenza asintotico.

Osservazione 4.2.3. Se lo zero α è semplice, ovvero $m = 1$, il metodo di Newton converge almeno quadraticamente sulla base di Proposizione 4.2.1. Al contrario, se lo zero α è multiplo ($m > 1$), il metodo di Newton converge solo linearmente sulla base di Proposizione 4.2.2. Osserviamo che, in generale, tanto più alto è l’ordine di convergenza p , tanto più basso sarà il numero di iterazioni necessario per raggiungere un valore desiderato dell’errore, ovvero il metodo risulterà essere più efficiente.

Esempi di convergenza lineare e quadratica sono graficamente illustrati nell’Esempio 4.1.3.

4.2.2 Metodo di Newton modificato

Assumiamo che $f \in C^m(I_\alpha)$, con $\alpha \in I_\alpha$ e $m \geq 1$ la molteplicità di α . L’iterata k -esima del *metodo di Newton modificato* si scrive come:

$$x^{(k+1)} = x^{(k)} - m \frac{f(x^{(k)})}{f'(x^{(k)})} \quad \text{per ogni } k \geq 0, \quad (4.4)$$

data l’iterata iniziale $x^{(0)}$ e a condizione che $f'(x^{(k)}) \neq 0$ per ogni $k \geq 0$. Sulla base dell’Algoritmo 4.3, si ottiene il seguente per il metodo di Newton modificato.

Algorithm 4.4: Metodo di Newton modificato

```

selezionare il valore  $m$ ;
porre  $k = 0$  e l’iterata iniziale  $x^{(0)}$ ;
while (il criterio d’arresto non è soddisfatto) do
     $x^{(k+1)} = x^{(k)} - m \frac{f(x^{(k)})}{f'(x^{(k)})};$ 
    porre  $k = k + 1$ ;
end
```

Le proprietà di convergenza del metodo di Newton modificato sono caratterizzate come segue.

Proposizione 4.2.3 (Ordine di convergenza del metodo di Newton modificato). Se $f \in C^2(I_\alpha) \cap C^m(I_\alpha)$, dove $m \geq 1$ è la molteplicità dello zero $\alpha \in I_\alpha$ e $x^{(0)}$ è “sufficientemente” vicino ad α , allora il metodo di Newton modificato è convergente con ordine 2 (quadraticamente) ad α , a condizione che $f'(x^{(k)}) \neq 0$ per ogni $k \geq 0$.

Esempio 4.2.3. Approssimiamo lo zero $\alpha = 0$ della funzione $f(x) = \sin^m(x)$ nell’intervallo $I = (-\frac{\pi}{2}, \frac{\pi}{2})$, dove $m = 1, 2, 3, \dots$. A tal fine, consideriamo i metodi di Newton e di Newton modificato. Osserviamo che $f'(x) = m \sin^{m-1}(x) \cos(x)$, per cui $f'(\alpha) = 1$ se $m = 1$ e $f'(\alpha) = 0$ per $m \geq 2$; lo zero α è semplice per $m = 1$, ma multiplo (con molteplicità m) per $m \geq 2$. Se poniamo l’iterata iniziale $x^{(0)} = \frac{\pi}{6}$, la prima iterata del metodo di Newton corrisponde a $x^{(1)} = \frac{\pi}{6} - \frac{1}{\sqrt{3m}}$, per cui $x^{(1)}$ è tanto più lontano da α (e quindi tanto più vicino ad $x^{(0)}$),

tanto più grande è m . Al contrario, per il metodo di Newton modificato si ottiene da Eq. (4.4) che $x^{(1)} = \frac{\pi}{6} - \frac{1}{\sqrt{3}}$, indipendentemente dal valore di $m \geq 1$.

Osservazione 4.2.4. *Il metodo di Newton modificato richiede la conoscenza a priori della molteplicità m dello zero α . In alternativa, quest'ultima può essere stimata sulla base di opportuni metodi numerici o con metodi di tipo adattivo.*

Un esempio è appunto fornito dal *metodo di Newton adattivo*, la cui generica iterata è:

$$x^{(k+1)} = x^{(k)} - m^{(k)} \frac{f(x^{(k)})}{f'(x^{(k)})} \quad \text{con } m^{(k)} = \frac{x^{(k-1)} - x^{(k-2)}}{2x^{(k-1)} - x^{(k)} - x^{(k-2)}} \quad \text{per ogni } k \geq 2; \quad (4.5)$$

$m^{(k)}$ rappresenta un'approssimazione della molteplicità m dello zero α . Osserviamo infatti che $\lim_{k \rightarrow +\infty} \frac{x^{(k)} - \alpha}{x^{(k-1)} - \alpha} = \lambda$, dove per il metodo di Newton $\lambda = 1 - \frac{1}{m}$ (dimostreremo tale risultato in Sez. 4.3.5). Si verifica che

$$\lim_{k \rightarrow +\infty} \frac{x^{(k)} - \alpha}{x^{(k-1)} - \alpha} = \lim_{k \rightarrow +\infty} \lambda^{(k)} = \lambda \quad \text{dove } \lambda^{(k)} = \frac{x^{(k)} - x^{(k-1)}}{x^{(k-1)} - x^{(k-2)}},$$

infatti

$$\begin{aligned} \lim_{k \rightarrow +\infty} \lambda^{(k)} &= \lim_{k \rightarrow +\infty} \frac{(x^{(k)} - \alpha) - (x^{(k-1)} - \alpha)}{(x^{(k-1)} - \alpha) - (x^{(k-2)} - \alpha)} = \lim_{k \rightarrow +\infty} \frac{(x^{(k)} - \alpha)/(x^{(k-1)} - \alpha) - 1}{1 - (x^{(k-2)} - \alpha)/(x^{(k-1)} - \alpha)} \\ &= \frac{\lim_{k \rightarrow +\infty} (x^{(k)} - \alpha)/(x^{(k-1)} - \alpha) - 1}{1 - 1/\lim_{k \rightarrow +\infty} (x^{(k-1)} - \alpha)/(x^{(k-2)} - \alpha)} = \frac{\lambda - 1}{1 - 1/\lambda} = \lambda. \end{aligned}$$

Posto $\lambda^{(k)} = 1 - \frac{1}{m^{(k)}}$, si deduce la scelta di $m^{(k)}$. Tipicamente, il metodo di Newton adattivo converge a uno zero multiplo più rapidamente del metodo di Newton, ma più lentamente del metodo di Newton modificato; $m^{(k)}$ fornisce inoltre una stima della molteplicità m dello zero α . L'algoritmo del metodo di Newton adattivo la cui iterata è riportata in Eq. (4.5) può però soffrire di instabilità numeriche; in pratica, per evitare tale inconveniente, l'aggiornamento di $m^{(k)}$ viene eseguito solo quando la differenza tra iterate successive è inferiore a una certa tolleranza.

4.2.3 Criteri d'arresto per i metodi di Newton

Consideriamo diversi criteri d'arresto per il metodo di Newton. Dato che lo zero α non è in generale noto, l'errore $e^{(k)} = |x^{(k)} - \alpha|$ risulta anch'esso non calcolabile; pertanto, è necessario introdurre un opportuno *stimatore dell'errore* (indicatore dell'errore) $\tilde{e}^{(k)}$ tale che $\tilde{e}^{(k)} \simeq e^{(k)}$. In riferimento per esempio agli Algoritmi 4.3 e 4.4 dei metodi di Newton e Newton modificato, le iterazioni vengono fermate per $k = k_{min}$ tale che $\tilde{e}^{(k_{min})} < tol$, dove tol è una tolleranza prescritta, oppure quando il massimo numero di iterazioni consentito viene raggiunto; si veda per esempio l'Algoritmo 4.2.

Consideriamo innanzitutto il criterio d'arresto basato sulla *differenza tra iterate successive*, per cui tale stimatore dell'errore è scelto come:

$$\tilde{e}^{(k)} = \begin{cases} |\delta^{(k-1)}| & \text{se } k \geq 1 \\ tol + 1 & \text{se } k = 0 \end{cases} \quad \text{con } \delta^{(k)} := x^{(k+1)} - x^{(k)} \quad \text{per } k \geq 0.$$

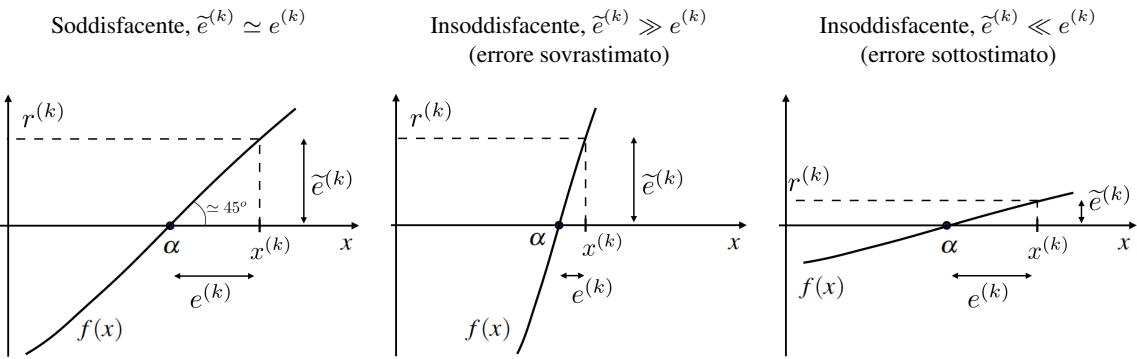
Tale criterio è *soddisfacente* se lo zero α è semplice; questo può essere mostrato interpretando il metodo di Newton come metodo delle iterazioni di punto fisso; si veda la Sez. 4.3.5.

Un altro criterio è basato sul *residuo* (assoluto), per cui:

$$\tilde{e}^{(k)} = |r^{(k)}| \quad \text{con } r^{(k)} := f(x^{(k)}) \quad \text{per } k \geq 0.$$

Questo criterio è *soddisfacente* se $|f'(x)| \simeq 1$ per $x \in I_\alpha$ un intorno di α : in tal caso si ha $\tilde{e}^{(k)} \simeq e^{(k)}$. Al contrario, il criterio è *insoddisfacente* se $|f'(x)| \gg 1$ oppure se $|f'(x)| \simeq 0$ per $x \in I_\alpha$. Nel caso specifico, se $|f'(x)| \gg 1$ per $x \in I_\alpha$, allora l'errore è *sovristimato* dallo stimatore dell'errore ($\tilde{e}^{(k)} \gg e^{(k)}$), per cui vengono effettuate più iterazioni di Newton del necessario; pertanto il criterio di rivela come inefficiente. Se invece $|f'(x)| \simeq 0$ per $x \in I_\alpha$, allora l'errore è *sottostimato* dallo stimatore dell'errore ($\tilde{e}^{(k)} \ll e^{(k)}$), per cui le iterazioni di Newton sono prematuramente terminate dato che l'errore risulta essere più grande di quanto indicato appunto dal suo stimatore.

Esempio 4.2.4. Gli esempi seguenti illustrano graficamente le situazioni per cui il criterio d'arresto basato sul *residuo* risulta essere soddisfacente o insoddisfacente.



4.2.4 Metodi di quasi–Newton e inesatti

I metodi di Newton e Newton modificato richiedono il calcolo e la valutazione della derivata prima della funzione $f(x)$; si vedano le Eq. (4.3) e (4.4). Tuttavia, in molti casi di interesse pratico, la valutazione di $f'(x)$ potrebbe essere "difficile" o computazionalmente costosa. Pertanto, facendo riferimento per esempio all'iterata di Newton (4.3), $f'(x^{(k)})$ può essere approssimato per mezzo di una quantità computazionalmente più conveniente $q^{(k)} \simeq f'(x^{(k)})$ che appunto approssima $f'(x^{(k)})$. I *metodi di quasi–Newton o inesatti*¹ sono basati sull'uso di valori approssimati di $f'(x^{(k)})$. L'iterata generica del metodo di quasi–Newton si scrive come:

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{q^{(k)}} \quad \text{per ogni } k \geq 0;$$

la scelta di $q^{(k)}$ determina il metodo specifico. Consideriamo i seguenti casi:

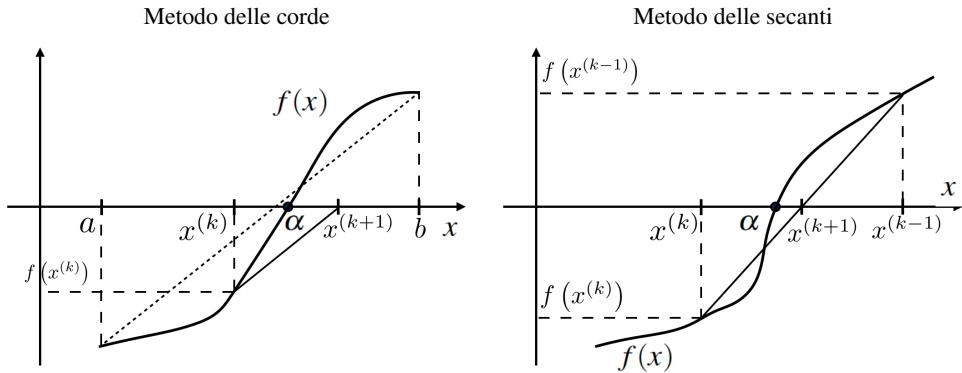
- per $q^{(k)} \equiv f'(x^{(k)})$, otteniamo il metodo di Newton;
- per $q^{(k)} \equiv \frac{f'(x^{(k)})}{m}$, otteniamo il metodo di Newton modificato (m è la molteplicità di α);
- per $q^{(k)} = q_C = \frac{f(b) - f(a)}{b - a}$ per ogni $k \geq 0$, con $\alpha \in (a, b)$, otteniamo il *metodo delle corde*;
- per $q^{(k)} = \frac{f(x^{(k)}) - f(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}$ per ogni $k \geq 1$, otteniamo il *metodo delle secanti*².

Il metodo delle corde converge linearmente ($p = 1$) se lo zero α è semplice e sotto certe condizioni su q_C , mentre può convergere come non convergere se lo zero α è multiplo. Il metodo delle secanti converge con ordine $p = \frac{1 + \sqrt{5}}{2} \simeq 1.6$ se lo zero α è semplice, mentre converge linearmente $p = 1$ se lo zero α è multiplo ($m > 1$).

¹La denominazione qui utilizzata dei metodi inesatti e di quasi–Newton non è del tutto precisa.

²Per il metodo delle secanti, $q^{(0)}$ può essere scelto come per il metodo delle corde.

Esempio 4.2.5. I seguenti esempi illustrano graficamente i metodi delle *corde* e delle *secanti* alla generica iterata $x^{(k)}$.



4.2.5 Metodo di Newton per sistemi di equazioni non lineari

Il metodo di Newton può essere utilizzato per approssimare la soluzione di sistemi di equazioni non lineari. Data $\mathbf{F} : \mathbb{R}^n \rightarrow \mathbb{R}^n$, per qualche $n \geq 1$, il problema consiste nel trovare il vettore $\alpha \in \mathbb{R}^n$ tale che $\mathbf{F}(\alpha) = 0$. Più specificamente, si hanno:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad \text{e} \quad \mathbf{F}(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_n(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} f_1(x_1, \dots, x_n) \\ \vdots \\ f_n(x_1, \dots, x_n) \end{pmatrix}.$$

Definizione 4.2.2. Sia $\mathbf{F} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ differenziabile in $I_{\mathbf{x}} \subseteq \mathbb{R}^n$ un intorno di $\mathbf{x} \in \mathbb{R}^n$, allora la Jacobiana di \mathbf{F} in \mathbf{x} è la matrice $J_{\mathbf{F}} : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times n}$ tale che $(J_{\mathbf{F}}(\mathbf{x}))_{ij} = \frac{\partial f_i}{\partial x_j}(\mathbf{x})$ per $i, j = 1, \dots, n$.

Il *metodo di Newton* è applicabile a un sistema di equazioni $\mathbf{F} \in C^0(I_{\alpha})$ che sia differenziabile in $I_{\alpha} \subseteq \mathbb{R}^d$, un intorno di α . Data l'iterata iniziale $x^{(0)} \in I_{\alpha}$, il metodo di Newton consiste nell'applicare sequenzialmente la seguente iterata di Newton:

$$\text{risolvere } J_{\mathbf{F}}(\mathbf{x}^{(k)}) \boldsymbol{\delta}^{(k)} = -\mathbf{F}(\mathbf{x}^{(k)}) \quad \text{e} \quad \text{porre } \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \boldsymbol{\delta}^{(k)} \quad \text{per ogni } k \geq 0, \quad (4.6)$$

a condizione che $\det(J_{\mathbf{F}}(\mathbf{x}^{(k)})) \neq 0$ per ogni $k \geq 0$. Osserviamo che l'iterata di Newton (4.6) può essere ottenuta come espansione in serie di Taylor al primo ordine di $\mathbf{F}(\mathbf{x})$ attorno a $\mathbf{x}^{(k)}$, ovvero come $\mathbf{F}(\mathbf{x}^{(k)}) + J_{\mathbf{F}}(\mathbf{x}^{(k)})(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) = 0$.

L'algoritmo del metodo di Newton deve impiegare un opportuno il criterio d'arresto. Come visto nel caso di una funzione non lineare, può essere utilizzato il criterio basato sul *residuo*, ovvero $\tilde{e}^{(k)} = \|\mathbf{F}(\mathbf{x}^{(k)})\| < tol$ per $k \geq 0$, oppure quello basato sulla *differenza tra iterate successive*, ovvero $\tilde{e}^{(k)} = \|\boldsymbol{\delta}^{(k-1)}\| < tol$ per $k \geq 1$, essendo tol una tolleranza assegnata.

Osservazione 4.2.5. Ad ogni iterata del metodo di Newton (4.6), è necessario assemblare e risolvere un sistema lineare, a meno che $n = 1$ per cui $J_{\mathbf{F}}(\mathbf{x}^{(k)}) \equiv f'(\mathbf{x}^{(k)})$. Per $n \gg 1$, tali operazioni possono rivelarsi computazionalmente costose e richiamano l'uso di metodi numerici per la risoluzione di sistemi lineari affrontati nel Cap. 2.

Riportiamo il seguente risultato inerente la convergenza del metodo di Newton per sistemi di equazioni non lineari; tale risultato rappresenta la generalizzazione del caso di una funzione non lineare di Prop. 4.2.1.

Proposizione 4.2.4. Siano $\mathbf{F} \in C^2(I_\alpha)$, con $I_\alpha \subseteq \mathbb{R}^n$ un intorno di α , $\mathbf{x}^{(0)} \in \mathbb{R}^n$ "sufficientemente" vicino ad α , e $\det(J_{\mathbf{F}}(\alpha)) \neq 0$, allora il metodo di Newton converge ad α con ordine $p = 2$, a condizione che $\det(J_{\mathbf{F}}(\mathbf{x}^{(k)})) \neq 0$ per ogni $k \geq 0$.

Esempio 4.2.6. Consideriamo il sistema di equazioni non lineari

$$\mathbf{F}(\mathbf{x}) = \begin{pmatrix} \sin(x_1 x_2) + x_2 \\ x_1 + x_2 - \frac{1}{2} e^{-x_1 x_2} \end{pmatrix}$$

e lo zero $\alpha = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$. La sua matrice Jacobiana è $J_{\mathbf{F}}(\mathbf{x}) = \begin{bmatrix} x_2 \cos(x_1 x_2) & x_1 \cos(x_1 x_2) + 1 \\ 1 + \frac{x_2}{2} e^{-x_1 x_2} & 1 + \frac{x_1}{2} e^{-x_1 x_2} \end{bmatrix}$.

Essendo $\det(J_{\mathbf{F}}(\alpha)) = -\frac{3}{2} \neq 0$, ci si attende una convergenza almeno quadratica ad α se $\mathbf{x}^{(0)}$ viene scelto "sufficientemente" vicino ad α .

4.3 Iterazioni di Punto Fisso

Consideriamo il metodo delle iterazioni di punto fisso, sia per trovare il *punto fisso* di una funzione di iterazione, sia come metodo per risolvere equazioni non lineari.

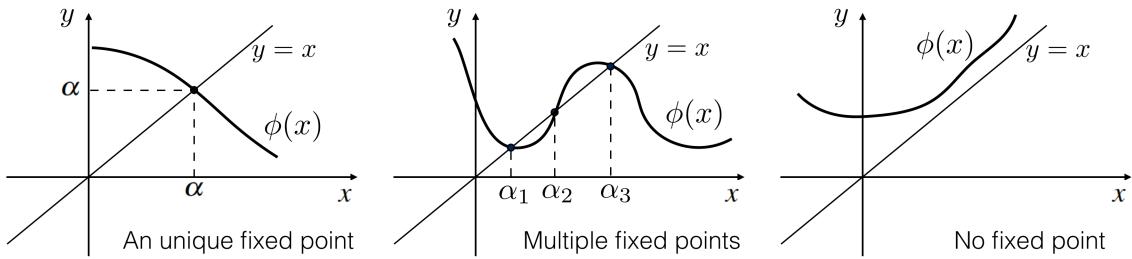
4.3.1 Equazioni non lineari, zeri, punti fissi e funzioni di iterazione

Data una funzione $f : \mathbb{R} \rightarrow \mathbb{R}$, l'obiettivo consiste nel determinare lo zero α (ovvero tale per cui $f(\alpha) = 0$). A questo scopo, trasformiamo il problema di trovare lo zero α di $f(x)$ in quello di risolvere un problema delle *iterazioni di punto fisso*.

Definizione 4.3.1. Data la funzione di iterazione $\phi : [a, b] \subseteq \mathbb{R} \rightarrow \mathbb{R}$, diciamo che $\alpha \in \mathbb{R}$ è un punto fisso di ϕ se e solo se $\phi(\alpha) \equiv \alpha$.

Esempio 4.3.1. Per $\phi(x) = \cos(x)$ nell'intervallo $[0.1, 1.1]$, si ha il punto fisso $\alpha = \cos(\alpha) \simeq 0.7391$.

Esempio 4.3.2. Illustriamo graficamente i punti fissi di alcune funzioni di iterazione.



L'obiettivo consiste nel trovare lo zero α della funzione non lineare $f(x)$. Trasformiamo questo problema in un problema di iterazioni di punto fisso selezionando in maniera opportuna una funzione di iterazione $\phi(x)$ tale che $f(\alpha) = 0$ se e solo se $\phi(\alpha) = \alpha$ per $\alpha \in [a, b]$. Osserviamo che esistono diverse funzioni di iterazione $\phi(x)$ che assolvono tale compito e molteplici modi per derivarle.

Il modo più semplice per ottenere $\phi(x)$ da $f(x)$ è basato sui seguenti passaggi. Dato che $f(\alpha) = 0$, abbiamo $f(\alpha) + \alpha = \alpha$, per cui possiamo porre $\phi(x) = f(x) + x$. Osserviamo però che questa è molto spesso una scelta inadeguata per la funzione di iterazione.

Esempio 4.3.3. Consideriamo $f(x) = 2x^2 - x - 1$ per cui siamo interessati allo zero $\alpha = 1$. Una possibilità consiste in porre $\phi_1(x) = f(x) + x = 2x^2 - 1$. Una seconda scelta può essere ricavata ponendo $f(x) = 0$, per cui si ottiene $x^2 = \frac{x+1}{2}$ e quindi $x = \pm\sqrt{\frac{x+1}{2}}$; in tal caso, possiamo prendere $\phi_2(x) = \sqrt{\frac{x+1}{2}}$.

4.3.2 Algoritmo delle iterazioni di punto fisso

Riportiamo l'algoritmo delle iterazioni di punto fisso, appunto basato dell'iterata di punto fisso:

$$x^{(k+1)} = \phi(x^{(k)}) \quad k \geq 0, \quad (4.7)$$

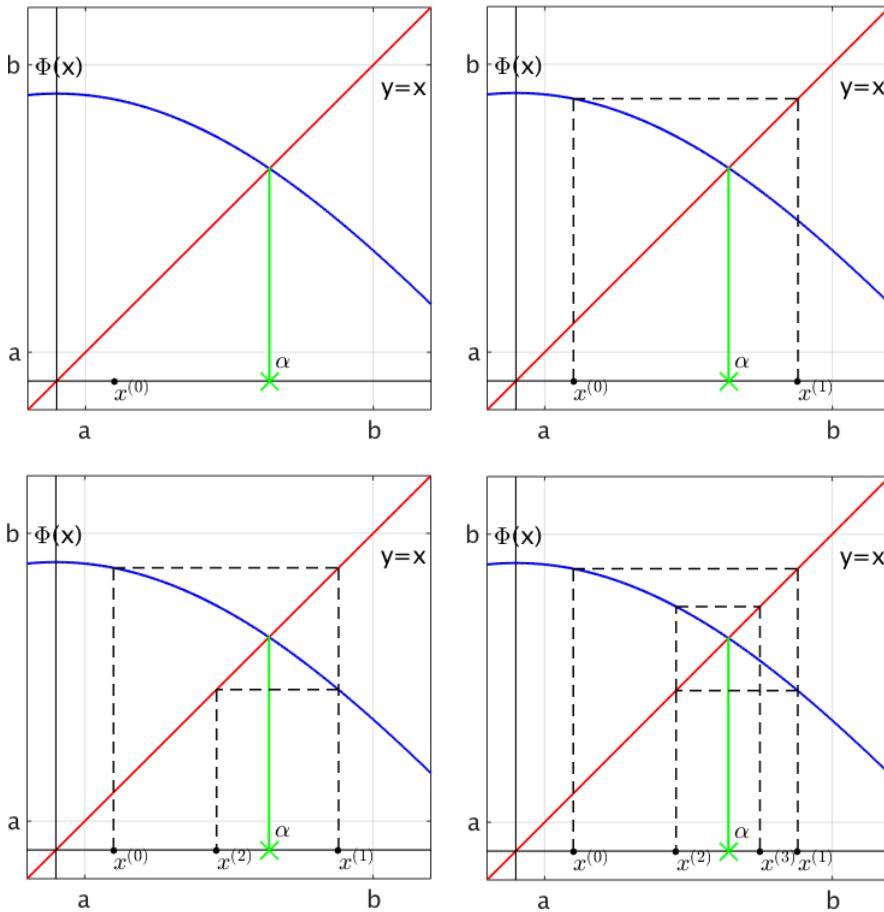
per una qualche iterata iniziale $x^{(0)}$.

Algorithm 4.5: Iterazioni di punto fisso

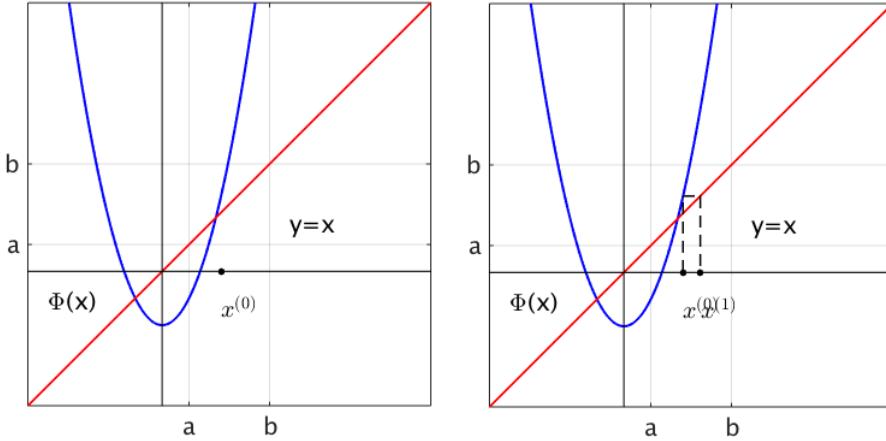
```

porre k = 0 e l'iterata iniziale x(0);
while (il criterio d'arresto non è soddisfatto) do
    | x(k+1) = φ(x(k));
    | porre k = k + 1;
end
```

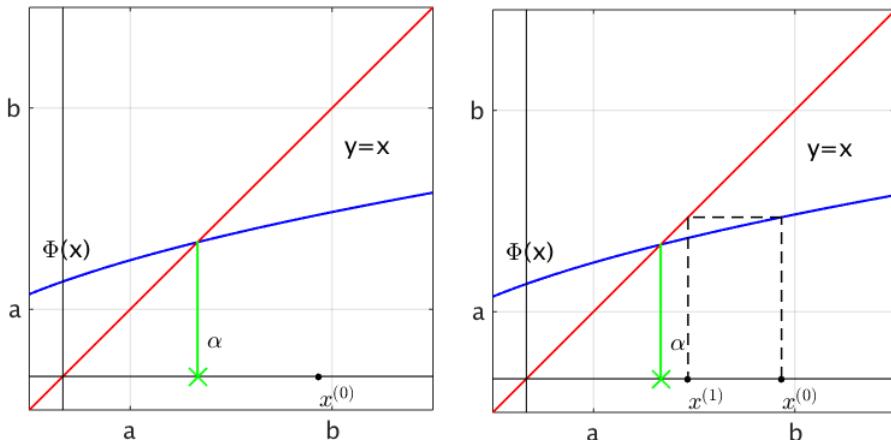
Esempio 4.3.4. Illustriamo graficamente l'algoritmo delle iterazioni di punto fisso nel seguito. Consideriamo il caso $\phi(x) = \cos(x)$, con $a = 0.1$, $b = 1.1$ e $x^{(0)} = 0.2$, per cui osserviamo che l'algoritmo è convergente a $\alpha = \cos(\alpha) \simeq 0.7391$.



Consideriamo ora $\phi(x) = 2x^2 - 1$, con $a = 0.5$, $b = 2$ e $x^{(0)} = 1.1$, per cui l'algoritmo risulta divergente dal punto fisso $\alpha = 1$; osserviamo che $\phi(x)$ corrisponde alla funzione di iterazione $\phi_1(x)$ dell'Esempio 4.3.3.



Infine, consideriamo $\phi(x) = \sqrt{\frac{1+x}{2}}$, con $a = 0.5$, $b = 2$ e $x^{(0)} = 1.9$, per cui l'algoritmo converge ad $\alpha = 1$; in tal caso, $\phi(x)$ corrisponde a $\phi_2(x)$ dell'Esempio 4.3.3.



4.3.3 Proprietà di convergenza del metodo delle iterazioni di punto fisso

Osserviamo che, a patto di richiedere che ϕ sia una funzione continua, se le iterazioni di punto fisso convergono allora il limite α della successione $\{x^{(k)}\}_{k \geq 0}$ è un punto fisso di ϕ . Infatti,

$$\alpha = \lim_{k \rightarrow +\infty} x^{(k)} = \lim_{k \rightarrow +\infty} \phi(x^{(k)}) = \phi \left(\lim_{k \rightarrow +\infty} x^{(k)} \right) = \phi(\alpha).$$

Riportiamo ora le proprietà che dobbiamo richiedere alla funzione di iterazione $\phi(x)$ per poter garantire l'esistenza e l'unicità del punto fisso α in un dato intervallo; inoltre, discutiamo le proprietà di convergenza del metodo delle iterazioni di punto fisso.

Proposizione 4.3.1 (Convergenza (globale) in un intervallo). *Consideriamo la funzione di iterazione $\phi : \mathbb{R} \rightarrow \mathbb{R}$ e le iterate di punto fisso di Eq. (4.7).*

1. *Se $\phi \in C^0([a, b])$ e $\phi(x) \in [a, b]$ per ogni $x \in [a, b]$, allora esiste almeno un punto fisso $\alpha \in [a, b]$ di $\phi(x)$.*
2. *Se, oltre alle ipotesi del punto (1), esiste una costante $L \in [0, 1)$ tale che $|\phi(x_1) - \phi(x_2)| \leq L |x_1 - x_2|$ per ogni x_1 e $x_2 \in [a, b]$, allora il punto fisso α è unico in $[a, b]$ e l'algoritmo delle iterazioni di punto fisso converge ($\lim_{k \rightarrow +\infty} x^{(k)} = \alpha$) per ogni iterata iniziale $x^{(0)} \in [a, b]$.*

Dimostrazione. (1) Mostriamo l'esistenza di $\alpha \in [a, b]$ sulla base delle ipotesi (1). Introduciamo una funzione $g(x) = \phi(x) - x$; si ha $g(\alpha) = 0$. Dato che $\phi \in C^0([a, b])$, anche $g \in C^0([a, b])$. In aggiunta, essendo $\phi(x) \in [a, b]$ per ogni $x \in [a, b]$, abbiamo $g(a) = \phi(a) - a \geq 0$ e $g(b) = \phi(b) - b \leq 0$, per cui $g(a) g(b) \leq 0$. Dato che $g(x)$ soddisfa le ipotesi del Teorema 4.1.1, esiste almeno uno zero α di $g(x)$ in $[a, b]$; quest'ultimo è anche un punto fisso di $\phi(x)$ (infatti, $g(\alpha) = \phi(\alpha) - \alpha = 0$).

(2) Mostriamo l'unicità di $\alpha \in [a, b]$ e la convergenza del metodo per ogni $x^{(0)} \in [a, b]$ sulla base delle ipotesi (2). Assumiamo per assurdo, che esistano due distinti punti fissi $\alpha_1 \neq \alpha_2$ tali che $\phi(\alpha_1) = \alpha_1$ e $\phi(\alpha_2) = \alpha_2$. Sulla base di questa assunzione, otteniamo $0 < |\alpha_1 - \alpha_2| = |\phi(\alpha_1) - \phi(\alpha_2)| \leq L |\alpha_1 - \alpha_2|$; dato che $L < 1$ per ipotesi, abbiamo $0 < |\alpha_1 - \alpha_2| < |\alpha_1 - \alpha_2|$, che è evidentemente assurdo. Pertanto, $\alpha_1 \equiv \alpha_2 = \alpha$, cioè il punto fisso è unico. Per quanto riguarda la convergenza del metodo, osserviamo che l'errore $e^{(k+1)} = |x^{(k+1)} - \alpha| = |\phi(x^{(k)}) - \phi(\alpha)| \leq |x^{(k)} - \alpha| = L e^{(k)}$. Per ricorsione, $e^{(k)} \leq L^k e^{(0)}$ per ogni $k \geq 0$; essendo $L < 1$, otteniamo $\lim_{k \rightarrow +\infty} e^{(k)} = 0$, ovvero che il metodo è convergente per ogni $x^{(0)} \in [a, b]$. \square

Sotto le ipotesi de risultato precedente (Proposizione 4.3.1) e per quanto riguarda la convergenza del metodo, osserviamo che l'errore

$$e^{(k+1)} = |x^{(k+1)} - \alpha| = |\phi(x^{(k)}) - \phi(\alpha)| \leq |x^{(k)} - \alpha| = L e^{(k)} \quad \text{per ogni } k \geq 0.$$

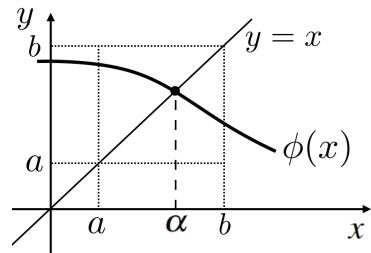
Per ricorsione, abbiamo dunque

$$e^{(k)} \leq L^k e^{(0)} \quad \text{per ogni } k \geq 0.$$

Essendo $L < 1$, otteniamo $\lim_{k \rightarrow +\infty} e^{(k)} = 0$, ovvero che il metodo è convergente per ogni $x^{(0)} \in [a, b]$. Se, oltre alle ipotesi di Proposizione 4.3.1 abbiamo $\phi(x) \in C^1([a, b])$, allora possiamo scegliere

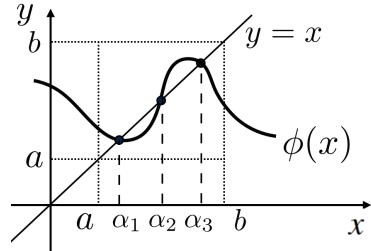
$$L = \max_{x \in [a, b]} |\phi'(x)|.$$

Esempio 4.3.5. Illustriamo i risultati di Proposizione 4.3.1 tramite i seguenti esempi.

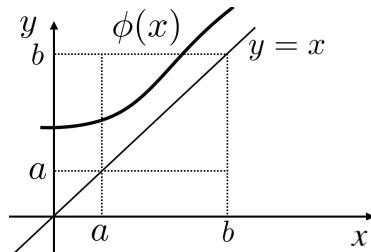


Le ipotesi (1) e (2) di Proposizione 4.3.1 sono soddisfatte; pertanto, esiste un unico punto fisso $\alpha \in [a, b]$ e il metodo converge ad α per ogni $x^{(0)} \in [a, b]$.

Le ipotesi (1) sono soddisfatte, ma non le ipotesi (2) di Proposizione 4.3.1; pertanto, possiamo solo garantire che esista almeno un punto fisso $\alpha \in [a, b]$.



Le ipotesi (1) e (2) di Proposizione 4.3.1 non sono soddisfatte, pertanto potrebbe non esistere alcun punto fisso $\alpha \in [a, b]$.



Consideriamo il seguente risultato a proposito della convergenza globale nell'intervallo $[a, b]$ che utilizza ipotesi più restrittive sulla funzione di iterazione $\phi(x)$ rispetto alla Proposizione 4.3.1.

Proposizione 4.3.2 (Convergenza globale in un intervallo). *Se $\phi \in C^1([a, b])$, $\phi(x) \in [a, b]$ per ogni $x \in [a, b]$ e $|\phi'(x)| < 1$ per ogni $x \in [a, b]$, allora esiste un unico punto fisso $\alpha \in [a, b]$ e il metodo delle iterazioni di punto fisso converge per ogni $x^{(0)} \in [a, b]$ con ordine almeno uguale a 1 (linearmente), ovvero:*

$$\lim_{k \rightarrow +\infty} \frac{x^{(k+1)} - \alpha}{x^{(k)} - \alpha} = \phi'(\alpha),$$

con $\phi'(\alpha)$ il fattore di convergenza asintotico.

Illustriamo alcuni risultati sulla *convergenza locale* al punto fisso α , ovvero in un intorno di α . A tal fine, richiamiamo il teorema di Lagrange.

Teorema 4.3.1 (Lagrange, valor medio). *Se la funzione $g \in C^1([a, b])$, allora esiste $\xi \in (a, b)$ tale che $g(a) - g(b) = g'(\xi)(a - b)$.*

Proposizione 4.3.3 (Ostrowski, convergenza locale in un intorno del punto fisso). *Se $\phi \in C^1(I_\alpha)$, con I_α un intorno del punto fisso α di $\phi(x)$, e $|\phi'(\alpha)| < 1$, allora, se l'iterata iniziale $x^{(0)}$ è “sufficientemente” vicino ad α , il metodo delle iterazioni di punto fisso converge con ordine almeno uguale a 1 (linearmente), ovvero:*

$$\lim_{k \rightarrow +\infty} \frac{x^{(k+1)} - \alpha}{x^{(k)} - \alpha} = \phi'(\alpha),$$

con $\phi'(\alpha)$ il fattore di convergenza asintotico.

Dimostrazione. Mostriamo solamente la proprietà di convergenza lineare del metodo. Sotto le ipotesi del Teorema di Lagrange 4.3.1, si ha $x^{(k+1)} - \alpha = \phi(x^{(k)}) - \phi(\alpha) = \phi'(\xi^{(k)})(x^{(k)} - \alpha)$, per qualche $\xi^{(k)}$ compreso tra α e $x^{(k)}$. Se $\lim_{k \rightarrow +\infty} x^{(k)} = \alpha$, anche $\lim_{k \rightarrow +\infty} \xi^{(k)} = \alpha$ e quindi

$$\lim_{k \rightarrow +\infty} \frac{x^{(k+1)} - \alpha}{x^{(k)} - \alpha} = \lim_{k \rightarrow +\infty} \phi'(\xi^{(k)}) = \phi'(\alpha).$$

□

Osservazione 4.3.1. Sulla base di Proposizione 4.3.3, osserviamo che per $\phi \in C^1(I_\alpha)$:

- se $|\phi'(\alpha)| < 1$, il metodo delle iterazioni di punto fisso converge ad α con ordine uguale ad almeno 1, se $x^{(0)}$ è “sufficientemente” vicino ad α ;
- se $|\phi'(\alpha)| \equiv 1$, la convergenza del metodo ad α dipende dalle proprietà di $\phi(x)$ in un intorno di I_α e la scelta dell’iterata iniziale $x^{(0)}$ (di fatto, il metodo può convergere o meno);
- se $|\phi'(\alpha)| > 1$, la convergenza del metodo ad α è impossibile, a meno che $x^{(0)} \equiv \alpha$.

Proposizione 4.3.4 (Convergenza locale in un intorno del punto fisso). Se $\phi \in C^2(I_\alpha)$, con I_α un intorno del punto fisso α di $\phi(x)$, $\phi'(\alpha) = 0$ e $\phi''(\alpha) \neq 0$, allora, se l’iterata iniziale $x^{(0)}$ è “sufficientemente” vicino ad α , il metodo delle iterazioni di punto fisso converge con ordine 2 (quadraticamente), ovvero:

$$\lim_{k \rightarrow +\infty} \frac{x^{(k+1)} - \alpha}{(x^{(k)} - \alpha)^2} = \frac{1}{2} \phi''(\alpha),$$

con $\frac{1}{2} \phi''(\alpha)$ il fattore di convergenza asintotico.

Il seguente risultato generalizza i precedenti.

Proposizione 4.3.5 (Convergenza locale in un intorno del punto fisso). Se $\phi \in C^p(I_\alpha)$ per $p \geq 1$, con I_α un intorno del punto fisso α di $\phi(x)$, $\phi^{(i)}(\alpha) = 0$ per ogni $i = 1, \dots, p-1$, e $\phi^{(p)}(\alpha) \neq 0$, allora, se l’iterata iniziale $x^{(0)}$ è “sufficientemente” vicino ad α , il metodo delle iterazioni di punto fisso converge con ordine p , ovvero:

$$\lim_{k \rightarrow +\infty} \frac{x^{(k+1)} - \alpha}{(x^{(k)} - \alpha)^p} = \frac{1}{p!} \phi^{(p)}(\alpha),$$

con $\frac{1}{p!} \phi^{(p)}(\alpha)$ il fattore di convergenza asintotica.

4.3.4 Criterio d’arresto per iterazioni di punto fisso

È necessario considerare un criterio d’arresto per terminare le iterazioni di punto fisso dell’Algoritmo 4.5. A questo scopo, introduciamo un opportuno *stimatore dell’errore* $\tilde{e}^{(k)}$ dell’errore $e^{(k)} := |x^{(k)} - \alpha|$. Tale stimatore dell’errore è basato sulla *differenza tra iterate successive*, ovvero:

$$\tilde{e}^{(k)} = \begin{cases} |\delta^{(k-1)}| & \text{se } k \geq 1 \\ tol + 1 & \text{if } k = 0 \end{cases} \quad \text{con } \delta^{(k)} := x^{(k+1)} - x^{(k)} \quad \text{per } k \geq 0;$$

$tol > 0$ è un’opportuna tolleranza. L’algoritmo delle iterazioni di punto fisso si arresta alla prima iterazione \bar{k} tale per cui $\tilde{e}^{(\bar{k})} < tol$ o quando $\bar{k} = k_{max}$, con k_{max} il massimo numero di iterazioni consentito.

Osserviamo che $\alpha - x^{(k+1)} = \alpha - x^{(k)} + x^{(k)} - x^{(k+1)} = (\alpha - x^{(k)}) - \delta^{(k)}$. Inoltre, se $\phi \in C^1(I_\alpha)$, abbiamo dal Teorema 4.3.1 che $\alpha - x^{(k+1)} = \phi'(\xi^{(k)}) (\alpha - x^{(k)})$ per qualche $\xi^{(k)}$ tra $x^{(k)}$ e α . Pertanto, otteniamo $x^{(k)} - \alpha = \phi'(\xi^{(k)}) (x^{(k)} - \alpha) - \delta^{(k)}$ e quindi:

$$x^{(k)} - \alpha = -\frac{1}{1 - \phi'(\xi^{(k)})} \delta^{(k)}, \tag{4.8}$$

per qualche $\xi^{(k)}$ tra $x^{(k)}$ e α . Utilizziamo il precedente risultato per determinare se il criterio d’arresto basato sulla differenza tra iterate successive è soddisfacente o meno. Se $\phi'(x) \simeq 0$ in un intorno di α ($\phi'(\alpha) \simeq 0$), il criterio è *soddisfacente* dato che $e^{(k)} \simeq \tilde{e}^{(k+1)}$. Se $\phi'(x) > -1$, ma $\phi'(x) \simeq -1$ in un intorno di α , il criterio è ancora *soddisfacente* dato che $e^{(k)} \simeq \frac{1}{2} \tilde{e}^{(k+1)}$ (l’errore è *sovristimato* dallo stimatore per un fattore 2). Al contrario, se $\phi'(x) < 1$, ma $\phi'(x) \simeq 1$ in un intorno di α , il criterio è *insoddisfacente* dato che $\tilde{e}^{(k+1)} \ll e^{(k)}$, ovvero l’errore è *sottostimato* dallo stimatore dell’errore.

4.3.5 Il metodo di Newton come metodo delle iterazioni di punto fisso

Il *metodo di Newton* (Sez. 4.2) può essere opportunamente utilizzato per approssimare lo zero α di una funzione generica $f(x)$, la cui iterata di Newton è specificata in Eq. (4.3). Il problema di trovare lo zero α di $f(x)$ con il metodo di Newton può essere ricondotto in un metodo delle iterazioni di punto fisso per mezzo della funzione di iterazione $\phi_N(x)$ tale che $\phi_N(\alpha) = \alpha$. Dalle Eq. (4.3) e (4.7) si deduce che la funzione di iterazione associata al metodo di Newton è:

$$\phi_N(x) = x - \frac{f(x)}{f'(x)}. \quad (4.9)$$

Ne consegue che le proprietà del metodo di Newton, inclusa la convergenza ad α , possono essere dedotte da quelle della funzione di iterazione $\phi_N(x)$.

Proposizione 4.3.6. Se $f \in C^m(I_\alpha)$, con I_α un intorno dello zero α e $m \geq 1$ è la molteplicità di α , la funzione di iterazione $\phi_N(x)$ di Eq. (4.9) è tale che $\phi'_N(\alpha) = 1 - \frac{1}{m}$.

Dimostrazione. La dimostrazione è basata sulla scrittura di $f(x)$ in serie di Taylor attorno allo zero α , per cui $f(x) = \frac{f^{(m)}(\xi)}{m!} (x - \alpha)^m$ per $x \in I_\alpha$ e per ξ tra α e x ; $f^{(m)}(\xi) \neq 0$ per $x \in I_\alpha$. Si ottiene quindi

$$\phi'_N(x) = \frac{f(x) f''(x)}{[f'(x)]^2} = \frac{m(m-1)}{m^2} = 1 - \frac{1}{m} \quad \text{per } x \in I_\alpha,$$

da cui il risultato. □

Corollario 4.3.1. Se $f \in C^2(I_\alpha)$, α è uno zero semplice ($m = 1$) e $x^{(0)}$ è “sufficientemente” vicino ad α , allora il metodo di Newton converge con ordine 2 (quadraticamente), infatti:

$$\lim_{k \rightarrow +\infty} \frac{x^{(k+1)} - \alpha}{(x^{(k)} - \alpha)^2} = \frac{1}{2} \phi''_N(\alpha) = \frac{1}{2} \frac{f''(\alpha)}{f'(\alpha)}.$$

Dimostrazione. Il risultato segue dalle Proposizioni 4.3.4 e 4.3.5, Eq. (4.9) e infine da Proposizione 4.3.6. □

Dalla Proposizione 4.3.6, se α è uno zero semplice ($m = 1$), abbiamo $\phi'_N(\alpha) = 1 - \frac{1}{m} \equiv 0$.

Corollario 4.3.2. Se $f \in C^m(I_\alpha)$, α è uno zero di molteplicità $m > 1$, e $x^{(0)}$ è “sufficientemente” vicino ad α , allora il metodo di Newton converge con ordine 1 (linearmente), infatti:

$$\lim_{k \rightarrow +\infty} \frac{x^{(k+1)} - \alpha}{x^{(k)} - \alpha} = \phi'_N(\alpha) = 1 - \frac{1}{m} \neq 0.$$

Dimostrazione. Il risultato segue da Proposizione 4.3.3, Eq. (4.9) e Proposizione 4.3.6. □

In maniera simile, al *metodo di Newton modificato* (Sez. 4.2.2) basato sull’iterata di Eq. (4.4), associamo la funzione di iterazione $\phi_{mN}(x)$ definita come:

$$\phi_N(x) = x - m \frac{f(x)}{f'(x)}, \quad (4.10)$$

dove m è la molteplicità dello zero α .

Proposizione 4.3.7. Se $f \in C^m(I_\alpha)$, con I_α un intorno dello zero α e $m \geq 1$ la molteplicità di α , otteniamo per la funzione di iterazione $\phi_{mN}(x)$ di Eq. (4.10) che $\phi'_{mN}(\alpha) = 1 - m \frac{1}{m} \equiv 0$ per ogni $m \geq 1$.

Dimostrazione. Il risultato segue analogamente a quello di Proposizione 4.3.6. \square

Corollario 4.3.3. Se $f \in C^2(I_\alpha) \cap C^m(I_\alpha)$, α è uno zero di molteplicità $m \geq 1$ e $x^{(0)}$ è “sufficientemente” vicino ad α , allora il metodo di Newton modificato converge con ordine 2 (quadraticamente).

Dimostrazione. Il risultato segue dalle Proposizioni 4.3.4 e 4.3.5, Eq. (4.10) e Proposizione 4.3.7. \square

Per quanto riguarda la qualità del *criterio d’arresto* basato sulla *differenza tra iterate successive* per il metodo di Newton (Sez. 4.2.3), richiamiamo le proprietà presentate in Sez. 4.3.4 per le iterazioni di punto fisso. Da Eq. (4.8) e Proposizione 4.3.6, osserviamo che:

$$e^{(k)} = |x^{(k)} - \alpha| \simeq \left| \frac{1}{1 - \phi'_N(\alpha)} \right| \tilde{e}^{(k+1)} = m \tilde{e}^{(k+1)},$$

dove lo stimatore dell’errore $\tilde{e}^{(k+1)} = \delta^{(k)} = |x^{(k+1)} - x^{(k)}|$ e $m \geq 1$ è la molteplicità dello zero α .

Pertanto, se lo zero α è semplice ($m = 1$), abbiamo $e^{(k)} \simeq \tilde{e}^{(k+1)}$ e il criterio d’arresto basato sulla differenza tra iterate successive è soddisfacente. Altrimenti, per uno zero di molteplicità $m > 1$, e soprattutto per $m \gg 1$, il criterio risulta insoddisfacente dal momento che l’errore viene sottostimato dallo stimatore dell’errore, ovvero $e^{(k)} \gg \tilde{e}^{(k+1)}$. Utilizzando argomenti del tutto simili per il *metodo di Newton modificato*, il criterio basato sulla differenza tra iterate successive risulta essere sempre soddisfacente, essendo $e^{(k)} \simeq \tilde{e}^{(k+1)}$ indipendentemente dalla molteplicità $m \geq 1$ dello zero.

4.3.6 Il metodo delle corde come metodo delle iterazioni di punto fisso

Come per il metodo di Newton, anche il metodo delle corde può essere interpretato come un metodo delle iterazioni di punto fisso con funzione di iterazione

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

dove $q_C = \frac{f(b) - f(a)}{b - a}$.

Utilizzando il risultato di Prop. 4.3.3, la convergenza del metodo è garantita se $|\phi'_C(\alpha)| < 1$, essendo $\phi'_C(x) = 1 - \frac{f'(x)}{q_C}$, per $x^{(0)}$ “sufficientemente” vicino ad α . Ovvero, si hanno le seguenti condizioni su q_C :

$$q_C > \frac{1}{2} f'(\alpha) \quad \text{se } f'(\alpha) > 0,$$

mentre

$$q_C < \frac{1}{2} f'(\alpha) \quad \text{se } f'(\alpha) < 0.$$

Si osserva che se lo zero α è multiplo, allora $\phi'_C(\alpha) = 1$ per cui la convergenza del metodo non è più garantita. In generale, se $|\phi'_C(\alpha)| < 1$ il metodo delle corde converge linearmente ad α con fattore di convergenza asintotico $1 - \frac{f'(\alpha)}{q_C}$, per $x^{(0)}$ “sufficientemente” vicino ad α , come si deduce dalla Proposizione 4.3.3. Infine, se $q_C = f'(\alpha)$ si ha $\phi'_C(\alpha) = 0$ per cui il metodo delle corde, sotto le ipotesi di Proposizione 4.3.4, converge ad α quadraticamente (con ordine $p = 2$) con fattore di convergenza asintotico $-\frac{f''(\alpha)}{2 q_C}$.

Osservazione 4.3.2. Il metodo delle secanti non può essere interpretato come metodo delle iterazioni di punto fisso.

4.3.7 Iterazioni di punto fisso per funzioni vettoriali

Il metodo delle iterazioni di punto fisso può essere usato con funzioni di iterazione vettoriali $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^n$, per $n \geq 1$. In tal caso, il problema consiste nel trovare il vettore $\alpha \in \mathbb{R}^n$, il punto fisso, tale che $\phi(\alpha) = \alpha$. Il metodo delle iterazioni di punto fisso consiste nell'applicare sequenzialmente la seguente iterata di punto fisso:

$$\boxed{\mathbf{x}^{(k+1)} = \phi(\mathbf{x}^{(k)}) \quad \text{per ogni } k \geq 0,}$$

data l'iterata iniziale $\mathbf{x}^{(0)} \in \mathbb{R}^n$; come criterio d'arresto si può utilizzare quello basato sulla *differenza tra iterate successive* in maniera analoga a Sez. 4.3.4, ovvero $\tilde{e}^{(k)} = \|\delta^{(k-1)}\| < tol$ per $k \geq 1$, con tol una tolleranza prescritta e $\delta^{(k)} = \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}$.

Sia $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^n$ differenziabile in $I_{\mathbf{x}} \subseteq \mathbb{R}^n$ un intorno di $\mathbf{x} \in \mathbb{R}^n$, allora la sua matrice *Jacobiana* in \mathbf{x} è $J_{\phi} : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times n}$ tale che $(J_{\phi}(\mathbf{x}))_{ij} = \frac{\partial \phi_i}{\partial x_j}(\mathbf{x})$ per $i, j = 1, \dots, n$. Indichiamo inoltre con $\rho(J_{\phi}(\mathbf{x}))$ il raggio spettrale di $J_{\phi}(\mathbf{x})$ in $\mathbf{x} \in \mathbb{R}^n$. Si ha quindi il seguente risultato di convergenza.

Proposizione 4.3.8. *Se $\phi \in C^1(I_{\alpha})$, con $I_{\alpha} \subseteq \mathbb{R}^n$ un intorno del punto fisso $\alpha \in \mathbb{R}^n$ di $\phi(\mathbf{x})$, l'iterata iniziale $\mathbf{x}^{(0)} \in \mathbb{R}^n$ è “sufficientemente” vicino ad α e $\rho(J_{\phi}(\alpha)) < 1$, allora il metodo delle iterazioni di punto fisso converge con ordine almeno uguale ad 1 (linearmente), ovvero:*

$$\lim_{k \rightarrow +\infty} \frac{\|\mathbf{x}^{(k+1)} - \alpha\|}{\|\mathbf{x}^{(k)} - \alpha\|} = \rho(J_{\phi}(\alpha)),$$

dove $\rho(J_{\phi}(\alpha))$ è il fattore di convergenza asintotico.

Capitolo 5

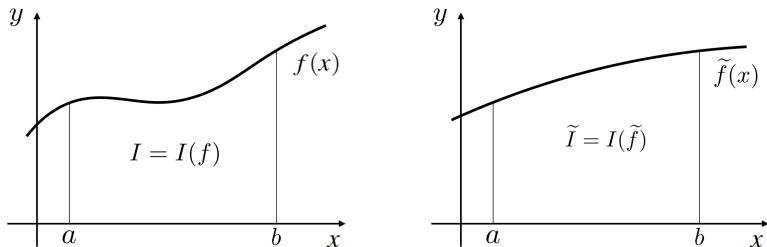
Approssimazione di Funzioni e Dati

Consideriamo l'approssimazione di funzioni e dati, in particolare per mezzo di metodi di *interpolazione* e approssimazione nel senso dei *minimi quadrati*.

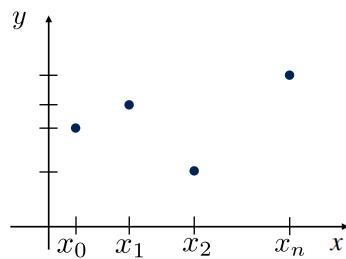
5.1 Motivazioni ed Esempi

Illustriamo attraverso esempi alcune delle motivazioni alla base dell'approssimazione di funzioni e dati.

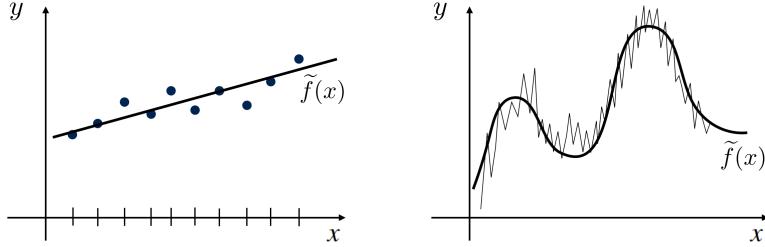
Esempio 5.1.1. Assumiamo di essere interessati a calcolare l'integrale I di una funzione $f(x)$ nell'intervallo $[a, b]$, ovvero $I = I(f) = \int_a^b f(x) dx$, ma di non essere in grado di determinare la primitiva della funzione $f(x)$. Una possibile alternativa consiste nell'*approssimare* $f(x)$ tramite un'altra funzione $\tilde{f}(x)$ di cui però possiamo determinare la funzione primitiva e quindi calcolarne l'integrale in forma chiusa (analiticamente) come $\tilde{I} = I(\tilde{f}) = \int_a^b \tilde{f}(x) dx$, tale per cui $\tilde{I} \simeq I$.



Esempio 5.1.2. Assumiamo che la funzione $f(x)$ sia nota solo tramite sue valutazioni in un insieme di $n + 1$ nodi $\{x_i\}_{i=0}^n$, ovvero siano note le *coppie di dati* $\{(x_i, f(x_i))\}_{i=0}^n$. Potremmo pertanto essere interessati a definire una funzione *approssimante* $\tilde{f}(x)$ della funzione incognita $f(x)$.



Esempio 5.1.3. Dato l'insieme di coppie di dati $\{(x_i, y_i)\}_{i=0}^n$, potremmo voler determinare valori intermedi alle coppie di dati o realizzare previsioni sul valore dei dati al di fuori dell'intervallo determinato dall'insieme di $n + 1$ nodi $\{x_i\}_{i=0}^n$.



5.1.1 Approssimazione di funzioni tramite polinomi di Taylor

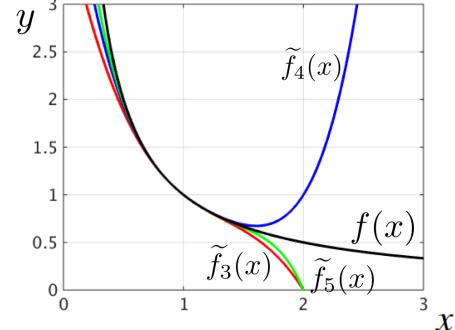
Un modo di approssimare una funzione $f \in C^n(I_{x_0})$ in un intorno di I_{x_0} di un valore $x_0 \in \mathbb{R}$ è basato sull'espansione in serie di Taylor (*polinomi di Taylor*) di ordine n . L'espansione di Taylor di ordine n della funzione $f(x)$ attorno a x_0 si scrive come:

$$\tilde{f}(x) = f(x_0) + \sum_{i=1}^n \frac{1}{i!} f^{(i)}(x_0) (x - x_0)^i.$$

Tuttavia, l'approssimazione di $f(x)$ tramite $\tilde{f}(x)$ presenta alcune problematiche. Innanzitutto è necessario calcolare e valutare n derivate della funzione $f(x)$, operazioni che possono essere computazionalmente costose. Inoltre, l'espansione di Taylor è accurata solo in un intorno I_{x_0} di x_0 , mentre è generalmente del tutto inaccurata al di fuori di tale intorno I_{x_0} .

Esempio 5.1.4. Consideriamo l'espansione di Taylor di ordine n della funzione $f(x) = \frac{1}{x}$, ovvero $\tilde{f}_n(x)$, attorno a $x_0 = 1$.

Dato che $f^{(i)}(x) = (-1)^i i! x^{-(i+1)}$ per $i = 0, 1, \dots, n$, abbiamo $\tilde{f}(x) = \tilde{f}_n(x) = 1 + \sum_{i=1}^n (-1)^i (x - 1)^i$. Come riportato in figura, l'approssimazione fornita da $\tilde{f}_n(x)$ risulta del tutto inaccurata "lontano" dal valore $x_0 = 1$.



5.2 Interpolazione

Introduciamo il concetto di interpolazione e classifichiamo i diversi metodi di interpolazione.

Definizione 5.2.1. Consideriamo un insieme di $n + 1$ coppie di dati $\{(x_i, y_i)\}_{i=0}^n$, dove $\{x_i\}_{i=0}^n$ sono $n + 1$ nodi distinti, ovvero tali che $x_i \neq x_j$ per ogni $i \neq j$ con $i, j = 0, \dots, n$; nel caso la funzione $f(x)$ sia nota, poniamo $y_i = f(x_i)$ per ogni $i = 0, \dots, n$. Interpolare le coppie di dati $\{(x_i, y_i)\}_{i=0}^n$ significa determinare la funzione approssimante $\tilde{f}(x)$ tale che $\tilde{f}(x_i) = y_i$ per ogni $i = 0, \dots, n$ oppure, se $f(x)$ è nota, tale che $\tilde{f}(x_i) = f(x_i)$ per ogni $i = 0, \dots, n$. La funzione $\tilde{f}(x)$ è detta interpolante dei dati ai nodi.

Esistono diversi tipi e metodi di interpolazione. Per esempio:

- interpolazione *polinomiale*, tale per cui $\tilde{f}(x) = a_0 + a_1 x + \cdots + a_n x^n$ per opportuni $n+1$ coefficienti $a_0, a_1, \dots \in \mathbb{R}$;
- interpolazione *razionale*, tale per cui $\tilde{f}(x) = \frac{a_0 + a_1 x + \cdots + a_k x^k}{a_{k+1} + a_{k+2} x + \cdots + a_{k+n+1} x^n}$ per opportuni coefficienti $a_0, a_1, \dots \in \mathbb{R}$ con $k, n \geq 0$;
- interpolazione *trigonometrica*, tale per cui $\tilde{f}(x) = \sum_{j=-M}^M a_j e^{\iota j x}$, essendo ι l'unità immaginaria ($\iota^2 = -1$) e $e^{\iota j x} = \cos(j x) + \iota \sin(j x)$, per qualche M e coefficienti complessi a_j .
- interpolazione *polinomiale a tratti* (composita);
- interpolazione ottenuta mediante funzioni *splines*.

5.2.1 Interpolazione polinomiale di Lagrange

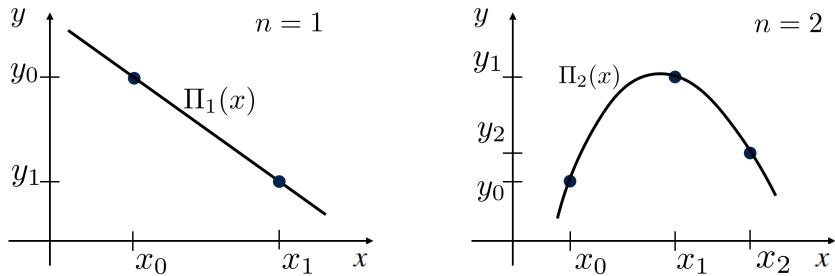
Consideriamo innanzitutto l'*interpolazione polinomiale*, che realizziamo specificamente mediante *polinomi interpolanti di Lagrange*. Ricordiamo che \mathbb{P}_n indica l'insieme dei polinomi di grado minore o uguale a n . L'interpolazione polinomiale è basata sul risultato seguente che determina la corrispondenza tra il numero di nodi distinti e il grado polinomiale dell'interpolante.

Proposizione 5.2.1. *Per ogni coppia di dati $\{(x_i, y_i)\}_{i=0}^n$, di cui $\{x_i\}_{i=0}^n$ sono $n+1$ nodi distinti, esiste un unico polinomio di grado minore o uguale a n , che indichiamo con $\Pi_n(x)$, tale per cui $\Pi_n(x_i) = y_i$ per ogni $i = 0, \dots, n$. $\Pi_n(x) \in \mathbb{P}_n$ è detto polinomio interpolante dei dati ai nodi $\{x_i\}_{i=0}^n$.*

Se invece $f(x)$ è una funzione continua assegnata, per cui $y_i = f(x_i)$ per ogni $i = 0, \dots, n$, allora $\Pi_n f(x) \in \mathbb{P}_n$ è detto polinomio interpolante della funzione $f(x)$ ai nodi $\{x_i\}_{i=0}^n$.

Dimostrazione. Dimostriamo l'*unicità* del polinomio interpolante. Per assurdo, si assuma che esista un polinomio $\tilde{\Pi}_n(x) \in \mathbb{P}_n$ tale che $\tilde{\Pi}_n(x_i) = y_i$ per ogni $i = 0, \dots, n$ e diverso da $\Pi_n(x)$; ovvero $\tilde{\Pi}_n(x) \neq \Pi_n(x)$ per qualche $x \in \mathbb{R}$, ma $\tilde{\Pi}_n(x_i) = \Pi_n(x_i)$ per ogni $i = 0, \dots, n$. Allora la funzione errore $p(x) := \Pi_n(x) - \tilde{\Pi}_n(x)$ è anch'essa un polinomio di grado minore o uguale a n , ovvero $p(x) \in \mathbb{P}_n$, e tale per cui $p(x_i) = 0$ per ogni $i = 0, \dots, n$. Si ottiene pertanto che $p(x)$ è un polinomio di grado n dotato di $n+1$ zeri distinti (i nodi $\{x_i\}_{i=0}^n$). Ne consegue che $p(x) \equiv 0$ per ogni $x \in \mathbb{R}$ e per ogni grado $n \geq 0$, ovvero che $\tilde{\Pi}_n(x) \equiv \Pi_n(x)$, cioè $\Pi_n(x)$ è unico. \square

Esempio 5.2.1. Illustriamo due casi per cui $n = 1$ (sinistra) e $n = 2$ (destra).



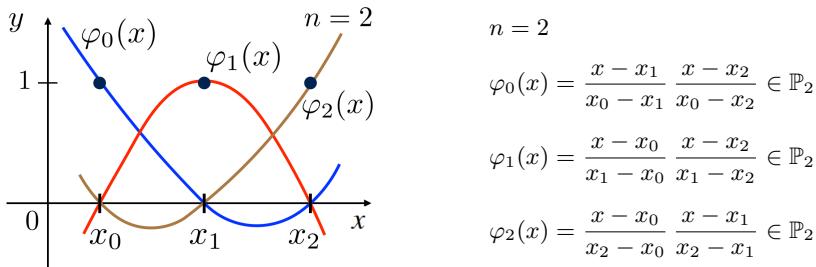
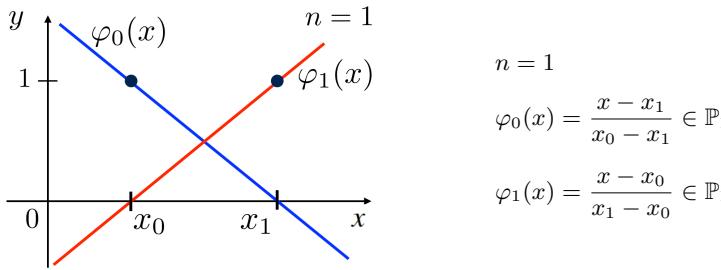
L'obiettivo consiste nel determinare l'espressione del polinomio interpolante $\Pi_n(x)$ (oppure $\Pi_n f(x)$), ovvero $\Pi_n(x) = a_0 + a_1 x + \cdots + a_n x^n$; a tal fine è necessario calcolare i coefficienti $\{a_i\}_{i=0}^n$ di tale polinomio di grado n . A tale scopo, consideriamo una famiglia speciale di polinomi associati agli $n+1$ nodi distinti $\{x_i\}_{i=0}^n$.

Definizione 5.2.2. Dato un insieme di $n + 1$ nodi distinti $\{x_i\}_{i=0}^n$, la funzione caratteristica di Lagrange associata al nodo x_k , detta $\varphi_k \in \mathbb{P}_n$, è un polinomio di grado n tale che $\varphi_k(x_i) = \delta_{ki}$ per ogni $i = 0, \dots, n$, dove $\delta_{ki} = \begin{cases} 0 & \text{se } i \neq k \\ 1 & \text{se } i = k \end{cases}$, e tale per cui:

$$\varphi_k(x) = \prod_{\substack{i=0 \\ i \neq k}}^n \frac{x - x_i}{x_k - x_i}.$$

L'insieme dei polinomi $\{\varphi_k(x)\}_{k=0}^n$ rappresenta la base dei polinomi caratteristici di Lagrange.

Esempio 5.2.2. Illustriamo le basi dei polinomi caratteristici di Lagrange per $n = 1$ e $n = 2$.



Definizione 5.2.3. Data la base dei polinomi caratteristici di Lagrange $\{\varphi_k(x)\}_{k=0}^n$ associati agli $n + 1$ nodi distinti $\{x_i\}_{i=0}^n$, il polinomio interpolante di Lagrange delle coppie di dati $\{(x_i, y_i)\}_{i=0}^n$ si esprime come:

$$\Pi_n(x) = \sum_{k=0}^n y_k \varphi_k(x).$$

Se la funzione $f(x)$ è data e continua, il polinomio interpolante di Lagrange della funzione $f(x)$ ai nodi $\{x_i\}_{i=0}^n$ si esprime come:

$$\Pi_n f(x) = \sum_{k=0}^n f(x_k) \varphi_k(x).$$

Osservazione 5.2.1. Il polinomio interpolante di Lagrange $\Pi_n(x)$ interpola i dati ai nodi; infatti,

$$\Pi_n(x_i) = \sum_{k=0}^n y_k \varphi_k(x_i) = \sum_{k=0}^n y_k \delta_{ki} = y_i \quad \text{per ogni } i = 0, \dots, n.$$

Analogamente, $\Pi_n f(x)$ interpola $f(x)$ ai nodi.

Il polinomio interpolante di Lagrange $\Pi_n(x) \in \mathbb{P}_n$ considera la base dei polinomi caratteristici di Lagrange $\{\varphi_k(x)\}_{k=0}^n$ per determinare i coefficienti $\{a_i\}_{i=0}^n$ di tale polinomio di grado n , ovvero

$$\Pi_n(x) = \sum_{k=0}^n y_k \varphi_k(x) = a_0 + a_1 x + \cdots + a_n x^n.$$

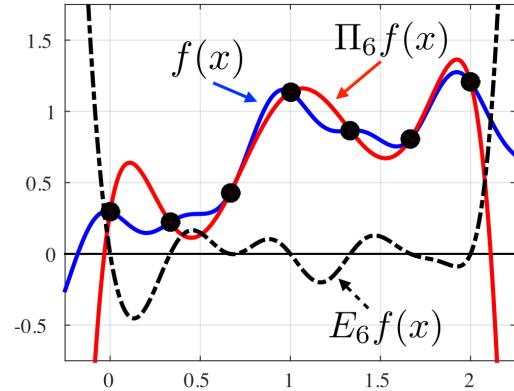
Esempio 5.2.3. Costruiamo l'interpolante polinomiale di Lagrange delle coppie di dati $\{(1, 3)\}, \{(2, 2)\}$ e $\{(4, 6)\}$, dove $n = 2$. Essendo $x_0 = 1, x_1 = 2$ e $x_2 = 4$, abbiamo $\varphi_0(x) = \frac{1}{3}(x-2)(x-4) = \frac{1}{3}x^2 - 2x + \frac{8}{3}$, $\varphi_1(x) = -\frac{1}{2}(x-1)(x-4) = -\frac{1}{2}x^2 + \frac{5}{2}x + 2$ e $\varphi_2(x) = \frac{1}{6}(x-1)(x-2) = \frac{1}{6}x^2 - \frac{1}{2}x + \frac{1}{3}$. Il polinomio interpolante di Lagrange di grado $n = 2$ è $\Pi_2(x) = y_0 \varphi_0(x) + y_1 \varphi_1(x) + y_2 \varphi_2(x) = x^2 - 4x + 6$, per $y_0 = 3$, $y_1 = 2$ e $y_3 = 6$.

Definizione 5.2.4. Per una funzione continua $f(x)$ e l'intervallo $I = [a, b]$ partizionato da $n+1$ nodi ordinati come $a = x_0 < x_1 < \cdots < x_n = b$, definiamo la funzione errore $E_n f(x) := f(x) - \Pi_n f(x)$ associata al polinomio interpolante $\Pi_n f(x)$. L'errore è $e_n(f) := \max_{x \in I} |E_n f(x)|$.

Dato che $\Pi_n f(x)$ interpola $f(x)$ ai nodi, $E_n f(x_i) = 0$ per ogni $i = 0, \dots, n$.

Esempio 5.2.4. Data la funzione $f(x) = \sin(x) + \frac{1}{4} \sin(2\pi x + \sqrt{3}) + \frac{1}{10} \sin(4\pi x + \sqrt{7})$, consideriamo la sua interpolazione polinomiale su $n+1$ nodi equispaziati in $I = [0, 2]$.

Definiamo il grado polinomiale $n = 6$ tale per cui otteniamo l'interpolante polinomiale $\Pi_6 f(x)$ di $f(x)$ ai nodi $x_0 = 0, x_1 = \frac{1}{3}, x_2 = \frac{2}{3}, x_3 = 1, x_4 = \frac{4}{3}, x_5 = \frac{5}{3}$ e $x_6 = 2$. Rappresentiamo graficamente la funzione errore $E_6 f(x) = f(x) - \Pi_6 f(x)$ e osserviamo che $E_6 f(x_i) = 0$ per ogni $i = 0, \dots, 6$.



Proposizione 5.2.2. Consideriamo $n+1$ nodi distinti $\{x_i\}_{i=0}^n$ in un intervallo $I = [a, b]$ tale che $a = x_0 < x_1 < \cdots < x_n = b$ e il polinomio interpolante $\Pi_n f(x)$ di una funzione $f(x)$ in tali nodi. Se $f \in C^{n+1}(I)$ per ogni $x \in I$ allora esiste $\xi = \xi(x) \in I$ tale che:

$$E_n f(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi(x)) \omega_n(x), \quad (5.1)$$

dove $\omega_n(x) := \prod_{i=0}^n (x - x_i)$. Inoltre, l'errore $e_n(f)$ è limitato dallo stimatore dell'errore $\tilde{e}_n(f)$ come:

$$e_n(f) \leq \tilde{e}_n(f) := \frac{1}{(n+1)!} \max_{x \in I} |f^{(n+1)}(x)| \max_{x \in I} |\omega_n(x)|. \quad (5.2)$$

Proposizione 5.2.3. Consideriamo $n+1$ nodi equispaziati $\{x_i\}_{i=0}^n$ nell'intervallo $I = [a, b]$ tale che $x_i = x_0 + i h$ per $i = 0, \dots, n$, con $x_0 = a$, $x_n = b$ e $h = \frac{b-a}{n}$, allora la funzione $\omega_n(x)$ di Proposizione 5.2.2 è tale che:

$$\max_{x \in I} |\omega_n(x)| \leq \frac{n!}{4} h^{n+1} = \frac{n!}{4} \left(\frac{b-a}{n} \right)^{n+1}.$$

Pertanto, deduciamo dall'Eq. (5.2) la stima dell'errore $e_n(f)$ seguente:

$$e_n(f) \leq \tilde{e}_n(f) := \frac{h^{n+1}}{4(n+1)} \max_{x \in I} |f^{(n+1)}(x)| = \frac{1}{4(n+1)} \left(\frac{b-a}{n} \right)^{n+1} \max_{x \in I} |f^{(n+1)}(x)|. \quad (5.3)$$

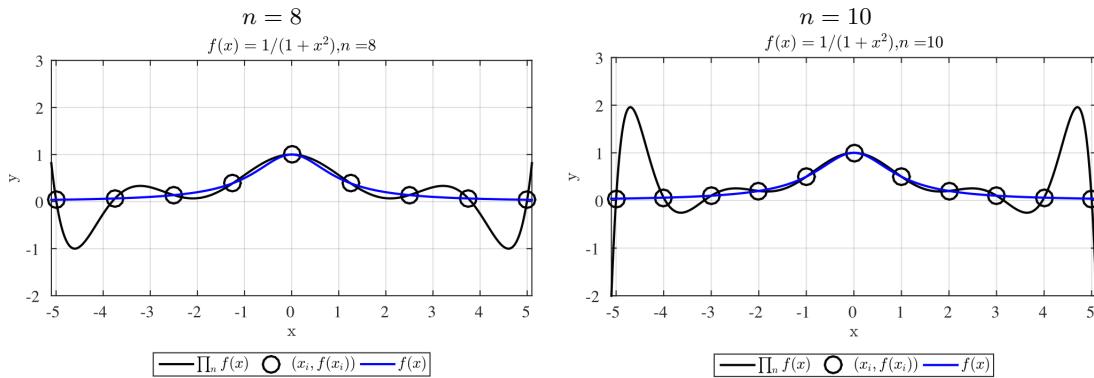
Corollario 5.2.1. Sotto le stesse ipotesi di Proposizione 5.2.3, abbiamo:

$$\max_{x \in I} |f'(x) - (\Pi_n f)'(x)| \leq C_n h^n \max_{x \in I} |f^{(n+1)}(x)|$$

per una costante positiva C_n .

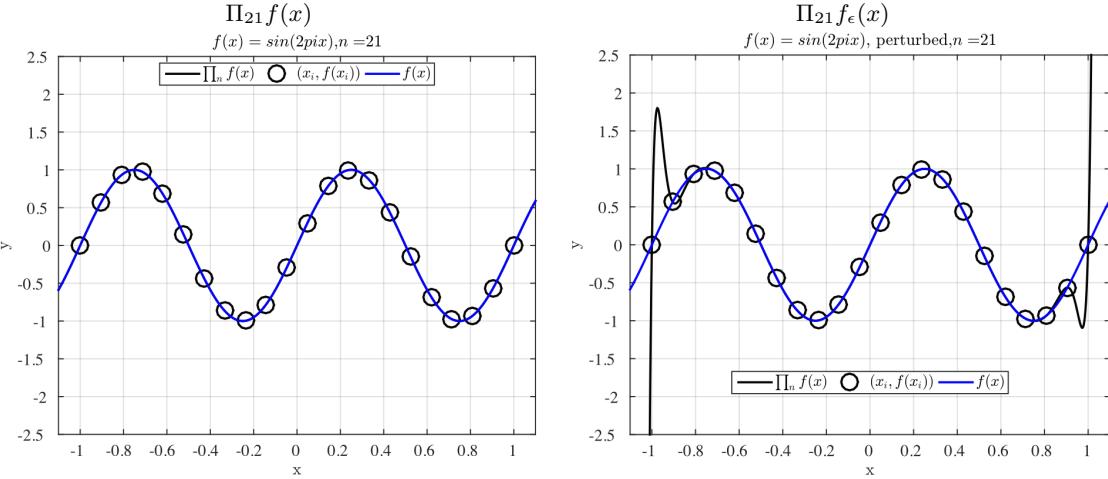
Osservazione 5.2.2. Se gli $n+1$ nodi sono equispaziati nell'intervallo I , l'errore $e_n(f)$ può tendere a zero o meno per $n \rightarrow +\infty$ a seconda della funzione $f(x)$ da interpolare. Osserviamo a partire dall'Eq. (5.3) che $\lim_{n \rightarrow +\infty} \frac{h^{n+1}}{4(n+1)} = 0$. Al contrario, $\max_{x \in I} |f^{(n+1)}(x)|$ potrebbe crescere al crescere di n ; infatti, esistono funzioni tali per cui $\lim_{n \rightarrow +\infty} \max_{x \in I} |f^{(n+1)}(x)| = +\infty$. In tali casi, la crescita di $\max_{x \in I} |f^{(n+1)}(x)|$ potrebbe non essere compensata dal decremento di $\frac{h^{n+1}}{4(n+1)}$ con n , da cui $\lim_{n \rightarrow +\infty} \tilde{e}_n(f) = +\infty$; pertanto, lo stimatore dell'errore $\tilde{e}_n(f)$ "esplode" e, tipicamente, l'errore $e_n(f)$ si comporta in maniera simile. Il cosiddetto **fenomeno di Runge** è un esempio di tale comportamento per cui la funzione errore $E_n f(x)$ tende a "esplodere" per valori crescenti di n in prossimità degli estremi dell'intervallo I quando nodi equispaziati vengono usati per l'interpolazione polinomiale.

Esempio 5.2.5. Consideriamo l'interpolazione polinomiale della *funzione di Runge* $f(x) = \frac{1}{1+x^2}$ su $n+1$ nodi equispaziati nell'intervallo $I = [-5, 5]$. In tal caso, il polinomio interpolante $\Pi_n f(x)$ di $f(x)$ esibisce il cosiddetto fenomeno di Runge per valori crescenti di n come è possibile osservare in prossimità degli estremi dell'intervallo I . Inoltre, $\lim_{n \rightarrow +\infty} e_n(f) = +\infty$.



Un altro aspetto importante da considerare con l'interpolazione polinomiale su nodi *equispaziati* concerne la *stabilità del polinomio interpolante*. Infatti, l'uso di $n+1$ nodi equispaziati nell'intervallo I potrebbe portare ad una significativa sensitività del polinomio interpolante $\Pi_n(x)$ (o $\Pi_n f(x)$ se $f(x)$ è nota) a *perturbazioni* sui dati.

Esempio 5.2.6. Evidenziamo il problema della stabilità considerando l’interpolazione polinomiale di $f(x) = \sin(\pi x)$ su $n + 1$ nodi equispaziati di $I = [-1, 1]$. Ponendo $n = 21$ otteniamo il polinomio interpolante $\Pi_{21}f(x)$ che qualitativamente risulta sovrapposto a $f(x)$. Applichiamo ora l’interpolazione polinomiale alla funzione perturbata $f_\epsilon(x) = f(x) + \epsilon(x)$, con $\epsilon(x)$ una funzione rappresentante rumore bianco tale che $|\epsilon(x)| < 10^{-3}$ per ogni $x \in I$. La sua interpolante polinomiale di grado $n = 21$, $\Pi_{21}f_\epsilon(x)$, risulta molto sensibile a questa pur “piccola” perturbazione, essendo molto distante da $\Pi_{21}f(x)$.



La stabilità dell’interpolazione polonomiale di Lagrange può essere analizzata come segue.

Definizione 5.2.5. Si consideri la base $\{\varphi_k(x)\}_{k=0}^n$ di polinomi caratteristici di Lagrange su $n + 1$ nodi distinti tali che $I = [x_0, x_n]$. La costante di Lebesgue associata a tale base polinomiale di Lagrange di grado n è:

$$\Lambda_n := \max_{x \in I} \sum_{k=0}^n |\varphi_k(x)|.$$

Sia $\Pi_n(x) = \sum_{k=0}^n f(x_k) \varphi_k(x)$ un polonomio interpolante di Lagrange di grado n e $\tilde{\Pi}_n(x) = \sum_{k=0}^n \tilde{f}(x_k) \varphi_k(x)$

il medesimo tipo di interpolante, ma sulla funzione $\tilde{f}(x)$ perturbata rispetto a $f(x)$. Si ha allora, sull’intervallo I contenente i nodi di interpolazione:

$$\max_{x \in I} |\Pi_n(x) - \tilde{\Pi}_n(x)| = \max_{x \in I} \left| \sum_{k=0}^n (f(x_k) - \tilde{f}(x_k)) \varphi_k(x) \right| \leq \Lambda_n \max_{x \in I} |f(x) - \tilde{f}(x)|.$$

Osservazione 5.2.3. La perturbazione sui dati influisce sulla perturbazione del polonomio interpolante e la costante di Lebesgue Λ_n quantifica la propagazione di tale perturbazione sul polinomio interpolante di Lagrange. Se Λ_n è “piccola”, allora la perturbazione sui dati ha un effetto contenuto sul polinomio interpolante, mentre, se Λ_n è “grande”, l’interpolazione polinomiale di Lagrange è molto sensibile a perturbazioni, pur piccole, sui dati.

Nel caso di polinomi caratteristici di Lagrange su nodi **equispaziati** si ha:

$$\Lambda_n \simeq \frac{2^{n+1}}{e n (\log n + \gamma)},$$

con $\gamma \simeq 0.5$. Dato che $\lim_{n \rightarrow +\infty} \Lambda_n = +\infty$ e tale crescita è di tipo esponenziale, si deduce che l’interpolazione polinomiale di Lagrange su nodi equispaziati non è **stabile** per n “grande”.

Un rimedio per mitigare il fenomeno di Runge e problemi di stabilità legati all'interpolazione polinomiale consiste nell'utilizzare nodi che *non* siano equispaziati nell'intervallo I . La seguente definizione fornisce una famiglia speciale di nodi che può essere utilizzata per l'interpolazione polinomiale.

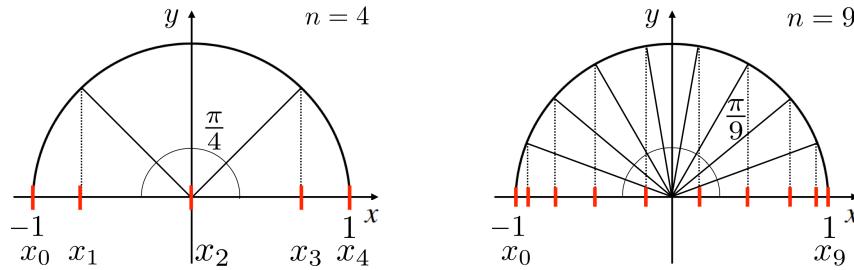
Definizione 5.2.6. *Dato $n \geq 1$, gli $n + 1$ nodi di Chebyshev–Gauss–Lobatto nell'intervallo di riferimento $\hat{I} = [-1, 1]$ sono:*

$$\hat{x}_i = -\cos\left(\frac{\pi}{n} i\right) \quad i = 0, \dots, n;$$

nel generico intervallo $I = [a, b]$, gli $n + 1$ nodi di Chebyshev–Gauss–Lobatto sono:

$$x_i = \frac{a+b}{2} + \frac{b-a}{2} \hat{x}_i \quad i = 0, \dots, n.$$

Esempio 5.2.7. Evidenziamo graficamente gli $n + 1$ nodi di Chebyshev–Gauss–Lobatto $\{\hat{x}_i\}_{i=0}^n$ nell'intervallo di riferimento $\hat{I} = [-1, 1]$ per $n = 4$ (a sinistra) e $n = 9$ (a destra).



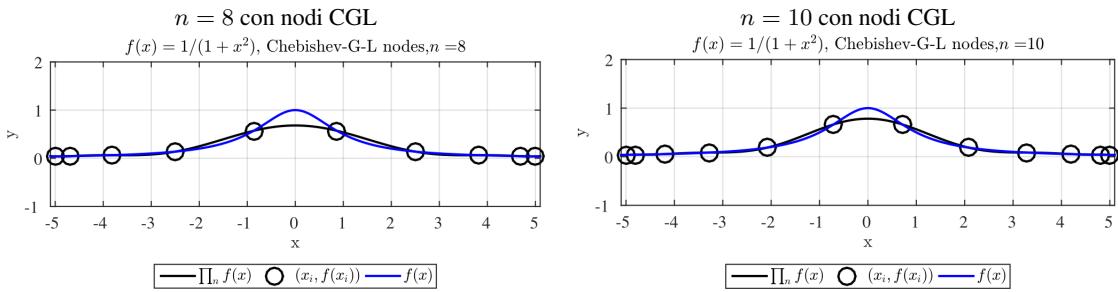
Proposizione 5.2.4. *Se $f \in C^{n+1}(I)$ e vengono usati gli $n + 1$ nodi di Chebyshev–Gauss–Lobatto in $I = [a, b]$, allora $\lim_{n \rightarrow +\infty} \Pi_n f(x) = f(x)$ per ogni $x \in I$, ovvero $\lim_{n \rightarrow +\infty} e_n(f) = 0$; inoltre, i problemi di stabilità sono mitigati.*

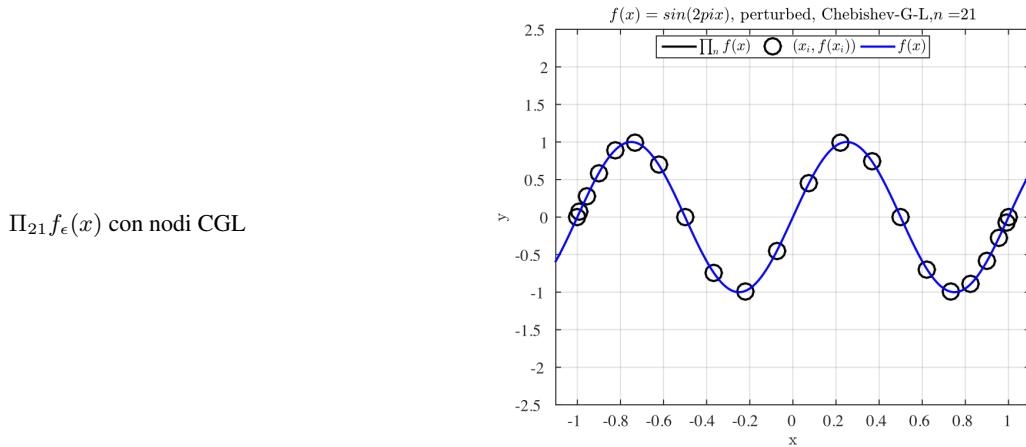
Nel caso di interpolazione polinomiale su nodi di Chebyshev–Gauss–Lobatto la costante di Lebesgue Λ_n^{CGL} è tale per cui

$$\Lambda_n^{CGL} \leq \frac{2}{\pi} (\log n + a) + \frac{\pi}{72 n^2} \quad \text{per } a > 0,$$

ovvero $\Lambda_n^{CGL} \sim \log n$ per n “grande”. Rispetto al caso dei nodi equispaziati, la crescita della costante di Lebesgue è pertanto significativamente contenuta, per cui la sensitività alle perturbazioni dei dati del polinomio interpolante su nodi di Chebyshev–Gauss–Lobatto è significativamente più contenuta.

Esempio 5.2.8. Richiamando gli Esempi 5.2.5 e 5.2.6, oltre a considerare gli stessi dati, mostriamo che l'uso dei nodi di Chebyshev–Gauss–Lobatto (CGL) evita l'insorgenza del fenomeno di Runge e controlla i problemi di stabilità.





Come anticipato, il polinomio interpolante di Lagrange $\Pi_n(x) \in \mathbb{P}_n$ usa la base dei polinomi caratteristici di Lagrange $\{\varphi_k(x)\}_{k=0}^n$ per determinare i coefficienti $\{a_i\}_{i=0}^n$ di questo polinomio, ovvero

$$\Pi_n(x) = \sum_{k=0}^n y_k \varphi_k(x) = a_0 + a_1 x + \cdots + a_n x^n.$$

Un approccio alternativo all'interpolazione polinomiale costruita tramite la base di Lagrange consiste nel determinare direttamente gli $n+1$ coefficienti $\mathbf{a} = (a_0, a_1, \dots, a_n)^T \in \mathbb{R}^{n+1}$ imponendo gli $n+1$ vincoli di interpolazione $\Pi_n(x_i) = y_i$ per ogni $i = 0, \dots, n$; ovvero, $\Pi_n(x_i) = a_0 + a_1 x_i + \cdots + a_n x_i^n = y_i$ per ogni $i = 0, \dots, n$. In questo modo, il problema si riconduce alla soluzione del seguente sistema lineare:

$$V \mathbf{a} = \mathbf{y} \quad (5.4)$$

dove $V \in \mathbb{R}^{(n+1) \times (n+1)}$ è la matrice di Vandermonde, con $V_{ij} = (x_{i-1})^{j-1}$ per $i, j = 1, \dots, n+1$, e $\mathbf{y} = (y_0, y_1, \dots, y_n)^T \in \mathbb{R}^{n+1}$. Il sistema lineare (5.4) ammette un'unica soluzione se e solo se $\det(V) \neq 0$, ovvero se e solo se gli $n+1$ nodi $\{x_i\}_{i=0}^n$ sono distinti. Tuttavia, per quanto intuitivo, questo approccio basato sulla soluzione del sistema lineare (5.4) potrebbe soffrire di problemi di stabilità già per valori di n relativamente "piccoli"; questo in ragione del fatto che il numero di condizionamento della matrice è V è generalmente molto elevato, ovvero $K_2(V) \gg 1$. La soluzione calcolata al calcolatore \mathbf{a} sarà pertanto generalmente affetta da significativi errori.

Osservazione 5.2.4. L'interpolazione polinomiale non è in generale adeguata a estrapolare informazioni al di fuori dell'intervallo I contenente i nodi (si veda l'Esempio 5.2.4).

5.2.2 Interpolazione trigonometrica

Consideriamo l'*interpolazione trigonometrica* che sfrutta funzioni di base di tipo trigonometrico. Tale interpolante è anche nota come *serie discreta di Fourier* e si usa per segnali e funzioni di tipo periodico.

Assumiamo dunque che sia data una funzione $f : [0, 2\pi] \rightarrow \mathbb{C}$ periodica, ovvero tale che $f(0) = f(2\pi)$. Dato che in generale consideriamo funzioni in campo complesso e lavoreremo con numeri complessi, ricordiamo che ι indica l'unità immaginaria, ovvero $\iota^2 = -1$. Anticipiamo inoltre che tale tipo di interpolante può essere applicato anche a insiemi di coppie di dati, sotto l'assunzione che tali dati siano generati da una funzione f non nota, ma comunque periodica.

Definizione 5.2.7. Dati $n + 1$ nodi $\{x_j\}_{j=0}^n$ tali che $x_j = j h$ per $j = 0, \dots, n$ con $h = \frac{2\pi}{n+1}$, l'interpolante trigonometrico della funzione periodica $f : [0, 2\pi] \rightarrow \mathbb{C}$, che indichiamo come $I_t f(x)$, è:

$$I_t f(x) = \sum_{k=-(M+\mu)}^{M+\mu} \tilde{c}_k e^{\iota k x},$$

dove:

$$M = \begin{cases} n/2 & \text{se } n \text{ è pari} \\ (n-1)/2 & \text{se } n \text{ è dispari} \end{cases}, \quad \mu = \begin{cases} 0 & \text{se } n \text{ è pari} \\ 1 & \text{se } n \text{ è dispari} \end{cases},$$

$$\tilde{c}_k = \begin{cases} c_k & \text{per } k = -M, \dots, M \\ c_k/2 & \text{se } k = -(M+1) \text{ o } M+1 \end{cases},$$

e

$$c_k = \frac{1}{n+1} \sum_{j=0}^n f(x_j) e^{-\iota k j h}.$$

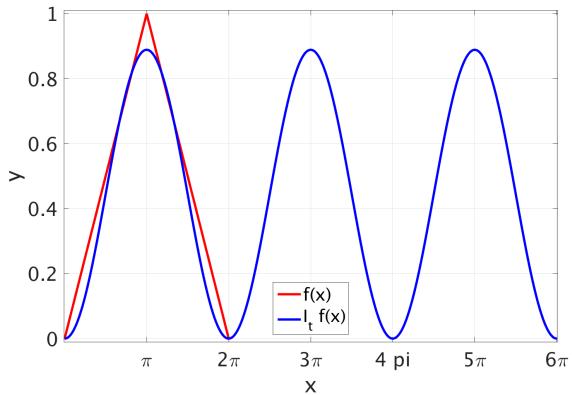
Osserviamo che, in generale, $\{c_k\}_{k=0}^n \in \mathbb{C}$ e $e^{\iota k x} = \cos(kx) + \iota \sin(kx)$. Se $f(x)$ è una funzione a valori reali (ovvero $f(x) \in \mathbb{R}$ per ogni $x \in \mathbb{R}$), allora anche l'interpolante trigonometrico assume valori reali (ovvero $I_t f(x) \in \mathbb{R}$ per ogni $x \in \mathbb{R}$); infatti, in tale caso, si mostra che $c_{-k} = \bar{c}_k$ per $k = 0, \dots, n$.

Osservazione 5.2.5. L'interpolante trigonometrico $I_t f(x)$ interpola $f(x)$ agli $n+1$ nodi $\{x_j\}_{j=0}^n$; infatti, $I_t f(x_j) = f(x_j)$ per ogni $j = 0, \dots, n$.

Il calcolo dei coefficienti $\{c_k\}_{k=0}^n$ sulla base dell'algoritmo precedente richiede un numero di operazioni dell'ordine $O(n^2)$. L'utilizzo invece della *Fast Fourier Transform* (FFT) consente di ridurre il numero di operazioni per il calcolo di tali coefficienti all'ordine $O(n \log n)$.

Esempio 5.2.9. Si consideri $f(x) : [0, 2\pi] \rightarrow \mathbb{R}$, con $f(x) = \begin{cases} x/\pi & x \in [0, \pi] \\ 2 - x/\pi & x \in (\pi, 2\pi] \end{cases}$, una funzione periodica (ovvero $f(0) = f(2\pi)$).

L'interpolante trigonometrico di $f(x)$, ovvero $I_t f(x)$, per $n = 2$ è $I_t f(x) = \frac{4}{9}(1 - \cos(x))$; infatti si hanno $h = \frac{2\pi}{n+1} = \frac{2}{3}$, $x_0 = 0$, $x_1 = h$, $x_2 = 2h$, $\tilde{c}_{-1} = c_{-1} = \tilde{c}_1 = c_1 = -\frac{2}{9}$ e $\tilde{c}_0 = c_0 = \frac{4}{9}$.



Osservazione 5.2.6. L'interpolazione trigonometrica di una funzione periodica potrebbe incorrere nel cosiddetto fenomeno di aliasing; in tal caso, un campionamento inadeguato di $f(x)$, con un valore di n troppo "basso", porta alla costruzione di un interpolante trigonometrico $I_t(x)$ che non approssima correttamente $f(x)$.

Esempio 5.2.10. Siano $f(x) = \sin(5x)$ e $n = 3$, allora si ottiene $I_t(x) = \sin(x)$ per via del fenomeno dell'aliasing. Infatti, si hanno $h = \frac{\pi}{2}$, $x_0 = 0$, $x_1 = \frac{\pi}{2}$, $x_2 = \pi$ e $x_3 = \frac{3}{2}\pi$, per cui $\sin(x_j) = \sin(5x_j)$ per ogni $j = 0, \dots, 3$.

5.2.3 Interpolazione polinomiale a tratti

L'interpolazione polinomiale a tratti, nota anche come *composita*, approssima una funzione $f(x)$ localmente tramite polinomi. L'interpolazione polinomiale a tratti è una buona alternativa all'interpolazione polinomiale con nodi equispaziati per estrarre informazioni all'interno di un intervallo contenente i nodi.

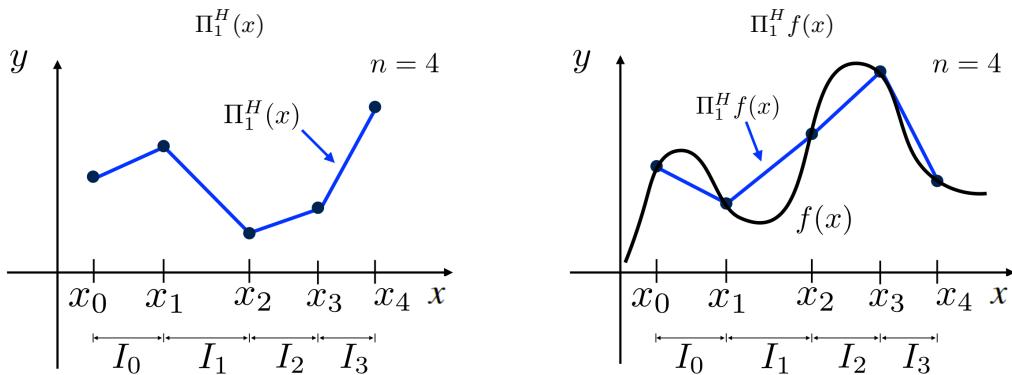
Definizione 5.2.8. Consideriamo $n + 1$ nodi distinti $\{x_i\}_{i=0}^n$ nell'intervallo $I = [a, b]$ tali che $a = x_0 < x_1 < \dots < x_n = b$ e che delimitano n sottointervalli $I_i = [x_i, x_{i+1}]$ per $i = 0, \dots, n - 1$; indichiamo con $H := \max_{i=0, \dots, n-1} |I_i| = \max_{i=0, \dots, n-1} (x_{i+1} - x_i)$ la dimensione caratteristica di tali sottointervalli. Dato l'insieme di coppie di dati $\{(x_i, y_i)\}_{i=0}^n$, l'interpolante polinomiale lineare a tratti $\Pi_1^H(x)$ dei dati è un polinomio interpolante di grado 1 a tratti tale che $\Pi_1^H(x) \in \mathbb{P}_1$ per ogni $x \in I_i$ e $i = 0, \dots, n - 1$ (ovvero $\Pi_1^H(x)|_{I_i} \in \mathbb{P}_1$ per ogni $i = 0, \dots, n - 1$), con:

$$\boxed{\Pi_1^H(x) = y_i + \frac{y_{i+1} - y_i}{x_{i+1} - x_i} (x - x_i) \quad \text{per ogni } i = 0, \dots, n - 1.}$$

Se la funzione $f \in C^0(I)$ è nota, allora il polinomio interpolante lineare a tratti $\Pi_1^H f(x)$ della funzione $f(x)$ ai nodi è $\Pi_1^H f(x)|_{I_i} \in \mathbb{P}_1$ per ogni $i = 0, \dots, n - 1$, con:

$$\boxed{\Pi_1^H f(x) = f(x_i) + \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} (x - x_i) \quad \text{per ogni } i = 0, \dots, n - 1.}$$

Esempio 5.2.11. Riportiamo gli interpolanti lineari a tratti delle $n + 1$ coppie di dati, ovvero $\Pi_1^H(x)$, e di una funzione $f(x)$, indicata come $\Pi_1^H f(x)$, agli $n + 1$ nodi nell'intervallo I ; nello specifico, consideriamo $n = 4$ (ovvero abbiamo $n + 1 = 5$ nodi). La dimensione caratteristica dei sottointervalli $\{I_i\}_{i=0}^3$ è $H = \max_{i=0,1,2,3} |I_i|$.



Definizione 5.2.9. Se la funzione $f \in C^0(I)$ è nota, definiamo l'errore associato all'interpolante polinomiale lineare a tratti $\Pi_1^H f(x)$ come $e_1^H(f) := \max_{x \in I} |f(x) - \Pi_1^H f(x)|$.

Proposizione 5.2.5. Se $f \in C^2(I)$, allora l'errore $e_1^H(f)$ associato all'interpolante polinomiale lineare a tratti $\Pi_1^H f(x)$ può essere limitato dallo stimatore dell'errore $\tilde{e}_1^H(f)$ come:

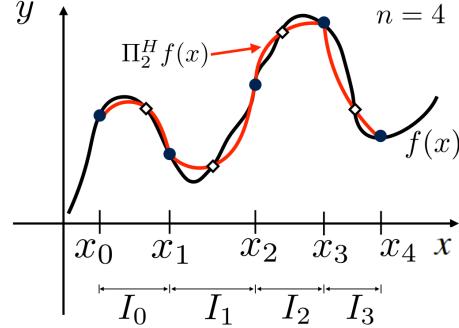
$$e_1^H(f) \leq \tilde{e}_1^H(f) := \frac{H^2}{8} \max_{x \in I} |f''(x)|;$$

si deduce che l'errore converge a zero con ordine 2 in H (quadraticamente).

Analogamente a $\Pi_1^H(x)$, è possibile definire il *polinomio interpolante quadratico a tratti* $\Pi_2^H(x)$ tale che $\Pi_2^H(x)|_{I_i} \in \mathbb{P}_2$ per ogni sottointervallo I_i di I per $i = 0, \dots, n - 1$; se $f \in C^0(I)$ è noto, allora utilizziamo la notazione $\Pi_2^H f(x)$. In maniera analoga, è possibile definire l'*interpolante polinomiale di grado $r \geq 1$ a tratti*, ovvero $\Pi_r^H(x)$, tale che $\Pi_r^H(x)|_{I_i} \in \mathbb{P}_r$ per ogni $i = 0, \dots, n - 1$ (oppure $\Pi_r^H f(x)$ se $f \in C^0(I)$ è noto).

Esempio 5.2.12. Consideriamo l'interpolante polinomiale quadratico a tratti di una funzione continua $f(x)$, detta $\Pi_2^H f(x)$, su $n + 1$ nodi nell'intervallo I ; nello specifico, poniamo $n = 4$.

L'interpolante quadratico a tratti $\Pi_2^H f(x)$ intercala $f(x)$ agli $n + 1 = 5$ nodi, oltre a punti intermedi ed interni a ciascun sottointervallo di I , come possono essere ad esempio i punti medi dei sottointervalli.



Proposizione 5.2.6. Se $f \in C^{r+1}(I)$, l'errore $e_r^H(f) := \max_{x \in I} |f(x) - \Pi_r^H f(x)|$ associato all'interpolante polinomiale di grado $r \geq 1$ a tratti, $\Pi_r^H f(x)$, è limitato dallo stimatore dell'errore $\tilde{e}_r^H(f)$ come:

$$e_r^H(f) \leq \tilde{e}_r^H(f) := C_r H^{r+1} \max_{x \in I} |f^{(r+1)}(x)|,$$

dove C_r è una costante positiva; deduciamo che l'errore converge a zero con ordine $r + 1$ in H .

Per l'interpolante polinomiale composito $\Pi_r^H(x)$, gli $r + 1$ nodi all'interno di ciascun sottointervallo I_i possono essere scelti equispaziati o come i nodi di Chebyshev–Gauss–Lobatto.

Osservazione 5.2.7. Gli interpolanti polinomiali a tratti $\Pi_r^H f(x)$ di qualsiasi grado $r \geq 1$ sono solo C^0 -continui tra un sottointervallo e l'altro (tra un nodo interno e l'altro), come è possibile vedere nell'Esempio 5.2.12.

5.2.4 Splines

Le funzioni splines, oppure semplicemente *splines*, sono polinomi interpolanti a tratti più regolari a cavallo dei sottointervalli rispetto agli interpolanti polinomiali a tratti $\Pi_r^H f(x)$. Per esempio, le *splines cubiche* interpolatorie sono polinomi interpolanti di grado $r = 3$ a tratti C^2 -continui a cavallo dei nodi interni che partizionano I . Le splines – così come loro generalizzazioni come B-splines e NURBS – sono ampiamente utilizzate in Computer Graphics e applicazioni industriali per cui è necessaria un'alta regolarità delle funzioni interpolanti o approssimanti.

Definizione 5.2.10. Una funzione spline cubica, indicata con $s_3(x)$, è un polinomio interpolante a tratti di grado 3 che risulta C^2 -continuo nei nodi interni dell'intervallo I ; dati gli $n+1$ nodi distinti in $I = [x_0, x_n]$ con $x_0 < x_1 < \dots < x_n$ e i sottointervalli $I_i = [x_i, x_{i+1}]$ con $i = 0, \dots, n - 1$, si ha che:

$$s_3(x)|_{I_i} \in \mathbb{P}_3 \quad \text{per ogni } i = 0, \dots, n - 1 \quad \text{and} \quad s_3''(x_i^-) = s_3''(x_i^+) \quad \text{per ogni } i = 1, \dots, n - 1.$$

Possiamo dunque scrivere $s_3(x)|_{I_i} = a_{0,i} + a_{1,i}x + a_{2,i}x^2 + a_{3,i}x^3$ per ogni $i = 0, \dots, n - 1$, ottenendo in tutto $4n$ coefficienti $\{a_{j,i}\}$ da determinare, con $j = 0, 1, 2, 3$ e $i = 0, \dots, n - 1$. Per ottenere i coefficienti

di $s_3(x)$ imponiamo i seguenti $4n - 2$ vincoli di interpolatori, in accordo alla Definizione 5.2.10:

$$s_3(x_i) = y_i \quad (\text{or } s_3(x_i) = f(x_i) \text{ se } f \text{ è nota}) \quad \text{per ogni } i = 0, \dots, n,$$

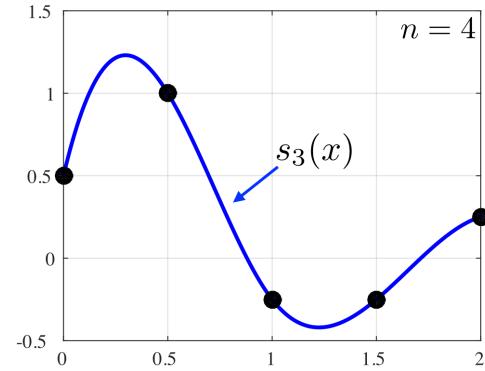
$$s_3(x_i^-) = s_3(x_i^+), \quad s'_3(x_i^-) = s'_3(x_i^+), \quad s''_3(x_i^-) = s''_3(x_i^+) \quad \text{per ogni } i = 1, \dots, n-1.$$

Per determinare univocamente $s_3(x)$, occorrono tuttavia 2 vincoli aggiuntivi, la cui scelta determina il tipo di spline cubica. In particolare:

Definizione 5.2.11. Imponendo che $s''_3(x_0) = s''_3(x_n) = 0$, $s_3(x)$ viene detta spline cubica interpolante naturale. Imponendo invece che $s'''_3(x_1^-) = s'''_3(x_1^+)$ e che $s'''_3(x_{n-1}^-) = s'''_3(x_{n-1}^+)$, $s_3(x)$ è detta spline cubica interpolante not-a-knot.

Il comando MATLAB `spline` permette di determinare spline cubiche interpolanti not-a-knot.

Esempio 5.2.13. Si consideri l'interpolazione di un insieme di $n + 1$ dati mediante una spline cubica interpolante not-a-knot $s_3(x)$; sia dunque $n = 4$.



La spline cubica $s_3(x)$ interpola i dati agli $n + 1 = 5$ nodi $\{x_i\}_{i=0}^4$ e risulta C^2 -continua in ogni nodo interno $\{x_i\}_{i=1}^3$.

Il seguente risultato consente di stimare l'errore commesso approssimando una funzione $f(x)$ data tramite una spline cubica interpolante.

Proposizione 5.2.7. Si considerino $n + 1$ nodi distinti $\{x_i\}_{i=0}^n$ sull'intervallo $I = [x_0, x_n]$, con $x_0 < x_1 < \dots < x_n$, e i corrispondenti n sottointervalli $I_i = [x_i, x_{i+1}]$, $i = 0, \dots, n - 1$, indicando con $H := \max_{i=0, \dots, n-1} |I_i|$ la lunghezza caratteristica di tali sottointervalli. Allora, se $f \in C^4(I)$ e $s_3(x)$ è la sua spline cubica interpolante ai nodi, valgono le seguenti stime dell'errore:

$$\max_{x \in I} \left| f^{(k)}(x) - s_3^{(k)}(x) \right| \leq C_k H^{4-k} \max_{x \in I} \left| f^{(4)}(x) \right| \quad \text{per } k = 0, 1, 2$$

e

$$\max_{x \in I \setminus \{x_1, \dots, x_{n-1}\}} |f'''(x) - s'''_3(x)| \leq C_3 H \max_{x \in I} \left| f^{(4)}(x) \right|,$$

con $C_k > 0$ costanti positive, da cui si deduce che l'ordine di convergenza dell'errore è pari a $4 - k$ in H , in funzione dell'ordine di derivazione $k = 0, 1, 2, 3$.

5.3 Metodo dei Minimi Quadrati

L'approssimazione nel senso dei minimi quadrati è ideale per estrarre informazioni da un insieme di dati relativamente grande, con dati eventualmente affetti da incertezza e rumore, oltre a realizzare predizioni al di fuori dell'intervallo in cui tali dati sono disponibili.

Definizione 5.3.1. Date le coppie di dati $\{(x_i, y_i)\}_{i=0}^n$ (oppure $\{(x_i, f(x_i))\}_{i=0}^n$ se la funzione $f(x)$ è nota) e un intero $m \geq 0$, cerchiamo il polinomio approssimante $\tilde{f}_m(x)$ di grado m tale che:

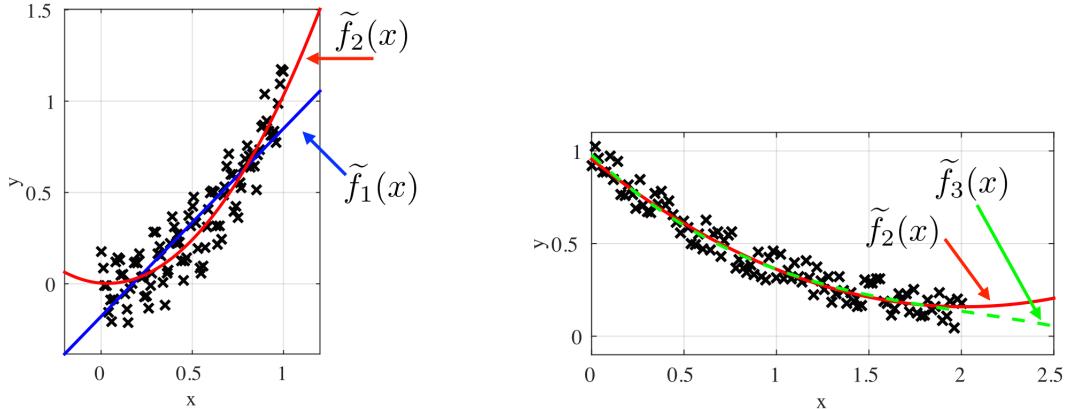
$$\sum_{i=0}^n (y_i - \tilde{f}_m(x_i))^2 \leq \sum_{i=0}^n (y_i - p_m(x_i))^2 \quad \text{per ogni } p_m \in \mathbb{P}_m.$$

Se $\tilde{f}_m \in \mathbb{P}_m$ esiste, allora è detto polinomio di grado m approssimante i dati nel senso dei minimi quadrati (oppure la funzione $f(x)$ se nota).

Per convenzione, assumiamo che i nodi $\{x_i\}_{i=0}^n$ siano distinti e $0 \leq m \leq n$. In uno scenario tipico d'utilizzo del metodo dei minimi quadrati, si ha $0 \leq m \ll n$.

Osservazione 5.3.1. Il polinomio approssimante nel senso dei minimi quadrati $\tilde{f}_m(x)$ non interpola in generale i dati (o la funzione $f(x)$) ai nodi. In particolare, solo se $m = n$ è possibile assicurare che $\tilde{f}_m(x_i) = y_i$ (oppure $\tilde{f}_m(x_i) = f(x_i)$) per ogni $i = 0, \dots, n$; infatti, in tal caso, $\tilde{f}_m(x)$ coincide con il polinomio interpolante $\Pi_n(x)$ di grado n (oppure $\Pi_n f(x)$).

Esempio 5.3.1. Illustriamo graficamente polinomi approssimanti nel senso dei minimi quadrati $\tilde{f}_m(x)$ per insiemi di dati relativamente grandi $\{(x_i, y_i)\}_{i=0}^n$, con $n = 100$. Consideriamo $m = 1$ e 2 (a sinistra) e $m = 2$ e 3 (a destra).



Come per l'interpolazione polinomiale, determinare polinomi approssimanti ai minimi quadrati $\tilde{f}_m(x)$ di grado m consiste nel determinare gli $m+1$ coefficienti $\{a_i\}_{i=0}^m$; infatti, $\tilde{f}_m(x) = a_0 + a_1 x + \dots + a_m x^m$. A tal fine, definiamo il vettore $\mathbf{a} = (a_0, a_1, \dots, a_m)^T \in \mathbb{R}^{m+1}$ e la funzione $\Phi : \mathbb{R}^{m+1} \rightarrow \mathbb{R}$ come:

$$\Phi(\mathbf{b}) = \sum_{i=0}^n [y_i - (b_0 + b_1 x_i + \dots + b_m x_i^m)]^2,$$

che è associata all'insieme di dati $\{(x_i, y_i)\}_{i=0}^n$ per un generico vettore $\mathbf{b} = (b_0, b_1, \dots, b_m)^T \in \mathbb{R}^{m+1}$. Il *metodo dei minimi quadrati* consiste nel determinare i coefficienti \mathbf{a} del polinomio $\tilde{f}_m(x)$ tali che:

$$\Phi(\mathbf{a}) = \min_{\mathbf{b} \in \mathbb{R}^{m+1}} \Phi(\mathbf{b}).$$

Dato Φ che è differenziabile e convessa, il problema di minimizzazione precedente corrisponde a risolvere il seguente problema differenziale:

$$\text{trovare } \mathbf{a} \in \mathbb{R}^{m+1} : \frac{\partial \Phi}{\partial b_j}(\mathbf{a}) = 0 \quad \text{per ogni } j = 0, \dots, m, \tag{5.5}$$

che porta alla soluzione del sistema lineare:

$$A \mathbf{a} = \mathbf{q}, \quad (5.6)$$

dove $A \in \mathbb{R}^{(m+1) \times (m+1)}$ e $\mathbf{q} \in \mathbb{R}^{m+1}$ sono rispettivamente:

$$A = \begin{bmatrix} (n+1) & \sum_{i=0}^n x_i & \cdots & \sum_{i=0}^n x_i^m \\ \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 & \cdots & \sum_{i=0}^n x_i^{m+1} \\ \vdots & \vdots & & \vdots \\ \sum_{i=0}^n x_i^m & \sum_{i=0}^n x_i^{m+1} & \cdots & \sum_{i=0}^n x_i^{2m} \end{bmatrix} \quad \text{e} \quad \mathbf{q} = \begin{bmatrix} \sum_{i=0}^n y_i \\ \sum_{i=0}^n x_i y_i \\ \vdots \\ \sum_{i=0}^n x_i^m y_i \end{bmatrix}.$$

Richiamando la matrice di Vandermonde $V \in \mathbb{R}^{(n+1) \times (m+1)}$, con $V_{ij} = (x_{i-1})^{j-1}$ per $i = 1, \dots, n+1$ e $j = 1, \dots, m+1$, e il vettore $\mathbf{y} = (y_0, y_1, \dots, y_n)^T \in \mathbb{R}^{n+1}$, osserviamo che:

$$A = V^T V \quad \text{e} \quad \mathbf{q} = V^T \mathbf{y}.$$

Il sistema lineare (5.6) rappresenta una generalizzazione del sistema lineare (5.4) usato per l'interpolazione polinomiale. Di fatto, se i nodi sono distinti e $m = n$, otteniamo che $\tilde{f}_n(x) = \Pi_n(x)$, per cui risolvere i due sistemi lineari (5.4) e (5.6) fornisce risultati equivalenti.

5.3.1 La retta di regressione

Definizione 5.3.2. Sulla base della Definizione 5.3.1, il polinomio approssimante nel senso dei minimi-quadrati $\tilde{f}_1(x)$ di grado $m = 1$ è detto retta di regressione o linea retta ai minimi-quadrati.

Illustriamo la derivazione del sistema (5.6) per $\tilde{f}_1(x)$ (ovvero per la retta di regressione, $m = 1$) con $n \geq 1$. In tal caso, la funzione $\Phi(\mathbf{b})$ si scrive come:

$$\Phi(\mathbf{b}) = \sum_{i=0}^n [y_i - (b_0 + b_1 x_i)]^2 = \sum_{i=0}^n [y_i^2 + b_0^2 + b_1^2 x_i^2 - 2b_0 y_i - 2b_1 x_i y_i + 2b_0 b_1 x_i].$$

Per riformulare il problema come in Eq. (5.5), calcoliamo le derivate parziali di Φ :

$$\frac{\partial \Phi}{\partial b_0}(\mathbf{b}) = \sum_{i=0}^n [2b_0 - 2y_i + 2b_1 x_i],$$

$$\frac{\partial \Phi}{\partial b_1}(\mathbf{b}) = \sum_{i=0}^n [2b_1 x_i^2 - 2x_i y_i + 2b_0 x_i].$$

A questo punto, il problema (5.5) può essere scritto come il sistema lineare (5.6) con $A \in \mathbb{R}^{2 \times 2}$ e $\mathbf{q} \in \mathbb{R}^2$, essendo:

$$A = \begin{bmatrix} (n+1) & \sum_{i=0}^n x_i \\ \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 \end{bmatrix} \quad \text{e} \quad \mathbf{q} = \begin{bmatrix} \sum_{i=0}^n y_i \\ \sum_{i=0}^n x_i y_i \end{bmatrix},$$

rispettivamente. Osserviamo che, in tal caso, la matrice di Vandermonde $V \in \mathbb{R}^{(n+1) \times 2}$ si scrive come:

$$V = \begin{bmatrix} 1 & x_0 \\ 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}.$$

Capitolo 6

Integrazione Numerica

Consideriamo in questo capitolo alcuni possibili modi per approssimare il calcolo di integrali definiti di funzioni reali a variable reale (in una sola variabile) mediante cosiddette *formule di quadratura (integrazione numerica)*. In molti casi, infatti, può accadere che non si conosca una primitiva della funzione che si desidera integrare, come per esempio

$$\int_0^1 e^{-x^2} dx, \quad \text{oppure} \quad \int_0^1 \cos(x^2) dx;$$

in ogni caso, risulta fondamentale approssimare integrali dato che, come vedremo in seguito, l'approssimazione di un problema descritto da equazioni alle derivate parziali con il metodo degli elementi finiti conduce a risolvere sistemi lineari in cui le componenti della matrice e del termine noto sono determinati da approssimazioni di integrali.

6.1 Scopo e Classificazione delle Formule di Quadratura

Data una funzione $f \in C^0([a, b])$, desideriamo approssimare numericamente il suo integrale (definito) sull'intervallo $[a, b]$, ovvero

$$I(f) = \int_a^b f(x) dx,$$

mediante opportune *formule di quadratura*, indicate con $I_q(f)$, tali che $I_q(f) \simeq I(f)$. Possiamo distinguere le formule di quadratura in due classi principali: formule *semplici* oppure *composite*.

- Le *formule di quadratura semplici* sono basate su un'approssimazione globale della funzione $f(x)$ nell'intervallo $[a, b]$ mediante funzioni $\tilde{f}(x)$ che risultino “semplici” da integrare in $[a, b]$ (ovvero, che siano integrabili esplicitamente, ovvero in forma chiusa); si ha cioè che

$$I_q(f) = I(\tilde{f}) = \int_a^b \tilde{f}(x) dx,$$

dove $\tilde{f}(x)$ è un'approssimazione di $f(x)$ per $x \in [a, b]$. Tipicamente, per formule di quadratura semplici, $\tilde{f}(x)$ è un polinomio di grado n che interpola¹ $f(x)$ in $n + 1$ nodi in $[a, b]$;

- Le *formule di quadratura composite* si basano sulla suddivisione dell'intervallo $[a, b]$ in M sottointervalli, eventualmente della stessa dimensione $H = \frac{b-a}{M}$, su cui la funzione $f(x)$ viene approssimata da una funzione a tratti $\tilde{f}(x)$. Indicando gli $M + 1$ nodi $\{x_k\}_{k=0}^M$ come $x_k = a + k H$ per

¹Si parla a tal proposito di *formule di quadratura interpolatorie*.

$k = 0, \dots, M$, con $x_0 = a$ e $x_M = b$, ricordiamo che

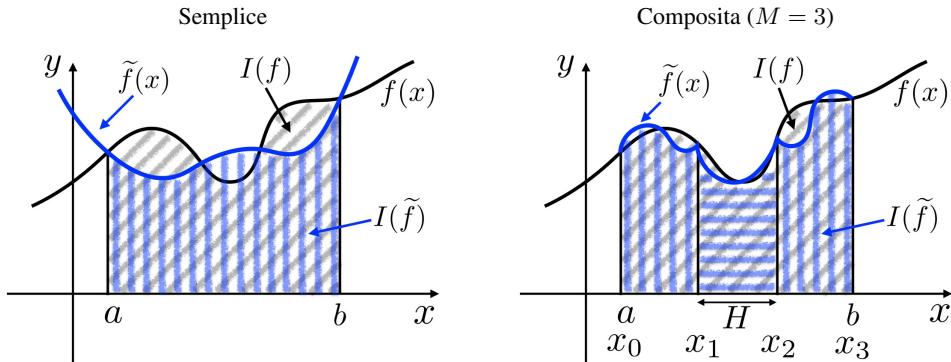
$$I(f) = \sum_{k=1}^M \int_{x_{k-1}}^{x_k} f(x) dx;$$

quindi, la formula di quadratura composita diviene

$$I_q(f) = \sum_{k=1}^M \int_{x_{k-1}}^{x_k} \tilde{f}_k(x) dx, \quad \text{dove } \tilde{f}_k(x) = \tilde{f}(x)|_{[x_{k-1}, x_k]} \quad \forall k = 1, \dots, M.$$

In generale, per le formule di quadratura composite, $\tilde{f}(x)$ è un polinomio a tratti di grado n che intercala $f(x)$ in $n + 1$ nodi in ciascuno dei sottointervalli $\{[x_{k-1}, x_k]\}_{k=1}^M$ di $[a, b]$.

Esempio 6.1.1. La differenza tra una formula di quadratura semplice e una composita per approssimare l'integrale $I(f) = \int_a^b f(x) dx$ di una generica funzione $f(x)$ nell'intervallo $[a, b]$ è mostrata nella seguente figura; con $I(\tilde{f})$ indichiamo l'integrale approssimato.



Le seguenti definizioni sono utili per caratterizzare le formule di quadratura.

Definizione 6.1.1. Il grado di esattezza di una formula di quadratura è il massimo numero intero $r \geq 0$ tale che tutti i polinomi di grado inferiore o uguale a r sono esattamente integrati dalla formula, vale a dire, tale per cui $I_q(p) \equiv I(p)$ per ogni $p \in \mathbb{P}_r$.

Definizione 6.1.2. L'ordine di convergenza di una formula di quadratura composita (chiamato anche ordine di accuratezza) è l'ordine di convergenza dell'errore associato rispetto ad H , la dimensione dei sottointervalli.

6.2 Formule di Quadratura del Punto Medio

Approssimando la funzione $f(x)$ con il suo interpolante polinomiale di Lagrange di grado 0 (ovvero con una funzione costante), $\tilde{f}(x) = \Pi_0 f(x)$, oppure con un interpolante composito di grado 0 (ovvero, una funzione costante a tratti) $\Pi_0^H f(x)$, l'integrale $\int_a^b \tilde{f}(x) dx$ conduce rispettivamente alle formule del punto medio semplice e composita.

Definizione 6.2.1. Sia $f(x) \in C^0([a, b])$. La formula di quadratura del punto medio semplice è definita come:

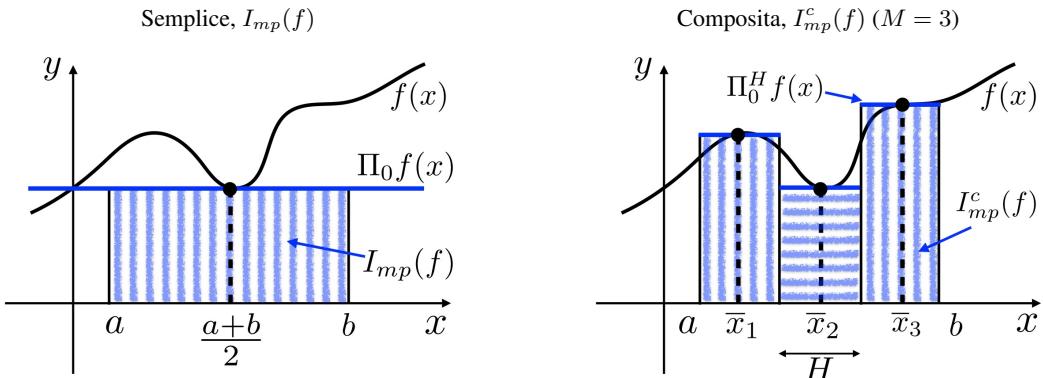
$$I_{mp}(f) := I(\Pi_0 f) = (b - a) f\left(\frac{a + b}{2}\right),$$

dove $\Pi_0 f(x)$ è il polinomio di grado 0 che interpola $f(x)$ nel punto medio $\bar{x} = \frac{a + b}{2}$ dell'intervallo $[a, b]$. La formula di quadratura del punto medio composita è definita come:

$$I_{mp}^c(f) := I(\Pi_0^H f) = H \sum_{k=1}^M f(\bar{x}_k), \quad (6.1)$$

dove $\Pi_0^H f(x)$ è il polinomio a tratti di grado 0 interpolante $f(x)$ nei punti medi $\{\bar{x}_k\}_{k=1}^M$ degli M sottointervalli di lunghezza $H = \frac{b - a}{M}$ in cui $[a, b]$ è suddiviso, essendo $\bar{x}_k = \frac{x_{k-1} + x_k}{2}$ per ogni $k = 1, \dots, M$.

Esempio 6.2.1. Illustriamo graficamente la formula del punto medio semplice (a sinistra) e composita (a destra) per l'approssimazione dell'integrale $I(f)$ di una generica funzione $f(x)$.



Proposizione 6.2.1. Se $f \in C^2([a, b])$, l'errore $e_{mp}(f)$ associato alla formula di quadratura del punto medio semplice risulta

$$e_{mp}(f) := I(f) - I_{mp}(f) = \frac{(b - a)^3}{24} f''(\xi) \quad \text{per un certo } \xi \in [a, b],$$

mentre l'errore $e_{mp}^c(f)$ associato alla formula di quadratura del punto medio composita risulta

$$e_{mp}^c(f) := I(f) - I_{mp}^c(f) = \frac{(b - a)}{24} H^2 f''(\xi) \quad \text{per un certo } \xi \in [a, b].$$

Dimostrazione. (Formula semplice). Indicando il punto medio di $[a, b]$ come $\bar{x} = (a + b)/2$, consideriamo l'espansione di Taylor di $f(x)$ intorno a \bar{x} , ovvero

$$f(x) = f(\bar{x}) + f'(\bar{x})(x - \bar{x}) + \frac{1}{2} f''(\eta(x))(x - \bar{x})^2 \quad \text{per un certo } \eta(x) \in [a, b].$$

Integrando tale espressione, otteniamo

$$I(f) = I_{mp}(f) + f'(\bar{x}) \int_a^b (x - \bar{x}) dx + \frac{1}{2} f''(\xi) \int_a^b (x - \bar{x})^2 dx \quad \text{per un certo } \eta(x) \in [a, b],$$

in virtù del teorema della media integrale². Poiché $\int_a^b (x - \bar{x}) dx = 0$ e $\int_a^b (x - \bar{x})^2 dx = \frac{(b-a)^3}{12}$, si ottiene il risultato cercato.

(Formula composita). Si scrive l'espansione di Taylor centrata nel punto medio \bar{x}_k di ciascuno dei sottointervalli $[x_{k-1}, x_k]$ per $k = 1, \dots, M$, ovvero $f(x) = f(\bar{x}_k) + f'(\bar{x}_k)(x - \bar{x}_k) + \frac{1}{2}f''(\eta_k(x))(x - \bar{x}_k)^2$ per qualche $\eta_k(x) \in [x_{k-1}, x_k]$. Scrivendo $I(f) = \sum_{k=1}^M \int_{x_{k-1}}^{x_k} f(x) dx$ e utilizzando l'espressione precedente, si ottiene, in analogia al caso semplice, che $I(f) = I_{pm}^c(f) + \frac{1}{24}H^3 \sum_{k=1}^M f''(\xi_k)$ per qualche $\xi_k \in [x_{k-1}, x_k]$. Infine, applicando il teorema della media integrale nel discreto³, esiste $\xi \in [a, b]$ tale che $\sum_{k=1}^M f''(\xi_k) = M f''(\xi)$, da cui segue il risultato. \square

Osservazione 6.2.1. Le formule di quadratura del punto medio hanno grado di esattezza 1; infatti, gli errori $e_{mp}(f)$ e $e_{mp}^c(f)$ sono identicamente zero per tutti i polinomi di grado inferiore o uguale a 1 (se $f \in \mathbb{P}_1$, $f''(\xi) = 0$ per ogni $\xi \in \mathbb{R}$).

Osservazione 6.2.2. La formula del punto medio composita ha ordine di convergenza (o ordine di accuratezza) 2; infatti, l'errore $e_{mp}^c(f)$ risulta proporzionale ad H^2 .

6.3 Formule di Quadratura del Trapezio

Approssimando la funzione $f(x)$ con il suo interpolante Lagrangiano di grado 1, $\tilde{f}(x) = \Pi_1 f(x)$, oppure con un interpolante composito di grado 1 (ovvero, un interpolante lineare a tratti) $\Pi_1^H f(x)$, l'integrale $\int_a^b \tilde{f}(x) dx$ conduce rispettivamente alle formule del trapezio semplice e composita.

Definizione 6.3.1. Sia $f(x) \in C^0([a, b])$. La formula di quadratura del trapezio semplice è definita come:

$$I_t(f) := I(\Pi_1 f) = (b-a) \frac{f(a) + f(b)}{2}, \quad (6.2)$$

dove $\Pi_1 f(x)$ è il polinomio di grado 1 che interpola $f(x)$ nei nodi a e b . La formula di quadratura del trapezio composita è definita come:

$$I_t^c(f) := I(\Pi_1^H f) = \frac{H}{2} \sum_{k=1}^M [f(x_{k-1}) + f(x_k)] = \frac{H}{2} [f(x_0) + f(x_M)] + H \sum_{k=1}^{M-1} f(x_k),$$

dove $\Pi_1^H f(x)$ è il polinomio di grado 1 a tratti interpolante $f(x)$ nei nodi $\{x_k\}_{k=0}^M$ degli M sottointervalli di lunghezza $H = \frac{b-a}{M}$ in cui $[a, b]$ è suddiviso.

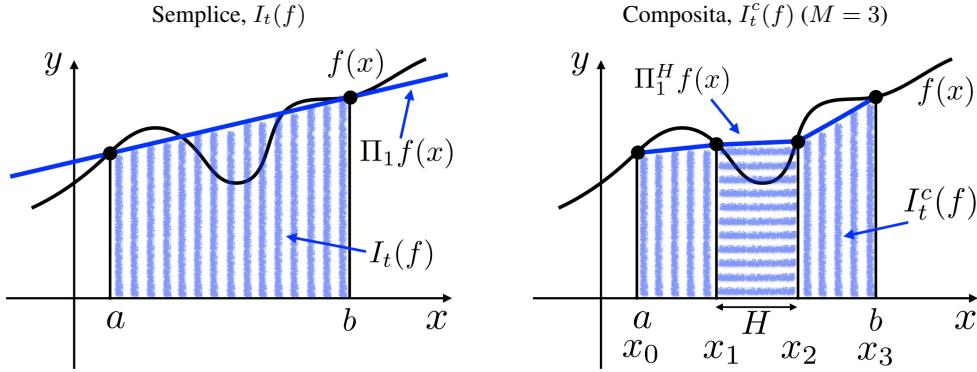
²Teorema della media integrale. Siano $f, g \in C^0([a, b])$ e sia $g(x)$ di segno costante in $[a, b]$. Allora esiste un punto $c \in [a, b]$ tale che

$$\int_a^b f(x)g(x)dx = f(c) \int_a^b g(x)dx.$$

³Se $f \in C^0([a, b])$ e sono dati $s+1$ punti $\{x_j\}_{j=0}^s$ in $[a, b]$ e $s+1$ costanti $\{\delta_j\}_{j=0}^s$, tutte dello stesso segno, esiste allora $c \in [a, b]$ tale che

$$\sum_{j=0}^s \delta_j f(x_j) = f(c) \sum_{j=0}^s \delta_j.$$

Esempio 6.3.1. Illustriamo graficamente la formula del trapezio semplice (a sinistra) e composita (a destra) per l'approssimazione dell'integrale $I(f)$ di una generica funzione $f(x)$.



Proposizione 6.3.1. Se $f \in C^2([a, b])$, l'errore $e_t(f)$ associato alla formula di quadratura del trapezio semplice risulta

$$e_t(f) := I(f) - I_t(f) = -\frac{(b-a)^3}{12} f''(\xi) \quad \text{per un certo } \xi \in [a, b],$$

mentre l'errore $e_t^c(f)$ associato alla formula di quadratura del trapezio composita risulta

$$e_t^c(f) := I(f) - I_t^c(f) = -\frac{(b-a)}{12} H^2 f''(\xi) \quad \text{per un certo } \xi \in [a, b].$$

Dimostrazione. (Formula semplice). Poiché $e_t(f) = I(f) - I(\Pi_1 f)$, abbiamo che $e_t(f) = \int_a^b (f(x) - \Pi_1 f(x)) dx$. Ricordando la funzione errore $E_1 f(x)$ dell'Eq. (5.1) nel caso dell'interpolazione polinomiale di grado 1 (Proposizione 5.2.2), si ha che $E_1 f(x) = \frac{1}{2} f''(\eta(x)) \omega_1(x)$ per un certo $\eta(x) \in [a, b]$, con $\omega_1(x) = (x-a)(x-b)$. Usando il teorema del valor medio (di Lagrange), si ha che

$$e_t(f) = \frac{1}{2} \int_a^b f''(\eta(x)) \omega_1(x) dx = \frac{1}{2} f''(\xi) \int_a^b \omega_1(x) dx, \quad \text{per un certo } \xi \in [a, b];$$

dunque, siccome $\int_a^b \omega_1(x) dx = -\frac{(b-a)^3}{6}$, si ottiene il risultato cercato. \square

Osservazione 6.3.1. Le formule di quadratura del trapezio hanno grado di esattezza $r = 1$; infatti, gli errori $e_t(f)$ e $e_t^c(f)$ sono identicamente zero per tutti i polinomi di grado inferiore o uguale a 1 (se $f \in \mathbb{P}_1$, $f''(\xi) = 0$ per ogni $\xi \in \mathbb{R}$).

Osservazione 6.3.2. La formula dei trapezi composita ha ordine di convergenza (o ordine di accuratezza) 2; infatti, l'errore $e_{mp}^c(f)$ risulta proporzionale ad H^2 .

Osservazione 6.3.3. Consideriamo le formule di quadratura del punto medio e del trapezio, per le quali gli errori $e_{mp}(f)$, $e_{mp}^c(f)$, $e_t(f)$ e $e_t^c(f)$ sono indicati nelle Proposizioni 6.2.1 e 6.4.1. Notiamo che, in generale, non si può garantire che $|e_{mp}(f)| = \frac{1}{2}|e_t(f)|$ e $|e_{mp}^c(f)| = \frac{1}{2}|e_t^c(f)|$ poiché $f''(\xi)$ può essere valutata in valori diversi di ξ in $[a, b]$ a seconda della formula in esame. Tuttavia, se $f \in \mathbb{P}_2$, si ha che $f''(x) = C$, un valore costante, per ogni $x \in \mathbb{R}$, per cui $|e_{mp}(f)| = \frac{1}{2}|e_t(f)|$ e $|e_{mp}^c(f)| = \frac{1}{2}|e_t^c(f)|$; quindi, in questo caso specifico, le formule del punto medio sono più accurate di quelle del trapezio (a parità di H).

6.4 Formule di Quadratura di Simpson

Approssimando la funzione $f(x)$ con il suo interpolante polinomiale di Lagrange di grado 2, $\tilde{f}(x) = \Pi_2 f(x)$, oppure con un interpolante composito di grado 2 (ovvero, una funzione quadratica a tratti) $\Pi_2^H f(x)$, l'integrale $\int_a^b \tilde{f}(x) dx$ conduce rispettivamente alle formule di Simpson semplice e composita.

Definizione 6.4.1. Sia $f(x) \in C^0([a, b])$. La formula di quadratura di Simpson semplice è definita come:

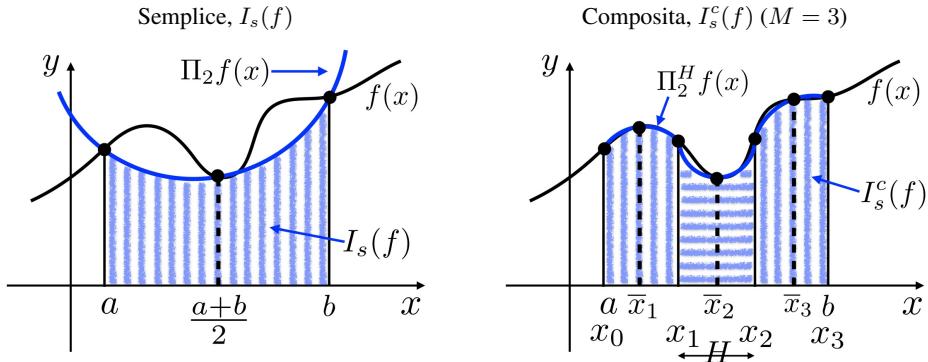
$$I_s(f) := I(\Pi_2 f) = \frac{b-a}{6} \left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right], \quad (6.3)$$

dove $\Pi_2 f(x)$ è il polinomio di grado 2 che interpola $f(x)$ nei nodi a, b e nel punto medio $(a+b)/2$. La formula di quadratura di Simpson composita è definita come:

$$I_s^c(f) := I(\Pi_2^H f) = \frac{H}{6} \sum_{k=1}^M [f(x_{k-1}) + 4f(\bar{x}_k) + f(x_k)],$$

dove $\Pi_2^H f(x)$ è il polinomio a tratti di grado 2 interpolante $f(x)$ nei nodi $\{x_k\}_{k=0}^M$ e nei punti medi $\{\bar{x}_k\}_{k=1}^M$ degli M sottointervalli di lunghezza $H = \frac{b-a}{M}$ in cui $[a, b]$ è suddiviso; i punti medi dei sottointervalli sono definiti come $\bar{x}_k = \frac{x_{k-1} + x_k}{2}$ per ogni $k = 1, \dots, M$.

Esempio 6.4.1. Illustriamo graficamente la formula di Simpson semplice (a sinistra) e composita (a destra) per l'approssimazione dell'integrale $I(f)$ di una generica funzione $f(x)$.



Proposizione 6.4.1. Se $f \in C^4([a, b])$, l'errore $e_s(f)$ associato alla formula di quadratura di Simpson semplice risulta

$$e_s(f) := I(f) - I_s(f) = -\frac{(b-a)^5}{2880} f^{(4)}(\xi) \quad \text{per un certo } \xi \in [a, b],$$

mentre l'errore $e_s^c(f)$ associato alla formula di quadratura di Simpson composita risulta

$$e_s^c(f) := I(f) - I_s^c(f) = -\frac{(b-a)}{2880} H^4 f^{(4)}(\xi) \quad \text{per un certo } \xi \in [a, b].$$

Osservazione 6.4.1. Le formule di quadratura di Simpson hanno grado di esattezza $r = 3$; infatti, gli errori $e_s(f)$ e $e_s^c(f)$ sono identicamente zero per tutti i polinomi di grado inferiore o uguale a 3 (se $f \in \mathbb{P}_3$, $f^{(4)}(\xi) = 0$ per ogni $\xi \in \mathbb{R}$).

Osservazione 6.4.2. La formula di Simpson composita ha ordine di convergenza (o ordine di accuratezza) 4; infatti, l'errore $e_s^c(f)$ risulta proporzionale ad H^4 .

Esempio 6.4.2. Si possono usare le stime dell'errore per le formule composite per determinare quale sia il numero minimo di intervalli necessari per raggiungere una certa accuratezza nell'approssimazione di un integrale. Si consideri ad esempio la formula dei trapezi composita per approssimare l'integrale $\int_{-2}^2 e^x dx$. Se vogliamo che l'errore di quadratura sia minore di una tolleranza $\varepsilon > 0$, imporremo che

$$|e_t^c(f)| = \left| -\frac{b-a}{12} H^2 f''(\xi) \right| \leq \frac{b-a}{12} H^2 \max_{x \in [a,b]} |f''(x)| < \varepsilon$$

da cui deve risultare, essendo $f''(x) = e^x$, $\max_{x \in [-2,2]} |f''(x)| = e^2$, che

$$\frac{b-a}{12} \left(\frac{b-a}{M} \right)^2 \max_{x \in [a,b]} |f''(x)| < \varepsilon \quad \Rightarrow \quad M^2 > \frac{(b-a)^3}{12\varepsilon} \max_{x \in [a,b]} |f''(x)|;$$

scegliendo ad esempio $\varepsilon = 10^{-4}$, si ottiene che $M > \sqrt{\frac{4^3}{12 \cdot 10^{-4}} e^2} \approx 627.76$, ovvero $M \geq 628$.

6.5 Formule di Quadratura Interpolatorie

Le formule di quadratura che abbiamo introdotto nelle sezioni precedenti sono i più semplici esempi di formule di quadratura interpolatorie. Possiamo infatti generalizzare quanto visto finora al caso in cui f sia approssimata sull'intervallo $[a, b]$ mediante un polinomio interpolante di grado n in $n+1$ nodi. Per semplicità, consideriamo il caso delle formule semplici, sebbene l'estensione al caso delle formule composite non rappresenti difficoltà evidenti.

Definizione 6.5.1. Sia $f(x) \in C^0([a, b])$. Una formula di quadratura interpolatoria (semplice) è definita come

$$I_{q,n}(f) := I(\tilde{f}) = \sum_{j=0}^n \alpha_j f(y_j), \quad (6.4)$$

dove $\tilde{f}(x)$ è una funzione interpolante $f(x)$ in $n+1$ nodi di quadratura $\{y_j\}_{j=0}^n$ in $[a, b]$ e $\{\alpha_j\}_{j=0}^n$ sono i corrispondenti pesi di quadratura, con $n \geq 0$.

La funzione interpolante $\tilde{f}(x)$ deve risultare “semplice” da integrare, e può essere ottenuta in molti modi diversi; la sua scelta determina il metodo (o la famiglia di formule) di quadratura numerica.

Osservazione 6.5.1. Se $\tilde{f}(x) = \Pi_n f(x)$, il polinomio interpolante di grado $n = 0, 1, 2$ su $n+1$ nodi equispaziati in $[a, b]$, si ottengono rispettivamente le formule di quadratura del punto medio, del trapezio e di Simpson (semplici). Effettivamente, scegliendo

$$\tilde{f}(x) = \Pi_0 f(x) \quad \text{con } n = 0, \quad \alpha_0 = b-a \quad \text{e} \quad y_0 = \frac{a+b}{2}$$

in Eq. (6.4), si ottiene la formula di quadratura del punto medio semplice (6.1). Per

$$\tilde{f}(x) = \Pi_1 f(x) \quad \text{con } n = 1, \quad \alpha_0 = \alpha_1 = \frac{b-a}{2} \quad \text{e} \quad y_0 = a, y_1 = b$$

in Eq. (6.4), si ottiene la formula di quadratura del trapezio semplice (6.2). Infine, per

$$\tilde{f}(x) = \Pi_2 f(x) \quad \text{con } n = 2, \quad \alpha_0 = \alpha_2 = \frac{b-a}{6}, \quad \alpha_1 = \frac{2(b-a)}{3} \quad e \quad y_0 = a, \quad y_1 = \frac{a+b}{2}, \quad y_2 = b$$

in Eq. (6.4), si ottiene la formula di quadratura di Simpson semplice (6.3).

In generale, se si sceglie

$$\tilde{f}(x) = \Pi_n f(x) = \sum_{k=0}^n f(x_k) \varphi_k(x),$$

polinomio interpolante di Lagrange di grado $n \geq 0$ su $n+1$ nodi $\{x_k\}_{k=0}^n$ in $[a, b]$, con $\{\varphi_k(x)\}_{k=0}^n$ le corrispondenti funzioni caratteristiche di Lagrange, la formula interpolatoria (6.4) si ottiene scegliendo i nodi di quadratura coincidenti a quelli di interpolazione, $y_j = x_j$, e pesi $\alpha_j = \int_a^b \varphi_j(x) dx$ per ogni $j = 0, \dots, n$. In tal caso, il grado di esattezza delle formule è $r \geq n$.

Definizione 6.5.2. Le formule di quadratura (6.4) si dicono di Newton–Cotes (semplici) se la funzione approssimante $f(x)$ è il suo polonomio interpolante di grado n , ovvero $\Pi_n f(x)$, su nodi equispaziati in $[a, b]$. Le formule di Newton–Cotes si dicono chiuse se $y_0 = a$ e $y_n = b$, con $h = \frac{b-a}{n}$; le formule si dicono aperte se $y_0 = a + h$ e $y_n = b - h$, con $h = \frac{b-a}{n+2}$.

Proposizione 6.5.1. Se n è pari e $f \in C^{n+2}([a, b])$, l'errore associato alla formula di Newton–Cotes è:

$$e_{q,n}(f) = I_{q,n}(f) - I(f) = \frac{M_n}{(n+2)!} h^{n+3} f^{(n+2)}(\xi) \quad \text{per qualche } \xi \in [a, b],$$

con $M_n \in \mathbb{R}$ (positivo o negativo) dipendente dall'uso di una formula chiusa o aperta. Se invece n è dispari e $f \in C^{n+1}([a, b])$, l'errore associato alla formula di Newton–Cotes è:

$$e_{q,n}(f) = I_{q,n}(f) - I(f) = \frac{K_n}{(n+1)!} h^{n+2} f^{(n+1)}(\xi) \quad \text{per qualche } \xi \in [a, b],$$

con $K_n \in \mathbb{R}$ (positivo o negativo) dipendente dall'uso di una formula chiusa o aperta.

Osservazione 6.5.2. Le formule di Newton–Cotes (semplici) hanno grado di esattezza $r = n+1$ se n è pari ($n = 0, 2, 4, \dots$), mentre hanno grado di esattezza $r = n$ se n è dispari ($n = 1, 3, 5, \dots$).

Osservazione 6.5.3. Le formule del trapezio e di Simpson sono formule di Newton–Cotes chiuse; la formula del punto medio è una formula di Newton–Cotes aperta.

Le formule di quadratura interpolatorie (6.4) sono specificate da n , i nodi di quadratura $\{y_j\}_{j=0}^n$ e i pesi di quadratura $\{\alpha_j\}_{j=0}^n$. Tuttavia, i pesi e i nodi in quadratura dipendono dall'intervallo $[a, b] \subset \mathbb{R}$ in questione. Per fornire formule di quadratura generali che possono essere applicate a funzioni $f(x)$ definite in un qualsiasi intervallo $[a, b]$, i nodi e i pesi di quadratura sono specificati per l'intervallo di riferimento $[-1, 1]$ e indicati come $\{\bar{y}_j\}_{j=0}^n$ e $\{\bar{\alpha}_j\}_{j=0}^n$, rispettivamente. I nodi di quadratura e i pesi per l'intervallo generale $[a, b]$ possono essere quindi ottenuti rispettivamente come⁴:

$$y_j = \frac{a+b}{2} + \frac{b-a}{2} \bar{y}_j \quad \text{per } j = 0, \dots, n,$$

e

$$\alpha_j = \frac{b-a}{2} \bar{\alpha}_j \quad \text{per } j = 0, \dots, n.$$

⁴La prima di queste due relazioni è analoga alla trasformazione usata in Definizione 5.2.6 per mappare i nodi di Chebyshev–Gauss–Lobatto dall'intervallo di riferimento $[-1, 1]$ al generico intervallo $[a, b]$.

6.5.1 Formule di quadratura Gaussiane

Le formule di Newton–Cotes possiedono grado di esattezza $r \geq n$; specificamente, per n pari, si ha $r = n + 1$, mentre per n dispari, si ha $r = n$. Risulta tuttavia possibile trovare, per un dato $n \geq 0$, la posizione ottimale dei nodi di quadratura $\{\bar{y}_j\}_{j=0}^n$ in $[-1, 1]$ e i valori dei corrispondenti pesi di quadratura $\{\bar{\alpha}_j\}_{j=0}^n$ tali che il *grado di esattezza* della formula di quadratura è *massimo*. Le corrispondenti formule di quadratura si ottengono considerando polinomi interpolanti costruiti usando come basi opportune famiglie di polinomi (detti polinomi ortogonali di *Legendre*) e prendono il nome di *formule di quadratura Gaussiane*. Il loro grado di esattezza risulta pari a $r = 2n + 1$ se gli estremi dell’intervallo $[-1, 1]$ sono esclusi (formule di *Gauss–Legendre*), oppure pari a $r = 2n - 1$ se sono inclusi (formule di *Gauss–Legendre–Lobatto*). Ricordiamo che tali formule di quadratura si possono estendere poi a qualsiasi intervallo $[a, b]$ come visto in precedenza.

6.5.2 Formule di quadratura di Gauss–Legendre

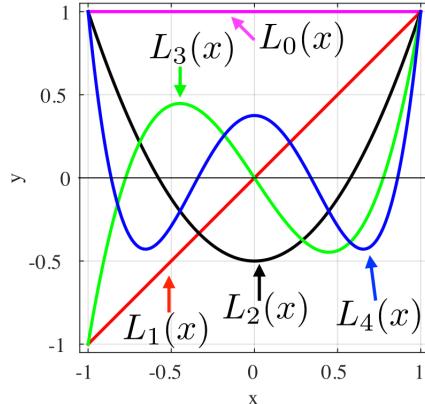
Le *formule di quadratura di Gauss–Legendre* indicano una famiglia di formule di quadratura interpolatorie ottenute approssimando la funzione integranda $f(x)$ mediante *polinomi di Legendre*. I polinomi di Legendre $\{L_k(x)\}_{k=0}^{n+1}$ sull’intervallo $[-1, 1]$ sono definiti ricorsivamente come segue:

$$L_0(x) = 1, \quad L_1(x) = x, \quad \text{e} \quad L_{k+1}(x) = \frac{2k+1}{k+1} x L_k(x) - \frac{k}{k+1} L_{k-1}(x) \quad \text{per } k = 1, \dots, n;$$

tali polinomi risultano ortogonali, nel senso che $\int_{-1}^1 L_{n+1}(x) L_k(x) dx = 0$ per ogni $k = 0, \dots, n$.

Esempio 6.5.1. Consideriamo i polinomi di Legendre sull’intervallo $[-1, 1]$ nel caso $n = 3$; si ha:

$$\begin{aligned} L_0(x) &= 1, \\ L_1(x) &= x, \\ L_2(x) &= \frac{3}{2} x L_1(x) - \frac{1}{2} L_0(x), \\ L_3(x) &= \frac{5}{3} x L_2(x) - \frac{2}{3} L_1(x), \\ L_4(x) &= \frac{7}{4} x L_3(x) - \frac{3}{4} L_2(x). \end{aligned}$$



Osservazione 6.5.4. Più in generale, se $w = w(x)$ è una funzione peso definita su $[-1, 1]$, non negativa, l’insieme $\{P_k(x)\}_{k=0}^n$ è una famiglia di polinomi ortogonali rispetto a w se

$$\int_{-1}^1 P_k(x) P_m(x) w(x) dx = 0 \quad \text{se } k \neq m$$

con $P_k \in \mathbb{P}_k$. I polinomi di Legendre sono polinomi ortogonali su $[-1, 1]$ rispetto alla funzione peso $w(x) = 1$; più precisamente,

$$\int_{-1}^1 L_k(x) L_m(x) dx = \frac{2}{2m+1} \delta_{km}$$

dove $L_m(x)$ indica il polinomio di Legendre di grado m .

Definizione 6.5.3. Sia $f(x) \in C^0([a, b])$, allora la formula di quadratura di Gauss–Legendre per $n \geq 0$ sull’intervallo di riferimento $[-1, 1]$ è

$$I_{GL,n} = \sum_{j=0}^n \bar{\alpha}_j^{GL} f(\bar{y}_j^{GL}),$$

dove i nodi e i pesi di quadratura sono dati, rispettivamente, da

$$\bar{y}_j^{GL} := \text{zeri di } L_{n+1}(x) \quad \text{per ogni } j = 0, \dots, n,$$

$$\bar{\alpha}_j^{GL} := \frac{2}{\left[1 - (\bar{y}_j^{GL})^2\right] \left[L'_{n+1}(\bar{y}_j^{GL})\right]^2} \quad \text{per ogni } j = 0, \dots, n.$$

Osservazione 6.5.5. Il grado di esattezza della formula di Gauss–Legendre è $r = 2n + 1$ per ogni $n \geq 0$.

Nella seguente tabella vengono riportati i nodi e i pesi di quadratura delle formule di Gauss–Legendre sull’intervallo $[-1, 1]$ per $n = 0, 1, 2$, e il corrispondente grado di esattezza r . Tale formula massimizza il grado di esattezza r per ogni dato $n \geq 0$.

n	$\{\bar{y}_j^{GL}\}_{j=0}^n$	$\{\bar{\alpha}_j^{GL}\}_{j=0}^n$	r
0	0	2	1 (formula del punto medio)
1	$\left\{-\frac{1}{\sqrt{3}}, +\frac{1}{\sqrt{3}}\right\}$	{1, 1}	3
2	$\left\{-\frac{\sqrt{15}}{5}, 0, +\frac{\sqrt{15}}{5}\right\}$	$\left\{\frac{5}{9}, \frac{8}{9}, \frac{5}{9}\right\}$	5

Osserviamo come la formula di quadratura di Gauss–Legendre nel caso $n = 0$ coincida con la formula del punto medio semplice.

Esempio 6.5.2. Possiamo verificare che la formula di Gauss–Legendre per $n = 1$ ha grado di esattezza $r = 3$, ovvero $I_{GL,1}(f) = I(f)$ per ogni $f \in \mathbb{P}_3$. Considerando una generica funzione $f(x) = c_0 + c_1x + c_2x^2 + c_3x^3$ per determinati c_0, c_1, c_2 , e $c_3 \in \mathbb{R}$, per esempio sull’intervallo di riferimento $[-1, 1]$, abbiamo $I(f) = \int_{-1}^1 f(x) dx = 2c_0 + \frac{2}{3}c_2$. Prendendo $n = 1$, abbiamo che $I_{GL,1}(f) = (\bar{\alpha}_0^{GL} + \bar{\alpha}_1^{GL}) c_0 + (\bar{\alpha}_0^{GL} \bar{y}_0^{GL} + \bar{\alpha}_1^{GL} \bar{y}_1^{GL}) c_1 + (\bar{\alpha}_0^{GL} (\bar{y}_0^{GL})^2 + \bar{\alpha}_1 (\bar{y}_1^{GL})^2) c_2 + (\bar{\alpha}_0 (\bar{y}_0^{GL})^3 + \bar{\alpha}_1 (\bar{y}_1^{GL})^3) c_3$. Imponendo che valgano i seguenti vincoli (cioè, imponendo che $I_{GL,1}(f) = I(f)$ per ogni c_0, c_1, c_2 , e $c_3 \in \mathbb{R}$):

$$\begin{cases} \bar{\alpha}_0^{GL} + \bar{\alpha}_1^{GL} = 2, \\ \bar{\alpha}_0^{GL} \bar{y}_0^{GL} + \bar{\alpha}_1^{GL} \bar{y}_1^{GL} = 0, \\ \bar{\alpha}_0^{GL} (\bar{y}_0^{GL})^2 + \bar{\alpha}_1 (\bar{y}_1^{GL})^2 = \frac{2}{3}, \\ \bar{\alpha}_0^{GL} (\bar{y}_0^{GL})^3 + \bar{\alpha}_1 (\bar{y}_1^{GL})^3 = 0, \end{cases}$$

otteniamo i nodi di quadratura $\bar{y}_0^{GL} = -\frac{1}{\sqrt{3}}$ e $\bar{y}_1^{GL} = +\frac{1}{\sqrt{3}}$, e i corrispondenti pesi $\bar{\alpha}_0^{GL} = \bar{\alpha}_1^{GL} = 1$; deduciamo quindi che la formula di Gauss–Legendre $I_{GL,1}(f)$ integra esattamente polinomi di grado 3 indipendentemente dal valore dei coefficienti c_0, c_1, c_2 , e $c_3 \in \mathbb{R}$, ovvero abbiamo verificato che la formula ha grado di esattezza pari a 3.

Per quale motivo il grado di esattezza di queste formule di quadratura è massimo? La ragione è fornita dal seguente risultato.

Teorema 6.5.1. Per un dato $m > 0$, la formula di quadratura

$$I_n(f) = \sum_{j=0}^n \bar{\alpha}_j f(\bar{y}_j)$$

ha grado di esattezza $n + m$ se e solo se è di tipo interpolatorio e il polinomio nodale

$$\omega_{n+1}(x) = \prod_{j=0}^n (x - \bar{y}_j)$$

associato ai nodi $\{\bar{y}_j\}_{j=0}^n$ è tale che

$$\int_{-1}^1 \omega_{n+1}(x) p(x) dx = 0 \quad \text{per ogni } p \in \mathbb{P}_{m-1}. \quad (6.5)$$

Si può mostrare che il valore massimo che m può assumere è $n + 1$ e che tale valore viene raggiunto quando $\omega_{n+1}(x)$ è proporzionale al polinomio di Legendre $L_{n+1}(x)$ di grado $n + 1$. Si può infatti verificare che L_{n+1} è ortogonale a tutti i polinomi di grado minore o uguale a n , ovvero

$$\int_{-1}^1 L_{n+1}(x) p(x) dx = 0 \quad \text{per ogni } p \in \mathbb{P}_n$$

dal momento che:

- un generico polinomio di grado n , $p \in \mathbb{P}_n$, può essere espresso come combinazione lineare dei polinomi L_0, L_1, \dots, L_n (per opportuni coefficienti), ovvero

$$p(x) = \beta_0 L_0(x) + \beta_1 L_1(x) + \dots + \beta_n L_n(x);$$

- $L_{n+1}(x)$ è ortogonale a $L_0(x), L_1(x), \dots, L_n(x)$, per definizione di polinomi di Legendre.

Dunque la relazione (6.5) è verificata a patto che $\omega_{n+1}(x)$ sia proporzionale a $L_{n+1}(x)$. In tal caso, $m - 1 = n$ da cui $m = n + 1$ e dunque il grado di esattezza della formula $I_{GL,n}(f)$ è $n + m = 2n + 1$.

Osserviamo inoltre che $\omega_{n+1}(x)$ è un polinomio monico (ovvero, il cui termine di grado massimo ha coefficiente pari a 1) di grado $n + 1$, ortogonale a tutti i polinomi di grado inferiore; tale polinomio è l'unico polinomio monico multiplo di $L_{n+1}(x)$; in particolare, le radici di $\omega_{n+1}(x)$, indicate con \bar{y}_j , coincidono con quelle di $L_{n+1}(x)$, ovvero

$$L_{n+1}(\bar{y}_j) = 0 \quad \text{per ogni } j = 0, \dots, n.$$

Le ascisse $\{\bar{y}_j\}_{j=0}^n$ si dicono *nodi di Gauss* (da cui il nome di formule di quadratura di Gauss) per una generica famiglia di polinomi ortogonali: usando i polinomi di Legendre si ottengono le formule di quadratura di Gauss-Legendre.

Osservazione 6.5.6. Una formula di Gauss-Legendre ha tutti pesi positivi, $\bar{\alpha}_j^{GL} > 0$ per ogni $j = 0, \dots, n$, e nodi \bar{y}_j^{GL} interni all'intervallo $[-1, 1]$, per ogni $j = 0, \dots, n$.

6.5.3 Formule di quadratura di Gauss-Legendre-Lobatto

Le formule di quadratura di Gauss-Legendre massimizzano il grado di esattezza r per ogni dato valore di $n \geq 0$, ma i corrispondenti nodi di quadratura sono tutti interni all'intervallo di riferimento $[-1, 1]$. Tuttavia, in alcune situazioni si vorrebbe includere gli estremi dell'intervallo $\{-1, 1\}$ nell'insieme dei nodi di quadratura. Le formule di quadratura di Gauss-Legendre-Lobatto permettono di massimizzare il grado di esattezza nel caso in cui gli estremi dell'intervallo $\{-1, 1\}$ siano inclusi nell'insieme dei nodi di quadratura; tali formule usano come nodi di quadratura le $n + 1$ radici del polinomio $\tilde{L}_{n+1}(x) = L_{n+1}(x) + aL_n(x) + bL_{n-1}(x)$, dove a e b sono scelti in modo tale che $\tilde{L}_{n+1}(-1) = \tilde{L}_{n+1}(1) = 0$.

Definizione 6.5.4. Sia $f(x) \in C^0([a, b])$, allora la formula di quadratura di Gauss–Legendre–Lobatto per $n \geq 1$ sull’intervallo di riferimento $[-1, 1]$ è

$$I_{GLL,n} = \sum_{j=0}^n \bar{\alpha}_j^{GLL} f(\bar{y}_j^{GLL}),$$

dove i nodi e i pesi di quadratura sono dati, rispettivamente, da

$$\bar{y}_0^{GLL} := -1, \quad \bar{y}_n^{GLL} := +1, \quad e \quad \bar{y}_j^{GLL} := zeri di L'_n(x) \quad per ogni j = 1, \dots, n-1,$$

$$\bar{\alpha}_j^{GLL} := \frac{2}{n(n+1)} \frac{1}{[L_n(\bar{y}_j^{GLL})]^2} \quad per ogni j = 0, \dots, n.$$

Osservazione 6.5.7. Il grado di esattezza della formula di Gauss–Legendre–Lobatto è $r = 2n - 1$ per ogni $n \geq 1$.

Nella seguente tabella vengono riportati i nodi e i pesi di quadratura delle formule di Gauss–Legendre–Lobatto sull’intervallo $[-1, 1]$ per $n = 1, 2, 3$, e il corrispondente grado di esattezza r . Osserviamo che tale formula non è definita per $n = 0$.

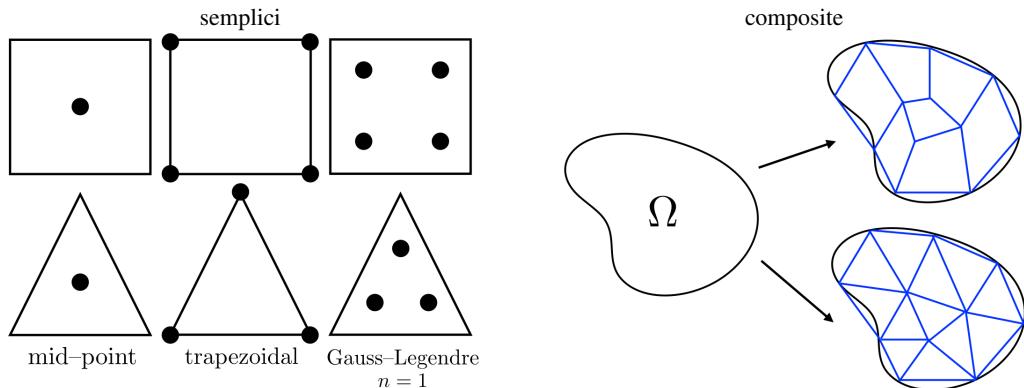
n	$\{\bar{y}_j^{GLL}\}_{j=0}^n$	$\{\bar{\alpha}_j^{GLL}\}_{j=0}^n$	r
1	$\{-1, +1\}$	$\{1, 1\}$	1 (formula del trapezio)
2	$\{-1, 0, +1\}$	$\left\{\frac{1}{3}, \frac{4}{3}, \frac{1}{3}\right\}$	3 (formula di Simpson)
3	$\left\{-1, -\frac{1}{\sqrt{5}}, +\frac{1}{\sqrt{5}}, +1\right\}$	$\left\{\frac{1}{6}, \frac{5}{6}, \frac{5}{6}, \frac{1}{6}\right\}$	5

Osserviamo come le formule di quadratura di Gauss–Legendre–Lobatto nei casi $n = 1$ e 2 coincidano con la formula del trapezio semplice e con la formula di Simpson semplice, rispettivamente.

6.6 Integrazione Numerica in Dimensione $d > 1$

L’integrazione numerica in più dimensioni, ovvero l’integrazione di funzioni continue $f : \Omega \rightarrow \mathbb{R}$, con $\Omega \subset \mathbb{R}^d$ per $d \geq 2$, si basa sulla generalizzazione delle formule di quadratura viste nelle sezioni precedenti. Le formule di quadratura semplici sono definite per domini specifici, come ad esempio trapezi e triangoli per $d = 2$ oppure tetraedri ed esaedri per $d = 3$. L’integrazione numerica su domini complessi si basa invece su formule composite. Tali formule sono di interesse fondamentale nei codici di calcolo ad elementi finiti per simulazioni numeriche di problemi differenziali su qualsiasi griglia di calcolo (in dimensione maggiore di 1); per il caso $d = 2$ viene riportata di seguito una rappresentazione schematica di alcune di queste formule di quadratura.

Esempio 6.6.1. Rappresentazione schematica di una formula di quadratura per $d = 2$.



Capitolo 7

Equazioni Differenziali Ordinarie

Affrontiamo in questo capitolo l'approssimazione numerica di Equazioni Differenziali Ordinarie (EDO), sia nel caso scalare che vettoriale, considerando in un primo momento equazioni del primo ordine, ed estendendo successivamente la trattazione al caso di equazioni del secondo ordine. L'approssimazione numerica di sistemi dinamici costituisce un aspetto fondamentale nel calcolo scientifico, dato il vastissimo insieme di applicazioni che conducono alla soluzione di tali problemi. Sebbene in questo corso ci concentriamo sui più semplici schemi numerici per approssimare numericamente la soluzione di un problema di Cauchy, le idee esposte in questo capitolo costituiscono la base di numerosi schemi numerici, implementati in molti codici di calcolo abitualmente usati.

7.1 Esempi Notevoli di EDO

Le EDO sono uno strumento imprescindibile per la formulazione matematica di leggi fisiche, a cominciare dalla legge di Newton

$$m \frac{d^2x}{dt^2}(t) = f(t),$$

che esprime il legame tra l'accelerazione di una massa puntiforme m e la forza f che agiscono su di essa. Se la forza dipende a sua volta dallo spostamento, come nel caso di una molla ($f = -kx$), l'equazione diviene

$$m \frac{d^2x}{dt^2}(t) = -kx(t)$$

ed esprime il fatto che la forza esercitata dalla molla è proporzionale allo spostamento rispetto alla posizione di riposo ed agisce nella direzione opposta ($k > 0$).

Un altro esempio notevole, sempre tratto dalla meccanica classica, è quello di un pendolo semplice che oscilla in un piano verticale per effetto della gravità g ed eventualmente dell'attrito del perno. In tal caso si ha un legame tra l'angolo θ formato dal pendolo con la verticale e le sue velocità e accelerazione angolare, espresso dall'equazione:

$$l \frac{d^2\theta}{dt^2}(t) + \alpha \frac{d\theta}{dt}(t) + g \sin \theta(t) = 0,$$

dove l è la lunghezza del filo. In questi casi si parla di equazioni differenziali ordinarie poiché le incognite dipendono da una sola variabile indipendente, che convenzionalmente indicheremo con t (facendo prevalere le applicazioni di tipo dinamico o evolutivo, in cui t è il tempo). Le equazioni differenziali alle derivate parziali esprimono relazioni tra incognite dipendenti da più variabili (ad esempio il tempo e una o più variabili spaziali) e quindi fanno intervenire le derivate parziali delle incognite. Entrambe le equazioni introdotte sono di ordine 2, essendo 2 l'ordine più alto delle derivate della funzione incognita.

Altri casi di interesse descrivono:

- il decadimento della radioattività di un materiale, mediante l'equazione

$$y'(t) = \lambda y(t) \quad \lambda \in \mathbb{R}, \lambda < 0, \quad (7.1)$$

dal momento che il tasso di riduzione della quantità y di una sostanza radioattiva è proporzionale, con $\lambda < 0$, alla quantità di sostanza stessa ($y'(t) = \frac{dy}{dt}(t)$). Essendo

$$\frac{d}{dt} e^{\lambda t} = \lambda e^{\lambda t} \quad \text{per } t \in \mathbb{R}, \quad (7.2)$$

valida per ogni $\lambda \in \mathbb{R}$, si ha che ogni soluzione dell'equazione (7.1) si scrive come

$$y(t) = ce^{\lambda t}, \quad \text{con } c \in \mathbb{R} \text{ arbitrario;}$$

tal problema, a cui ci riferiremo in seguito come *problema modello*, sarà ampiamente utilizzato anche per derivare alcune proprietà degli schemi numerici che introdurremo nel Capitolo 8;

- la crescita di una popolazione di batteri (con un tasso $\alpha > 0$) in un ambiente in cui non possono coesistere più di β elementi (equazione logistica), è descritta da una EDO del tipo:

$$y'(t) = \alpha y(t) \left(1 - \frac{y(t)}{\beta}\right); \quad (7.3)$$

- l'evoluzione di due popolazioni, y_1 and y_2 , di prede e predatori, rispettivamente, corrisponde a un sistema di EDO del primo ordine:

$$\begin{cases} y'_1(t) = C_1 y_1(t) [1 - b_1 y_1(t) - d_2 y_2(t)], \\ y'_2(t) = -C_2 y_2(t) [1 - b_2 y_2(t) - d_1 y_1(t)], \end{cases} \quad (7.4)$$

dove C_1 e C_2 rappresentano i tassi di crescita delle due popolazioni, d_1 e d_2 governano il tipo di mutua interazione mentre b_1 e b_2 sono relativi alla quantità disponibile di nutrimento (equazioni di Lotka-Volterra);

- come vedremo nel seguito, anche la discretizzazione in spazio di un'Equazione alle Derivate Parziali parabolica, come l'equazione del calore seguente

$$\begin{cases} \frac{\partial u}{\partial t} - D \frac{\partial^2 u}{\partial x^2} = f & x \in (a, b), t \in (0, T), \\ u(a, t) = u(b, t) = 0 & t > 0, \\ u(x, 0) = u_0(x), & x \in [a, b], \end{cases} \quad (7.5)$$

mediante il metodo degli elementi finiti conduce alla soluzione di un sistema di EDO lineare e del primo ordine, nella forma

$$\begin{cases} \mathbf{u}'(t) + A\mathbf{u}(t) = \mathbf{f}(t) & \text{per } t \in (0, T), \\ \mathbf{u}(t_0) = \mathbf{u}_0, \end{cases}$$

per il vettore \mathbf{u} dei gradi di libertà della soluzione approssimata.

7.2 Il Problema di Cauchy: Buona Posizione

Consideriamo più in dettaglio il caso di EDO scalari (vale a dire in una sola variabile incognita) nella forma:

$$\mathcal{F}(t, y(t), y'(t), y''(t), \dots, y^{(p)}(t)) = 0 \quad \text{per } t \in I = (t_0, t_f), \quad (7.6)$$

dove t è una variabile indipendente, che gioca spesso il ruolo di variabile temporale, $y(t)$ la soluzione del problema differenziale, e p l'ordine dell'equazione differenziale, ovvero l'ordine più alto delle derivate di y che intervengono nella (7.6). \mathcal{F} indica un generico legame funzionale tra tutti i suoi argomenti.

7.2.1 Il problema di Cauchy (o ai valori iniziali)

Per soluzione (in senso classico) dell'equazione differenziale in un intervallo $I \subseteq \mathbb{R}$, intendiamo una funzione $y : I \rightarrow \mathbb{R}$, derivabile p volte in I , tale che valga la relazione (7.6). Spesso è possibile esprimere mediante la (7.6) la derivata di ordine massimo in funzione di t e delle derivate di ordine inferiore, in modo da scrivere l'equazione differenziale nella forma (detta normale)

$$y^{(p)}(t) = f(t, y(t), y'(t), y''(t), \dots, y^{(p-1)}(t)) \quad \text{per } t \in I. \quad (7.7)$$

Tratteremo il caso di EDO del primo ordine ($p = 1$), nella forma:

$$y'(t) = f(t, y(t)) \quad \text{per } t \in I, \quad (7.8)$$

dove $f : I \times \mathbb{R} \rightarrow \mathbb{R}$ è funzione di due argomenti; di EDO del secondo ordine ($p = 2$):

$$y''(t) = f(t, y(t), y'(t)) \quad \text{per } t \in I, \quad (7.9)$$

dove $f : I \times \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ è funzione di tre argomenti; e di sistemi di EDO del primo ordine:

$$\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)) \quad \text{per } t \in I, \quad (7.10)$$

dove, in quest'ultimo caso, $\mathbf{y}(t) = (y_1(t), \dots, y_n(t))^T$ and $\mathbf{f}(t, \mathbf{y}) = (f_1(t, \mathbf{y}), \dots, f_n(t, \mathbf{y}))^T$, per $\mathbf{f} : I \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ funzione di due argomenti.

Osservazione 7.2.1. Risulta sempre possibile scrivere una EDO di ordine n (in una sola variabile $y(t)$) come un sistema di n EDO (in n variabili) del primo ordine, ponendo

$$y_i(t) = y^{(i-1)}(t), \quad \text{per } t \in I, \quad \forall i = 1, \dots, n,$$

ovvero:

$$\begin{aligned} y_1(t) &= y(t) \\ y_2(t) &= y'(t) = y'_1(t) \\ y_3(t) &= y''(t) = (y'(t))' = y'_2(t) \\ &\vdots \\ y_n(t) &= y^{((n-1)}(t) = (y^{(n-2)}(t))' = y'_{n-1}(t), \end{aligned}$$

per $t \in I$; l'EDO originale risulta allora

$$y'_n(t) = f(t, y_1(t), \dots, y_n(t)) \quad \text{per } t \in I,$$

ovvero dà luogo al seguente sistema di EDO del primo ordine:

$$\left\{ \begin{array}{l} y'_1(t) = y_2(t) \\ \vdots \\ y'_{n-1}(t) = y_n(t) \\ y'_n(t) = f(t, y_1(t), \dots, y_n(t)) \end{array} \right. \begin{array}{l} \text{per } t \in I, \\ \text{per } t \in I, \\ \text{per } t \in I. \end{array}$$

Una EDO ha, in generale, infinite soluzioni; in generale, tali soluzioni dipendono, oltre che da t , da p costanti arbitrarie c_1, \dots, c_p , dove p indica l'ordine della EDO. Un modo molto naturale per selezionare una soluzione particolare è quello di imporre che la soluzione ad un certo istante $t_0 \in I$ (solitamente l'*istante iniziale* di $I = (t_0, t_f)$) assuma un valore assegnato $y_0 \in \mathbb{R}$, e lo stesso accada per le prime $p - 1$ derivate. Consideriamo cioè il problema di trovare $y : I \subset \mathbb{R} \rightarrow \mathbb{R}$ tale che

$$\boxed{\left\{ \begin{array}{l} y^{(p)}(t) = f(t, y(t), y'(t), \dots, y^{(p-1)}(t)) \quad \text{per } t \in I, \\ y(t_0) = y_0, \\ y'(t_0) = v_0, \\ \dots \\ y^{(p-1)}(t_0) = z_0. \end{array} \right.}$$

Questo problema prende il nome di *problema di Cauchy* (o problema ai valori iniziali) per l'equazione differenziale. Nel caso particolare in cui $p = 1$, si ha una singola equazione e il problema di Cauchy diviene: trovare $y : I \subset \mathbb{R} \rightarrow \mathbb{R}$ tale che

$$\begin{cases} y'(t) = f(t, y(t)) & \text{per } t \in I, \\ y(t_0) = y_0, \end{cases} \quad (7.11)$$

dove $f : I \times \mathbb{R} \rightarrow \mathbb{R}$ è una funzione data di due argomenti e y_0 è il dato iniziale.

Se f è continua rispetto alla prima variabile, il problema si può trasformare nel problema integrale seguente:

$$\int_{t_0}^t \frac{dy}{d\tau}(\tau) d\tau = \int_{t_0}^t f(\tau, y(\tau)) d\tau \quad \text{per ogni } t \in I,$$

da cui, se richiediamo che $y \in C^1(I)$, allora:

$$y(t) = y(t_0) + \int_{t_0}^t f(\tau, y(\tau)) d\tau = y_0 + \int_{t_0}^t f(\tau, y(\tau)) d\tau.$$

Si noti come il valore di $y(t_0) = y_0$ assegnato permette di identificare una soluzione tra le infinite possibili, tutte altrimenti definite a meno di una costante. Osserviamo inoltre, che la precedente consente di determinare l'espressione analitica del problema di Cauchy solo a condizione di poter determinare la primitiva di $f(t, y(t))$, operazione in generale tutt'altro che banale.

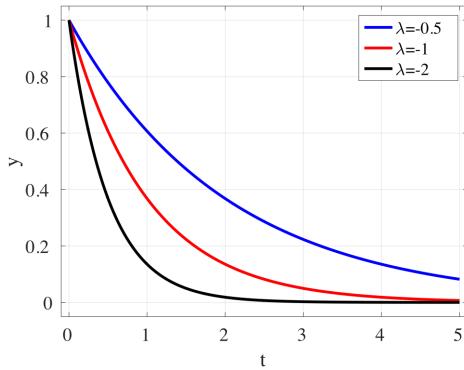
Nel caso di un sistema di n EDO di ordine $p = 1$, il problema di Cauchy diviene: trovare $\mathbf{y} : I \subset \mathbb{R} \rightarrow \mathbb{R}^m$ tale che

$$\begin{cases} \mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)) & \text{per } t \in I, \\ \mathbf{y}(t_0) = \mathbf{y}_0. \end{cases} \quad (7.12)$$

Esempio 7.2.1. Problema Modello. Il *problema modello* è un problema di Cauchy (7.11) (Eq. (7.1)) tale per cui $f(t, y) = \lambda y$ per $\lambda \in \mathbb{R}$ e $\lambda < 0$. Tale problema ammette la soluzione in forma chiusa:

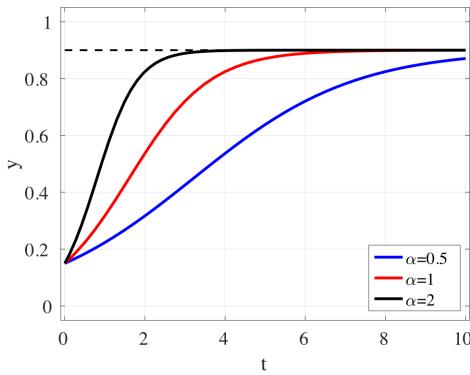
$$y(t) = y_0 e^{\lambda(t-t_0)} \quad \text{per ogni } t \in [t_0, t_f]; \quad (7.13)$$

spesso il problema modello è definito nell'intervallo $I = (t_0, +\infty)$.



Soluzioni del problema modello per $y_0 = 1$, $t_0 = 0$, $t_f = 5$ e differenti valori di λ ; $\lambda = -0.5$ (blu), -1.0 (rosso), and -2.0 (nero).

Esempio 7.2.2. Equazione della logistica. Modelli biologici spesso coinvolgono EDO. Un semplice modello per descrivere l'evoluzione della concentrazione $y(t)$ di batteri in una soluzione corrisponde al problema di Cauchy (7.11) (Eq. (7.3)) con $f(t, y) = \alpha y \left(1 - \frac{1}{\beta}\right)$, dove α e $\beta > 0$, dove l'ultimo parametro rappresenta la massima concentrazione di batteri.



Soluzione del problema per $y_0 = 0.15$, $\beta = 0.9$, $t_0 = 0$, $t_f = 10$ e diversi valori di α ; $\alpha = 0.5$ (blu), 1.0 (rosso) e 2.0 (nero).

Esempio 7.2.3. Modello epidemiologico SEIR. Consideriamo un modello epidemiologico compartimentale di tipo SEIR per lo studio della dinamica di una malattia infettiva in un gruppo di individui (popolazione).

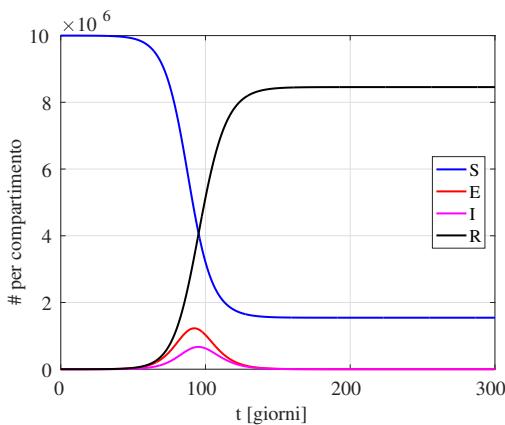
https://it.wikipedia.org/wiki/Modelli_matematici_in_epidemiologia

Il modello SEIR considera in una popolazione di N individui i seguenti compartimenti: $S(t)$, il numero di individui suscettibili al tempo t ; $I(t)$, il numero di individui infettivi; $E(t)$, il numero di individui esposti (per tener conto del periodo di incubazione per individui infetti, ma non ancora infettivi) e $R(t)$ il numero di recuperati o rimossi, ovvero individui che non rientrano più nella dinamica dell'epidemia (guariti, immuni, isolati, deceduti, etc...). Il modello SEIR corrisponde a un sistema di EDO del primo ordine nella forma seguente:

$$\left\{ \begin{array}{ll} \frac{dS}{dt} = -\beta \frac{S I}{N} & t \in I, \\ \frac{dE}{dt} = +\beta \frac{S I}{N} - \alpha E & t \in I, \\ \frac{dI}{dt} = +\alpha E - \gamma I & t \in I, \\ \frac{dR}{dt} = +\gamma I & t \in I, \\ S(t_0) = S_0, E(t_0) = E_0, I(t_0) = I_0, R(t_0) = R_0, \end{array} \right.$$

dove $I = (t_0, t_f)$, S_0 , E_0 , I_0 e R_0 sono i valori iniziali, β rappresenta il tasso di infezione o contagio, α il tasso di incubazione (α^{-1} rappresenta il tempo di incubazione medio), γ il tasso di rimozione (γ^{-1} rappresenta il tempo infettivo medio). Il valore del numero di riproduzione di base è $\mathcal{R}_0 = \frac{\beta}{\gamma}$, ovvero il numero di infetti secondari

medi che ogni infezione produce. Il numero di riproduzione effettivo è invece $\mathcal{R}_t = \mathcal{R}_0 \frac{S}{N}$, per il modello SEIR in considerazione, e rappresenta il numero di infetti secondari medi che ogni infezione produce al tempo t in una popolazione parziale di suscettibili. Osserviamo che, essendo $S(t) + E(t) + I(t) + R(t) = N$ per ogni $t \in [t_0, t_f]$, il precedente sistema di 4 equazioni può essere ridotto a un sistema di EDO nella forma di Eq. (7.12) con $m = 3$ equazioni in S , E ed I , da cui poi ricavare R , ovvero in cui $\mathbf{y}(t) = (S(t), E(t), I(t))^T$.



Esempio di evoluzione epidemica tramite modello SEIR per $N = 10^7$, $E_0 = 100$, $I_0 = 30$ con $\mathcal{R}_0 = 2.2$.

Piattaforma per l'analisi di dati dell'epidemia Covid-19 in Italia con simulazioni predittive basate su un modello epidemiologico compartimentale:

<https://www.epimox.polimi.it/>

Riportiamo le seguenti osservazioni.

- Un problema di Cauchy è *lineare*, quando $f(t, y)$ è una funzione lineare (affine) in entrambe le variabili. Ad esempio, il seguente problema di Cauchy:

$$\begin{cases} y'(t) = 3y(t) - 3t & \text{per } t > 0, \\ y(0) = 1, \end{cases} \quad (7.14)$$

con $f(t, y) = 3y - 3t$ è lineare; la sua soluzione risulta $y(t) = (1 - 1/3)e^{3t} + t + 1/3$.

- Nel caso in cui $f(t, y)$ non sia una funzione lineare (affine) in entrambe le variabili, il problema di Cauchy è *non lineare*. Ad esempio, il seguente problema di Cauchy:

$$\begin{cases} y'(t) = \sqrt[3]{y(t)} & \text{per } t > 0, \\ y(0) = 0, \end{cases}$$

con $f(t, y) = \sqrt[3]{y}$ è non lineare. Pur avendo assegnato un unico dato iniziale, questo problema ammette *tre soluzioni*: $y(t) = 0$, $y(t) = \sqrt{8t^3/27}$, $y(t) = -\sqrt{8t^3/27}$.

- Se $f = f(y)$, ovvero è funzione solo di un argomento $y \in \mathbb{R}$, allora l'EDO si dice *autonoma* e si scrive come

$$\begin{cases} y'(t) = f(y(t)) & \text{per } t \in I, \\ y(t_0) = y_0. \end{cases} \quad (7.15)$$

Esempi di EDO autonome sono il problema modello e l'equazione della logistica degli Esempi 7.2.1 e 7.2.2. Le soluzioni \bar{y} di un'EDO autonoma del primo ordine si dicono stazionarie se $\bar{y}' = 0$, ovvero $f(\bar{y}) = 0$. Tali soluzioni stazionarie, dette anche di equilibrio, possono essere instabili, stabili o asintoticamente stabili.

- La soluzione di un problema di Cauchy non è necessariamente definita per ogni $t \in I$. Ad esempio, nel caso seguente:

$$\begin{cases} y'(t) = 1 + y^2(t) & \text{per } t > 0, \\ y(0) = 0, \end{cases}$$

una soluzione risulta data dalla funzione $y(t) = \tan(t)$ dove $0 < t < \frac{\pi}{2}$, ovvero essa risulta una *soluzione locale*.

Non affrontiamo, in questa breve introduzione, la questione di determinare l'intervallo massimale di esistenza della soluzione, vale a dire il più grande intervallo I contenente t_0 in cui il problema è risolubile. Ci limitiamo a precisare, nella seguente sezione, quali ipotesi su f garantiscano che ogni soluzione di un problema di Cauchy esista su tutto l'intervallo I ; si parla, in tal caso, di *esistenza globale*¹.

7.2.2 Esistenza e unicità (globale) per il problema di Cauchy

Ci occupiamo ora della *buona posizione* del problema di Cauchy (7.11), ovvero di stabilire sotto quali ipotesi sui dati esso *ammetta un'unica soluzione*, che dipenda con continuità dai dati. Per semplicità consideriamo il caso di una equazione scalare ($n = 1$) di ordine $p = 1$; i risultati enunciati si estendono facilmente al caso dei sistemi di EDO (caso $n > 1$) e dunque al caso di EDO di ordine superiore al primo.

Diamo innanzitutto delle condizioni su $f(t, y)$ che garantiscono la risolubilità locale, ovvero almeno in un intorno di t_0 , del problema di Cauchy.

Iniziamo con l'osservare che la sola continuità di $f(t, y)$ rispetto a t e y è sufficiente a garantire l'*esistenza di (almeno) una funzione* (di classe C^1 in un intorno di t_0) che risulta soluzione del problema di Cauchy (7.12), ma non la sua unicità. Sotto la sola ipotesi di continuità di f possono esistere infinite soluzioni del problema di Cauchy.

¹Analogamente, si parla di risolubilità locale di un problema di Cauchy qualora si fosse interessati a trovare una soluzione del problema in un intorno di t_0 , ma non necessariamente per ogni $t \in I$.

Esempio 7.2.4. Il problema di Cauchy:

$$\begin{cases} y'(t) = \frac{3}{2}y^{1/3}(t) & \text{per } t > 0, \\ y(0) = 0, \end{cases}$$

ammette tanto la soluzione costante $y(t) = 0$ quanto la soluzione $y(t) = t^{3/2}$; addirittura, ammette infinite soluzioni, alcune delle quali sono date da:

$$y_\alpha(t) = \begin{cases} 0 & \text{se } t \leq \alpha, \\ \sqrt{(t - \alpha)^3} & \text{se } t > \alpha, \end{cases} \quad \text{per } \alpha \geq 0.$$

Un'ipotesi aggiuntiva su $f(t, y)$ che permette di garantire l'unicità della soluzione del problema di Cauchy è quella di essere una funzione lipschitziana rispetto alla seconda variabile.

Definizione 7.2.1. Sia $f : I \times \mathbb{R} \rightarrow \mathbb{R}$. Si dice che $f(t, y)$ è Lipschitz-continua (o lipschitziana) in I rispetto alla variabile y , uniformemente rispetto alla variabile t , se esiste una costante $L \geq 0$ tale che:

$$|f(t, y_1) - f(t, y_2)| \leq L |y_1 - y_2| \quad \forall y_1, y_2 \in \mathbb{R}, \quad \forall t \in I. \quad (7.16)$$

Osservazione 7.2.2. Se la funzione $f : I \times \mathbb{R} \rightarrow \mathbb{R}$ è continua di classe C^1 nel secondo argomento y , allora è anche Lipschitz-continua nel secondo argomento. Si ha infatti che:

$$|f(t, y_1) - f(t, y_2)| \leq \max_{y \in \mathbb{R}, t \in I} \left| \frac{\partial f}{\partial y}(y, t) \right| |y_1 - y_2|$$

$$\text{per cui si può considerare } L = \max_{y \in \mathbb{R}, t \in I} \left| \frac{\partial f}{\partial y}(y, t) \right|.$$

Esempio 7.2.5. Consideriamo la funzione $f(t, y) = y^2$, definita e continua da $I \subseteq \mathbb{R} \times \mathbb{R}$ in \mathbb{R} . Si ha che, per ogni $y_1, y_2 \in \mathbb{R}$ e per ogni $t \in \mathbb{R}$:

$$|f(t, y_1) - f(t, y_2)| = |y_1^2 - y_2^2| = |y_1 + y_2| |y_1 - y_2|.$$

In questo caso, la costante di Lipschitz varrebbe $L = 2M$ a patto che $y_1, y_2 \in (-M, M)$. Tale funzione non è dunque Lipschitz-continua (o lipschitziana) in I rispetto alla variabile y , uniformemente rispetto alla variabile t , nel senso della definizione, dal momento che $|y_1 + y_2| \rightarrow +\infty$ se $y_1, y_2 \rightarrow +\infty$ mantenendo lo stesso segno (ovvero per ogni $M > 0$).

Esempio 7.2.6. Consideriamo ora la funzione $f(t, y) = \sin(ty)$, definita e continua da $I \subseteq \mathbb{R} \times \mathbb{R}$ in \mathbb{R} . In tal caso, si ha che per ogni $y_1, y_2 \in \mathbb{R}$ e per ogni $t \in \mathbb{R}$:

$$|\sin(ty_1) - \sin(ty_2)| \leq |(ty_1) - (ty_2)| = |t| |y_1 - y_2|,$$

essendo la funzione $\theta \rightarrow \sin(\theta)$ lipschitziana su \mathbb{R} con costante di Lipschitz uguale a 1. In questo caso, la costante di Lipschitz vale $L = \max_{y \in \mathbb{R}, t \in I} \left| \frac{\partial f}{\partial y}(y, t) \right| = |t_f - t_0|$ e risulta dunque dipendente dall'ampiezza dell'intervallo I . Dunque $f(t, y)$ è Lipschitz-continua (o lipschitziana) in I rispetto alla variabile y , uniformemente rispetto alla variabile t , nel senso della definizione, a patto di considerare un intervallo I di ampiezza limitata.

Esempio 7.2.7. Consideriamo infine la funzione affine $f(t, y) = a(t)y + b(t)$, dove $a(t)$ e $b(t)$ sono funzioni definite e continue su un intervallo aperto $I \subseteq \mathbb{R}$. Dunque $f(t, y)$ è definita e continua da $I \times \mathbb{R}$ in \mathbb{R} . Per ogni $y_1, y_2 \in \mathbb{R}$ e per ogni $t \in I$, si ha:

$$|f(t, y_1) - f(t, y_2)| = |a(t)y_1 - a(t)y_2| = |a(t)(y_1 - y_2)| \leq |a(t)| |y_1 - y_2|,$$

Se $a(t)$ è limitata su I , allora f è lipschitziana rispetto a y per ogni $t \in I$, con costante di Lipschitz $L = \max_{t \in I} |a(t)|$.

L'assunzione di Lipschitz-continuità della funzione $f(y, t)$ che compare nel problema di Cauchy è fondamentale a garantire che esista e sia unica la sua soluzione $y(t)$ del problema di Cauchy (7.11), globalmente in $I = (t_0, t_f)$. Vale infatti il seguente risultato fondamentale:

Teorema 7.2.1 (Cauchy-Lipschitz, esistenza e unicità globale della soluzione). *Se $f(t, y) : I \times \mathbb{R} \rightarrow \mathbb{R}$ è:*

- 1. continua in entrambi gli argomenti,*
- 2. Lipschitz-continua in I rispetto alla variabile y , uniformemente rispetto alla variabile t ,*

allora esiste un'unica $y(t) : I \rightarrow \mathbb{R}$ soluzione del problema di Cauchy (7.11), ed è tale che $y \in C^1(I)$.

Osserviamo come il Teorema 7.2.1 permetta di stabilire se il problema di Cauchy sia ben posto (nel senso di ammettere un'unica soluzione $y(t)$ in I), ma non dica nulla su come determinarla in forma esplicita (analiticamente).

Esempio 7.2.8. Il problema (7.14) ammette un'unica soluzione, definita per ogni $t \in I$. Infatti, in questo caso $f(t, y) = 3y - 3t$ è continua rispetto a entrambi gli argomenti e si ha che:

$$|f(t, y_1) - f(t, y_2)| = |3y_1 - 3t - (3y_2 - 3t)| = |3y_1 - 3y_2| \leq 3|y_1 - y_2|$$

ovvero $|f(t, y_1) - f(t, y_2)| \leq L|y_1 - y_2|$ per ogni $y_1, y_2 \in \mathbb{R}$, e per ogni $t > 0$, con $L = 3$. La funzione f soddisfa dunque le ipotesi del teorema di Cauchy-Lipschitz 7.2.1 e possiamo dunque dedurre che il problema (7.14) ammette un'unica soluzione $y(t)$, definita per ogni $t \in I$.

Esempio 7.2.9. Nel caso del problema dell'Esempio 7.2.4, si ha che $f(t, y) = \frac{3}{2}y^{1/3}$. Tale funzione non è lipschitziana in qualsiasi intervallo della forma $(0, T)$ (sebbene essa risulti continua), dal momento che essa non è derivabile per $t = 0$, ovvero:

$$\frac{\partial f}{\partial t}(t, y) = \frac{1}{2}y^{-2/3} = \frac{1}{2\sqrt[3]{y^2}} \quad \text{e} \quad L = \max_{y \in \mathbb{R}, t \in (0, T)} \left| \frac{\partial f}{\partial t}(t, y) \right| \rightarrow \infty.$$

7.2.3 Stabilità secondo Liapunov del problema di Cauchy

Sotto le stesse ipotesi del Teorema di Cauchy-Lipschitz, la soluzione $y(t)$ del problema di Cauchy (7.14) dipende con continuità dai dati. La sensibilità della soluzione di tale problema a perturbazioni sui dati in $I = (t_0, t_f)$, per $t_f < +\infty$ si traduce nella definizione di *stabilità* (del problema di Cauchy) secondo *Lyapunov*. Siano $(\delta_0, \delta(t))$ due perturbazioni, dove $\delta_0 \in \mathbb{R}$ e $\delta : I \rightarrow \mathbb{R}$; indichiamo con $z = z(t)$ la soluzione del seguente problema di Cauchy *perturbato*: trovare $z : I \subset \mathbb{R} \rightarrow \mathbb{R}$ tale che

$$\begin{cases} z'(t) = f(t, z(t)) + \delta(t) & \text{per } t \in I, \\ z(t_0) = y_0 + \delta_0. \end{cases} \quad (7.17)$$

Definizione 7.2.2. *Il problema di Cauchy (7.11) si dice stabile (secondo Lyapunov) sull'intervallo I se per ogni perturbazione $(\delta_0, \delta(t))$ tale che*

$$|\delta_0| < \varepsilon \quad \text{e} \quad |\delta(t)| < \varepsilon \quad \forall t \in I, \quad (7.18)$$

con $\varepsilon > 0$, e tale da garantire la soluzione $z(t)$ del problema perturbato (7.17), allora esiste una costante $C > 0$ tale che

$$|y(t) - z(t)| \leq C\varepsilon \quad \forall t \in I.$$

Il problema di Cauchy (7.11) si dice asintoticamente stabile se per $\lim_{t \rightarrow +\infty} |\delta(t)| = 0$ si ottiene che $\lim_{t \rightarrow +\infty} |y(t) - z(t)| = 0$.

La stabilità (secondo Lyapunov) del problema di Cauchy è conseguenza diretta delle ipotesi del Teorema di Cauchy-Lipschitz. Vale il seguente risultato.

Proposizione 7.2.1. *Supponiamo che $f(t, y)$ verifichi le ipotesi del Teorema 7.2.1 di Cauchy-Lipschitz. Allora la stabilità (secondo Lyapunov) del problema di Cauchy (7.11) è condizione necessaria e sufficiente per l'esistenza di una sua unica soluzione. Inoltre, per ogni $t \in I$, con $t > t_0$, si ha:*

$$|y(t) - z(t)| \leq |\delta_0| + L \int_{t_0}^t |y(\tau) - z(\tau)| d\tau + \int_{t_0}^t |\delta(\tau)| d\tau,$$

dove $y : I \rightarrow \mathbb{R}$ e $z : I \rightarrow \mathbb{R}$ sono rispettivamente soluzioni del problema di Cauchy (7.11) e del problema di Cauchy perturbato (7.17). In particolare, per ogni perturbazione $(\delta_0, \delta(t))$ tale da soddisfare le condizioni (7.18) con $\varepsilon > 0$, vale la seguente stima di stabilità:

$$|y(t) - z(t)| \leq (1 + |t - t_0|) e^{L|t-t_0|} \varepsilon \quad \forall t \in I = (t_0, t_f), \quad (7.19)$$

dove appunto $C = (1 + |t - t_0|) e^{L|t-t_0|}$.

Quest'ultima proprietà esprime la dipendenza continua della soluzione del problema di Cauchy dai dati: una perturbazione di ampiezza ε nei dati si traduce in una perturbazione di ampiezza al più $(1 + |t - t_0| e^{L|t-t_0|}) \varepsilon$ sulla soluzione al tempo $t > t_0$. In altri termini, la distanza tra due traiettorie può crescere al più di un fattore

$$C = (1 + |t - t_0|) e^{L|t-t_0|}$$

nel passaggio da t_0 a t . Si noti tuttavia il carattere esponenziale di tale fattore, la cui grandezza dipende non solo dalla distanza $|t - t_0|$ – e dunque dall'ampiezza di I – ma anche dalla grandezza della costante di Lipschitz L della funzione f . Tanto maggiore è la costante di Lipschitz L di f , tanto maggiore è l'effetto delle perturbazioni sulla soluzione del problema di Cauchy.

Osservazione 7.2.3. Sottolineiamo come la (7.19) fornisca una limitazione di ugnato le soluzioni corrispondenti a dati iniziali (e a funzioni f) diverse possano allontanarsi nell'intervallo I . La presenza del fattore esponenziale indica che la convergenza è uniforme in intervalli limitati, mentre non si può escludere un allontanamento esponenziale all'infinito anche per soluzioni corrispondenti a dati iniziali molto vicini. Ad esempio, nel caso in cui $f(t, y) = y$, $y_0 = 1$ e $\delta_0 = \varepsilon$, $\delta(t) = 0 \forall t > 0$, le soluzioni sono date da $y(t) = e^t$ e $z(t) = (1 + \varepsilon)e^t$, ma $y(t) - z(t) = \varepsilon e^t \rightarrow \infty$ per $t \rightarrow \infty$.

I risultati precedenti ci assicurano dunque che il problema di Cauchy (7.11) è ben posto in I sotto le ipotesi su $f(t, y)$ che abbiamo fatto. Risultati analoghi si possono ottenere per il sistema di EDO (7.12).

Per dimostrare il risultato della Proposizione 7.2.1, riportiamo il seguente lemma.

Lemma 7.2.1 (di Gronwall). *Siano date due funzioni $\varphi, g \in C^0([t_0, t_f])$ e una funzione $p(t) \geq 0$ per ogni $t \in (t_0, t_f)$. Se*

$$\varphi(t) \leq g(t) + \int_{t_0}^t p(\tau) \varphi(\tau) d\tau,$$

allora

$$\varphi(t) \leq g(t) e^{\int_{t_0}^t p(\tau) d\tau} \quad \text{per ogni } t \in (t_0, t_f).$$

Dimostrazione. (Proposizione 7.2.1). Mostriamo solo una delle due implicazioni, quella per cui l'esistenza di un'unica soluzione del problema di Cauchy implica la stabilità secondo Lyapunov di tale problema. Supponiamo dunque che $f(t, y)$ verifichi le ipotesi del Teorema 7.2.1 di Cauchy-Lipschitz.

Poniamo $w(t) = z(t) - y(t)$; risulta allora $w'(t) = z'(t) - y'(t) = f(t, z(t)) + \delta(t) - f(t, y(t))$, con $w(t_0) = z(t_0) - y(t_0) = y_0 + \delta_0 - y_0 = \delta_0$. Integrando tra t_0 e t otteniamo

$$w(t) - w(t_0) = \int_{t_0}^t w'(\tau) d\tau = \int_{t_0}^t (f(\tau, z(\tau)) - f(\tau, y(\tau))) d\tau + \int_{t_0}^t \delta(\tau) d\tau$$

ovvero, portando a secondo membro $w(t_0)$, prendendo i valori assoluti e sfruttando la diseguaglianza triangolare, si ha che

$$|w(t)| \leq |\delta_0| + \left| \int_{t_0}^t (f(\tau, z(\tau)) - f(\tau, y(\tau))) d\tau \right| + \left| \int_{t_0}^t \delta(\tau) d\tau \right|.$$

Risulta dunque, sfruttando la proprietà dell'integrale definito secondo cui

$$\left| \int_a^b f(t) dt \right| \leq \int_a^b |f(t)| dt,$$

che

$$|w(t)| \leq |\delta_0| + \int_{t_0}^t |f(\tau, z(\tau)) - f(\tau, y(\tau))| d\tau + \int_{t_0}^t |\delta(\tau)| d\tau.$$

Poiché f è Lipschitziana rispetto al secondo argomento, si ha che

$$|w(t)| \leq |\delta_0| + L \int_{t_0}^t |z(\tau) - y(\tau)| d\tau + \int_{t_0}^t |\delta(\tau)| d\tau$$

da cui, essendo la perturbazione minore o uguale a ε ,

$$|w(t)| \leq (1 + |t - t_0|) \varepsilon + L \int_{t_0}^t |w(\tau)| d\tau.$$

Dall'ultima diseguaglianza, si ha che l'entità della perturbazione sulla soluzione è dunque maggiorata dalla perturbazione sui dati (primo termine) e dalla storia pregressa della perturbazione, moltiplicata per la costante di Lipschitz L (secondo termine); entrambi i termini dipendono inoltre dal tempo trascorso $|t - t_0|$. Per trovare una stima per $|w(t)|$ che non dipenda da $w(t)$ stessa, sfruttiamo il Lemma 7.2.1 di Gronwall, ponendo

$$\varphi(t) = |w(t)|, \quad g(t) = (1 + |t - t_0|) \varepsilon, \quad p(t) = L \geq 0$$

da cui otteniamo direttamente che

$$|w(t)| \leq (1 + |t - t_0|) \varepsilon e^{L \int_{t_0}^t d\tau}$$

ovvero, per ogni $t \in (t_0, t_f)$,

$$|w(t)| \leq (1 + |t - t_0|) \varepsilon e^{L|t-t_0|}.$$

□

7.3 Derivazione Numerica

I metodi numerici per l'approssimazione della soluzione di un problema di Cauchy si basano sull'idea di suddividere l'intervallo temporale $I = (t_0, t_f)$ in sottointervalli introducendo un insieme (finito) di istanti di tempo in cui *collocare* l'equazione differenziale, approssimando le *derivate* che vi compaiono con opportuni rapporti incrementali; infatti, l'operazione di derivazione di una funzione (coinvolgendo un passaggio al limite) non può essere affrontata da un calcolatore. In questa sezione consideriamo brevemente i più comuni metodi per l'approssimazione delle derivate di una funzione, un argomento tipicamente indicato come *derivazione numerica*.

Data una funzione $f \in C^1((a, b))$, l'obiettivo consiste nell'approssimare per esempio $f'(\bar{x})$ per qualche $\bar{x} \in (a, b) \subseteq \mathbb{R}$. Indipendentemente dalla necessità di approssimare derivate per costruire schemi numerici per le EDO, data una funzione $f(x)$ potrebbe essere preferibile evitare di calcolare esplicitamente $f'(\bar{x})$ per qualche $\bar{x} \in (a, b)$, un'operazione computazionalmente costosa in molti casi. Le stesse considerazioni valgono per l'approssimazione di $f''(\bar{x})$.

Osservazione 7.3.1. *La derivazione numerica è invece indispensabile se solamente l'insieme delle coppie di dati $\{(x_i, y_i)\}_{i=0}^n$ è disponibile, mentre non è fornita la funzione $f(x)$ da cui tali dati sono stati generati. In tal caso, potrebbe comunque essere di interesse fornire informazioni sulla derivata prima (o seconda) della funzione incognita $f(x)$ in uno o più dei nodi $\{x_i\}_{i=0}^n$.*

7.3.1 Schemi alle differenze finite per la derivata prima

Consideriamo alcuni schemi alle *differenze finite* per l'approssimazione di $f'(\bar{x})$ per qualche $\bar{x} \in (a, b)$, a partire dai più semplici schemi basati su differenze finite in avanti e all'indietro.

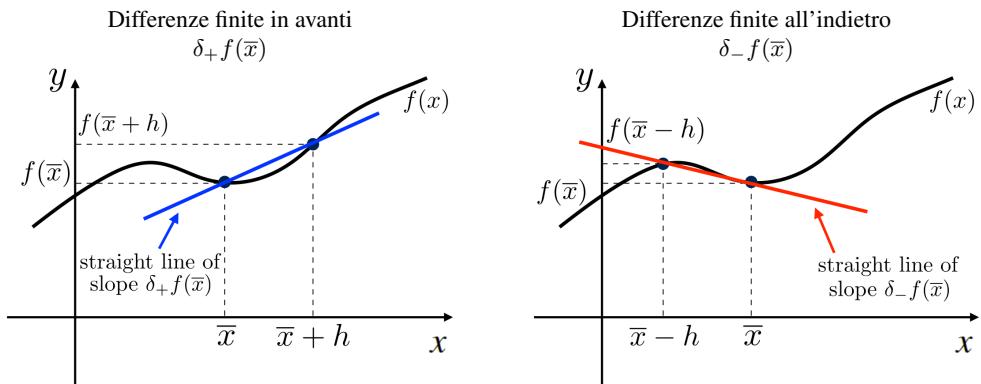
Definizione 7.3.1. Data una funzione $f(x)$ e il passo di ampiezza $h > 0$, l'approssimazione di $f'(\bar{x})$ in qualche $\bar{x} \in (a, b) \subseteq \mathbb{R}$ tramite lo schema delle differenze finite in avanti è definita come:

$$\delta_+ f(\bar{x}) := \frac{f(\bar{x} + h) - f(\bar{x})}{h},$$

mentre tramite lo schema delle differenze finite all'indietro come:

$$\delta_- f(\bar{x}) := \frac{f(\bar{x}) - f(\bar{x} - h)}{h}.$$

Esempio 7.3.1. Illustriamo graficamente gli schemi alle differenze finite in avanti (a sinistra) e all'indietro (a destra) per l'approssimazione di $f'(\bar{x})$ per qualche $h > 0$; a tale scopo tracciamo le rette passanti per i punti $(\bar{x}, f(\bar{x}))$ e $(\bar{x}, f(\bar{x} \pm h))$ e aventi pendenza $\delta_+ f(\bar{x})$ e $\delta_- f(\bar{x})$, rispettivamente.



Proposizione 7.3.1. Se $f \in C^2((a, b))$ e $\bar{x} \in (a, b)$, l'errore $E_+ f(\bar{x})$ associato allo schema delle differenze finite in avanti è:

$$E_+ f(\bar{x}) := f'(\bar{x}) - \delta_+ f(\bar{x}) = -\frac{1}{2} h f''(\xi_+) \quad \text{per qualche } \xi_+ \in [\bar{x}, \bar{x} + h],$$

mentre l'errore $E_- f(\bar{x})$ associato allo schema delle differenze finite all'indietro è:

$$E_- f(\bar{x}) := f'(\bar{x}) - \delta_- f(\bar{x}) = \frac{1}{2} h f''(\xi_-) \quad \text{per qualche } \xi_- \in [\bar{x} - h, \bar{x}].$$

Dimostrazione. (Schema alle differenze finite in avanti). Consideriamo l'espansione in serie di Taylor di $f(\bar{x} + h)$ attorno a \bar{x} , ovvero $f(\bar{x} + h) = f(\bar{x}) + f'(\bar{x})h + \frac{1}{2} f''(\xi_+)h^2$ per qualche $\xi_+ \in [\bar{x}, \bar{x} + h]$. Si ottiene il risultato usando la definizione dell'errore $E_+ f(\bar{x})$. In maniera del tutto analoga si ottiene la dimostrazione dello schema delle differenze finite all'indietro. \square

Gli schemi delle differenze finite in avanti e all'indietro sono metodi di *ordine* di accuratezza 1; infatti, gli errori $E_+ f(\bar{x})$ e $E_- f(\bar{x})$ convergono a zero con *ordine* 1 rispetto al passo h .

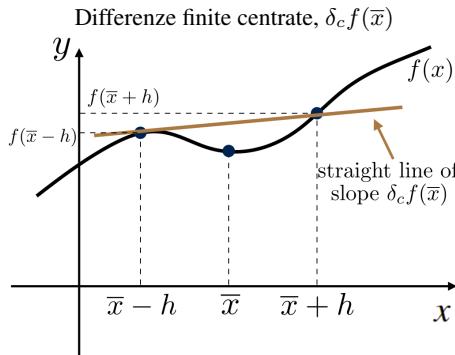
Osservazione 7.3.2. Per $f \in \mathbb{P}_1$, si ottiene $\delta_+ f(\bar{x}) \equiv \delta_- f(\bar{x}) \equiv f'(\bar{x})$ per ogni $\bar{x} \in \mathbb{R}$.

Un ulteriore schema per approssimare la derivata prima di una funzione è quello basato su differenze finite centrate.

Definizione 7.3.2. Data una funzione $f(x)$ e il passo di ampiezza $h > 0$, l'approssimazione di $f'(\bar{x})$ in qualche $\bar{x} \in (a, b) \subseteq \mathbb{R}$ per mezzo dello schema delle differenze finite centrate è definita come:

$$\delta_c f(\bar{x}) := \frac{f(\bar{x} + h) - f(\bar{x} - h)}{2h}.$$

Esempio 7.3.2. Illustriamo graficamente lo schema delle differenze finite centrate per l'approssimazione di $f'(\bar{x})$ per qualche $h > 0$; nel caso specifico, rappresentiamo graficamente la retta di pendenza $\delta_c f(\bar{x})$.



Proposizione 7.3.2. Se $f \in C^3((a, b))$ e $\bar{x} \in (a, b)$, l'errore $E_c f(\bar{x})$ associato allo schema delle differenze finite centrate è:

$$E_c f(\bar{x}) := f'(\bar{x}) - \delta_c f(\bar{x}) = -\frac{1}{12} h^2 [f'''(\xi_+) + f'''(\xi_-)]$$

per qualche $\xi_+ \in [\bar{x}, \bar{x} + h]$ e $\xi_- \in [\bar{x} - h, \bar{x}]$.

Dimostrazione. Consideriamo l'espansione in serie di Taylor di $f(\bar{x} + h)$ in \bar{x} per cui si ottiene $f(\bar{x} + h) = f(\bar{x}) + f'(\bar{x})h + \frac{1}{2}f''(\bar{x})h^2 + \frac{1}{6}f'''(\xi_+)h^3$ per qualche $\xi_+ \in [\bar{x}, \bar{x} + h]$; in maniera del tutto simile, l'espansione di Taylor di $f(\bar{x} - h)$ in \bar{x} è $f(\bar{x} - h) = f(\bar{x}) - f'(\bar{x})h + \frac{1}{2}f''(\bar{x})h^2 - \frac{1}{6}f'''(\xi_-)h^3$ per qualche $\xi_- \in [\bar{x} - h, \bar{x}]$. Applicando la definizione dell'errore $E_c f(\bar{x})$, si ottiene il risultato. \square

Lo schema delle differenze finite centrate è un metodo con ordine di accuratezza 2; infatti, l'errore $E_c f(\bar{x})$ converge a zero con ordine 2 rispetto al passo h .

Osservazione 7.3.3. Osserviamo che, per $f \in \mathbb{P}_2$, si ottiene $\delta_c f(\bar{x}) \equiv f'(\bar{x})$ per ogni $\bar{x} \in \mathbb{R}$.

Osservazione 7.3.4. È possibile mostrare che $\delta_c f(\bar{x}) = (\Pi_{2,\{\bar{x}-h, \bar{x}, \bar{x}+h\}} f)'(\bar{x})$, dove $\Pi_{2,\{\bar{x}-h, \bar{x}, \bar{x}+h\}} f(x)$ è il polinomio di grado 2 interpolante la funzione $f(x)$ ai nodi $\bar{x} - h$, \bar{x} e $\bar{x} + h$.

Supponiamo ora di essere interessati ad approssimare la derivata prima di una funzione $f(x)$ in molti nodi che siano equispaziati nell'intervallo $[a, b]$, ovvero $x_i = a + ih$ per ogni $i = 0, \dots, n$, con $h = \frac{b-a}{n}$; $x_0 = a$ e $x_n = b$ sono i nodi agli estremi dell'intervallo $[a, b]$. A questo scopo, consideriamo lo schema alle differenze finite centrate. Tuttavia, tale schema può solamente essere utilizzato ai nodi interni all'intervallo $[a, b]$ come:

$$\delta_c f(x_i) = \frac{f(x_{i+1}) - f(x_{i-1})}{2h} \quad \text{per ogni } i = 1, \dots, n-1;$$

infatti per l'approssimazione di $f'(x)$ nei nodi x_0 e x_n , sono necessarie valutazioni di $f(x)$ ai nodi x_{-1} e x_{n+1} , le quali non sono generalmente definite.

A tale scopo, si possono però utilizzare gli schemi alle differenze finite in avanti e all'indietro per approssimare rispettivamente $f'(x_0)$ e $f'(x_n)$; in maniera corrispondente si ottengono pertanto le approssimazioni $\delta_+ f(x_0) = \frac{f(x_1) - f(x_0)}{h}$ e $\delta_- f(x_n) = \frac{f(x_n) - f(x_{n-1})}{h}$. Tuttavia, in tal caso, l'approssimazione di f' coinvolge uno schema con ordine di accuratezza 2 per i nodi interni $\{x_i\}_{i=1}^{n-1}$, mentre schemi di ordine 1 ai nodi estremi x_0 e x_n . Dato che è auspicabile considerare metodi dello stesso ordine di accuratezza per tutti i nodi (interni ed esterni) di $[a, b]$, ordine 2 in questo caso, è possibile considerare rispettivamente le seguenti approssimazioni ai nodi estremi x_0 e x_n :

$$\delta_{c,0} f(x_0) = \frac{-3f(x_0) + 4f(x_1) - f(x_2)}{2h} \quad (7.20)$$

e

$$\delta_{c,n} f(x_n) = \frac{f(x_{n-2}) - 4f(x_{n-1}) + 3f(x_n)}{2h}; \quad (7.21)$$

tali schemi alle differenze finite, pur essendo decentrati, garantiscono un ordine di accuratezza pari a 2 in h .

Osservazione 7.3.5. È semplice mostrare che $\delta_{c,0} f(x_0) = (\Pi_{2,\{x_0,x_1,x_2\}} f)'(x_0)$ e, in maniera simile, $\delta_{c,n} f(x_n) = (\Pi_{2,\{x_{n-2},x_{n-1},x_n\}} f)'(x_n)$, dove $\Pi_{2,\{x_0,x_1,x_2\}} f(x)$ e $\Pi_{2,\{x_{n-2},x_{n-1},x_n\}} f(x)$ sono i polinomi di grado 2 interpolanti la funzione $f(x)$ rispettivamente ai nodi $\{x_i\}_{i=0}^2$ e $\{x_i\}_{i=n-2}^n$.

Lo stesso risultato si può ottenere procedendo come segue. Si consideri $\delta_{c,0} f(x_0)$ da determinarsi sulla base dei dati $f(x_0)$, $f(x_1)$ e $f(x_2)$. Nel caso generale, l'espressione di $\delta_{c,0} f(x_0)$ è:

$$\delta_{c,0} f(x_0) = a f(x_0) + b f(x_1) + c f(x_2),$$

ovvero espressa in termini delle incognite a , b e $c \in \mathbb{R}$. Per determinare tali coefficienti, si scriva l'espansione di Taylor di $f(x)$ in x_0 , ovvero:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2 + \frac{1}{6} f'''(x_0)(x - x_0)^3 + \dots,$$

da cui si hanno $f(x_0) \equiv f(x_0)$, $f(x_1) = f(x_0) + h f'(x_0) + \frac{h^2}{2} f''(x_0) + \frac{h^3}{6} f'''(x_0) + \dots$ e $f(x_2) = f(x_0) + 2h f'(x_0) + 2h^2 f''(x_0) + \frac{4}{3} h^3 f'''(x_0) + \dots$, essendo $x_1 - x_0 = h$ e $x_2 - x_0 = 2h$ per nodi equispaziati. Sostituendo gli sviluppi precedenti nell'espressione di $\delta_{c,0} f(x_0)$, si ottiene:

$$\delta_{c,0} f(x_0) = (a + b + c) f(x_0) + (b + 2c) h f'(x_0) + \left(\frac{b}{2} + 2c\right) h^2 f''(x_0) + \left(\frac{b}{6} + \frac{4}{3} c\right) h^3 f'''(x_0) + \dots$$

Dato che $\delta_{c,0} f(x_0)$ deve approssimare $f'(x_0)$, è necessario richiedere che $(a + b + c) = 0$ e $(b + 2c)h = 1$; tale scelta garantisce un'accuratezza di ordine 1 e infinite scelte possibili per i coefficienti a , b e c . Se inoltre si richiede che $\left(\frac{b}{2} + 2c\right) h^2 = 0$, allora si ottengono i coefficienti:

$$a = -\frac{3}{2h}, \quad b = \frac{2}{h} \quad \text{e} \quad c = -\frac{1}{2h},$$

da cui il risultato (7.20) che è accurato all'ordine 2 in h ; infatti, si ha $f'(x_0) - \delta_{c,0} f(x_0) = \frac{h^2}{3} f'''(x_0) + O(h^3)$ se $f \in C^3([x_0, x_2])$.

7.3.2 Schema alle differenze finite per la derivata seconda

Consideriamo lo schema delle *differenze finite centrate* per l'approssimazione della derivata seconda $f''(\bar{x})$ per $\bar{x} \in (a, b)$.

Definizione 7.3.3. Data una funzione $f(x)$ e il passo di dimensione $h > 0$, l'approssimazione di $f''(\bar{x})$ in $\bar{x} \in (a, b) \subseteq \mathbb{R}$ tramite le differenze finite centrate è definita come:

$$\delta_c^2 f(\bar{x}) := \frac{f(\bar{x} + h) - 2f(\bar{x}) + f(\bar{x} - h)}{h^2}.$$

Proposizione 7.3.3. Se $f \in C^4((a, b))$ e $\bar{x} \in (a, b)$, l'errore $E_c^2 f(\bar{x})$ associato allo schema alle differenze finite centrate è:

$$E_c^2 f(\bar{x}) := f''(\bar{x}) - \delta_c^2 f(\bar{x}) = -\frac{1}{24} h^2 [f^{(4)}(\xi_+) + f^{(4)}(\xi_-)]$$

per qualche $\xi_+ \in [\bar{x}, \bar{x} + h]$ e $\xi_- \in [\bar{x} - h, \bar{x}]$.

Dimostrazione. Si consideri l'espansione di Taylor per $f(\bar{x} + h)$ in \bar{x} , per cui $f(\bar{x} + h) = f(\bar{x}) + f'(\bar{x})h + \frac{1}{2}f''(\bar{x})h^2 + \frac{1}{6}f'''(\bar{x})h^3 + \frac{1}{24}f^{(4)}(\xi_+)h^4$ per qualche $\xi_+ \in [\bar{x}, \bar{x} + h]$; in maniera simile, l'espansione di Taylor di $f(\bar{x} - h)$ in \bar{x} è $f(\bar{x} - h) = f(\bar{x}) - f'(\bar{x})h + \frac{1}{2}f''(\bar{x})h^2 - \frac{1}{6}f'''(\bar{x})h^3 + \frac{1}{24}f^{(4)}(\xi_-)h^4$ per qualche $\xi_- \in [\bar{x} - h, \bar{x}]$. A questo punto, applicando la definizione dell'errore $E_c^2 f(\bar{x})$ e sostituendo in $\delta_c^2 f(\bar{x})$ le espansioni di Taylor si ottiene il risultato. \square

Lo schema alle differenze finite centrate per l'approssimazione di $f''(\bar{x})$ è un metodo di *ordine* 2.

Osservazione 7.3.6. Osserviamo che, se $f \in \mathbb{P}_3$, si ha $\delta_c^2 f(\bar{x}) \equiv f''(\bar{x})$ per ogni $\bar{x} \in \mathbb{R}$.

7.4 Approssimazione Numerica di EDO del Primo Ordine

Consideriamo ora l'approssimazione e la soluzione numerica di un problema di Cauchy per un'equazione differenziale ordinaria nella forma

$$\begin{cases} y'(t) = f(t, y(t)) & \text{per } t \in I, \\ y(t_0) = y_0, \end{cases} \quad (7.22)$$

dove $I = (t_0, t_f)$ e $f(t, y)$ è funzione di due argomenti tale che $f : I \times \mathbb{R} \rightarrow \mathbb{R}$. Successivamente, estenderemo i metodi numerici qui sviluppati al caso dei sistemi di EDO.

Consideriamo la partizione di $\bar{I} = [t_0, t_f]$ in N_h sottointervalli di uguale ampiezza $h = \frac{t_f - t_0}{N_h}$ tali che $t_n = t_0 + nh$ per $n = 0, 1, \dots, N_h$ ($t_{N_h} \equiv t_f$). L'idea di base è quella di imporre che l'equazione differenziale sia valida per ogni t_n , $n = 0, 1, \dots, N_h$, sostituire alle derivate che vi compaiono un'approssimazione basata su schemi alle differenze finite e, infine, riscrivere il problema risultante per un'opportuna approssimazione della soluzione (ottenendo così la soluzione numerica).

Osservazione 7.4.1. Indichiamo con $y_n = y(t_n)$ la valutazione della soluzione (esatta) $y(t)$ a $t = t_n$. Al contrario, indichiamo con u_n l'approssimazione di $y(t_n)$, ovvero la soluzione numerica a $t = t_n$.

7.4.1 Metodo di Eulero in avanti

Il metodo di *Eulero in avanti* applicato al problema di Cauchy (7.22) approssima $\frac{dy}{dt}(t_n)$ con il metodo delle differenze finite in avanti a $t = t_n$ per cui:

$$f(t_n, y_n) = \frac{dy}{dt}(t_n) \approx \frac{y(t_n + h) - y(t_n)}{h} = \frac{y_{n+1} - y_n}{h};$$

imponendo l'uguaglianza del primo e ultimo membro della precedente, otteniamo:

$$\frac{u_{n+1} - u_n}{h} = f(t_n, u_n) \quad \text{per } n = 0, 1, \dots, N_h - 1.$$

Il metodo di Eulero in avanti si scrive dunque come: trovare $\{u_n\}_{n=0}^{N_h}$ tale che

$$\begin{cases} u_{n+1} = u_n + h f(t_n, u_n) & \text{per } n = 0, 1, \dots, N_h - 1, \\ u_0 = y_0. \end{cases} \quad (7.23)$$

Il metodo di Eulero in avanti è un metodo di tipo *esplicito*, dal momento che risulta possibile determinare u_{n+1} esplicitamente in funzione di quantità note dal passo di tempo precedente. Spesso tale metodo è pertanto indicato come metodo di Eulero Esplicito.

Esempio 7.4.1. Per il *problema modello* dell'Esempio 7.2.1 ($f(t, y) = \lambda y$), il metodo di Eulero in avanti si scrive come:

$$\begin{cases} u_{n+1} = (1 + h \lambda) u_n & \text{per } n = 0, 1, \dots, N_h - 1, \\ u_0 = y_0. \end{cases}$$

Esempio 7.4.2. Per il problema di Cauchy (7.22) con $f(t, y) = e^{-t} (1 - y^\alpha)$, per qualche $\alpha > 0$, l'approssimazione con il metodo di Eulero in avanti si scrive come:

$$\begin{cases} u_{n+1} = u_n + h e^{-t_n} (1 - u_n^\alpha) & \text{per } n = 0, 1, \dots, N_h - 1, \\ u_0 = y_0. \end{cases}$$

7.4.2 Metodo di Eulero all'indietro

Il metodo di *Eulero all'indietro* per il problema di Cauchy (7.22) approssima $\frac{dy}{dt}(t_{n+1})$ con le differenze finite all'indietro a $t = t_{n+1}$, per cui:

$$f(t_{n+1}, y_{n+1}) = \frac{dy}{dt}(t_{n+1}) \approx \frac{y(t_{n+1}) - y(t_{n+1} - h)}{h} = \frac{y_{n+1} - y_n}{h};$$

da cui abbiamo:

$$\frac{u_{n+1} - u_n}{h} = f(t_{n+1}, u_{n+1}) \quad \text{per } n = 0, 1, \dots, N_h - 1.$$

Il metodo di Eulero all'indietro si scrive come: trovare $\{u_n\}_{n=0}^{N_h}$ tale che

$$\begin{cases} u_{n+1} = u_n + h f(t_{n+1}, u_{n+1}) & \text{per } n = 0, 1, \dots, N_h - 1, \\ u_0 = y_0. \end{cases} \quad (7.24)$$

Il metodo di Eulero all'indietro è un metodo *implicito*, dal momento che u_{n+1} è definita in funzione di u_n stessa; infatti, tale metodo è spesso chiamato anche metodo di Eulero Implicito.

Esempio 7.4.3. Per il *problema modello* dell'Esempio 7.2.1, il metodo di Eulero all'indietro si scrive come:

$$\begin{cases} u_{n+1} = \frac{u_n}{1 - h \lambda} & \text{per } n = 0, 1, \dots, N_h - 1, \\ u_0 = y_0. \end{cases}$$

Dato che il metodo di Eulero all'indietro è un metodo implicito, è necessario, in linea di principio², risolvere un'equazione *non lineare* per ogni $n = 0, 1, \dots, N_h - 1$, ovvero.:

$$\text{trovare } u_{n+1} : F_n^{EI}(u_{n+1}) = 0 \quad \text{per ogni } n = 0, 1, \dots, N_h - 1, \quad (7.25)$$

con $u_0 = y_0$, dove

$$F_n^{EI}(y) := y - u_n - h f(t_{n+1}, y).$$

Dato che l'Eq. (7.25) raramente può essere risolta analiticamente, una soluzione approssimata può essere ottenuta per mezzo del metodo di Newton, il metodo delle iterazioni di punto fisso oppure un altro metodo numerico per equazioni non lineari. Per esempio, considerando il metodo di Newton (4.3) per risolvere l'Eq. (7.25), è necessario disporre della derivata prima di $F_n^{EI}(y)$ nel suo argomento y , ovvero

$$(F_n^{EI})'(y) = 1 - h \frac{\partial f}{\partial y}(t_{n+1}, y);$$

a questo punto, è possibile utilizzare l'Algoritmo 7.1 per risolvere il problema di Cauchy con il metodo di Eulero all'indietro in combinazione con il metodo di Newton.

Algoritmo 7.1: Metodo di Eulero all'indietro con metodo di Newton

```

porre  $u_0 = y_0$ ;
for  $n = 0, 1, \dots, N_h - 1$  do
    porre  $u_{n+1}^{(0)} = u_n$ ;
    porre  $F_n^{EI}(y) = y - u_n - h f(t_{n+1}, y)$  e  $(F_n^{EI})'(y) = 1 - h \frac{\partial f}{\partial y}(t_{n+1}, y)$ ;
    for  $k = 0, 1, \dots$  fino a che un criterio d'arresto è soddisfatto do
         $u_{n+1}^{(k+1)} = u_{n+1}^{(k)} - \frac{F_n^{EI}\left(u_{n+1}^{(k)}\right)}{(F_n^{EI})'\left(u_{n+1}^{(k)}\right)}$ ;
    end
    porre  $u_{n+1} = u_{n+1}^{(k+1)}$ ;
end
```

Esempio 7.4.4. Per il problema di Cauchy (7.22) con $f(t, y) = e^{-t} (1 - y^\alpha)$, per qualche $\alpha > 0$, abbiamo da Eq. (7.25): $F_n^{EI}(y) = y - u_n - h e^{-t_{n+1}} (1 - y^\alpha)$ e $(F_n^{EI})'(y) = 1 + \alpha h e^{-t_{n+1}} y^{\alpha-1}$, qualora fossimo interessati a usare il metodo di Newton.

Alternativamente al metodo di Newton, è possibile utilizzare il metodo delle *iterazioni di punto fisso* (Eq. 4.7) con funzione di iterazione

$$\phi_n^{EI}(y) = u_n + h f(t_{n+1}, y),$$

per esempio. Si veda l'Algoritmo 7.2.

Le proprietà di convergenza dei metodi di Newton e iterazioni di punto fisso dipendono da $F_n^{EI}(y)$ e $\phi_n^{EI}(y)$ e, in ultima analisi dalla funzione f e da h . In generale, riducendo h è possibile garantire e/o controllare le proprietà di convergenza dei metodi per risolvere l'equazione non lineare ad ogni t_n per $n = 0, 1, \dots, N_h - 1$.

²Ciò è vero se $f = f(t, y)$ è non lineare nel suo secondo argomento; se invece f è lineare (affine) in y , l'equazione che compare in (7.24) risulta lineare nell'incognita u_{n+1} .

Algorithm 7.2: Metodo di Eulero all'indietro con metodo delle iterazioni di punto fisso

```

porre  $u_0 = y_0$ ;
for  $n = 0, 1, \dots, N_h - 1$  do
    porre  $u_{n+1}^{(0)} = u_n$ ;
    porre  $\phi_n^{EI}(y) = u_n + h f(t_{n+1}, y)$ ;
    for  $k = 0, 1, \dots$  fino a che un criterio d'arresto è soddisfatto do
         $u_{n+1}^{(k+1)} = \phi_n^{EI}(u_{n+1}^{(k)})$ ;
    end
    porre  $u_{n+1} = u_{n+1}^{(k+1)}$ ;
end

```

7.4.3 Metodo di Crank-Nicolson

Il metodo di Crank-Nicolson si può ricavare a partire dalla forma integrale del problema di Cauchy, riscritta sul generico intervallo $[t_n, t_{n+1}]$, ovvero:

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

Tale metodo deriva dall'approssimazione dell'integrale che compare nel membro di destra di quest'ultima equazione mediante la formula del trapezio (semplice), in base alla quale (essendo $t_{n+1} - t_n = h$)

$$y_{n+1} - y_n = \int_{t_n}^{t_{n+1}} f(\tau, y(\tau)) d\tau \approx \frac{h}{2} [f(t_n, y_n) + f(t_{n+1}, y_{n+1})]$$

In tal modo, il metodo di Crank-Nicolson per l'approssimazione del problema di Cauchy (7.22) considera la seguente approssimazione:

$$\frac{u_{n+1} - u_n}{h} = \frac{1}{2} [f(t_n, u_n) + f(t_{n+1}, u_{n+1})] \quad \text{per } n = 0, 1, \dots, N_h - 1;$$

il metodo si scrive dunque come: trovare $\{u_n\}_{n=0}^{N_h}$ tale che

$$\begin{cases} u_{n+1} = u_n + \frac{h}{2} [f(t_n, u_n) + f(t_{n+1}, u_{n+1})] & \text{per } n = 0, 1, \dots, N_h - 1, \\ u_0 = y_0. \end{cases} \quad (7.26)$$

Anche il metodo di Crank-Nicolson è un metodo di tipo **Implicito**. Per questa ragione, è necessario, in linea di principio, risolvere la seguente equazione *non lineare* per ogni $n = 0, 1, \dots, N_h - 1$:

$$\text{trovare } u_{n+1} : F_n^{CN}(u_{n+1}) = 0 \quad \text{per ogni } n = 0, 1, \dots, N_h - 1,$$

con $u_0 = y_0$, dove

$$F_n^{CN}(y) := y - u_n - \frac{h}{2} [f(t_n, u_n) + f(t_{n+1}, y)].$$

Usando il metodo the Newton in maniera simile all'Algoritmo 7.1 per risolvere tale problema non lineare allora è necessario calcolare la derivata prima di $F_n^{CN}(y)$, ovvero

$$(F_n^{CN})'(y) = 1 - \frac{h}{2} \frac{\partial f}{\partial y}(t_{n+1}, y).$$

Se invece si considera il metodo delle iterazioni di punto fisso in analogia all'Algoritmo 7.2, allora può essere utilizzata la funzione di iterazione

$$\phi_n^{CN}(y) := u_n + \frac{h}{2} [f(t_n, u_n) + f(t_{n+1}, y)].$$

7.4.4 Metodo di Heun

Il metodo di *Heun* per l'approssimazione del problema di Cauchy (7.22) si scrive come: trovare $\{u_n\}_{n=0}^{N_h}$ tali che

$$\begin{cases} u_{n+1}^* = u_n + h f(t_n, u_n) \\ u_{n+1} = u_n + \frac{h}{2} [f(t_n, u_n) + f(t_{n+1}, u_{n+1}^*)] \\ u_0 = y_0. \end{cases} \quad \text{per } n = 0, 1, \dots, N_h - 1, \quad (7.27)$$

Il metodo di Heun può essere interpretato come un metodo di Crank–Nicolson modificato con il termine $f(t_{n+1}, u_{n+1})$, che rende implicito il metodo, sostituito da $f(t_{n+1}, u_{n+1}^*)$, essendo u_{n+1}^* la soluzione estrapolata per mezzo del metodo di Eulero in avanti. Il metodo di Heun risulta dunque un metodo **esplicito**.

7.4.5 Analisi dell'errore dei metodi

I metodi introdotti finora differiscono non solo per la loro natura esplicita o implicita, ma anche – come vedremo in questa sezione – per la velocità con cui l'errore di approssimazione tende a zero quando $h \rightarrow 0$ (assumendo fissato l'intervallo $I = (t_0, t_f)$).

Definizione 7.4.1. L'errore associato all'approssimazione numerica del problema di Cauchy (7.22) a t_n è

$$e_n := |y_n - u_n| \quad \text{per } n = 0, 1, \dots, N_h.$$

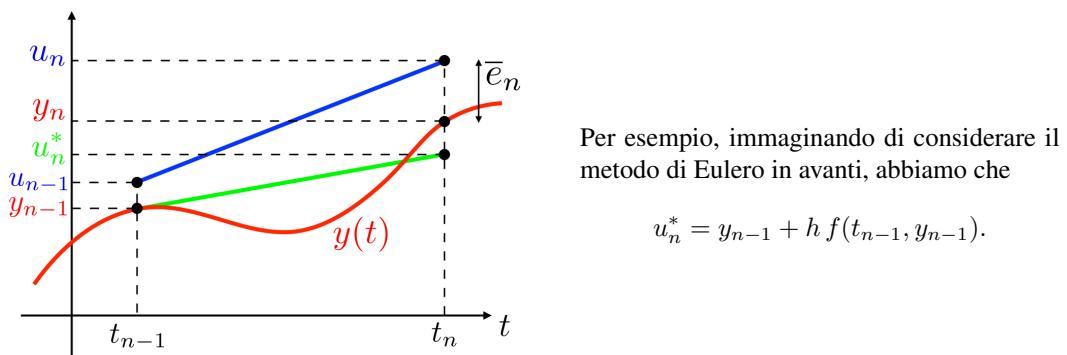
Risulta conveniente indicare con $\bar{e}_n := y_n - u_n$, tale per cui $e_n = |\bar{e}_n|$. L'errore è spesso definito come

$$E_h := \max_{n=0,1,\dots,N_h} e_n.$$

Osserviamo dalla seguente figura che l'errore \bar{e}_n può essere interpretato come la somma di due contributi. In particolare, abbiamo:

$$\bar{e}_n = y_n - u_n = \bar{e}_n^{(I)} + \bar{e}_n^{(II)} = (y_n - u_n^*) + (u_n^* - u_n), \quad (7.28)$$

dove u_n^* indica la soluzione *estrapolata* a t_n applicando il metodo numerico alla soluzione esatta a t_{n-1} .



Osserviamo che il contributo all'errore indicato come $\bar{e}_n^{(I)} = (y_n - u_n^*)$ rappresenta l'errore generato dall'applicazione di un passo del metodo numerico ed è “controllato” dalla *consistenza* del metodo. In altre parole, $\bar{e}_n^{(I)}$ dipende dall'errore di troncamento locale, ovvero dall'errore generato forzando la soluzione esatta a soddisfare il metodo numerico.

Definizione 7.4.2. L'errore di troncamento locale è:

$$\tau_n(h) = \frac{y_n - u_n^*}{h} \quad \text{per } n = 0, \dots, N_h,$$

mentre l'errore di troncamento globale è definito come:

$$\tau(h) = \max_{n=0, \dots, N_h} |\tau_n(h)|.$$

Definizione 7.4.3. Un metodo numerico per l'approssimazione di una EDO si dice consistente se $\lim_{h \rightarrow 0} \tau(h) = 0$ e consistente con ordine $p \geq 1$ se $\tau(h) = O(h^p)$.

Il contributo dell'errore $\bar{e}_n^{(II)} = (u_n^* - u_n)$ in Eq. (7.28) dipende invece dalla propagazione dell'errore accumulato ai passi (temporali) precedenti; tale termine è sostanzialmente legato alla (zero) *stabilità* del metodo numerico, concetto, quest'ultimo, che introdurremo nel seguito. Osserviamo infine dal Teorema 1.6.1 di equivalenza di Lax–Richtmeyer che un problema (metodo) numerico consistente e stabile è anche *convergente*.

Definizione 7.4.4. Se l'errore associato a un metodo numerico per EDO è tale per cui

$$e_n \leq \tilde{e}_n := C h^p \quad \text{per } n = 0, \dots, N_h,$$

con \tilde{e}_n lo stimatore dell'errore e C una costante positiva indipendente da h , oppure

$$E_h = \max_{n=0,1,\dots,N_h} e_n \leq C h^p,$$

allora il metodo ha ordine di convergenza $p \geq 1$ (ordine di accuratezza del metodo).

Stimiamo per esempio l'errore e_n associato al metodo di Eulero in avanti. Al tal fine iniziamo stimando separatamente i due contributi $\bar{e}_n^{(I)}$ e $\bar{e}_n^{(II)}$ di Eq. 7.28 per $n = 1, \dots, N_h$.

- (I) Dalla definizione di errore di troncamento locale, dalla definizione di u_n^* e da una espansione in serie di Taylor di y_n attorno a y_{n-1} , abbiamo:

$$h \tau_n(h) = y_n - u_n^* = \left(y_{n-1} + h y'_{n-1} + \frac{h^2}{2} y''(\eta_n) \right) - (y_{n-1} + h f(t_{n-1}, y_{n-1})),$$

per qualche $\eta_n \in [t_{n-1}, t_n]$. Posto per semplicità $M_n = y''(\eta_n)$, si ha che

$$\tau_n(h) = \frac{M_n}{2} h \quad \text{e} \quad \tau(h) = \frac{M}{2} h$$

essendo $M = \max_{n=0, \dots, N_h} M_n$. Il metodo di Eulero in avanti risulta pertanto consistente di ordine 1. Infine, si ha che

$$\bar{e}_n^{(I)} \leq h \tau(h) = \frac{M}{2} h^2.$$

- (II) Dalla definizione di u_n^* e applicazione del metodo di Eulero in avanti, si ottiene:

$$\begin{aligned} \bar{e}_n^{(II)} &= u_n^* - u_n = (y_{n-1} + h f(t_{n-1}, y_{n-1})) - (u_{n-1} + h f(t_{n-1}, u_{n-1})) \\ &\quad (y_{n-1} - u_{n-1}) + h (f(t_{n-1}, y_{n-1}) - f(t_{n-1}, u_{n-1})). \end{aligned}$$

Dalla definizione di errore e per una funzione f Lipschitz continua nel secondo argomento con costante $L \geq 0$, si ottiene:

$$|\bar{e}_n^{(II)}| \leq |y_{n-1} - u_{n-1}| + h |f(t_{n-1}, y_{n-1}) - f(t_{n-1}, u_{n-1})|,$$

ovvero che

$$\left| \bar{e}_n^{(II)} \right| \leq (1 + h L) |\bar{e}_{n-1}|.$$

Combinando i contributi (I) e (II) precedenti, si ottiene:

$$e_n \leq h \tau(h) + \leq (1 + h L) |\bar{e}_{n-1}| \quad \text{per } n = 1, \dots, N_h,$$

ovvero procedendo ricorsivamente:

$$e_n \leq \left[\sum_{k=0}^{n-1} (1 + h L)^k \right] h \tau(h) = \left[\frac{(1 + h L)^n - 1}{h L} \right] h \tau(h) \quad \text{per } n = 1, \dots, N_h.$$

Dato che $(1 + h L) \leq e^{h L}$ e $n h = t_n - t_0$, si può stimare l'errore precedente anche come

$$e_n \leq \frac{e^{L(t_n-t_0)} - 1}{L} \tau(h) \quad \text{per } n = 1, \dots, N_h.$$

Infine, essendo $\tau(h) = \frac{M}{2} h$, si ottiene per il metodo di Eulero in avanti:

$$e_n \leq \left(\frac{M}{2} \frac{e^{L(t_n-t_0)} - 1}{L} \right) h \quad \text{per } n = 1, \dots, N_h, \quad (7.29)$$

da cui **si deduce che tale metodo è convergente con ordine $p = 1$** in h sulla base della Definizione 7.4.4 con costante $C = \left(\frac{M}{2} \frac{e^{L(t_n-t_0)} - 1}{L} \right)$ **indipendente da h** . Osserviamo in particolare che l'errore globale E_h tende a zero con lo stesso ordine dell'errore di troncamento locale, e che la costante di Lipschitz di f gioca un ruolo fondamentale nella stima precedente: maggiore è L , maggiore risulta la costante C . Abbiamo dunque provato il seguente risultato.

Proposizione 7.4.1. *Se la soluzione del problema di Cauchy (7.22) è $y \in C^2(I)$, allora il metodo di Eulero in avanti converge con ordine $p = 1$ in h .*

Un risultato analogo vale per il metodo di Eulero all'indietro.

Proposizione 7.4.2. *Se la soluzione del problema di Cauchy (7.22) è $y \in C^2(I)$, allora il metodo di Eulero all'indietro converge con ordine $p = 1$ in h .*

I metodi di Crank-Nicolson e di Heun risultano invece di ordine $p = 2$.

Proposizione 7.4.3. *Se la soluzione del problema di Cauchy (7.22) è $y \in C^3(I)$, allora i metodi di Crank-Nicolson e Heun convergono con ordine $p = 2$ in h .*

Per determinare numericamente l'ordine di convergenza p di un metodo di approssimazione per ODE, si può risolvere il problema di Cauchy (7.22) per valori diversi di $h = h_1, h_2, \dots$ e, conoscendo la soluzione esatta $y(t)$, calcolare gli errori per i valori di $n = n_1, n_2, \dots$ corrispondenti allo stesso $\bar{t} \in (t_0, t_f]$, ovvero

$$n_i = \frac{\bar{t} - t_0}{h_i}$$

per alcuni $i = 1, 2, \dots$, se $n_i \in \mathbb{N}$. Quindi, si può stimare l'ordine di convergenza come

$$p \simeq \log_{h_1/h_2} \frac{e_{n_1}}{e_{n_2}},$$

per h_1 e h_2 “sufficientemente piccoli”. Infatti, se $e_n \sim Ch^p$, otteniamo che

$$\frac{e_{n_1}}{e_{n_2}} \sim \left(\frac{h_1}{h_2}\right)^p \Rightarrow \log \frac{e_{n_1}}{e_{n_2}} \sim p \log_{h_1/h_2} \Rightarrow p \simeq \frac{\log(e_{n_1}/e_{n_2})}{\log(h_1/h_2)},$$

da cui la relazione precedente. Gli stessi risultati valgono considerando l’errore globale E_h .

7.4.6 Stabilità dei metodi numerici: zero-stabilità e stabilità assoluta

Riguardo alla stabilità dei metodi numerici per l’approssimazione di EDO, è possibile introdurre due concetti di stabilità, a seconda delle caratteristiche dell’intervallo temporale I . In entrambi i casi, il concetto di stabilità ha a che fare con il controllo delle *perturbazioni* sulla soluzione numerica.

Zero-stabilità

La *zero-stabilità* è una proprietà di un metodo numerico riguardante la possibilità di controllare l’effetto delle *perturbazioni numeriche* sulla soluzione in intervalli limitati I , tali che $|I| < +\infty$, o comunque *brevi*; si tratta cioè della controparte numerica del concetto di stabilità secondo Lyapunov del problema di Cauchy (si veda Definizione 7.2.2). In generale, se h è “sufficientemente” piccolo, la zero-stabilità di un metodo è garantita.

A titolo di esempio, consideriamo il metodo di Eulero in avanti, perturbato introducendo un termine addizionale ρ_n a ogni passo di tempo, che rappresenti ad esempio l’effetto di errori di troncamento: indicando con $\{z_n\}_{n=0}^{N_h}$ la soluzione ottenuta con tale metodo, risulta:

$$\begin{cases} z_{n+1} = z_n + h(f(t_n, z_n) + \rho_{n+1}), & n = 0, \dots, N_h - 1 \\ z_0 = y_0 + \rho_0. \end{cases} \quad (7.30)$$

Un metodo numerico è zero-stabile se la differenza tra la soluzione ottenuta con il metodo numerico perturbato $\{z_n\}_{n=0}^{N_h}$, e quella ottenuta con il metodo non perturbato $\{u_n\}_{n=0}^{N_h}$, si può controllare in funzione della perturbazione stessa $\{\rho_n\}_{n=0}^{N_h}$ su un intervallo di tempo I limitato.

Definizione 7.4.5. Un metodo numerico per l’approssimazione del problema di Cauchy (7.22) è zero-stabile se esistono $h_0 > 0$, $C > 0$ e $\varepsilon_0 > 0$ tali che, per ogni $h \in (0, h_0]$ e per ogni $\varepsilon \in (0, \varepsilon_0]$,

$$|\rho_n| \leq \varepsilon \quad \text{per ogni } n = 0, 1, \dots, N_h,$$

implica che

$$|z_n - u_n| \leq C\varepsilon \quad \text{per ogni } n = 0, 1, \dots, N_h,$$

dove ρ_n indica la perturbazione introdotta a t_n , z_n e u_n sono rispettivamente le soluzioni ottenute con il metodo numerico perturbato e non perturbato (a t_n); C è una costante indipendente da h , ma dipendente dalla lunghezza dell’intervallo $|I| = t_f - t_0$, mentre ε rappresenta la massima ampiezza della perturbazione introdotta.

Sottolineiamo come la zero-stabilità sia una proprietà intrinseca del metodo numerico, dal momento che il problema di Cauchy risulta infatti stabile a patto che f sia Lipschitziana rispetto al secondo argomento. Tale proprietà garantisce tuttavia che anche il metodo numerico sia zero-stabile:

Teorema 7.4.1. Si consideri l’approssimazione numerica della soluzione del problema di Cauchy (7.22). A patto che $f(t, y)$ sia Lipschitz-continua rispetto a y , uniformemente in t , allora i metodi numerici precedenti risultano zero-stabili. Inoltre, si ha che $C = (1 + T)e^{TL}$ essendo L la costante di Lipschitz e $T = |I| = |t_f - t_0|$.

Vale inoltre il seguente teorema:

Teorema 7.4.2 (Equivalenza, Lax-Richtmyer). *Nelle stesse ipotesi del Teorema 7.4.1, si ha che:*

$$e_n = |y_n - u_n| \leq [|y_0 - u_0| + n h \tau(h)] e^{n h L} \quad \text{per ogni } n = 0, 1, \dots, N_h.$$

In particolare, se il metodo numerico è consistente, e $|y_0 - u_0| \rightarrow 0$ per $h \rightarrow 0$, allora il metodo numerico risulta convergente. Inoltre, se $|y_0 - u_0| = O(h^p)$ e il metodo è consistente con ordine p , allora esso è anche convergente con ordine p .

Se in particolare $u_0 = y_0$ e il metodo è consistente con ordine p , allora esso è anche convergente con ordine p , e vale

$$E_h = \max_{n=0, \dots, N_h} |y_n - u_n| \leq C \tau(h).$$

Pertanto, un metodo consistente e zero-stabile risulta anche convergente; viceversa, un metodo convergente è anche zero-stabile a patto che sia consistente.

Stabilità assoluta (stabilità su intervalli illimitati)

La *stabilità assoluta* si riferisce al controllo delle perturbazioni su intervalli illimitati, per i quali $t_f = +\infty$ o ad intervalli per i quali $|I| < +\infty$, ma molto “lunghi”; si tratta cioè della controparte numerica del concetto di asintotica stabilità del problema di Cauchy (secondo Lyapunov). In tali casi, anche se h è fissato, allora $\lim_{t_f \rightarrow +\infty} N_h = +\infty$; in questi casi, non risulta possibile scegliere h piccolo a piacere per garantire la stabilità. Ciò nonostante, siamo comunque interessati a controllare la propagazione delle perturbazioni numeriche per $t_f \rightarrow +\infty$.

Osserviamo innanzitutto come la nozione di zero-stabilità non risulti utile in questo caso: poiché la costante $C = (1 + T)e^{TL}$ dipende dalla lunghezza dell’intervallo I , si ha che $C \rightarrow +\infty$ se $T \rightarrow +\infty$. Inoltre, se il problema di Cauchy non fosse asintoticamente stabile secondo Lyapunov, l’instabilità dello schema numerico sui tempi lunghi sarebbe intrinseca. Possiamo cioè analizzare l’assoluta stabilità dei metodi numerici introdotti qualora essi siano applicati a problemi di Cauchy asintoticamente stabili, tali per cui $C \not\rightarrow +\infty$ se $T \rightarrow +\infty$ (ovvero, tali per cui la costante C è indipendente da t); in altri termini, nel caso di un problema di Cauchy che non risulta assolutamente stabile, non possiamo controllare l’effetto delle perturbazioni sui dati, su un intervallo di tempo infinito (indipendentemente dal fatto che il metodo numerico sia assolutamente stabile).

Richiamiamo e generalizziamo il *problema modello* (Esempio 7.2.1):

$$\begin{cases} y'(t) = \lambda y(t) & t \in (0, +\infty), \\ y(0) = y_0, \end{cases} \quad \text{con } \lambda \in \mathbb{C}. \quad (7.31)$$

La sua soluzione risulta $y(t) = y_0 e^{\lambda t}$ e, se $\Re\{\lambda\} < 0$, necessariamente

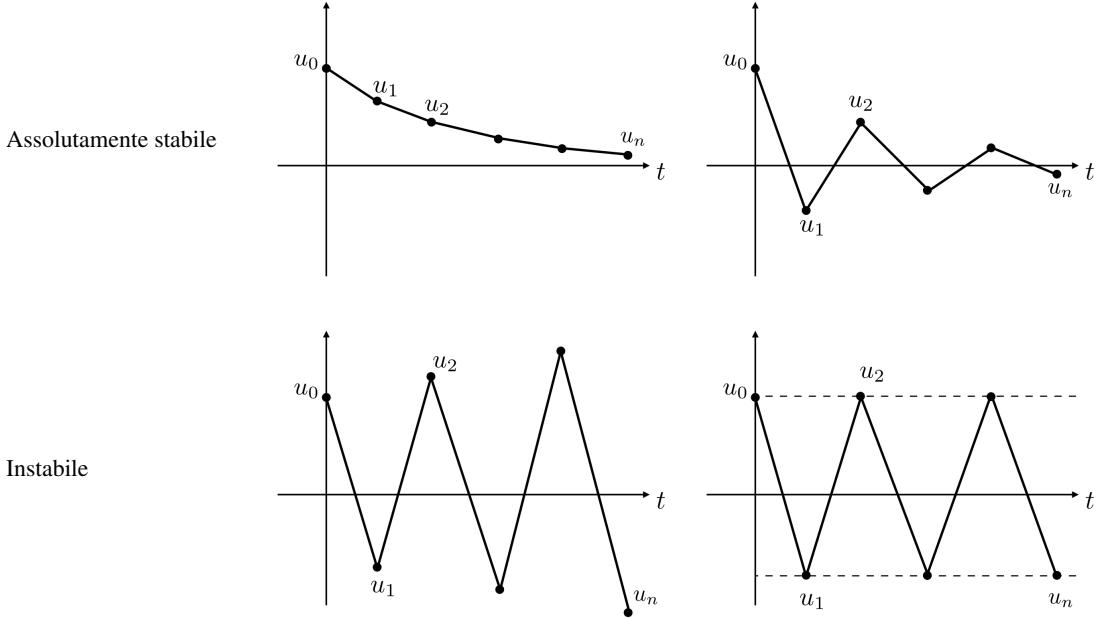
$$y(t) = y_0 e^{\lambda t} \rightarrow 0 \quad \text{per } t \rightarrow +\infty.$$

Definizione 7.4.6. Un metodo numerico è assolutamente stabile se, applicato al problema modello (7.31) per $\lambda \in \mathbb{R}$, con $\lambda < 0$ fornisce una soluzione numerica tale che:

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

Il metodo si dice incondizionatamente assolutamente stabile se ciò accade per ogni $h > 0$, oppure condizionatamente assolutamente stabile se ciò accade per ogni h tale che $0 < h < h_{\max}$, per un opportuno $h_{\max} > 0$. In generale, il concetto di stabilità assoluta si può estendere al caso in cui $\lambda \in \mathbb{C}$.

Esempio 7.4.5. Mostriamo alcuni esempi di soluzioni approssimate con metodi numerici assolutamente stabili oppure instabili:



Definizione 7.4.7. La funzione di stabilità associata a un metodo numerico per la soluzione del problema modello (7.31) è la funzione (in campo complesso) $R(z) : \mathbb{C} \rightarrow \mathbb{C}$ tale che

$$u_n = [R(h\lambda)]^n y_0 \quad \text{per } n = 0, 1, \dots \quad (7.32)$$

In particolare, si ha che:

- per il metodo di Eulero in avanti,

$$R^{EA}(z) = 1 + z;$$

infatti, scrivendo lo schema (7.23) nel caso del problema modello dell'Esempio 7.2.1, si ha che $u_{n+1} = (1 + h\lambda)u_n$ per $n = 0, 1, \dots$, da cui il risultato segue scegliendo $z = h\lambda$;

- per il metodo di Eulero all'indietro,

$$R^{EI}(z) = \frac{1}{1 - z};$$

infatti, scrivendo lo schema (7.24) nel caso del problema modello, $u_{n+1} = \frac{u_n}{1 - h\lambda}$ per $n = 0, 1, \dots$;

- per il metodo di Crank–Nicolson,

$$R^{CN}(z) = \frac{1 + z/2}{1 - z/2};$$

infatti, scrivendo lo schema (7.26) nel caso del problema modello, $u_{n+1} = \frac{1 + (h\lambda)/2}{1 - (h\lambda)/2} u_n$ per $n = 0, 1, \dots$;

- per il metodo di Heun,

$$R^H(z) = 1 + z + \frac{z^2}{2},$$

poiché scrivendo lo schema (7.27) nel caso del problema modello, $u_{n+1} = \left[1 + h\lambda + \frac{(h\lambda)^2}{2}\right] u_n$ per $n = 0, 1, \dots$

Dall'Eq. (7.32) risulta evidente come un metodo risulti *assolutamente stabile* se e solo se $|R(h\lambda)| < 1$. Più precisamente, un metodo è *incondizionatamente assolutamente stabile* se, per $\lambda \in \mathbb{R}$, con $\lambda < 0$:

$$|R(h\lambda)| < 1 \quad \text{per ogni } h > 0;$$

invece, un metodo è *condizionatamente assolutamente stabile* se:

$$|R(h\lambda)| < 1 \quad \text{per } 0 < h < h_{max}.$$

Osservazione 7.4.2. Il metodo di Eulero in avanti è condizionatamente assolutamente stabile, ovvero risulta assolutamente stabile a patto che:

- nel caso $\lambda \in \mathbb{R}$ e $\lambda < 0$, si prenda $0 < h < h_{max}$, dove

$$h_{max} = \frac{2}{|\lambda|};$$

si ha infatti che

$$|R^{EA}(h\lambda)| < 1 \iff |1 + h\lambda| < 1 \iff 0 < h < \frac{2}{|\lambda|}.$$

- nel caso $\lambda \in \mathbb{C}$ e $\operatorname{Re}\{\lambda\} < 0$, si prenda $0 < h < h_{max}$, con

$$h_{max} = -\frac{2\operatorname{Re}\{\lambda\}}{|\lambda|^2};$$

in questo caso il modulo è da intendersi nel senso dei numeri complessi, per cui

$$|R^{EA}(h\lambda)| < 1 \iff |1 + h\lambda| < 1 \iff (1 + h\operatorname{Re}\{\lambda\})^2 + h^2(\operatorname{Im}\{\lambda\})^2 < 1,$$

ovvero

$$1 + h^2(\operatorname{Re}\{\lambda\})^2 + 2h\operatorname{Re}\{\lambda\} + h^2(\operatorname{Im}\{\lambda\})^2 < 1 \iff h|\lambda|^2 + 2\operatorname{Re}\{\lambda\} < 0,$$

da cui il risultato precedente.

Osservazione 7.4.3. Il metodo di Eulero all'indietro è incondizionatamente assolutamente stabile, come si può dedurre facilmente imponendo che $|R^{EI}(h\lambda)| < 1$.

Osservazione 7.4.4. Il metodo di Crank–Nicolson è incondizionatamente assolutamente stabile, come si può dedurre facilmente imponendo che $|R^{CN}(h\lambda)| < 1$.

Osservazione 7.4.5. Il metodo di Heun è condizionatamente assolutamente stabile; in particolare, se $\lambda \in \mathbb{R}$, $\lambda < 0$, deve risultare $0 < h < h_{max}$, con $h_{max} = \frac{2}{|\lambda|}$, come si deduce imponendo che $|R^H(h\lambda)| < 1$.

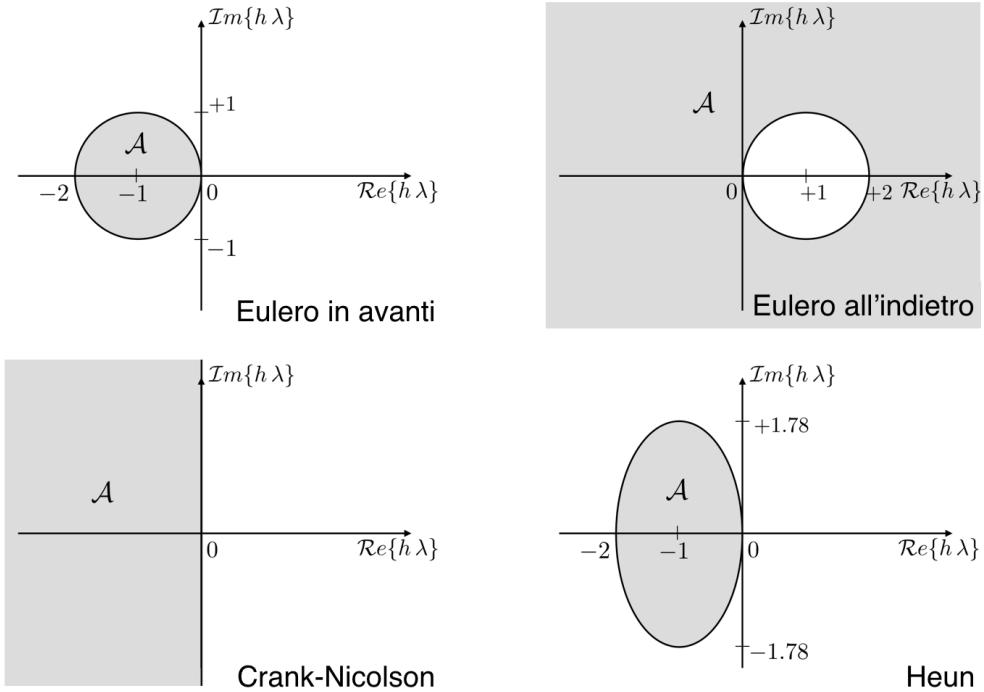
Il concetto di stabilità assoluta si generalizza al caso di $\lambda \in \mathbb{C}$ e non necessariamente per $\operatorname{Re}\{\lambda\} < 0$, a prescindere dunque dal significato fisico della soluzione del problema modello. Si introduce a tal proposito la seguente definizione.

Definizione 7.4.8. La regione di assoluta stabilità di un metodo numerico applicato al problema modello (7.31) è l'insieme nel piano complesso

$$\mathcal{A} := \{z \in \mathbb{C} : |R(z)| < 1\},$$

dove $R(z) : \mathbb{C} \rightarrow \mathbb{C}$ è la funzione di stabilità.

Evidenziamo nella seguente Figura le regioni \mathcal{A} di assoluta stabilità dei metodi numerici incontrati finora.



Definizione 7.4.9. Un metodo numerico si dice \mathcal{A} -stabile se risulta incondizionatamente assolutamente stabile per il problema modello (7.31), per ogni $\lambda \in \mathbb{C}$ tale che $\operatorname{Re}\{\lambda\} < 0$.

Osservazione 7.4.6. I metodi di Eulero all'indietro e Crank–Nicolson sono \mathcal{A} -stabili.

Per un problema di Cauchy generico, la stabilità assoluta garantisce il controllo delle perturbazioni sui tempi lunghi. Nel caso di un generico problema di Cauchy (7.22) per cui f sia differenziabile con continuità rispetto al secondo argomento, ed esistano $\lambda_{min}, \lambda_{max} \in \mathbb{R}$ tali che $-\infty < \lambda_{max} < \lambda_{min} < 0$ e

$$\lambda_{max} < \frac{\partial f}{\partial y}(t, y(t)) < \lambda_{min} \quad \forall t \in I,$$

allora a patto che

$$0 < h < h_{max} = \frac{2}{|\lambda_{max}|} = \frac{2}{\max_{t \in I} \left| \frac{\partial f}{\partial y}(t, y(t)) \right|},$$

i metodi di *Eulero in avanti* e *Heun* sono condizionatamente assolutamente stabili, ovvero

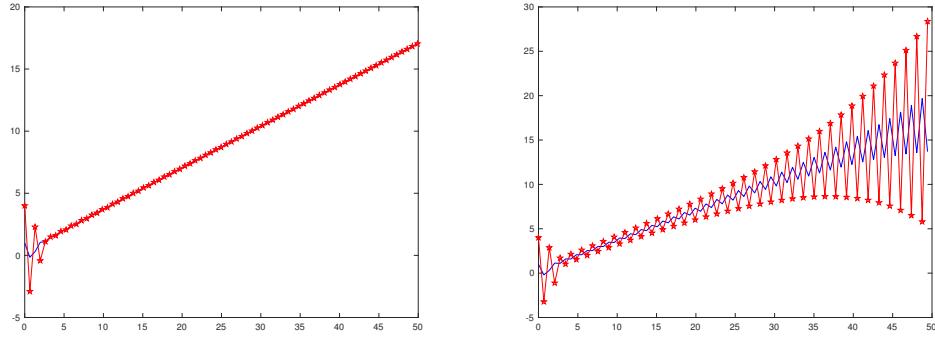
$$\lim_{n \rightarrow \infty} |z_n - u_n| \leq \frac{\rho_{max}}{|\lambda_{min}|}, \quad (7.33)$$

dove $\rho_{max} = \max_{n=0,1,\dots,N_h} |\rho_n|$ e z_n è la soluzione ottenuta perturbando il metodo numerico con una perturbazione ρ_n al tempo t_n . In particolare, la differenza $z_n - u_n$ resta limitata per ogni $n = 0, 1, \dots$ indipendentemente da n e da h .

Esempio 7.4.6. Si consideri il seguente problema di Cauchy:

$$\begin{cases} y'(t) = \arctan(3y) - 3y + t, & t \in (0, +\infty), \\ y(0) = 1. \end{cases} \quad (7.34)$$

Essendo in questo caso $f_y(t, y) = \partial f / \partial y(t, y) = 3/(1 + 9y^2) - 3$, si ha $\lambda_{\max} < f_y(t, y) < \lambda_{\min} < 0$ per ogni $t > 0$, dove $\lambda_{\max} = -\sup_{t>0} |f_y(t, y(t))| = -3$ e $\lambda_{\min} = -\inf_{t>0} |f_y(t, y(t))| \simeq -2.24$. È infatti possibile mostrare che y non può annullarsi, dunque è lecito assumere che sia $y > 0$; inoltre $y > 0.57$ per $t > 0$. Possiamo quindi aspettarci che le perturbazioni nel metodo di Eulero in avanti restino controllate purché $h < h_{\max} = 2/3$. Considerando $y_0 = 1$ e, successivamente, $y_0 = 1 + \rho_0$, dove $\rho_0 = 3$ è una perturbazione sul dato iniziale, risolviamo il problema ottenendo rispettivamente le soluzioni numeriche u_n e z_n per $n = 0, 1, \dots$. Nel caso in cui $h = 2/3 - 0.01 < 2/3$ (metodo assolutamente stabile, grafico a sinistra) la perturbazione resta controllata, dal momento che la differenza $z_n - u_n$ tra la soluzione perturbata e quella non perturbata è controllata da un valore costante anche sui tempi lunghi. Viceversa, scegliendo $h = 2/3 + 0.02 > 2/3$ (metodo non assolutamente stabile, grafico a destra), non è possibile controllare le perturbazioni sui tempi lunghi.



7.4.7 Metodi Runge–Kutta

I metodi *Runge–Kutta* sono metodi a un passo per l'approssimazione di EDO in cui l'approssimazione u_{n+1} della soluzione al tempo t_{n+1} è determinata valutando $f(t, y)$ in $s \geq 1$ stadi sull'intervallo $[t_n, t_{n+1}]$. In generale un metodo Runge–Kutta per l'approssimazione del problema di Cauchy (7.22) consiste nel: trovare $\{u_n\}_{n=0}^{N_h}$ tale che

$$\boxed{\begin{cases} u_{n+1} = u_n + h \sum_{i=1}^s b_i K_i & \text{per } n = 0, 1, \dots, N_h - 1, \\ u_0 = y_0, & \\ \text{dove } K_i := f \left(t_n + c_i h, u_n + h \sum_{j=1}^s a_{ij} K_j \right) & \text{per } i = 1, \dots, s, \end{cases}} \quad (7.35)$$

per opportuni coefficienti $\mathbf{c} = (c_1, \dots, c_s)^T \in \mathbb{R}^s$, $\mathbf{b} = (b_1, \dots, b_s)^T \in \mathbb{R}^s$, e $A \in \mathbb{R}^{s \times s}$, con $(A)_{ij} = a_{ij}$ for $i, j = 1, \dots, s$. Questi coefficienti, che determinano il particolare tipo di metodo di Runge–Kutta, vengono memorizzati nel cosiddetto *array di Butcher* come:

$$\begin{array}{c|c} \mathbf{c} & A \\ \hline & \mathbf{b}^T \end{array}$$

Osserviamo che se la matrice A è memorizzata a partire dall'angolo in basso a sinistra, il metodo Runge–Kutta è *esplicito* se $a_{ij} = 0$ per $j \geq i$, per ogni $i = 1, \dots, s$; altrimenti, il metodo Runge–Kutta è *implicito*. Nel primo caso, ogni K_i dipende solo dai valori K_1, \dots, K_{i-1} ; nel secondo caso, infatti, per determinare i coefficienti K_i , $i = 1, \dots, s$, occorre risolvere un sistema di equazioni non lineari di dimensione s . Si assume inoltre che:

$$c_i = \sum_{j=1}^s a_{ij}$$

e che, al fine di garantire la consistenza di un metodo Runge-Kutta, risulti

$$\sum_{j=1}^s b_j = 1.$$

Infatti, definendo $u_{n+1}^* = y_n + h \sum_{i=1}^s b_i K_i^*$ la soluzione al tempo t_{n+1} ottenuta con un passo del metodo

Runge-Kutta a partire dalla soluzione esatta al tempo t_n , essendo $K_i^* := f \left(t_n + c_i h, y_n + h \sum_{j=1}^s a_{ij} K_j \right)$

si ha che l'*errore di troncamento locale* è:

$$\tau_{n+1}(h) = \frac{y_{n+1} - u_{n+1}^*}{h} = \sum_{i=1}^s b_i K_i^* - f(t_n, y_n) - \frac{h}{2} y''(\eta_n).$$

Poiché f è continua rispetto a entrambe le variabili, si ha che $\lim_{h \rightarrow 0} K_i^* = f(t_n, y_n)$ per ogni i , e dunque

$$\lim_{h \rightarrow 0} \tau_{n+1}(h) = \lim_{h \rightarrow 0} \sum_{i=1}^s b_i f(t_n, y_n) - f(t_n, y_n)$$

e dunque possiamo concludere che $\lim_{h \rightarrow 0} \tau_{n+1}(h) = 0$ se e solo se $\sum_{j=1}^s b_j = 1$. Se f è Lipschitz continua

rispetto alla seconda variabile, i metodi Runge-Kutta sono zero-stabili, e quindi anche convergenti. In particolare, un metodo Runge-Kutta esplicito a s stadi non può avere ordine maggiore di s , e non esistono metodi Runge-Kutta espliciti a s stadi di ordine s se $s \geq 5$.

Inoltre, all'aumentare di s , aumenta l'ordine dei metodi Runge-Kutta (sia esplicativi che impliciti), e cresce in estensione la regione di assoluta stabilità. In particolare, nel caso dei metodi RK esplicativi di ordine $s = 1, \dots, 4$, si trova che la funzione di stabilità è data da

$$R(h\lambda) = 1 + h\lambda + \frac{1}{2}(h\lambda)^2 + \dots + \frac{1}{s!}(h\lambda)^s$$

e di conseguenza $|R(h\lambda)|$ non può mai essere minore di 1 per tutti i valori di $h\lambda$, quindi la regione di assoluta stabilità non può mai essere illimitata per un metodo RK esplicito.

Esempio 7.4.7. Consideriamo alcuni esempi di metodi Runge-Kutta esplicativi con $s = 1, 2, e 4$ stadi.

- $s = 1$, RK1. In questo caso, Eq. (7.35) diventa:

$$\begin{cases} u_{n+1} = u_n + h b_1 K_1 & \text{per } n = 0, 1, \dots, N_h - 1, \\ u_0 = y_0, \end{cases}$$

dove $K_1 = f(t_n + c_1 h, u_n + h a_{11} K_1)$; scegliendo a questo punto $c_1 = 0$, $b_1 = 1$, e $a_{11} = 0$, ovvero usando il seguente array di Butcher,

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

otteniamo il metodo RK1, che coincide con il metodo di *Eulero in avanti* (7.23), essendo $K_1 = f(t_n, u_n)$.

- $s = 2$, RK2. In questo caso, Eq. (7.35) diventa:

$$\begin{cases} u_{n+1} = u_n + h b_1 K_1 + h b_2 K_2 & \text{per } n = 0, 1, \dots, N_h - 1, \\ u_0 = y_0, \end{cases}$$

dove $K_1 = f(t_n + c_1 h, u_n + h a_{11} K_1 + h a_{12} K_2)$ e $K_2 = f(t_n + c_2 h, u_n + h a_{21} K_1 + h a_{22} K_2)$. Considerando il seguente array di Butcher,

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \hline 1 & 1 & 0 \\ \hline & 1/2 & 1/2 \end{array}$$

si ottiene il metodo RK2, che coincide con il metodo di *Heun* (7.27), essendo $K_1 = f(t_n, u_n)$ e $K_2 = f(t_{n+1}, u_n + h K_1)$.

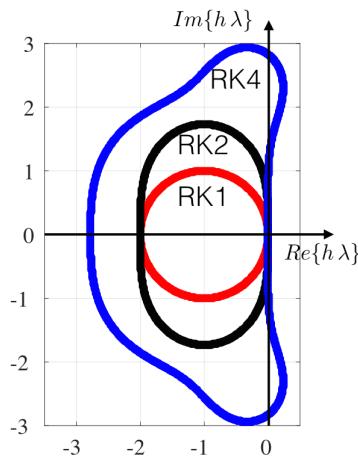
- $s = 4$, RK4. Considerando il seguente array di Butcher:

0	0	0	0	0
1/2	1/2	0	0	0
1/2	0	1/2	0	0
1	0	0	1	0
				1/6 1/3 1/3 1/6

si ottiene il seguente metodo (RK4):

$$\begin{cases} u_{n+1} = u_n + h \left(\frac{1}{6} K_1 + \frac{1}{3} K_2 + \frac{1}{3} K_3 + \frac{1}{6} K_4 \right) & \text{per } n = 0, 1, \dots, N_h - 1, \\ u_0 = y_0, \end{cases}$$

dove $K_1 = f(t_n, u_n)$, $K_2 = f\left(t_n + \frac{h}{2}, u_n + \frac{h}{2} K_1\right)$, $K_3 = f\left(t_n + \frac{h}{2}, u_n + \frac{h}{2} K_2\right)$, e infine $K_4 = f(t_{n+1}, u_n + h K_3)$, essendo $t_n + h \equiv t_{n+1}$. Il metodo RK4 ha ordine di accuratezza pari a 4.



Regioni di assoluta stabilità \mathcal{A} per i metodi RK1, RK2, e RK4 nel caso del problema modello (7.2.1). Le regioni \mathcal{A} sono contenute all'interno delle curve che ne rappresentano il contorno.

Osservazione 7.4.7. Osserviamo come i metodi impliciti ad un passo sin qui considerati siano incondizionatamente assolutamente stabili, mentre quelli esplicativi siano condizionatamente assolutamente stabili. Questa non è tuttavia una regola generale: possono infatti esistere schemi impliciti instabili o solo condizionatamente stabili. Al contrario, non esistono schemi esplicativi incondizionatamente assolutamente stabili.

7.4.8 Metodi multipasso

I metodi multipasso (*multistep*) indicano una famiglia di metodi per cui la soluzione approssimata u_{n+1} è ottenuta utilizzando u_n, \dots, u_{n-p} per qualche $p \geq 0$, essendo $p+1$ il numero dei *passi*. Un metodo multipasso per l'approssimazione del problema di Cauchy problem (7.22) si scrive come: trovare $\{u_n\}_{n=0}^{N_h}$ tale che

$$u_{n+1} = \sum_{j=0}^p a_j u_{n-j} + h \sum_{j=-1}^p b_j f(t_{n-j}, u_{n-j}) \quad \text{per } n = p, \dots, N_h - 1, \quad (7.36)$$

dati u_0, \dots, u_p , per alcuni coefficienti $\{a_j\}_{j=0}^p$ e $\{b_j\}_{j=-1}^p$ che determinano il metodo. Se $b_{-1} = 0$, il metodo è *esplicito*, altrimenti è *implicito*. Un metodo multipasso è *consistente* se si verificano le seguenti condizioni:

$$\sum_{j=0}^p a_j = 1 \quad \text{e} \quad - \sum_{j=0}^p j a_j + \sum_{j=-1}^p b_j = 1.$$

Esempio 7.4.8. Consideriamo alcuni metodi a un passo, cioè tali per cui $p = 0$. Essi si scrivono a partire dall'Eq. (7.36) come:

$$u_{n+1} = a_0 u_n + h b_{-1} f(t_{n+1}, u_{n+1}) + h b_0 f(t_n, u_n) \quad \text{per } n = 0, 1, \dots, N_h - 1,$$

con u_0 assegnato. Se $a_0 = 1, b_{-1} = 0$ e $b_0 = 1$, otteniamo il metodo di *Eulero in avanti* (7.23); se $a_0 = 1, b_{-1} = 1$ e $b_0 = 0$, otteniamo il metodo di *Eulero all'indietro* (7.24). Infine, per $a_0 = 1, b_{-1} = \frac{1}{2}$ e $b_0 = \frac{1}{2}$, otteniamo il metodo di *Crank–Nicolson* (7.26).

Esempio 7.4.9. Consideriamo due metodi multipasso molto utilizzati.

- *AB3* indica il metodo di *Adam–Bashforth*, un metodo esplicito avente ordine di accuratezza 3 e a 3 passi ($p = 2$). Dall'Eq. (7.36), abbiamo:

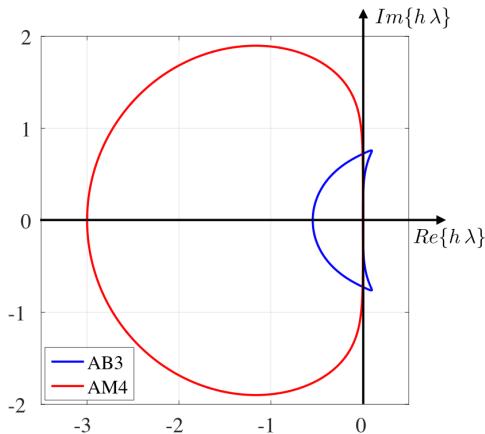
$$u_{n+1} = u_n + \frac{h}{12} [23 f(t_n, u_n) - 16 f(t_{n-1}, u_{n-1}) + 5 f(t_{n-2}, u_{n-2})] \quad \text{per } n = 2, \dots, N_h - 1;$$

i coefficienti sono dunque $a_0 = 1, a_1 = a_2 = 0, b_{-1} = 0, b_0 = \frac{23}{12}, b_1 = -\frac{16}{12}$ e $b_2 = \frac{5}{12}$.

- *AM4* indica il metodo di *Adam–Moulton*, un metodo implicito con ordine di accuratezza 4 e a 3 passi ($p = 2$). Dall'Eq. (7.36) otteniamo:

$$u_{n+1} = u_n + \frac{h}{24} [9 f(t_{n+1}, u_{n+1}) + 19 f(t_n, u_n) - 5 f(t_{n-1}, u_{n-1}) + f(t_{n-2}, u_{n-2})],$$

per $n = 2, \dots, N_h - 1$, con $a_0 = 1, a_1 = a_2 = 0, b_{-1} = \frac{9}{24}, b_0 = \frac{19}{24}, b_1 = -\frac{5}{24}$ e $b_2 = \frac{1}{24}$.



Regioni di assoluta stabilità \mathcal{A} per i metodi *AB3* e *AM4* riferiti al problema modello (7.2.1). Le regioni \mathcal{A} sono contenute all'interno delle linee che rappresentano i loro bordi.

7.5 Approssimazione Numerica di Sistemi di EDO del Primo Ordine

Consideriamo l'approssimazione di sistemi di EDO del primo ordine ($p = 1$), ovvero di problemi di Cauchy nella forma di (7.12). Procediamo richiamando il problema di Cauchy, per poi passare a considerare metodi numerici tra cui il θ -metodo.

7.5.1 Il problema di Cauchy nel caso vettoriale

Dato l'intervallo $I = (t_0, t_f) \subset \mathbb{R}$, il problema di Cauchy vettoriale è: trovare $\mathbf{y} : I \subset \mathbb{R} \rightarrow \mathbb{R}^m$ tale che

$$\begin{cases} \frac{d\mathbf{y}}{dt}(t) = \mathbf{f}(t, \mathbf{y}(t)) & \text{per } t \in I, \\ \mathbf{y}(t_0) = \mathbf{y}_0, \end{cases} \quad (7.37)$$

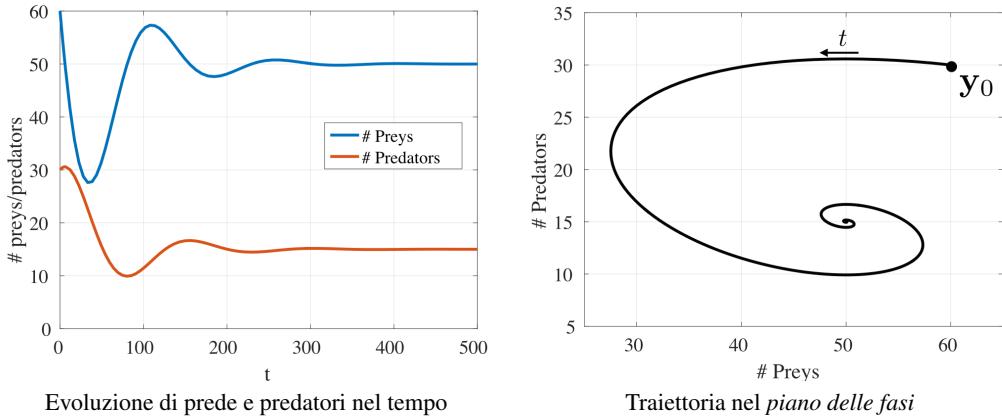
dove $m \geq 1$, $\mathbf{f}(t, \mathbf{y}) : I \times \mathbb{R}^m \rightarrow \mathbb{R}^m$ e \mathbf{y}_0 è il dato iniziale. Ricordiamo che $\mathbf{y}(t) = (y_1(t), \dots, y_m(t))^T$ e $\mathbf{f}(t, \mathbf{y}) = (f_1(t, \mathbf{y}), \dots, f_m(t, \mathbf{y}))^T$; dunque, se $m = 1$, ritorniamo al problema di Cauchy scalare affrontato in Sez. 7.4. Infine, assumiamo che $\mathbf{f}(t, \mathbf{y})$ sia continua in entrambi gli argomenti.

Definizione 7.5.1. Se $\mathbf{f}(t, \mathbf{y}) : I \times \mathbb{R}^m \rightarrow \mathbb{R}^m$ è nella forma:

$$\mathbf{f}(t, \mathbf{y}) = A\mathbf{y} + \mathbf{g}(t) \quad \text{con } A \in \mathbb{R}^{m \times m} \text{ e } \mathbf{g}(t) : I \rightarrow \mathbb{R}^m,$$

allora il sistema di EDO si dice in forma non omogenea a coefficienti costanti; è invece in forma omogenea se $\mathbf{g}(t) = \mathbf{0}$ per ogni $t \in (t_0, t_f]$.

Esempio 7.5.1. Modello preda-predatore di Lotka-Volterra. In riferimento a Eq. (7.37), prendiamo $\mathbf{f}(t, \mathbf{y}) = A\mathbf{y} + \mathbf{g}(t)$, ovvero un sistema di EDO in forma non omogenea a coefficienti costanti già introdotto in Eq. (7.4). Abbiamo: $m = 2$, $A = \begin{bmatrix} -b_1 C_1 & -d_2 C_1 \\ d_1 C_2 & b_2 C_2 \end{bmatrix}$ e $\mathbf{g} = (C_1, -C_2)^T$, con b_1, b_2, d_1, d_2, C_1 e $C_2 \in \mathbb{R}$. Il modello determina l'evoluzione in tempo t delle popolazioni di prede $y_1(t)$ e di predatori $y_2(t)$ in un ambiente chiuso. Abbiamo $b_1 = \frac{1}{100}$, $b_2 = 0$, $d_1 = \frac{1}{50}$, $d_2 = \frac{1}{30}$, $C_1 = 3$, $C_2 = 1$, $\mathbf{y}_0 = (60, 30)^T$, $t_0 = 0$ e $t_f = 500$.



7.5.2 Metodi numerici per sistemi di EDO del primo ordine

In analogia al caso del problema di Cauchy scalare ($m = 1$), consideriamo la partizione di $\bar{I} = [t_0, t_f]$ in N_h sottointervalli di ugual ampiezza $h = \frac{t_f - t_0}{N_h}$, da cui $t_n = t_0 + n h$ per $n = 0, 1, \dots, N_h$. Indichiamo con $\mathbf{y}_n = \mathbf{y}(t_n)$ la valutazione della soluzione esatta $\mathbf{y}(t)$ del problema di Cauchy (7.37) a $t = t_n$, mentre con \mathbf{u}_n la corrispondente approssimazione numerica a $t = t_n$. Osserviamo dunque che tutti i metodi numerici presentati in Sez. 7.4 possono essere estesi con minimi accorgimenti e applicati a sistemi di EDO del primo ordine per $m > 1$.

7.5.3 θ -metodo

Si tratta di una famiglia di metodi numerici molto utilizzata per sistemi di EDO del primo ordine. Introduciamo un parametro $\theta \in \mathbb{R}$ tale che $\theta \in [0, 1]$; allora il θ -metodo per l'approssimazione del problema di Cauchy (7.37) si scrive come: trovare $\{\mathbf{u}_n\}_{n=0}^{N_h}$ tale che

$$\left\{ \begin{array}{l} \mathbf{u}_{n+1} = \mathbf{u}_n + h [(1-\theta)\mathbf{f}(t_n, \mathbf{u}_n) + \theta\mathbf{f}(t_{n+1}, \mathbf{u}_{n+1})] \\ \mathbf{u}_0 = \mathbf{y}_0. \end{array} \right. \quad \text{per } n = 0, 1, \dots, N_h - 1, \quad (7.38)$$

Il θ -metodo è *esplicito* per $\theta = 0$, mentre è un metodo *implicito* per $\theta \in (0, 1]$. In questo ultimo caso è necessario risolvere un sistema di equazioni *non lineari* per ogni $n = 0, 1, \dots, N_h - 1$, ovvero:

$$\text{trovare } \mathbf{u}_{n+1} : \mathbf{F}_n^\theta(\mathbf{u}_{n+1}) = \mathbf{0} \quad \text{per ogni } n = 0, 1, \dots, N_h - 1,$$

con $\mathbf{u}_0 = \mathbf{y}_0$, dove:

$$\mathbf{F}_n^\theta(\mathbf{w}) := \mathbf{w} - \mathbf{u}_n - h [(1 - \theta) \mathbf{f}(t_n, \mathbf{u}_n) + \theta \mathbf{f}(t_{n+1}, \mathbf{w})].$$

La soluzione del problema non lineare può essere ottenuta utilizzando il metodo di Newton o delle iterazioni di punto fisso.

Osservazione 7.5.1. Se invece il sistema di EDO è in forma non omogenea a coefficienti costanti come specificato in Definizione 7.5.1, allora il θ -metodo si scrive come: trovare $\{\mathbf{u}_n\}_{n=0}^{N_h}$ tale che

$$\begin{cases} (I - h \theta A) \mathbf{u}_{n+1} = [I + h(1 - \theta) A] \mathbf{u}_n + h [(1 - \theta) \mathbf{g}(t_n) + \theta \mathbf{g}(t_{n+1})] & \text{per } n = 0, 1, \dots, N_h - 1, \\ \mathbf{u}_0 = \mathbf{y}_0. \end{cases}$$

In tal caso, il metodo comporta la soluzione di un sistema lineare con matrice $(I - h \theta A)$ per ogni $n = 0, 1, \dots, N_h - 1$, a meno che $\theta = 0$ (ovvero quando il metodo è esplicito).

Per $\theta = 0$, il θ -metodo coincide con il metodo di *Eulero in avanti* per sistemi di EDO: trovare $\{\mathbf{u}_n\}_{n=0}^{N_h}$ tale che

$$\begin{cases} \mathbf{u}_{n+1} = \mathbf{u}_n + h \mathbf{f}(t_n, \mathbf{u}_n) & \text{per } n = 0, 1, \dots, N_h - 1, \\ \mathbf{u}_0 = \mathbf{y}_0. \end{cases}$$

Per $\theta = 1$, il θ -metodo corrisponde al metodo di *Eulero all'indietro*: trovare $\{\mathbf{u}_n\}_{n=0}^{N_h}$ tale che

$$\begin{cases} \mathbf{u}_{n+1} = \mathbf{u}_n + h \mathbf{f}(t_{n+1}, \mathbf{u}_{n+1}) & \text{per } n = 0, 1, \dots, N_h - 1, \\ \mathbf{u}_0 = \mathbf{y}_0. \end{cases}$$

Infine per $\theta = \frac{1}{2}$ in Eq. (8.39), otteniamo il metodo di *Crank–Nicolson*: trovare $\{\mathbf{u}_n\}_{n=0}^{N_h}$ tale che

$$\begin{cases} \mathbf{u}_{n+1} = \mathbf{u}_n + \frac{h}{2} [\mathbf{f}(t_n, \mathbf{u}_n) + \mathbf{f}(t_{n+1}, \mathbf{u}_{n+1})] & \text{per } n = 0, 1, \dots, N_h - 1, \\ \mathbf{u}_0 = \mathbf{y}_0. \end{cases}$$

Definizione 7.5.2. L'errore associato all'approssimazione del problema di Cauchy (7.37) è:

$$E_h := \max_{n=0,1,\dots,N_h} e_n \quad \text{con } e_n := \|\mathbf{y}_n - \mathbf{u}_n\| \quad \text{per } n = 0, 1, \dots, N_h.$$

Se:

$$E_h \leq \tilde{E}_h := C h^p,$$

con C una costante positiva indipendente da h , allora il metodo ha ordine di convergenza $p > 0$ (ordine di accuratezza del metodo).

Se la soluzione del problema di Cauchy (7.37) è $\mathbf{y} \in C^2(I)$, allora i metodi di *Eulero in avanti* ($\theta = 0$) e *all'indietro* ($\theta = 1$), come in generale tutti i θ -metodi per $\theta \neq \frac{1}{2}$, convergono con *ordine* $p = 1$ in h . Se invece $\mathbf{y} \in C^3(I)$, il metodo di *Crank–Nicolson* ($\theta = \frac{1}{2}$) converge con *ordine* $p = 2$.

Osservazione 7.5.2. Anche se non si tratta di un θ -metodo, in accordo all'Eq. (8.39), scriviamo il metodo di Heun per sistemi di EDO applicando Eq. (7.27) a Eq. (7.37), ottenendo: trovare $\{\mathbf{u}_n\}_{n=0}^{N_h}$ tale che

$$\begin{cases} \mathbf{u}_{n+1}^* = \mathbf{u}_n + h \mathbf{f}(t_n, \mathbf{u}_n), \\ \mathbf{u}_{n+1} = \mathbf{u}_n + \frac{h}{2} [\mathbf{f}(t_n, \mathbf{u}_n) + \mathbf{f}(t_{n+1}, \mathbf{u}_{n+1}^*)] & \text{per } n = 0, 1, \dots, N_h - 1, \\ \mathbf{u}_0 = \mathbf{y}_0. \end{cases}$$

Consideriamo il problema di Cauchy con $\mathbf{f}(t, \mathbf{y}) = A\mathbf{y}$ (ovvero dalla Definizione 7.5.1 $\mathbf{g}(t) = \mathbf{0}$). Indicati con $\{\lambda_i\}_{i=1}^m \in \mathbb{C}$ gli autovalori di A e con $\{\mathbf{v}_i\}_{i=1}^m \in \mathbb{C}^m$ i corrispondenti autovettori, allora la soluzione del problema di Cauchy assume la seguente forma:

$$\mathbf{y}(t) = \sum_{i=1}^m C_i e^{\lambda_i(t-t_0)} \mathbf{v}_i \quad \text{per } t \geq t_0,$$

dove $\{C_i\}_{i=1}^m \in \mathbb{C}$ sono opportune costanti tali che $\mathbf{y}(t_0) = \mathbf{y}_0$. Se $\operatorname{Re}\{\lambda_i(A)\} < 0$ per ogni $i = 1, \dots, m$, allora $\lim_{t \rightarrow +\infty} \mathbf{y}(t) = \mathbf{0}$. È dunque lecito domandarsi quando un metodo numerico per l'approssimazione di tale problema di Cauchy è *assolutamente stabile*, ovvero se $\lim_{n \rightarrow +\infty} \mathbf{u}_n = \mathbf{0}$ e sotto quali condizioni su h . In particolare, un metodo numerico per approssimare il problema di Cauchy con $f(t, \mathbf{y}) = A\mathbf{y}$ è assolutamente stabile se $(h\lambda_i) \in \mathcal{A}$ per ogni $i = 1, \dots, m$ in accordo con quanto visto nella Sez. 7.4.6 e nella Definizione 7.4.8.

Esempio 7.5.2. Consideriamo il problema di Cauchy (7.37) con $\mathbf{f}(t, \mathbf{y}) = A\mathbf{y}$, dove $A = \begin{bmatrix} -3 & 3 \\ -3 & -3 \end{bmatrix}$, e $\mathbf{y}_0 = \mathbf{1}$. La matrice A ha autovalori $\lambda_{1,2} = -3 \pm 3i$. Il metodo di Eulero in avanti è pertanto assolutamente stabile solo se $0 < h < \frac{1}{3}$; in tal caso abbiamo infatti $(h\lambda_1) \in \mathcal{A}$, essendo $\mathcal{A} = \{z \in \mathbb{C} : |1+z| < 1\}$. Il metodo di Crank-Nicolson per esempio è invece assolutamente stabile per ogni $h > 0$; infatti, $(h\lambda_1) \in \mathcal{A}$ per ogni $h > 0$, essendo $\mathcal{A} = \left\{ z \in \mathbb{C} : \left| \frac{1+z/2}{1-z/2} \right| < 1 \right\}$.

Esempio 7.5.3. Per il problema di Cauchy (7.37) con $\mathbf{f}(t, \mathbf{y}) = A\mathbf{y}$, dove $A = \begin{bmatrix} -2 & -1 & -1 \\ 0 & -4 & 11 \\ 0 & 0 & -5 \end{bmatrix}$, e $\mathbf{y}_0 = \mathbf{1}$. Osserviamo che gli autovalori di A sono $\lambda_1 = -5$, $\lambda_2 = -4$ e $\lambda_3 = -2$. Pertanto, i metodi di Eulero in avanti e Heun sono assolutamente stabili per $0 < h < \frac{2}{|\lambda_1|} = \frac{2}{5}$, mentre i metodi di Euler all'indietro e Crank-Nicolson sono assolutamente stabili per ogni $h > 0$.

Per un problema di Cauchy generale tale per cui $\operatorname{Re}\{\lambda_i(J(t))\} < 0$ per ogni $i = 1, \dots, m$ e per ogni $t \geq t_0$, dove $J(t) = \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(t, \mathbf{y}(t))$, ha senso parlare di assoluta stabilità di un metodo numerico; questa in particolare si verifica se $(h\lambda_i(J(t))) \in \mathcal{A}$ per ogni $i = 1, \dots, m$ e per ogni $t \geq t_0$, essendo \mathcal{A} la regione di assoluta stabilità del metodo.

Esempio 7.5.4. Per il problema di Cauchy (7.37) con $m = 2$, poniamo $\mathbf{f}(t, \mathbf{y}) = (-(1+2e^{-t})y_1 - y_2, -y_2)^T$, $\mathbf{y}_0 = \mathbf{1}$, e $t_0 = 0$. Otteniamo dunque la matrice Jacobiana $J(t) = \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(t, \mathbf{y}(t)) = \begin{bmatrix} -(1+2e^{-t}) & -1 \\ 0 & -1 \end{bmatrix}$ i cui autovalori sono $\lambda_1(t) = -(1+2e^{-t})$ e $\lambda_2(t) = -1$ per ogni $t \geq 0$, entrambi reali. I metodi di Eulero in avanti e Heun sono pertanto assolutamente stabili se $0 < h < \frac{2}{\sup_{t \geq 0} |\lambda_1(t)|} = \frac{2}{3}$, mentre i metodi di Euler all'indietro e Crank-Nicolson sono assolutamente stabili per ogni $h > 0$.

7.6 Approssimazione Numerica di EDO del Secondo Ordine

Abbiamo considerato finora l'approssimazione numerica di EDO del primo ordine o di sistemi di EDO del primo ordine, vale a dire i problemi di Cauchy (7.22) e (7.37), rispettivamente. EDO (e sistemi di EDO) di *ordine maggiore di 1*, che sono spesso usati come modelli matematici per problemi di interesse fisico, possono essere riscritti come sistemi di ordine superiore. Pertanto, la soluzione numerica di EDO di alto ordine può essere ottenuta semplicemente riscrivendo tali problemi come sistemi di EDO di primo ordine

(a patto di aumentare il numero delle funzioni incognite), che vengono poi approssimati per mezzo dei metodi numerici considerati per questa classe di problemi, come ad esempio il θ -metodo introdotto nella Sez. 7.5.3. Un caso particolare è quello delle EDO di ordine 2, per le quali esistono anche metodi numerici *ad hoc*, a cui accenniamo brevemente. Tra questi vi sono il metodo *Leap Frog* e il metodo di *Newmark*.

Consideriamo l'intervallo $I = (t_0, t_f) \subset \mathbb{R}$; il problema di Cauchy per una EDO del *secondo ordine* risulta formulato come segue:

$$\text{trovare } y : I \subset \mathbb{R} \rightarrow \mathbb{R} \quad : \quad \begin{cases} \frac{d^2y}{dt^2}(t) = f\left(t, y(t), \frac{dy}{dt}(t)\right) & \text{per ogni } t \in I, \\ \frac{dy}{dt}(t_0) = y_0, \\ y(t_0) = w_0, \end{cases} \quad (7.39)$$

dove $f(t, y, w_2) : I \times \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ è una data funzione di tre argomenti, mentre y_0 e w_0 rappresentano i dati iniziali.

7.6.1 Riscrittura dell'EDO del secondo ordine come sistema di EDO del primo ordine

Possiamo riscrivere una EDO del secondo ordine come un equivalente sistema di due EDO di ordine 1. A tale scopo, introduciamo una variabile ausiliaria $w_2(t) : I \rightarrow \mathbb{R}$ tale che $w_2(t) = \frac{dy}{dt}(t)$ per ogni $t \in I$. Quindi, il problema (7.39) può essere riscritto come il seguente sistema di EDO (avendo espresso i dati iniziali y_0 e w_0 come $y_{0,0}$ e $y_{1,0}$, rispettivamente):

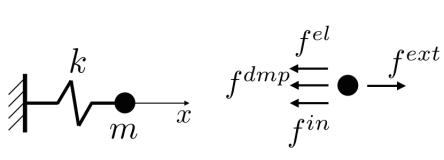
$$\text{trovare } y, w_2 : I \subset \mathbb{R} \rightarrow \mathbb{R} \quad : \quad \begin{cases} \frac{dw_2}{dt}(t) = f(t, y(t), w_2(t)) & \text{per ogni } t \in I, \\ \frac{dy}{dt}(t) = w_2(t) & \text{per ogni } t \in I, \\ w_2(t_0) = y_{1,0}, \\ y(t_0) = y_{0,0}, \end{cases}$$

e, in generale, come il problema di Cauchy (7.37) per $m = 2$, con

$$\mathbf{y}(t) = \begin{bmatrix} w_2(t) \\ y(t) \end{bmatrix}, \quad \mathbf{f}(t, \mathbf{y}) = \begin{bmatrix} f(t, y, w_2) \\ w_2 \end{bmatrix}, \quad \mathbf{y}_0 = \begin{bmatrix} y_{1,0} \\ y_{0,0} \end{bmatrix};$$

osserviamo che $\mathbf{y} : I \rightarrow \mathbb{R}^2$ e che $\mathbf{f}(t, \mathbf{y}) : I \times \mathbb{R}^2 \rightarrow \mathbb{R}^2$. Una volta riscritto in questa forma, il sistema di EDO del primo ordine può essere risolto con uno dei metodi indicati in Sez. 7.5.

Esempio 7.6.1. Consideriamo ad esempio la dinamica di una massa concentrata (m) vincolata a una molla.



Consideriamo la forza esterna $f^{ext}(t)$, la forza inerziale $f^{in}(t) = m \ddot{x}(t)$, essendo m la massa, la forza elastica $f^{el}(t) = k x(t)$, essendo k la costante elastica della molla e la forza di smorzamento $f^{dmp}(t) = c \dot{x}(t)$, con $c \geq 0$; $x(t)$ rappresenta la posizione della massa concentrata nel tempo t .

Il problema corrisponde alla seguente EDO di ordine 2 sull'intervallo di tempo $I = (t_0, t_f)$:

$$\text{trovare } x : I \subset \mathbb{R} \rightarrow \mathbb{R} \quad : \quad \begin{cases} m \ddot{x}(t) + c \dot{x}(t) + k x(t) = f^{ext}(t) & \text{per ogni } t \in I, \\ \dot{x}(t_0) = v_0, \\ x(t_0) = x_0, \end{cases}$$

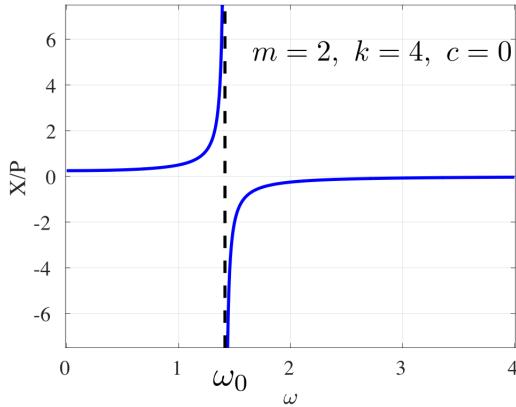
dove v_0 è la velocità iniziale della massa, mentre x_0 è la sua posizione iniziale. Questa EDO di ordine 2 può essere riscritta sotto forma di sistema di 2 EDO del primo ordine, ovvero sotto forma di problema di Cauchy (7.37) con due variabili, introducendo una variabile ausiliaria $v : I \rightarrow \mathbb{R}$ che in questo caso gioca il ruolo di velocità. Si ha cioè che

$$\mathbf{y}(t) = \begin{Bmatrix} v(t) \\ x(t) \end{Bmatrix}, \quad \mathbf{f}(t, \mathbf{y}) = \begin{bmatrix} -\frac{c}{m} v - \frac{k}{m} x + \frac{1}{m} f^{ext}(t) \\ v \end{bmatrix}, \quad \mathbf{y}_0 = \begin{bmatrix} v_0 \\ x_0 \end{bmatrix}.$$

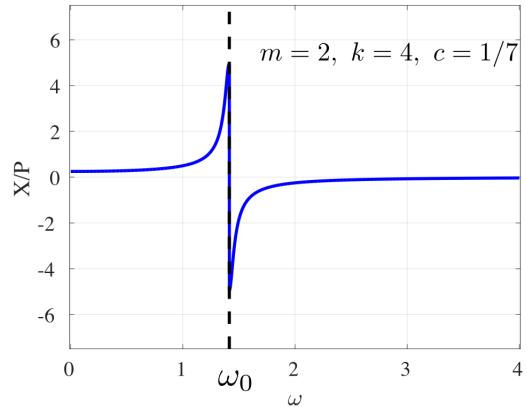
Osserviamo che questo sistema di EDO è in forma non omogenea, a coefficienti costanti, in base alla Definizione 7.5.1, ovvero si ha $\mathbf{f}(t, \mathbf{y}) = A\mathbf{y} + \mathbf{g}(t)$, con

$$A = \begin{bmatrix} -\frac{c}{m} & -\frac{k}{m} \\ 1 & 0 \end{bmatrix}, \quad \mathbf{g}(t) = \begin{bmatrix} \frac{1}{m} f^{ext}(t) \\ 0 \end{bmatrix}.$$

Assumiamo ora che $f^{ext}(t) = P \sin(\omega t)$ per qualche $P > 0$ e $\omega > 0$. Risulta allora possibile mostrare che, dopo un transitorio iniziale, la soluzione $x(t)$ assume la forma $\tilde{x}(t) = X \sin(\omega t - \phi)$, dove $\tan(\phi) = \frac{c\omega}{k - \omega^2 m}$ e $X = \frac{P}{(k - \omega^2 m) \cos(\phi) + c\omega \sin(\phi)}$. Se $c = 0$, abbiamo che $\phi = 0$ e $\tilde{x}(t) = X \sin(\omega t)$, con $X = \frac{P}{(k - \omega^2 m)}$; $\omega_0 = \sqrt{\frac{k}{m}}$ è chiamata *frequenza angolare naturale* del sistema, mentre $f_0 = \frac{\omega_0}{2\pi}$ è la frequenza naturale. Se $\omega = \omega_0$ (o sufficientemente prossimo a ω_0), il sistema si trova in una condizione di *risonanza*.



rapporto $\frac{X}{P}$ vs. ω per $m = 2, k = 4$, e $c = 0$.



rapporto $\frac{X}{P}$ vs. ω per $m = 2, k = 4$, e $c = \frac{1}{7}$.

7.6.2 Metodo Leap Frog (non svolto a lezione)

Il metodo Leap Frog è un metodo ad hoc per la soluzione di una EDO del secondo ordine; è basato sull'approssimazione della derivata prima e della derivata seconda di $y(t)$ con schemi alle differenze finite centrali. La differenza finita centrale per approssimare $y''_n = y''(t_n)$ è data da

$$y''_n \approx \frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} \quad \text{per } n = 1, \dots, N_h - 1$$

mentre la differenza finita centrale per approssimare $y'_n = y'(t_n)$ risulta

$$y'_n \approx \frac{y_{n+1} - y_{n-1}}{2h} \quad \text{per } n = 1, \dots, N_h - 1;$$

entrambi questi schemi hanno ordine di accuratezza pari a 2. Indicando con u_n l'approssimazione di y_n e con v_n l'approssimazione di y'_n , il metodo Leap Frog si ottiene collocando l'equazione differenziale negli istanti $\{t_n\}_{n=0}^{N_h}$ dati dalla partizione dell'intervallo $I = (t_0, t_f)$ in N_h sottointervalli di lunghezza h :

$$\begin{cases} \frac{u_{n+1} - 2u_n + u_{n-1}}{h^2} = f(t_n, u_n, v_n) \\ \frac{u_{n+1} - u_{n-1}}{2h} = v_n \end{cases} \quad \text{per } n = 1, \dots, N_h - 1. \quad (7.40)$$

Questa versione del metodo non risulta tuttavia fruibile, in quanto vi compare ancora la dipendenza delle soluzioni u_{n+1} e v_n da u_{n-1} . Osservando però che dalla seconda equazione è possibile esprimere

$$u_{n-1} = u_{n+1} - 2hv_n,$$

si deduce che

$$\begin{aligned} u_{n+1} &= 2u_n - u_{n-1} + h^2 f(t_n, u_n, v_n) = 2u_n - (u_{n+1} - 2hv_n) + h^2 f(t_n, u_n, v_n) \\ &= -u_{n+1} + 2hv_n + 2u_n + h^2 f(t_n, u_n, v_n), \end{aligned}$$

da cui, dividendo per 2 dopo aver portato a sinistra $-u_{n+1}$, otteniamo:

$$u_{n+1} = u_n + hv_n + \frac{h^2}{2} f(t_n, u_n, v_n);$$

possiamo dunque calcolare u_{n+1} solo a partire dai valori precedentemente calcolati di u_n e v_n .

Possiamo ora determinare v_{n+1} riscrivendo la seconda delle equazioni che compaiono in (7.42) e sostituendo l'espressione per u_{n+1} appena ottenuta, riscritta per u_{n+2} :

$$\begin{aligned} v_{n+1} &= \frac{u_{n+2} - u_n}{2h} = \frac{u_{n+1} + hv_{n+1} + \frac{h^2}{2} f(t_{n+1}, u_{n+1}, v_{n+1})}{2h} \\ &= \frac{\left(u_n + hv_n + \frac{h^2}{2} f(t_n, u_n, v_n) \right) + hv_{n+1} + \frac{h^2}{2} f(t_{n+1}, u_{n+1}, v_{n+1})}{2h}, \end{aligned}$$

da cui si deduce, portando v_{n+1} a sinistra e moltiplicando tutto per 2, che

$$v_{n+1} = v_n + \frac{h}{2} [f(t_n, u_n, v_n) + f(t_{n+1}, u_{n+1}, v_{n+1})].$$

Il metodo di *Leap Frog* risulta pertanto: trovare $\{u_n\}_{n=0}^{N_h}$ e $\{v_n\}_{n=0}^{N_h}$ tali che

$$\begin{cases} u_{n+1} = u_n + hv_n + \frac{h^2}{2} f(t_n, u_n, v_n) \\ v_{n+1} = v_n + \frac{h}{2} [f(t_n, u_n, v_n) + f(t_{n+1}, u_{n+1}, v_{n+1})] & \text{per } n = 1, \dots, N_h - 1, \\ u_0 = y_0, \\ v_0 = w_0, \end{cases} \quad (7.41)$$

Il metodo *Leap Frog* è un metodo a un passo, *Implicito*, dal momento che la seconda equazione dipende, in linea di principio, non linearmente da v_{n+1} . Esso fornisce l'approssimazione numerica non solo della soluzione $y(t_n)$, per $n = 1, \dots, N_h$, ma anche della sua derivata $y'(t_n)$, per $n = 1, \dots, N_h$.

Il suo ordine di accuratezza è pari a 2 in h , e risulta condizionatamente assolutamente stabile quando applicato al problema modello, che per una EDO del secondo ordine corrisponde al sistema massa-molla-smorzatore dell'Esempio 7.6.1.

7.6.3 Metodo di Newmark (non svolto a lezione)

Si tratta di una famiglia di metodi a due parametri θ e $\xi \geq 0$ che generalizza il metodo *Leap Frog* e consiste in: trovare $\{u_n\}_{n=0}^{N_h}$ e $\{v_n\}_{n=0}^{N_h}$ tali che

$$\begin{cases} u_{n+1} = u_n + hv_n + \frac{h^2}{2} [(1 - 2\xi) f(t_n, u_n, v_n) + 2\xi f(t_{n+1}, u_{n+1}, v_{n+1})] \\ v_{n+1} = v_n + \frac{h}{2} [(1 - \theta) f(t_n, u_n, v_n) + \theta f(t_{n+1}, u_{n+1}, v_{n+1})] & \text{per } n = 1, \dots, N_h - 1, \\ u_0 = y_0, \\ v_0 = w_0. \end{cases} \quad (7.42)$$

Le proprietà del metodo (esplicito/Implicito, ordine di convergenza, stabilità assoluta) dipendono dai parametri θ e ξ ; in particolare, tali parametri permettono di controllare le oscillazioni numeriche e rimuovere quelle non fisiche dalla soluzione numerica.

7.7 Approssimazione Numerica di EDO di Ordine Superiore a Due

Come anticipato, tali EDO di ordine $p \geq 2$ possono essere ricondotte a sistemi di EDO del primo ordine, per la cui approssimazione possono essere utilizzati i metodi numerici già indicati.

Consideriamo nuovamente l'intervallo $I = (t_0, t_f) \subset \mathbb{R}$; una EDO di ordine $p \geq 2$ assume la seguente forma:

$$\boxed{\text{trovare } y : I \subset \mathbb{R} \rightarrow \mathbb{R} : \begin{cases} \frac{d^p y}{dt^p}(t) = f\left(t, y(t), \frac{dy}{dt}(t), \dots, \frac{d^{p-1}y}{dt^{p-1}}(t)\right) & \text{per ogni } t \in I, \\ \frac{d^{p-1}y}{dt^{p-1}}(t_0) = y_{p-1,0}, \\ \vdots \\ y(t_0) = y_{0,0}, \end{cases}} \quad (7.43)$$

dove $f(t, y, w_2, \dots, w_p) : I \times \mathbb{R} \times \mathbb{R} \times \dots \times \mathbb{R} \rightarrow \mathbb{R}$ è una data funzione di $p+1$ argomenti, essendo $\{y_{k,0}\}_{k=0}^{p-1}$ i dati iniziali. Introducendo ora le variabili ausiliarie $w_k(t) : I \rightarrow \mathbb{R}$ tali che $w_k(t) = \frac{d^{k-1}y}{dt^{k-1}}(t)$ per ogni $t \in I$ e per ogni $k = 2, \dots, p$, il problema (7.43) può essere riscritto come il seguente sistema di EDO:

$$\boxed{\text{trovare } y, w_2, \dots, w_p : I \subset \mathbb{R} \rightarrow \mathbb{R} : \begin{cases} \frac{dw_p}{dt}(t) = f(t, y(t), w_2(t), \dots, w_p(t)) & \text{per ogni } t \in I, \\ \frac{dw_{p-1}}{dt}(t) = w_p(t) & \text{per ogni } t \in I, \\ \vdots \\ \frac{dy}{dt}(t) = w_2(t) & \text{per ogni } t \in I, \\ w_p(t_0) = y_{p-1,0}, \\ \vdots \\ w_2(t_0) = y_{1,0}, \\ y(t_0) = y_{0,0}. \end{cases}}$$

Ancora una volta, il sistema precedente può essere ricondotto alla forma generale del problema di Cauchy (7.37) indicando con

$$\mathbf{y}(t) = \begin{bmatrix} w_p(t) \\ \vdots \\ w_2(t) \\ y(t) \end{bmatrix}, \quad \mathbf{f}(t, \mathbf{y}) = \begin{bmatrix} f(t, y, w_2, \dots, w_p) \\ w_p \\ \vdots \\ w_2 \end{bmatrix}, \quad \mathbf{y}_0 = \begin{bmatrix} y_{p-1,0} \\ \vdots \\ y_{0,0} \end{bmatrix};$$

osserviamo che $\mathbf{y} : I \rightarrow \mathbb{R}^m$ e $\mathbf{f}(t, \mathbf{y}) : I \times \mathbb{R}^m \rightarrow \mathbb{R}^m$, dove $m = p$.

Osservazione 7.7.1. *I sistemi di EDO di alto ordine possono essere riscritti sotto forma di sistemi di EDO del primo ordine, seguendo la procedura sopra descritta. Se n è la dimensione del sistema di EDO di ordine p , allora il sistema corrispondente di EDO del primo ordine (7.37) ha dimensione $m = pn$.*

Capitolo 8

Problemi ai Limiti e ai Valori Iniziali

In questo capitolo consideriamo la soluzione numerica di semplici problemi ai limiti e ai valori iniziali, mediante il metodo delle differenze finite. Sebbene non si tratti di un metodo ampiamente utilizzato in tempi recenti, e di difficile estensione a casi di natura applicata viste le rigide assunzioni sotto le quali viene definita, il metodo delle differenze finite può essere costruito in maniera elementare, e risulta utile per comprendere buona parte delle questioni che sorgono nel momento in cui un problema alle derivate parziali viene approssimato numericamente. In particolare, vedremo come dall'approssimazione numerica di un problema ai limiti (o ai valori al contorno) per un'equazione a derivate parziali stazionaria lineare si giunga a un sistema lineare da risolvere; analogamente, a partire da un problema ai limiti e ai valori iniziali per un'equazione a derivate parziali tempo-dipendente lineare si giunge a un sistema di equazioni differenziali ordinarie lineari da risolvere. Ancora una volta, gli strumenti numerici introdotti nei capitoli precedenti risultano utili anche (e soprattutto) in vista della soluzione di problemi ancora più avanzati da un punto di vista matematico e utili da un punto di vista ingegneristico.

8.1 Definizioni ed Esempi

Se in un'*equazione differenziale ordinaria* tutte le derivate sono prese rispetto a una singola variabile indipendente, nel momento in cui siano presenti derivate parziali – ovvero, derivate della funzione incognita u rispetto a più variabili indipendenti (spaziali e/o temporali) – si parla di *equazione a derivate parziali* (EDP). Date dunque $d + 1$ variabili indipendenti, sotto forma di d variabili spaziali $\mathbf{x} = (x_1, \dots, x_d)$ (con $d = 1, 2, 3$ a seconda del problema in esame) e della variabile temporale t , una EDP è un legame tra le derivate della funzione incognita:

$$\mathcal{P}(u, g) = \mathcal{F}\left(\mathbf{x}, t, u, \frac{\partial u}{\partial t}, \frac{\partial u}{\partial x_1}, \dots, \frac{\partial u}{\partial x_d}, \dots, \frac{\partial^{p_1+\dots+p_d+p_t}}{\partial x_1^{p_1} \dots \partial x_d^{p_d} \partial t^{p_t}} u, g\right) = 0,$$

valido in un dominio $\Omega \subset \mathbb{R}^d$, per $t \in (0, T)$, dove $p_1, \dots, p_d, p_t \in \mathbb{N}$ e il massimo ordine di derivazione che compare nell'equazione, $q = p_1 + \dots + p_d + p_t$, è detto *ordine* dell'equazione, e $T > 0$ è il tempo finale. In particolare, distinguiamo tra:

- *problemi ai limiti (o ai valori al contorno)*: si tratta di EDP stazionarie (o indipendenti dal tempo t) per cui il valore di u (o delle sue derivate) sul bordo di Ω , indicato con $\partial\Omega$, è assegnato;
- *problemi ai limiti e ai valori iniziali (o ai valori iniziali e al contorno)*: si tratta di EDP non più indipendenti dal tempo t , valide per $\mathbf{x} \in \Omega$ e $t \in (0, T)$ in cui compaiano anche derivate parziali rispetto a t oltre che rispetto alle variabili spaziali. In questo caso, oltre ai valori di u (o delle sue derivate) sul bordo $\partial\Omega$ di Ω , per ogni $t \in (0, T)$, occorre specificare anche una condizione iniziale (prescrivendo il valore di u) in ogni $\mathbf{x} \in \Omega$, all'istante iniziale $t = 0$.

Esempio 8.1.1. Il problema di Poisson in dimensione $d = 1$

$$\begin{cases} -u''(x) = f(x) & x \in (a, b), \\ +\text{valori al bordo} & \text{in } x = a \text{ e } x = b \end{cases} \quad (8.1)$$

dove $u''(x) = \frac{d^2u}{dx^2}(x)$, è un problema ai limiti definito su un intervallo $(a, b) \subset \mathbb{R}$ nel quale il valore della soluzione sconosciuta (o della sua derivata) viene prescritto nei punti finali a e b dell'intervallo. Nel caso $d = 1$, in cui si ha una sola variabile indipendente x e compaiano nell'equazione solo derivate dell'incognita rispetto alla sola variabile indipendente x , si considera tale problema nella classe delle EDP per la natura delle condizioni imposte sulla sua soluzione (che risultano di natura *globale*, riguardando non più soltanto un singolo punto del dominio, e non locale).

Questa equazione modella un fenomeno stazionario (non compare infatti la variabile temporale) e rappresenta il più semplice modello di diffusione, come la diffusione di un inquinante lungo un canale monodimensionale (a, b) o lo spostamento verticale di un filo elastico (detto anche *linea elastica*) fissato ai suoi estremi. Nel primo caso $f = f(x)$ indica la sorgente dell'inquinante lungo il flusso, mentre nel secondo caso f è la forza trasversale che agisce sul filo elastico, nell'ipotesi di massa trascurabile e piccoli spostamenti.

Esempio 8.1.2. Il problema di Poisson in dimensione $d > 1$ L'estensione del problema (8.1) in $d > 1$ dimensioni è data da

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ +\text{valori al bordo} & \text{su } \partial\Omega, \end{cases} \quad (8.2)$$

dove

$$\Delta u = \sum_{i=1}^d \frac{\partial^2 u}{\partial x_i^2}$$

è l'*operatore di Laplace (o Laplaciano)* e $\Omega \subset \mathbb{R}^d$ è un dominio il cui contorno è indicato con $\partial\Omega$. In questo caso, nell'equazione compaiono (sommate tra loro) le derivate parziali seconde della funzione incognita rispetto alle coordinate spaziali x_1, \dots, x_d .

Tale equazione permette di modellare, ad esempio, la diffusione di particelle in un fluido (in assenza di termini convettivi), oppure lo spostamento verticale di una membrana elastica; risulta infine un problema di assoluto rilievo anche in elettrostatica.

Esempio 8.1.3. Un esempio notevole di problema ai limiti e ai valori iniziali è costituito dall'equazione del calore, che per semplicità consideriamo in dimensione (spaziale) $d = 1$; in questo caso, la soluzione $u = u(x, t)$ soddisfa

$$\begin{cases} \frac{\partial u}{\partial t} - \mu \frac{\partial^2 u}{\partial x^2} = f & x \in (a, b), t > 0, \\ +\text{valori al bordo} & \text{in } x = a \text{ e } x = b, \text{ per } t > 0 \\ +\text{condizione iniziale su } u & \text{in } [a, b], \text{ per } t = 0 \end{cases} \quad (8.3)$$

In questo caso, $u(x, t)$ descrive la temperatura nel punto x e all'istante t di una sbarra di metallo monodimensionale che occupa l'intervallo $\Omega = (a, b)$. Il coefficiente di diffusione μ rappresenta la risposta termica del materiale, ovvero $\mu = \kappa/\rho c_p$, dove $\kappa > 0$ è la conducibilità termica, ρ è la densità e c_p la capacità termica per unità di massa.

Se ad esempio imponiamo che $u(a, t) = u(b, t) = 0$, per ogni $t > 0$, le estremità della sbarra sono mantenute ad una temperatura di riferimento (zero gradi in questo caso). Con la condizione iniziale $u(x, 0) = u_0(x)$ per ogni $x \in [a, b]$ assegniamo invece la temperatura in ciascun punto $x \in [a, b]$

Esempio 8.1.4. Un ulteriore esempio di problema ai limiti e ai valori iniziali è fornito dall'equazione delle onde, che descrive la propagazione di un'onda in un mezzo omogeneo; se c è la velocità di propagazione (costante, nel caso di un mezzo omogeneo) che dipende dalle caratteristiche meccaniche del mezzo, l'ampiezza della perturbazione $u = u(x, t)$ soddisfa

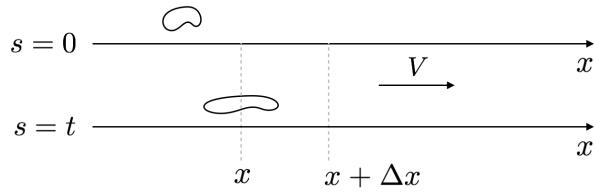
$$\begin{cases} \frac{\partial^2 u}{\partial t^2} - c \frac{\partial^2 u}{\partial x^2} = f & x \in (a, b), t > 0, \\ +\text{valori al bordo} & \text{in } x = a \text{ e } x = b, \text{ per } t > 0 \\ +\text{condizione iniziale su } u \text{ e } \partial u / \partial x & \text{in } [a, b], \text{ per } t = 0. \end{cases} \quad (8.4)$$

il primo problema in cui è stata derivata è stato quello della corda vibrante di uno strumento musicale, studiato tra gli altri da Eulero, Daniel Bernoulli e Joseph-Louis Lagrange. Versioni più generali di questo modello vengono utilizzate in acustica, elettromagnetismo e fluidodinamica per descrivere, ad esempio, la propagazione di onde sonore ed elettromagnetiche.

8.1.1 Da dove viene un modello? Leggi fisiche e leggi costitutive

In un caso di interesse, descriviamo sinteticamente quali siano gli ingredienti fondamentali di un modello matematico, mostrando come un modello basato su EDP possa originare dalla combinazione di principi primi basati sulla fisica (in primo luogo le leggi di conservazione) e da leggi costitutive che permettono di caratterizzare ciascuna situazione in esame.

Consideriamo per semplicità il caso della diffusione e del trasporto di una sostanza inquinante lungo un canale, di cui trascuriamo la profondità – supponendo che l'inquinante galleggi – e la dimensione trasversale – immaginando che il canale sia piuttosto stretto in rapporto alla sua lunghezza. Supponiamo inoltre che una corrente d'acqua con velocità costante $V > 0$ trasporti l'inquinante lungo la direzione positiva dell'asse x .



Vogliamo determinare un modello matematico per descrivere l'evoluzione della concentrazione $u = u(x, t)$ ($[massa] \times [lunghezza]^{-1}$) di inquinante in un punto x lungo il canale e al tempo t . L'integrale

$$\int_x^{x+\Delta x} u(y, t) dy \quad (8.5)$$

rappresenta la massa presente nell'intervallo $[x, x + \Delta x]$ al tempo t ; supponiamo, per il momento, che non vi siano sorgenti (o pozzi) di inquinante. Sfruttiamo la legge di conservazione della massa per definire il modello matematico, in base alla quale il tasso di variazione della massa nell'intervallo $[x, x + \Delta x]$ è pari al flusso netto di inquinante attraverso gli estremi di tale intervallo. Dall'Eq. (8.5), il tasso di variazione della massa contenuta in $[x, x + \Delta x]$ è dato da

$$\frac{d}{dt} \int_x^{x+\Delta x} u(y, t) dy = \int_x^{x+\Delta x} \frac{\partial u}{\partial t}(y, t) dy, \quad (8.6)$$

assumendo di poter scambiare le operazioni di derivata e integrale. Indicando con $q = q(x, t)$ ($[massa] \times [tempo]^{-1}$) il flusso di massa *entrante* nell'intervallo $[x, x + \Delta x]$, attraverso il punto x al tempo t e con $q = q(x + \Delta x, t)$ il flusso uscente attraverso il punto $x + \Delta x$, abbiamo che il flusso di massa netto attraverso gli estremi dell'intervallo $[x, x + \Delta x]$ è dato da

$$q(x, t) - q(x + \Delta x, t). \quad (8.7)$$

Imponendo che (8.6) e (8.7) siano uguali, la legge di conservazione della massa diviene

$$\int_x^{x+\Delta x} \frac{\partial u}{\partial t}(y, t) dy = q(x, t) - q(x + \Delta x, t).$$

Dividendo per Δx e prendendo il limite¹ per $\Delta x \rightarrow 0$ troviamo la seguente legge di bilancio,

$$\frac{\partial u}{\partial t} = -\frac{\partial q}{\partial x}. \quad (8.8)$$

Arrivati a questo punto, occorre decidere quale flusso di massa modelli il fenomeno in esame, ovvero necessitiamo di una *legge costitutiva* per q . Vi sono a tal proposito varie possibilità, ad esempio possiamo assumere che:

¹Sfruttiamo il teorema della media integrale, secondo cui $\int_x^{x+\Delta x} \frac{\partial u}{\partial t}(y, t) dy = \frac{\partial u}{\partial t}(\bar{x}, t)$ per un opportuno $\bar{x} \in [x, x + \Delta x]$; prendendo poi il limite per $\Delta x \rightarrow 0$ si ha che $\bar{x} \rightarrow x$, e quindi $\int_x^{x+\Delta x} \frac{\partial u}{\partial t}(y, t) dy \rightarrow \frac{\partial u}{\partial t}(x, t)$.

1. *Diffusione.* L'inquinante si espande da regioni a concentrazione più elevata a quelle a concentrazione più bassa. Si tratta di ciò che accade nell'ambito della trasmissione del calore; secondo la legge di Fourier, il flusso di calore risulta infatti proporzionale e opposto alla differenza di temperatura. Nel caso di una sostanza inquinante, possiamo adottare una legge simile (nota come *legge di Fick*), secondo la quale

$$q(x, t) = -\mu \frac{\partial u}{\partial x}(x, t),$$

dove la costante μ dipende dall'inquinante, rappresenta una diffusività e ha per dimensioni $[\mu] = [\text{lunghezza}]^2 \times [\text{tempo}]^{-1}$.

2. *Trasporto.* Il flusso di inquinante è determinato in questo caso solo dalla sola corrente d'acqua, ovvero un agglomerato di inquinante viene trasportato ma non si espande né si deforma. In termini quantitativi,

$$q(x, t) = Vu(x, t),$$

dove V indica la velocità della corrente.

Se sovrapponiamo gli effetti della diffusione e del trasporto, ovvero consideriamo come flusso

$$q(x, t) = Vu(x, t) - \mu \frac{\partial u}{\partial x}(x, t),$$

dalla (8.8) otteniamo

$$\frac{\partial u}{\partial t} - \mu \frac{\partial^2 u}{\partial x^2} + V \frac{\partial u}{\partial x} = 0, \quad (8.9)$$

equazione che costituisce il modello di *diffusione-trasporto* desiderato. In presenza di una sorgente distribuita di inquinante lungo il canale, di intensità $f = f(x, t)$ con dimensioni $[f] = [\text{massa}] \times [\text{lunghezza}]^{-1} \times [\text{tempo}]^{-1}$, la legge di bilancio fornirebbe

$$\frac{d}{dt} \int_x^{x+\Delta x} u(y, t) dy = -q(u(x + \Delta x, t)) + q(u(x, t)) + \int_x^{x+\Delta x} f(y, t) dt,$$

da cui, procedendo in modo analogo a quanto fatto in precedenza, otterremmo

$$\frac{\partial u}{\partial t} - \mu \frac{\partial^2 u}{\partial x^2} + V \frac{\partial u}{\partial x} = f.$$

Se infine sono presenti anche fenomeni di degradazione (con tasso $\sigma > 0$) il modello include anche un termine di reazione $-\sigma u(x, t)$ (in aggiunta ad $f(x, t)$), rappresentante reazioni chimiche dell'inquinante; ne risulta pertanto un modello di *diffusione-trasporto-reazione* nella forma seguente:

$$\frac{\partial u}{\partial t} - \mu \frac{\partial^2 u}{\partial x^2} + V \frac{\partial u}{\partial x} + \sigma u = f. \quad (8.10)$$

Anticipiamo che l'equazione precedente deve essere accompagnata da opportune condizioni al bordo da specificarsi agli estremi del canale e da una condizione iniziale, in cui appunto viene specificata la concentrazione dell'inquinante al tempo iniziale.

8.2 Differenze Finite per Problemi ai Limiti 1D

Consideriamo e formuliamo problemi di diffusione in una dimensione spaziale, ovvero tali per cui $u = u(x, t)$, con $x \in \mathbb{R}$. Dato l'intervallo $\Omega = (a, b) \in \mathbb{R}$ introduciamo l'equazione di ordine 2 (detta *equazione di Poisson*) corredata di due condizioni al contorno che assegnano il valore di u nel punti a e b :

$$\begin{cases} -u''(x) = f(x), & x \in (a, b) \\ u(a) = \alpha \\ u(b) = \beta; \end{cases} \quad (8.11)$$

Il problema precedente prende il nome di *problema di Dirichlet* per l'equazione di Poisson monodimensionale, e costituisce uno dei più semplici esempi di problemi ai limiti - avevamo già introdotto tale problema

nel Capitolo 1 a titolo di esempio notevole. Osserviamo come, nel caso di un problema di Dirichlet, agli estremi del dominio $x = a$ e $x = b$ sia stato assegnato il valore della funzione incognita, rispettivamente $u(a)$ e $u(b)$. Tale problema risulta ben posto:

Proposizione 8.2.1. *Sia $f \in C^0([a, b])$; allora la soluzione $u = u(x)$ del problema di Dirichlet per l'equazione di Poisson (8.11) esiste ed è unica, ed è tale che $u \in C^2([a, b])$.*

Consideriamo per semplicità $\alpha = \beta = 0$. Nel caso del problema (8.11), sappiamo trovare una soluzione esplicita nel caso in cui ad esempio $f(x) = c \in \mathbb{R}$ sia una costante. In questo caso, è immediato verificare che

$$u(x) = \frac{c}{2}(x - a)(b - x) \quad (8.12)$$

soddisfa sia l'equazione differenziale che le condizioni al contorno.

Più in generale, se consideriamo per semplicità il caso in cui $(a, b) = (0, 1)$, si ha che, a patto di considerare $f \in C^0([0, 1])$, l'unica soluzione del problema (8.11) è data da

$$u(x) = x \int_0^1 (1-s)f(s)ds - \int_0^x (x-s)f(s)ds. \quad (8.13)$$

Anche in questo caso, è possibile trovare la soluzione esplicitamente a patto di saper determinare le primitive delle funzioni che compaiono sotto il segno di integrale. Infatti, integrando due volte l'equazione differenziale, si ha che

$$-u'(x) + c_2 = \int_0^x f(y)dy \quad \Rightarrow \quad -u(x) + c_2 x + c_1 = \int_0^x \left(\int_0^s f(y)dy \right) ds$$

e dunque

$$u(x) = c_1 + c_2 x - \int_0^x \left(\int_0^s f(y)dy \right) ds.$$

Integrando per parti, otteniamo

$$\begin{aligned} \int_0^x \left(\int_0^s f(y)dy \right) ds &= \int_0^x 1 \cdot \left(\int_0^s f(y)dy \right) ds = \left[s \int_0^s f(y)dy \right]_{s=0}^{s=x} - \int_0^x s \left[\int_0^s f(y)dy \right]' ds \\ &= x \int_0^x f(y)dy - \int_0^x s f(s)ds = \int_0^x (x-s)f(s)ds \end{aligned}$$

da cui risulta che

$$u(x) = c_1 + c_2 x - \int_0^x (x-s)f(s)ds.$$

Dovendo poi risultare $u(0) = 0$ ricaviamo che $c_1 = 0$; dalla condizione $u(1) = 0$ si ha che

$$u(1) = c_1 + c_2 - \int_0^1 (1-s)f(s)ds = 0 \quad \Rightarrow \quad c_2 = \int_0^1 (1-s)f(s)ds$$

e dunque, infine,

$$u(x) = x \int_0^1 (1-s)f(s)ds - \int_0^x (x-s)f(s)ds.$$

Sebbene abbiamo determinato la soluzione u del problema in forma analitica, calcolare gli integrali che compaiono nella sua espressione richiede a priori di utilizzare delle formule di quadratura numeriche, nel caso in cui f sia una funzione la cui primitiva non sia semplice da ottenere. Per questo motivo, è opportuno introdurre uno schema numerico per calcolare la soluzione di un problema ai limiti, e per farlo sfrutteremo le differenze finite per l'approssimazione numerica delle derivate di una funzione.

Osservazione 8.2.1. Potremmo anche assegnare il valore di $u'(a)$ e $u'(b)$ agli estremi dell'intervallo (a, b) , al posto del valore di $u(a)$ e $u(b)$. In questo caso otteniamo un problema di Neumann per l'equazione di Poisson monodimensionale,

$$\begin{cases} -u''(x) = f(x), & x \in (a, b) \\ u'(a) = \gamma \\ u'(b) = \delta; \end{cases} \quad (8.14)$$

In questo caso, tuttavia, la soluzione del problema non è unica, ma è definita a meno di una costante; infatti, è semplice verificare che se $u(x)$ è soddisfa l'equazione differenziale ed è tale che $u'(a) = \gamma$ e $u'(b) = \delta$, anche $u(x) + C$, per una qualsiasi costante $C \in \mathbb{R}$, soddisfa sia l'equazione che le condizioni al contorno. Inoltre, deve essere rispettata una condizione di compatibilità sui dati f, γ, δ : infatti,

$$\int_a^b f(x)dx = - \int_a^b (u'(x))' dx = - u'(x)|_{x=a}^{x=b} = -u'(b) + u'(a) = \gamma - \delta,$$

ovvero deve essere verificata la condizione

$$\int_a^b f(x)dx = \gamma - \delta$$

affinché il problema abbia senso. Tuttavia, anche sotto questa condizione, la soluzione del problema di Neumann per l'equazione di Poisson è definita a meno di una costante.

8.2.1 Differenze finite per il problema di Poisson con condizioni di Dirichlet

Per costruire un metodo numerico che approssimi la soluzione del problema di Poisson monodimensionale con condizioni di Dirichlet, sfruttiamo le differenze finite introdotte nella Sezione 7.3 per approssimare le derivate di una funzione. Definiamo dunque $N + 2$ nodi $\{x_j\}_{j=0}^{N+1}$ equispaziati nell'intervallo $[a, b]$, a distanza

$$h = \frac{b - a}{N + 1},$$

con $x_j = a + jh$, $j = 0, \dots, N + 1$. Osserviamo che $x_0 = a$ e che $x_{N+1} = b$, ovvero risultano N i nodi interni all'intervallo (a, b) , x_1, \dots, x_N . Collochiamo il problema di Poisson nei nodi, supponendo che l'equazione differenziale sia soddisfatta in qualsiasi nodo x_j interno a (a, b) , ovvero

$$\begin{cases} u(x_0) = \alpha \\ -u''(x_j) = f(x_j) & \text{per } j = 1, \dots, N. \\ u(x_{N+1}) = \beta \end{cases}$$

Approssimiamo $u(x_j)$ con u_j e $u''(x_j)$ con la differenza finita centrale $\delta_c^2 u(x_j)$, ovvero

$$u''(x_j) \approx \frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{h^2} \quad \text{per } j = 1, \dots, N;$$

sostituendo la formula per la differenza finita centrale nel problema di Dirichlet collocato nei nodi, otteniamo il seguente problema numerico: al posto di cercare una funzione $u(x)$ che soddisfi il problema ai valori ai limiti, con il metodo delle differenze finite cerchiamo un insieme di $N + 2$ valori reali $\{u_j\}_{j=0}^{N+1}$ tali che

$$\begin{cases} u_0 = \alpha \\ -\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} = f(x_j) & \text{per ogni } j = 1, \dots, N. \\ u_{N+1} = \beta \end{cases} \quad (8.15)$$

Ovviamente, u_j costituirà un'approssimazione di $u(x_j)$; sottolineiamo come l'approssimazione della soluzione $u(x)$ che stiamo cercando non sia una funzione, ma un vettore di numeri, che approssimino i valori

di u nei nodi. Le equazioni (8.15) riconducono a un *sistema lineare* di $N + 2$ equazioni in $N + 2$ incognite,

$$\tilde{A}\tilde{\mathbf{u}}_h = \tilde{\mathbf{b}} \quad \Leftrightarrow \quad \begin{cases} u_0 = \alpha, \\ -\frac{u_0 - 2u_1 + u_2}{h^2} = f(x_1), \\ -\frac{u_1 - 2u_2 + u_3}{h^2} = f(x_2), \\ \vdots \\ -\frac{u_{N-1} - 2u_N + u_{N+1}}{h^2} = f(x_N), \\ u_{N+1} = \beta. \end{cases} \quad (8.16)$$

dove

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ -1/h^2 & 2/h^2 & -1/h^2 & 0 & \cdots & \cdots & 0 \\ 0 & -1/h^2 & 2/h^2 & -1/h^2 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & & & \vdots \\ 0 & 0 & 0 & \cdots & -1/h^2 & 2/h^2 & -1/h^2 \\ 0 & 0 & 0 & 0 & \cdots & \cdots & 1 \end{bmatrix}}_{\tilde{A}} \underbrace{\begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \\ u_N \\ u_{N+1} \end{bmatrix}}_{\tilde{\mathbf{u}}_h} = \underbrace{\begin{bmatrix} \alpha \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_N) \\ \beta \end{bmatrix}}_{\tilde{\mathbf{b}}}. \quad (8.17)$$

La matrice \tilde{A} è tridiagonale ma non risulta simmetrica. Possiamo tuttavia *condensare* la prima e l'ultima incognita del sistema, ottenendo un sistema lineare di N equazioni in N incognite, i valori u_1, \dots, u_N che rappresentano l'approssimazione della soluzione nei nodi interni – la prima e l'ultima equazione del sistema precedente corrispondono alle condizioni al bordo di Dirichlet, e permettono di determinare automaticamente i valori di u_0 e u_{N+1} . Infatti, per il nodo x_1 si ha che

$$-\frac{u_0 - 2u_1 + u_2}{h^2} = f(x_1)$$

ma siccome $u_0 = \alpha$ risulta che

$$\frac{2}{h^2}u_1 - \frac{1}{h^2}u_2 = f(x_1) + \frac{\alpha}{h^2}.$$

Analogamente, nel nodo x_N ,

$$-\frac{u_{N-1} - 2u_N + u_{N+1}}{h^2} = f(x_N)$$

ma siccome $u_{N+1} = \beta$, risulta che

$$-\frac{1}{h^2}u_{N-1} + \frac{2}{h^2}u_N = f(x_N) + \frac{\beta}{h^2}.$$

Complessivamente, otteniamo dunque le seguenti N equazioni:

$$\begin{cases} \frac{1}{h^2}(2u_1 - u_2) = f(x_1) + \frac{\alpha}{h^2} & \text{nel nodo } x_1 \\ -\frac{u_{j-1} - 2u_j + u_{j+1}}{h^2} = f(x_j) & \text{nei nodi interni } x_j, j = 2, \dots, N \\ \frac{1}{h^2}(-u_{N-1} + 2u_N) = f(x_N) + \frac{\beta}{h^2} & \text{nel nodo } x_N \end{cases}$$

ovvero abbiamo trovato il seguente sistema lineare da risolvere,

$$A\mathbf{u}_h = \mathbf{b} \quad (8.18)$$

dove $A \in \mathbb{R}^{N \times N}$ è detta *matrice di rigidezza* (o stiffness), $\mathbf{u}_h \in \mathbb{R}^N$ è il vettore delle incognite, e $\mathbf{b} \in \mathbb{R}^N$ è il termine noto (che ora include anche i dati al bordo):

$$A = \frac{1}{h^2} \begin{bmatrix} 2 & -1 & 0 & \dots & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ 0 & \dots & \dots & 0 & -1 & 2 \end{bmatrix} = \frac{1}{h^2} \text{tridiag}_N(-1, 2, -1)$$

$$\mathbf{u}_h = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \\ u_N \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{N-1}) \\ f(x_N) \end{bmatrix} + \frac{1}{h^2} \begin{bmatrix} \alpha \\ 0 \\ \vdots \\ 0 \\ \beta \end{bmatrix}.$$

La matrice A risulta simmetrica e tridiagonale, e dunque sparsa; risulta conveniente risolvere il sistema (8.18) con l'algoritmo di Thomas (Sezione 2.2.3), che richiede solo $O(8N)$ flops². La matrice A risulta anche definita positiva: si ha infatti che $\mathbf{v}^T A \mathbf{v} > 0$ per ogni $\mathbf{v} \in \mathbb{R}^N$, $\mathbf{v} \neq \mathbf{0}$. Si ha infatti che

$$\begin{aligned} \mathbf{v}^T A \mathbf{v} &= \sum_{i=1}^N \sum_{j=1}^N v_i a_{ij} v_j = \sum_{i=1}^N \frac{2}{h^2} v_i^2 - \sum_{i=2}^N \frac{1}{h^2} v_{i-1} v_i - \sum_{i=1}^{N-1} \frac{1}{h^2} v_{i+1} v_i \\ &= \frac{1}{h^2} (v_1^2 + v_N^2 + \sum_{j=2}^N (v_j - v_{j-1})^2) > 0 \end{aligned}$$

per ogni $\mathbf{v} \in \mathbb{R}^N$, dal momento che $a_{ij} \neq 0$ solo per $j = \{i-1, i, i+1\}$. Essendo A simmetrica e definita positiva, essa risulta non singolare, dunque il sistema lineare (8.18) ammette un'unica soluzione.

Mediante un'approssimazione alle differenze finite, abbiamo convertito un problema ai limiti per un'equazione differenziale in un problema algebrico, costituito da un sistema lineare di equazioni.

Osservazione 8.2.2. All'aumentare di N (e quindi al diminuire di h , ovvero considerando partizioni dell'intervallo $[a, b]$ sempre più fini) non solo la dimensione del sistema lineare (8.18) aumenta, comportando costi sempre maggiori per la sua risoluzione, ma aumenta anche il numero di condizionamento della matrice A . Siccome gli autovalori di A sono dati da

$$\lambda_j(A) = \frac{4}{h^2} \sin^2 \left(\frac{\pi}{2} \frac{j}{N+1} \right) \quad \text{per } j = 1, \dots, N,$$

si ha che

$$K_2(A) = K(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \approx \frac{C}{h^2},$$

per una costante $C > 0$ indipendente da h . In particolare, quando $h \rightarrow 0$, $K(A) \rightarrow \infty$ come $1/h^2$, ovvero per valori piccoli di h (N grandi) la matrice di rigidezza A è mal condizionata.

Riguardo invece alla convergenza del metodo delle differenze finite applicato al problema di Dirichlet per l'equazione di Poisson (8.11), in dimensione 1, vale il seguente risultato:

²Ecco spiegata la ragione della costruzione di un algoritmo specifico per le matrici tridiagonali...

Proposizione 8.2.2. Se $f \in C^2([a, b])$, ovvero, se $u \in C^4([a, b])$, vale la seguente stima per l'errore e_h tra la soluzione esatta e la soluzione approssimata ottenuta risolvendo il sistema lineare (8.18) derivante dall'approssimazione con differenze finite:

$$e_h = \max_{j=0, \dots, N+1} |u(x_j) - u_j| \leq \frac{h^2}{96} \max_{x \in [a, b]} |f''(x)|;$$

ovvero il metodo delle differenze finite converge con ordine 2 rispetto ad h quando si approssima il problema di Dirichlet per l'equazione di Poisson (8.11).

Osserviamo pertanto che (i) l'errore decade con un tasso proporzionale a h^2 (minore è $h < 1$, minore è l'errore), e che tale ordine di convergenza coincide con quello dall'approssimazione di una derivata seconda con le differenze finite centrate (quest'ultimo viene di fatto ereditato dal metodo delle differenze finite per un problema ai limiti) e che (ii) i dati del problema matematico entrano in gioco in una stima dell'errore anche per l'errore di discretizzazione.

8.2.2 Differenze finite per problemi di diffusione-trasporto-reazione con condizioni di Dirichlet

Generalizziamo ora il metodo introdotto nella sezione precedente al caso di un problema di diffusione-trasporto-reazione 1D con condizioni di Dirichlet, della forma

$$\begin{cases} -\mu u''(x) + \eta u'(x) + \sigma u(x) = f(x), & x \in (a, b) \\ u(a) = \alpha \\ u(b) = \beta; \end{cases} \quad (8.19)$$

dove $\mu > 0$, $\eta \in \mathbb{R}$ (il segno di η determina la direzione del trasporto) e $\sigma > 0$ sono i coefficienti di diffusione, trasporto e reazione, rispettivamente. Nel caso $\eta = \sigma = 0$ si ritrova il problema di Poisson considerato nella sezione precedente.

Usando le differenze finite centrate per approssimare $u'(x_j)$ e $u''(x_j)$, dove $x_j = a + jh$ e

$$h = \frac{b - a}{N + 1},$$

si ha che in un generico nodo x_j

$$u'(x_j) \approx \frac{u(x_{j+1}) - u(x_{j-1})}{2h} + O(h^2), \quad j = 1, \dots, N,$$

e

$$u''(x_j) \approx \frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{h^2} + O(h^2), \quad j = 1, \dots, N.$$

Collocando l'equazione differenziale nei nodi, approssimando le derivate con le differenze finite centrate e la soluzione $u(x_j)$ con u_j , otteniamo il seguente sistema di equazioni:

$$\begin{cases} -\mu \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + \eta \frac{u_{j+1} - u_{j-1}}{2h} + \sigma u_j = f(x_j) & \text{per } j = 1, \dots, N, \\ u_0 = \alpha \\ u_{N+1} = \beta. \end{cases} \quad (8.20)$$

ovvero otteniamo il *sistema lineare* di $N + 2$ equazioni in $N + 2$ incognite seguente:

$$\tilde{A}\tilde{\mathbf{u}}_h = \tilde{\mathbf{b}} \quad (8.21)$$

dove la matrice $\tilde{A} \in \mathbb{R}^{(N+2) \times (N+2)}$ è data da

$$\begin{aligned} \tilde{A} = & \begin{bmatrix} 1 & 0 & 0 & \dots & \dots & 0 \\ -\frac{\mu}{h^2} & 2\frac{\mu}{h^2} & -\frac{\mu}{h^2} & 0 & \dots & 0 \\ 0 & -\frac{\mu}{h^2} & 2\frac{\mu}{h^2} & -\frac{\mu}{h^2} & 0 & \dots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & -\frac{\mu}{h^2} & 2\frac{\mu}{h^2} & -\frac{\mu}{h^2} \\ 0 & & & \dots & 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & \dots & \dots & 0 \\ -\frac{\eta}{2h} & 0 & \frac{\eta}{2h} & 0 & \dots & 0 \\ 0 & -\frac{\eta}{2h} & 0 & \frac{\eta}{2h} & 0 & \dots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & -\frac{\eta}{2h} & 0 & \frac{\eta}{2h} \\ 0 & & & \dots & 0 & 0 \end{bmatrix} \\ & + \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & \sigma & 0 & \dots & 0 \\ 0 & \sigma & 0 & \dots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \sigma & 0 \\ 0 & & \dots & 0 & 0 \end{bmatrix}, \end{aligned}$$

mentre

$$\tilde{\mathbf{u}}_h = \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \\ u_N \\ u_{N+1} \end{bmatrix}, \quad \tilde{\mathbf{b}} = \begin{bmatrix} \alpha \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_N) \\ \beta \end{bmatrix}.$$

Analogamente al problema di Poisson, possiamo condensare le informazioni sulle condizioni al contorno di Dirichlet nel termine noto del sistema, ottenendo un sistema lineare di N equazioni in N incognite,

$$A\mathbf{u}_h = \mathbf{b} \quad (8.22)$$

dove $A \in \mathbb{R}^{N \times N}$ è data da

$$A = \underbrace{\frac{\mu}{h^2} \begin{bmatrix} 2 & -1 & 0 & \dots & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & -1 & 2 \end{bmatrix}}_{A_{diff}} + \underbrace{\frac{\eta}{2h} \begin{bmatrix} 0 & 1 & 0 & \dots & \dots & 0 \\ -1 & 0 & 1 & 0 & \dots & 0 \\ 0 & -1 & 0 & 1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & -1 & 0 \end{bmatrix}}_{A_{transp}} + \underbrace{\sigma I_{N \times N}}_{A_{react}}$$

mentre il vettore delle incognite $\mathbf{u}_h \in \mathbb{R}^N$ e il termine noto $\mathbf{b} \in \mathbb{R}^N$ sono dati, rispettivamente, da

$$\mathbf{u}_h = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \\ u_N \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{N-1}) \\ f(x_N) \end{bmatrix} + \begin{bmatrix} \left(\frac{\mu}{h^2} + \frac{\eta}{2h}\right)\alpha \\ 0 \\ \vdots \\ 0 \\ \left(\frac{\mu}{h^2} - \frac{\eta}{2h}\right)\beta \end{bmatrix}.$$

Osservazione 8.2.3. Nel caso in cui

$$\frac{|\eta| h}{2\mu} > 1$$

il trasporto è dominante (localmente) sulla diffusione; nel caso in cui

$$\frac{\sigma h^2}{6\mu} > 1$$

la reazione è dominante (localmente) sulla diffusione. In questi casi, possono insorgere oscillazioni numeriche nella soluzione ottenuta con il metodo delle differenze finite, senza ulteriori accorgimenti (come ad esempio quelli a cui accenneremo nella sezione successiva, in merito ai problemi di diffusione e trasporto).

Avendo usato differenze finite centrate per approssimare sia u'' che u' , l'approssimazione ottenuta risulta complessivamente di ordine 2 in h ; vale cioè il seguente risultato, analogo a quello ottenuto nel caso del problema di Dirichlet per l'equazione di Poisson:

Proposizione 8.2.3. Se $f \in C^2([a, b])$, ovvero $u \in C^4([a, b])$, vale la seguente stima per l'errore e_h tra la soluzione esatta e la soluzione approssimata ottenuta risolvendo il sistema lineare (8.22) derivante dall'approssimazione con differenze finite:

$$e_h = \max_{j=0, \dots, N+1} |u(x_j) - u_j| \leq Ch^2$$

per un'opportuna costante $C > 0$ indipendente da h , ovvero il metodo delle differenze finite converge con ordine 2 rispetto ad h quando si approssima il problema di Dirichlet per l'equazione di diffusione, trasporto e reazione, in dimensione 1.

Osservazione 8.2.4. Iniziamo a osservare che approssimare la derivata $u'(x_j)$ pesando ugualmente l'informazione per $x = x_{j-1}$ (a monte) e per $x = x_{j+1}$ (a valle), come fa una differenza finita centrata per la derivata prima, non è fisicamente sensato in presenza di un termine di trasporto con $\eta \gg 1$ oppure $\eta \ll 1$. Avrà senso sbilanciare questa valutazione, come vedremo nella prossima sezione.

8.2.3 Differenze finite per problemi di diffusione-trasporto e schema Upwind

Problemi ai limiti di diffusione-trasporto rappresentano prototipi dei modelli utilizzati per descrivere fenomeni della dinamica dei fluidi. Dal punto di vista computazionale, tali problemi presentano caratteristiche simili con problematiche comuni. In questa sezione vedremo come i cosiddetti problemi a trasporto dominante possono dare luogo a soluzioni numeriche affette da instabilità (*oscillazioni numeriche*), a meno che la griglia di calcolo sia sufficientemente fine. In alternativa, è possibile utilizzare le cosiddette tecniche di *stabilizzazione numerica*, tra cui appunto la tecnica *Upwind*.

Consideriamo il seguente problema di diffusione-trasporto con condizioni al contorno di Dirichlet

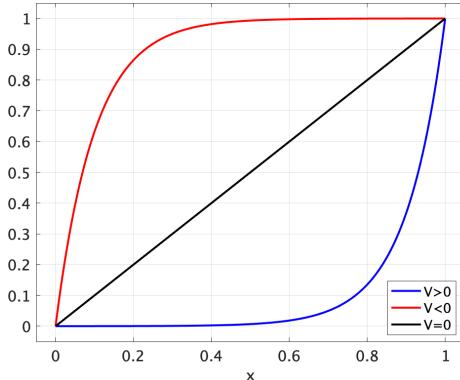
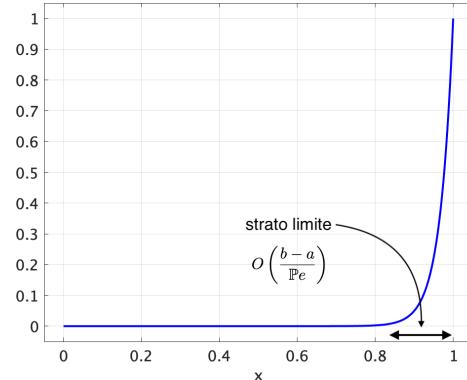
$$\begin{cases} -\mu u''(x) + \eta u'(x) = 0, & x \in (a, b) \\ u(a) = \alpha \\ u(b) = \beta, \end{cases} \quad (8.23)$$

dove $\mu > 0$ e $\eta \in \mathbb{R}$; per semplicità, abbiamo posto il termine forzante $f(x) = 0$. Per caratterizzare il problema precedente si introduce il numero adimensionale seguente, detto *numero di Péclet* (globale):

$$\text{Pe} := \frac{|\eta|(b-a)}{2\mu}. \quad (8.24)$$

Il numero di Péclet misura la dominanza del fenomeno di trasporto sul processo di diffusione. Se $\text{Pe} = 1$ allora i due fenomeni si equivalgono quantitativamente. Se $0 \leq \text{Pe} < 1$ significa che il processo diffusivo domina quello di trasporto. Viceversa, se $\text{Pe} > 1$, il fenomeno di trasporto domina sulla diffusione. Problemi a trasporto dominante di interesse fisico possono essere caratterizzati da valori di $\text{Pe} \gg 1$ (anche sopra a valori di 10^6). In problemi di fluidodinamica descritti dalle equazioni di Navier-Stokes, il corrispondente del numero di Péclet è il numero di Reynolds Re ; un aeromobile civile si muove nel fluido aria in condizioni di trasporto dominante sui fenomeni viscosi, raggiungendo valori di Re attorno a 10^7 .

Esempio 8.2.1. Consideriamo il caso in cui $a = 0$, $b = 1$, $\alpha = 0$ e $\beta = 1$ per il problema (8.23), con $\mu > 0$. Otteniamo soluzioni simili a quelle raffigurate di seguito. Osserviamo che, tanto maggiore è $\mathbb{P}e$, tanto più stretto sarà il cosiddetto strato limite, ovvero la porzione di dominio laddove si riscontrano valori "grandi" di $u'(x)$; in particolare, l'ampiezza dello strato limite è proporzionale a $\frac{1}{\mathbb{P}e}$.

soluzioni $u(x)$ per $\eta > 0$, $\eta < 0$ e $\eta = 0$  $\eta > 0$, rappresentazione dello strato limite

Per semplicità, consideriamo ora il problema (8.23) nel caso $a = 0$, $b = 1$, $\alpha = 0$, $\beta = 1$, ovvero

$$\begin{cases} -\mu u''(x) + \eta u'(x) = 0, & x \in (0, 1) \\ u(0) = 0 \\ u(1) = 1. \end{cases} \quad (8.25)$$

Cerchiamo la soluzione analitica del problema precedente. Esso ammette soluzioni generali nella forma $u_k(x) = e^{\lambda_k x}$ per qualche $k = 1, 2, \dots$, dove $\lambda_k \in \mathbb{R}$. Osserviamo che $(u_k)'(x) = \lambda_k e^{\lambda_k x} = \lambda_k u_k(x)$ e che $(u_k)''(x) = \lambda_k^2 e^{\lambda_k x} = \lambda_k^2 u_k(x)$. Sostituendo le precedenti in Eq. (8.25), dobbiamo soddisfare la seguente:

$$(-\mu \lambda_k^2 + \eta \lambda_k) e^{\lambda_k x} = 0 \quad \text{per ogni } x \in (0, 1),$$

che risulta soddisfatta se $\lambda_1 = 0$ e $\lambda_2 = \frac{\eta}{\mu}$. Le soluzioni del tipo $u_1(x) = 1$ e $u_2(x) = e^{\eta/\mu x}$ sono entambe ammissibili, per cui la soluzione analitica può essere scritta come combinazione lineare di queste, ovvero come

$$u(x) = C_1 u_1(x) + C_2 u_2(x) = C_1 + C_2 e^{\eta/\mu x},$$

per due costanti $C_1, C_2 \in \mathbb{R}$ da determinarsi. Sfruttiamo ora le condizioni al contorno di Dirichlet, tali per cui $u(0) = 0$ e $u(1) = 1$. Ciò implica che

$$\begin{cases} u(0) = C_1 + C_2 e^0 = C_1 + C_2 = 0, \\ u(1) = C_1 + C_2 e^{\eta/\mu 1} = C_1 + C_2 e^{\eta/\mu} = 1, \end{cases}$$

da cui si ottiene che $C_1 = -\frac{1}{e^{\eta/\mu} - 1}$ e $C_2 = \frac{1}{e^{\eta/\mu} - 1}$. La soluzione analitica di Eq. (8.25) è pertanto

$$u(x) = \frac{e^{\eta/\mu x} - 1}{e^{\eta/\mu} - 1} \quad \text{per } x \in [0, 1].$$

Nel limite $\frac{\eta}{\mu} \rightarrow 0$, la soluzione analitica diventa $u(x) = x$. Soluzioni analitiche del problema (8.25) sono rappresentate nell'Esempio 8.2.1.

Pur essendo la soluzione analitica nota, risulta interessante determinare il risultato dell'approssimazione numerica tramite differenze finite del precedente problema prototipo. In particolare, consideriamo in

maniera del tutto analoga alla Sezione 8.2.2 il metodo delle *differenze finite centrate*, che fornisce l'approssimazione u_j di $u(x_j)$ nel nodo j -esimo della griglia di calcolo (uniforme) con elementi di ampiezza $h = \frac{1}{N+1}$. Otteniamo:

$$\begin{cases} u_0 = 0 \\ -\mu \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + \eta \frac{u_{j+1} - u_{j-1}}{2h} = 0 & \text{per } j = 1, \dots, N, \\ u_{N+1} = 1 \end{cases}$$

da cui

$$\begin{cases} \left(-\frac{\mu}{h^2} + \frac{\eta}{2h} \right) u_{j+1} + \frac{2\mu}{h^2} u_j + \left(-\frac{\mu}{h^2} - \frac{\eta}{2h} \right) u_{j-1} = 0 & \text{per } j = 1, \dots, N, \\ u_{N+1} = 1 \end{cases} \quad (8.26)$$

Introduciamo ora un nuovo numero adimensionale, detto *numero di Péclet locale*

$$\boxed{\mathbb{P}e_h := \frac{|\eta| h}{2\mu}.} \quad (8.27)$$

Tale numero indica la dominanza del fenomeno di trasporto sulla diffusione a livello del singolo elemento della griglia di ampiezza h , ovvero *localmente*. Se $\mathbb{P}e_h > 1$, allora il fenomeno di trasporto domina quello diffusivo localmente. Se invece $\mathbb{P}e_h < 1$, il fenomeno diffusivo domina su quello di trasporto localmente; osserviamo che in generale

$$\mathbb{P}e_h = \frac{h}{b-a} \mathbb{P}e,$$

dunque è possibile avere $\mathbb{P}e_h < 1$ anche se $\mathbb{P}e > 1$ scegliendo il passo h "sufficientemente" piccolo, ovvero $h < \mathbb{P}e(b-a)$. Vedremo tra poco che il valore assunto da $\mathbb{P}e_h$ avrà delle implicazioni notevoli sulla soluzione approssimata del problema (8.23) (e (8.25)).

Consideriamo nuovamente l'Eq. (8.26), che moltiplichiamo per $\frac{h^2}{\mu}$, ottenendo, supponendo che $\eta > 0$ per semplicità:

$$\begin{cases} u_0 = 0 \\ (\mathbb{P}e_h - 1) u_{j+1} + 2u_j + (-\mathbb{P}e_h - 1) u_{j-1} = 0 & \text{per } j = 1, \dots, N, \\ u_{N+1} = 1 \end{cases}$$

Si osserva che la precedente è un'equazione alle differenze che ammette soluzioni del tipo $u_j = \rho^j$ per qualche $\rho \in \mathbb{R}$. Osservando che $u_{j+1} = \rho^{j+1}$, $u_{j-1} = \rho^{j-1}$, abbiamo, sempre per $\eta > 0$:

$$\begin{cases} u_0 = 0 \\ [(\mathbb{P}e_h - 1) \rho^2 + 2\rho + (-\mathbb{P}e_h - 1)] \rho^{j-1} = 0 & \text{per } j = 1, \dots, N, \\ u_{N+1} = 1 \end{cases}$$

Dato che la precedente deve essere soddisfatta per ogni $j = 1, \dots, N$, risolviamo la seguente equazione di secondo grado:

$$(\mathbb{P}e_h - 1) \rho^2 + 2\rho + (-\mathbb{P}e_h - 1) = 0,$$

dotata di due soluzioni generali $\rho_1 = 1$ e $\rho_2 = \frac{1 + \mathbb{P}e_h}{1 - \mathbb{P}e_h}$. La soluzione nodale del problema approssimato tramite differenze finite centrate è pertanto nella forma seguente

$$u_j = C_1 \rho_1^j + C_2 \rho_2^j = C_1 + C_2 \left(\frac{1 + \mathbb{P}e_h}{1 - \mathbb{P}e_h} \right)^j \quad \text{per } j = 1, \dots, N,$$

con $C_1, C_2 \in \mathbb{R}$ due costanti da determinarsi. Determiniamo tali costanti imponendo il soddisfacimento delle condizioni al contorno, ovvero

$$\begin{cases} u_0 = C_1 + C_2 \left(\frac{1 + \mathbb{P}e_h}{1 - \mathbb{P}e_h} \right)^0 = C_1 + C_2 = 0, \\ u_1 = C_1 + C_2 \left(\frac{1 + \mathbb{P}e_h}{1 - \mathbb{P}e_h} \right)^{N+1} = 1. \end{cases}$$

In tal modo si ottiene la seguente soluzione approssimata nei nodi

$$u_j = \frac{1 - \left(\frac{1 + \mathbb{P}e_h}{1 - \mathbb{P}e_h} \right)^j}{1 - \left(\frac{1 + \mathbb{P}e_h}{1 - \mathbb{P}e_h} \right)^{N+1}} \quad \text{per } j = 0, 1, \dots, N, N+1.$$

(8.28)

Osserviamo che tale soluzione è equivalente a quella ottenibile dalla risoluzione del sistema lineare $A\mathbf{u}_h = \mathbf{b}$ (in aritmetica esatta), dove

$$A = \frac{\mu}{h^2} \begin{bmatrix} 2 & -1 & 0 & \dots & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & -1 & 2 \end{bmatrix} + \frac{\eta}{2h} \begin{bmatrix} 0 & +1 & 0 & \dots & \dots & 0 \\ -1 & 0 & +1 & 0 & \dots & 0 \\ 0 & -1 & 0 & +1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & -1 & 0 \end{bmatrix}$$

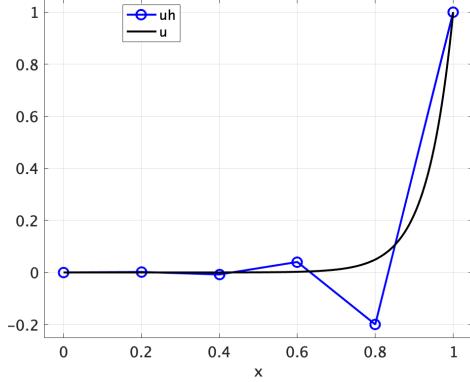
$$\mathbf{u}_h = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \\ u_N \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \left(\frac{\mu}{h^2} - \frac{\eta}{2h} \right) \end{bmatrix}.$$

Consideriamo la soluzione approssimata tramite le differenze finite centrate (8.28). Possiamo distinguere due casi:

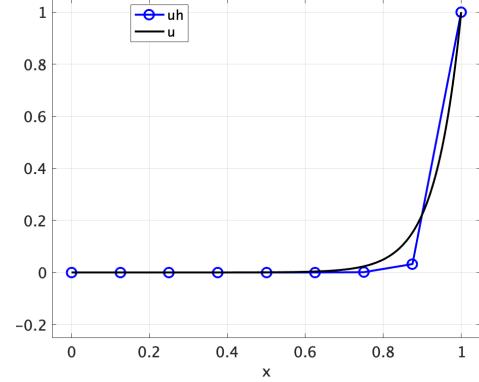
- Se $\mathbb{P}e_h < 1$, allora $\left(\frac{1 + \mathbb{P}e_h}{1 - \mathbb{P}e_h} \right)^j > 1$; dunque $\left(\frac{1 + \mathbb{P}e_h}{1 - \mathbb{P}e_h} \right)^j > 1$ per ogni $j = 1, 2, \dots$ e inoltre $\left(\frac{1 + \mathbb{P}e_h}{1 - \mathbb{P}e_h} \right)^{N+1} > 1$. Nel caso in cui $\mathbb{P}e_h > 1$, il numeratore e il denominatore di Eq. (8.28) hanno pertanto segno concorde e dunque otteniamo che $u_j > 0$ per ogni $j = 1, 2, 3, \dots, N$.
- Se $\mathbb{P}e_h > 1$, allora $\left(\frac{1 + \mathbb{P}e_h}{1 - \mathbb{P}e_h} \right)^j < -1$. Dunque abbiamo $\left(\frac{1 + \mathbb{P}e_h}{1 - \mathbb{P}e_h} \right)^j > 1$ se j è pari, mentre $\left(\frac{1 + \mathbb{P}e_h}{1 - \mathbb{P}e_h} \right)^j < -1$ se j è dispari; in maniera simile, $\left(\frac{1 + \mathbb{P}e_h}{1 - \mathbb{P}e_h} \right)^{N+1} > 1$ se $N+1$ è pari, mentre $\left(\frac{1 + \mathbb{P}e_h}{1 - \mathbb{P}e_h} \right)^{N+1} < -1$ se $N+1$ è dispari. Se $\mathbb{P}e_h > 1$, il numeratore e il denominatore di Eq. (8.28) posseggono segno discorde a seconda dell'indice j ; dunque la soluzione approssimata u_j è alternativamente positiva o negativa per $j = 1, 2, 3, \dots, N$.

Abbiamo dunque visto che nel caso in cui $\mathbb{P}e_h > 1$, la soluzione approssimata u_j di Eq. (8.28) presenta le cosiddette *oscillazioni numeriche*, ovvero evidenzia un fenomeno numerico spesso noto come *instabilità numerica*.

Esempio 8.2.2. Consideriamo i dati $a = 0$, $b = 1$, $\alpha = 0$, $\beta = 1$, $\mu = 1$ e $\eta = 15$ per il problema (8.23). Abbiamo $\text{Pe} = 7.5 > 1$, ovvero un problema a trasporto dominante. Riportiamo di seguito le approssimazioni numeriche ottenute con il metodo delle differenze finite centrate con diversi valori del passo di griglia $h > 0$.



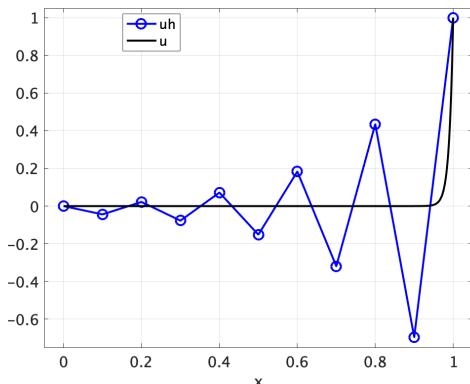
$h = 0.2, N = 4, \text{Pe}_h = 1.5 > 1$
presenza di oscillazioni numeriche



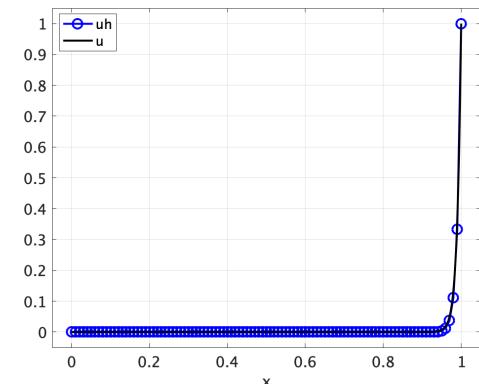
$h = 0.125, N = 7, \text{Pe}_h = 0.9375 < 1$

Osserviamo come la soluzione approssimata ottenuta in corrispondenza di un numero di Péclet locale minore di uno sia esente da oscillazioni e instabilità numeriche.

Consideriamo ora il medesimo problema con $\eta = 100$, ovvero tale per cui $\text{Pe} = 50 > 1$. Ottieniamo per diversi valori di $h > 0$ i seguenti risultati, che confermano le considerazioni precedenti.



$h = 0.1, N = 9, \text{Pe}_h = 5 > 1$
presenza di oscillazioni numeriche



$h = 0.01, N = 99, \text{Pe}_h = 0.5 < 1$

Il precedente esempio illustra come per ovviare all'insorgenza di oscillazioni numeriche sia necessario scegliere un passo di griglia $h > 0$ tale che il numero di Péclet locale sia inferiore ad uno ($\text{Pe}_h < 1$, ovvero:

$$h < \frac{2\mu}{|\eta|} = \frac{b-a}{\text{Pe}}.$$

Tale scelta garantisce che la soluzione approssimata di Eq. (8.28) sia esente da oscillazioni numeriche. Tuttavia, in problemi con numero di Péclet globale Pe molto maggiore di 1, ovvero in cui il fenomeno di trasporto risulti largamente dominante sulla diffusione, la scelta del passo h può risultare molto restrittiva e portare dunque a costi computazionali molto elevati (la dimensione N del sistema lineare cresce) e a problemi di condizionamento della matrice A . Un approccio computazionalmente più conveniente consiste

nel considerare un passo di griglia h non uniforme nel dominio, riservando il passo più piccolo in corrispondenza dello strato limite, anche se in tal caso si deve presupporre come nota a priori la sua posizione o da definirsi utilizzando opportune tecniche di adattività di griglia.

Ricordiamo che per h "sufficientemente" piccolo, ovvero tale per cui la soluzione numerica del problema di diffusione-trasporto sia esente da oscillazioni numeriche, il metodo delle differenze finite centrate presenta un errore convergente di ordine 2 rispetto ad h , secondo la Proposizione 8.2.3.

Metodo delle differenze finite con tecnica Upwind

Un metodo alternativo per contrastare le oscillazioni numeriche laddove la riduzione del passo di griglia h in maniera uniforme sia proibitiva, consiste nell'accompagnare il metodo numerico delle differenze finite con un'opportuna tecnica di *stabilizzazione numerica*, quale proprio la tecnica *Upwind*.

Tale tecnica consiste nell'approssimare il problema di diffusione-trasporto utilizzando le differenze finite centrate per l'approssimazione del termine di diffusione, legato a $u''(x)$, mentre le differenze finite in avanti o all'indietro per il termine di trasporto, legato a $u'(x)$ (Sezione 7.3.1). In particolare, immaginando di aver già provveduto a introdurre la discretizzazione del dominio con passo h , l'approssimazione del termine di trasporto avviene secondo i due casi seguenti:

- Se $\eta > 0$, allora l'approssimazione del termine di trasporto fa uso delle *differenze finite all'indietro*, ovvero, nel generico nodo interno x_j , si ha:

$$\eta u'(x_j) \approx \eta \frac{u(x_j) - u(x_{j-1})}{h}.$$

- Se $\eta < 0$, allora l'approssimazione del termine di trasporto fa uso delle *differenze finite in avanti*, ovvero, nel generico nodo interno x_j , si ha:

$$\eta u'(x_j) \approx \eta \frac{u(x_{j+1}) - u(x_j)}{h}.$$

Osserviamo che tale approccio prefigura ordini di accuratezza diversi, rispettivamente di ordine 2 ed 1, per il termine di diffusione e quello di trasporto per via dell'uso delle differenze finite centrate per il primo e in avanti/all'indietro per il secondo.

Consideriamo nuovamente per esempio il problema (8.25) prendendo in considerazione per semplicità il caso in cui $\eta > 0$. Faremo dunque uso delle differenze finite all'indietro per approssimare il termine di trasporto. Otteniamo dunque:

$$\begin{cases} u_0 = 0 \\ -\mu \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + \eta \frac{u_j - u_{j-1}}{h} = 0 & \text{per } j = 1, \dots, N, \\ u_{N+1} = 1 \end{cases} \quad (8.29)$$

Osserviamo ora che

$$\frac{u_j - u_{j-1}}{h} = \frac{u_{j+1} - u_{j-1}}{2h} - \frac{h}{2} \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2},$$

ovvero abbiamo riscritto l'approssimazione del termine di trasporto tramite le differenze finite all'indietro come combinazione di approssimazioni alle differenze finite centrate rispettivamente dei termini di trasporto e diffusione³. Sostituendo la precedente identità nell'Eq. (8.29), otteniamo

$$\begin{cases} u_0 = 0 \\ -\left(\mu + \frac{\eta h}{2}\right) \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + \eta \frac{u_{j+1} - u_{j-1}}{2h} = 0 & \text{per } j = 1, \dots, N, \\ u_{N+1} = 1 \end{cases}$$

³Analogamente, avremmo avuto per le differenze finite in avanti la seguente identità:

$$\frac{u_{j+1} - u_j}{h} = \frac{u_{j+1} - u_{j-1}}{2h} + \frac{h}{2} \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2}.$$

da cui, sfruttando la definizione di numero di Péclet locale $\mathbb{P}e_h$ (8.26), abbiamo infine:

$$\begin{cases} u_0 = 0 \\ -\mu(1 + \mathbb{P}e_h) \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + \eta \frac{u_{j+1} - u_{j-1}}{2h} = 0 & \text{per } j = 1, \dots, N. \\ u_{N+1} = 1 \end{cases}$$

Il precedente metodo numerico è noto come *schemi alle differenze finite centrate* con tecnica *Upwind* con viscosità artificiale e si scrive anche come

$$\begin{cases} u_0 = 0 \\ -\mu_h \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + \eta \frac{u_{j+1} - u_{j-1}}{2h} = 0 & \text{per } j = 1, \dots, N. \\ u_{N+1} = 1 \end{cases}$$

(8.30)

dove la *viscosità artificiale* (viscosità numerica) è:

$$\mu_h = \mu(1 + \mathbb{P}e_h). \quad (8.31)$$

Osserviamo che lo schema Upwind (8.30) si applica a entrambi i casi in cui $\eta > 0$ e $\eta < 0$ e può essere ricondotto alla soluzione di un sistema lineare in forma matriciale del tipo

$$A \mathbf{u}_h = \mathbf{b}.$$

Inoltre, lo schema Upwind è immune da oscillazioni numeriche per ogni scelta del passo di griglia $h > 0$. Infatti, allo schema Upwind (8.30) possiamo associare un nuovo numero di Péclet locale dipendente dalla viscosità artificiale, ovvero:

$$\mathbb{P}e_h^* := \frac{|\eta| h}{2 \mu_h} = \frac{|\eta| h}{2 \mu (1 + \mathbb{P}e_h)} = \frac{\mathbb{P}e_h}{1 + \mathbb{P}e_h},$$

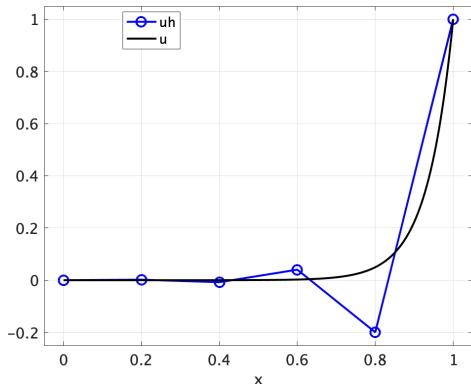
ancora una volta un numero adimensionale. Osserviamo che il numero Péclet locale $\mathbb{P}e_h^*$ associato allo schema Upwind (8.30) è strettamente inferiore ad uno per ogni $h > 0$, infatti

$$\mathbb{P}e_h^* = \frac{|\eta| h}{2 \mu (1 + \mathbb{P}e_h)} = \frac{\mathbb{P}e_h}{1 + \mathbb{P}e_h} < 1,$$

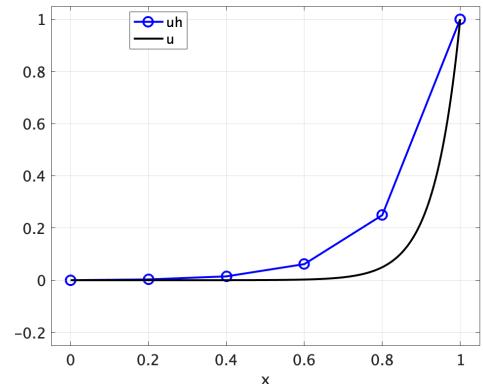
dato che $\mathbb{P}e_h > 0$, e dunque immune da oscillazioni numeriche per ogni scelta del passo di griglia $h > 0$.

Esempio 8.2.3. Consideriamo nuovamente l'Esempio 8.2.2 e osserviamo l'efficacia dello schema Upwind nel rimuovere le oscillazioni numeriche.

Per i dati $a = 0$, $b = 1$, $\alpha = 0$, $\beta = 1$, $\mu = 1$ e $\eta = 15$, abbiamo $\mathbb{P}e = 7.5 > 1$.

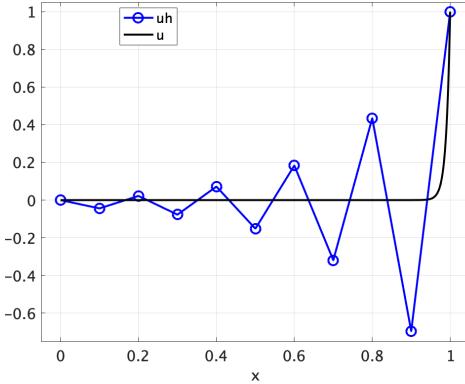


$h = 0.2$, $N = 4$, $\mathbb{P}e_h = 1.5 > 1$
schema differenze finite centrate

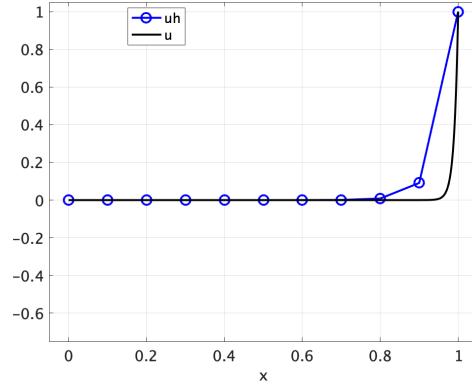


$h = 0.2$, $N = 4$, $\mathbb{P}e_h = 1.5 > 1$, $\mathbb{P}e_h^* = 0.6 < 1$
schema differenze finite centrate–Upwind

Per $\eta = 100$, abbiamo $\mathbb{P}e = 50$ e il risultato seguente:



$h = 0.1, N = 9, \mathbb{P}e_h = 5 > 1$
schema differenze finite centrate



$h = 0.1, N = 9, \mathbb{P}e_h^* = 0.8333 < 1$
schema differenze finite centrate–Upwind

Come anticipato, lo schema alle differenze finite con tecnica Upwind combina metodi di approssimazione delle derivate basati su schemi accurati di ordine 1 e 2. La ridotta accuratezza associata agli schemi alle differenze finite in avanti o all'indietro per la derivata prima comporta una riduzione complessiva dell'accuratezza del metodo all'ordine 1. Possiamo formulare il seguente risultato.

Proposizione 8.2.4. *Per l'approssimazione del problema di diffusione-trasporto (8.23) tramite differenze finite centrate con tecnica Upwind, si ha:*

$$e_h = \max_{j=0, \dots, N+1} |u(x_j) - u_j| \leq C h,$$

per un'opportuna costante positiva C , ovvero lo schema converge con ordine 1 rispetto ad h .

È possibile sviluppare metodi di stabilizzazione numerica più accurati dello schema Upwind, per esempio modificando l'espressione della viscosità artificiale μ_h di Eq. (8.31), ma sempre utilizzando schemi alle differenze finite centrate per l'approssimazione di u'' e u' , nel metodo di Eq (8.30). In particolare, è possibile scegliere μ_h come segue

$$\mu_h = \mu (1 + \phi(\mathbb{P}e_h)),$$

dove ϕ è un'opportuna funzione. Se $\phi(\mathbb{P}e_h) = \mathbb{P}e_h$, ritroviamo lo schema delle differenze finite centrate–Upwind, accurato di ordine 1. Se invece, $\phi(\mathbb{P}e_h) = \mathbb{P}e_h - 1 + \frac{2\mathbb{P}e_h}{e^{2\mathbb{P}e_h} - 1}$, otteniamo il cosiddetto schema alle differenze finite centrate di Scharfetter–Gummel, ovvero un metodo accurato di ordine 2.

8.2.4 Differenze finite per il problema di Poisson con condizioni miste di Dirichlet/Neumann

Consideriamo ora un problema di Poisson 1D con condizioni al contorno miste, di tipo Dirichlet su un bordo e di Neumann sull'altro, nella forma

$$\begin{cases} -\mu u''(x) = f(x), & x \in (a, b) \\ u(a) = \alpha \\ \mu u'(b) = \gamma; \end{cases} \quad (8.32)$$

dove $\mu > 0$; γ indica il dato al bordo di Neumann. Osserviamo che in questo caso il valore di $u(b)$ è incognito dato che invece è assegnato il valore della derivata prima $u'(b)$ (secondo la condizione al contorno di Neumann).

Procediamo all'approssimazione del precedente problema in maniera analoga alle Sezioni 8.2.1 e 8.2.2 introducendo la discretizzazione del dominio (a, b) con una griglia compresa dei nodi $x_j = a + j h$ per $j = 0, \dots, N + 1$, essendo il passo di griglia nuovamente $h = \frac{b - a}{N + 1} > 0$. Utilizziamo nuovamente le differenze centrate per l'approssimazione di $u''(x_j)$ in un generico nodo interno, ovvero

$$u''(x_j) \approx \frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{h^2} \quad \text{per } j = 1, \dots, N;$$

ricordiamo che tale schema è accurato di ordine 2 rispetto ad h . Per l'approssimazione di $u'(b) = u'(x_{N+1})$ possiamo considerare la *differenza finita all'indietro* (Sezione 7.3.1), ovvero

$$u'(b) = u'(x_{N+1}) \approx \delta_- u(x_{N+1}) = \frac{u(x_{N+1}) - u(x_N)}{h},$$

che ricordiamo essere uno schema accurato di ordine 1 rispetto ad h . Introducendo l'approssimazione u_j di $u(x_j)$, otteniamo dunque il seguente schema alla differenze finite

$$\begin{cases} u_0 = \alpha \\ -\mu \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} = f(x_j) \quad \text{per } j = 1, \dots, N, \\ \mu \frac{u_{N+1} - u_N}{h} = \gamma, \end{cases} \quad (8.33)$$

ovvero otteniamo il *sistema lineare* di $N + 2$ equazioni in $N + 2$ incognite $\tilde{A}\tilde{\mathbf{u}}_h = \tilde{\mathbf{b}}$, dove la matrice $\tilde{A} \in \mathbb{R}^{(N+2) \times (N+2)}$, $\tilde{\mathbf{u}}_h, \tilde{\mathbf{b}} \in \mathbb{R}^{(N+2)}$ sono

$$\tilde{A} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ -\frac{\mu}{h^2} & 2\frac{\mu}{h^2} & -\frac{\mu}{h^2} & 0 & \dots & 0 \\ 0 & -\frac{\mu}{h^2} & 2\frac{\mu}{h^2} & -\frac{\mu}{h^2} & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \\ 0 & \dots & 0 & -\frac{\mu}{h^2} & 2\frac{\mu}{h^2} & -\frac{\mu}{h^2} \\ 0 & \dots & & 0 & -\frac{\mu}{h} & \frac{\mu}{h} \end{bmatrix}, \quad \tilde{\mathbf{u}}_h = \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \\ u_N \\ u_{N+1} \end{bmatrix}, \quad \tilde{\mathbf{b}} = \begin{bmatrix} \alpha \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_N) \\ \gamma \end{bmatrix}.$$

Possiamo *condensare* le informazioni sulla condizione al contorno di Dirichlet nel termine noto del sistema, ottenendo un sistema lineare di $N + 1$ equazioni in $N + 1$ incognite $A\mathbf{u}_h = \mathbf{b}$, dove $A \in \mathbb{R}^{(N+1) \times (N+1)}$, $\mathbf{u}_h, \mathbf{b} \in \mathbb{R}^{N+1}$ sono dati da

$$A = \begin{bmatrix} 2\frac{\mu}{h^2} & -\frac{\mu}{h^2} & 0 & \dots & 0 \\ -\frac{\mu}{h^2} & 2\frac{\mu}{h^2} & -\frac{\mu}{h^2} & 0 & \dots \\ 0 & \ddots & \ddots & \ddots & \\ \vdots & 0 & -\frac{\mu}{h^2} & 2\frac{\mu}{h^2} & -\frac{\mu}{h^2} \\ 0 & \dots & 0 & -\frac{\mu}{h} & \frac{\mu}{h} \end{bmatrix}, \quad \mathbf{u}_h = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \\ u_N \\ u_{N+1} \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} f(x_1) + \frac{\mu}{h^2} \alpha \\ f(x_2) \\ \vdots \\ f(x_{N-1}) \\ f(x_N) \\ \gamma \end{bmatrix}.$$

La matrice A non è simmetrica, ma rimane tridiagonale e non singolare.

Dato che lo schema alla differenze finite (8.33) usa metodi di diverso ordine di accuratezza, e ammesso che la soluzione u del problema (8.32) sia sufficientemente regolare, allora l'ordine di accuratezza atteso rispetto ad h è solo pari ad 1, ovvero

$$e_h = \max_{j=0, \dots, N+1} |u(x_j) - u_j| \leq C h.$$

È possibile ripristinare l'ordine di convergenza pari a 2 rispetto ad h utilizzando uno schema alle differenze finite (decentrato) per l'approssimazione di $u'(b)$, in particolare per il caso in questione la formula di Eq. (7.21) accurata di ordine 2 rispetto ad h . Consideriamo dunque la seguente approssimazione

$$u'(b) = u'(x_{N+1}) \approx \delta_{c,N+1} u(x_{N+1}) = \frac{3u(x_{N+1}) - 4u(x_N) + u(x_{N-1})}{2h}.$$

In tal caso, otteniamo il seguente schema alla differenze finite

$$\left\{ \begin{array}{l} u_0 = \alpha \\ -\mu \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} = f(x_j) \quad \text{per } j = 1, \dots, N, \\ \mu \frac{3u_{N+1} - 4u_N + u_{N-1}}{2h} = \gamma, \end{array} \right. \quad (8.34)$$

da cui il sistema lineare condensato $A\mathbf{u}_h = \mathbf{b}$, dove $A \in \mathbb{R}^{(N+1) \times (N+1)}$, $\mathbf{u}_h, \mathbf{b} \in \mathbb{R}^{N+1}$ sono dati da

$$A = \begin{bmatrix} 2\frac{\mu}{h^2} & -\frac{\mu}{h^2} & 0 & \dots & 0 \\ -\frac{\mu}{h^2} & 2\frac{\mu}{h^2} & -\frac{\mu}{h^2} & 0 & \dots \\ 0 & \ddots & \ddots & \ddots & \\ \vdots & 0 & -\frac{\mu}{h^2} & 2\frac{\mu}{h^2} & -\frac{\mu}{h^2} \\ \dots & 0 & \frac{\mu}{2h} & -\frac{2\mu}{h} & \frac{3\mu}{2h} \end{bmatrix}, \quad \mathbf{u}_h = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \\ u_N \\ u_{N+1} \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} f(x_1) + \frac{\mu}{h^2}\alpha \\ f(x_2) \\ \vdots \\ f(x_{N-1}) \\ f(x_N) \\ \gamma \end{bmatrix}.$$

Osserviamo che la matrice precedente A non solo non è simmetrica, ma non è neppure tridiagonale, pur rimanendo non singolare. Otteniamo dunque il seguente risultato di convergenza per lo schema (8.34).

Proposizione 8.2.5. *Per l'approssimazione del problema (8.32) tramite differenze finite centrate e accurate di ordine 2 per l'approssimazione della condizione di Neumann di Eq. (8.34), se $u \in C^4([a, b])$, si ottiene:*

$$e_h = \max_{j=0, \dots, N+1} |u(x_j) - u_j| \leq C h^2,$$

per un'opportuna costante positiva C , ovvero lo schema converge con ordine 2 rispetto ad h .

Osservazione 8.2.5. *Risultati del tutto simili si ottengono per il problema di Poisson o di diffusione-trasporto-reazione con condizioni miste di Neumann (sul bordo sinistro, $x = a$) e Dirichlet (sul bordo destro, $x = b$).*

8.3 Differenze Finite per il Problema di Poisson 2D

Consideriamo il problema di Poisson in 2D, come rappresentato in Eq. (8.2) per $d = 2$, specificatamente con condizioni al contorno di Dirichlet omogenee, ovvero:

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = 0 & \text{su } \partial\Omega, \end{cases}$$

dove $\Omega \subset \mathbb{R}^2$. Osserviamo che in questo caso $u = u(\mathbf{x}) = u(x, y)$. Per semplicità consideriamo ora il problema precedente riferito a un dominio *rettangolare*, ovvero tale per cui $\Omega = (a, b) \times (c, d)$ (Figura 8.1). Il problema di Poisson si può dunque scrivere nel seguente modo:

$$\begin{cases} -\Delta u(x, y) = f(x, y) & (x, y) \in \Omega = (a, b) \times (c, d), \\ u(x, c) = u(x, d) = 0 & \text{per } x \in [a, b], \\ u(a, y) = u(b, y) = 0 & \text{per } y \in (c, d). \end{cases} \quad (8.35)$$

Osserviamo che per il problema precedente $\partial\Omega = \{\mathbf{x} = (x, y) \in \mathbb{R}^2 : x \in [a, b] \text{ se } y = c \text{ o } y = d, \text{ oppure } y \in (c, d) \text{ se } x = a \text{ o } x = b\}$.

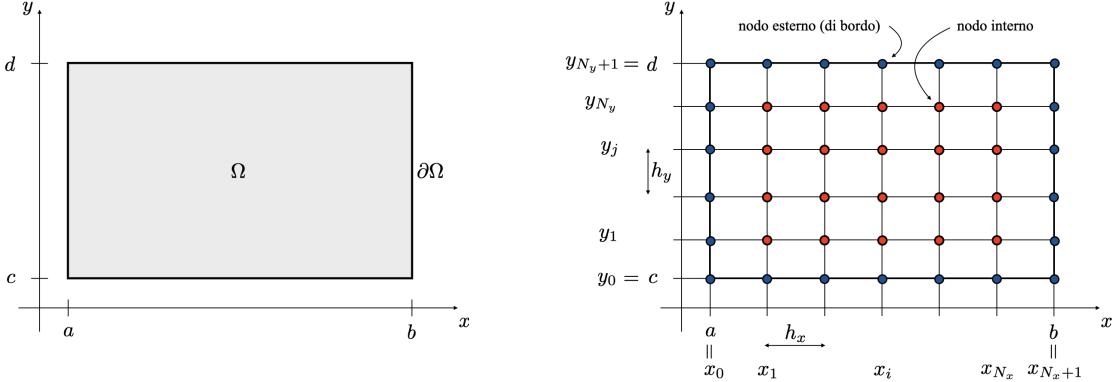


Figura 8.1: Dominio e griglia computazionale

Consideriamo il metodo delle differenze finite centrate per l'approssimazione del problema di Poisson (8.35). Iniziamo, come nel caso 1D, suddividendo l'intervallo $[a, b]$ in $N_x + 1$ sottointervalli di ampiezza

$$h_x = \frac{b - a}{N_x + 1}$$

e delimitati dai nodi $x_i = a + i h_x$ per $i = 0, \dots, N_x + 1$. In maniera analoga procediamo lungo la direzione y , suddividendo l'intervallo $[c, d]$ in $N_y + 1$ sottointervalli di ampiezza

$$h_y = \frac{d - c}{N_y + 1}$$

e delimitati dai nodi $y_j = c + j h_y$ per $j = 0, \dots, N_y + 1$. I nodi precedenti costituiscono gli insiemi $\Delta_x = \{x_0, x_1, \dots, x_{N_x}, x_{N_x+1}\}$ e $\Delta_y = \{y_0, y_1, \dots, y_{N_y}, y_{N_y+1}\}$. Il prodotto cartesiano di tali due insiemi costituisce la *griglia computazionale*

$$\Delta_h = \Delta_x \times \Delta_y,$$

caratterizzata da elementi di griglia (rettangoli) di *ampiezza*

$$h = \max \{h_x, h_y\}.$$

La griglia Δ_h è dotata di $(N_x + 1)(N_y + 1)$ nodi, di cui $N = N_x N_y$ nodi interni; si veda la Figura 8.1.

Collochiamo l'operatore Laplaciano in un nodo interno della griglia precedente (x_i, y_j) e consideriamo la sua approssimazione tramite le *differenze finite centrate*, ovvero

$$\begin{aligned} \Delta u(x, i, y_j) &= \frac{\partial^2 u}{\partial x^2}(x_i, y_j) + \frac{\partial^2 u}{\partial y^2}(x_i, y_j) \\ &\approx \frac{u(x_{i+1}, y_j) - 2u(x_i, y_j) + u(x_{i-1}, y_j)}{h_x^2} + \frac{u(x_i, y_{j+1}) - 2u(x_i, y_j) + u(x_i, y_{j-1})}{h_y^2} \end{aligned}$$

per ogni $i = 1, \dots, N_x$, $j = 1, \dots, N_y$.

Indichiamo ora con $u_{i,j}$ l'approssimazione di $u(x_i, y_j)$, ovvero nel generico nodo di griglia (x_i, y_j) , e sfruttiamo l'approssimazione precedente per ottenere lo schema alla differenze finite centrate per il problema di Poisson-Dirichlet:

$$\left\{ \begin{array}{l} -\frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h_x^2} - \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{h_y^2} = f(x_i, y_j) \\ \quad \text{per } i = 1, \dots, N_x, j = 1, \dots, N_y, \\ u_{i,0} = u_{i,N_y+1} = 0 \quad \text{per } i = 0, \dots, N_x + 1, \\ u_{0,j} = u_{N_x+1,j} = 0 \quad \text{per } j = 1, \dots, N_y + 1, \end{array} \right.$$

da cui otteniamo la formulazione dello schema noto come *schema a cinque punti*

$$\left\{ \begin{array}{l} -\frac{u_{i+1,j} + u_{i-1,j}}{h_x^2} + 2u_{i,j} \left(\frac{1}{h_x^2} + \frac{1}{h_y^2} \right) - \frac{u_{i,j+1} + u_{i,j-1}}{h_y^2} = f(x_i, y_j) \\ \quad \text{per } i = 1, \dots, N_x, j = 1, \dots, N_y, \\ u_{i,0} = u_{i,N_y+1} = 0 \quad \text{per } i = 0, \dots, N_x + 1, \\ u_{0,j} = u_{N_x+1,j} = 0 \quad \text{per } j = 1, \dots, N_y + 1. \end{array} \right.$$

Il precedente sistema lineare può essere riscritto in forma matriciale

$$A \mathbf{u}_h = \mathbf{b},$$

dove $A \in \mathbb{R}^{N \times N}$, $\mathbf{u}_h \in \mathbb{R}^N$ e $\mathbf{b} \in \mathbb{R}^N$ riordinando i nodi nel cosiddetto ordine *lessicografico* (Figura 8.2), in particolare tale per cui

$$\mathbf{u}_h = (u_{1,1}, \dots, u_{N_x,1}, u_{1,2}, \dots, u_{N_x,2}, \dots, u_{1,N_y}, \dots, u_{N_x,N_y})^T \in \mathbb{R}^N$$

e

$$\mathbf{b} = (f(x_1, y_1), \dots, f(x_{N_x}, y_1), f(x_1, y_2), \dots, f(x_{N_x}, y_2), \dots, f(x_1, y_{N_y}), \dots, f(x_{N_x}, y_{N_y}))^T \in \mathbb{R}^N.$$

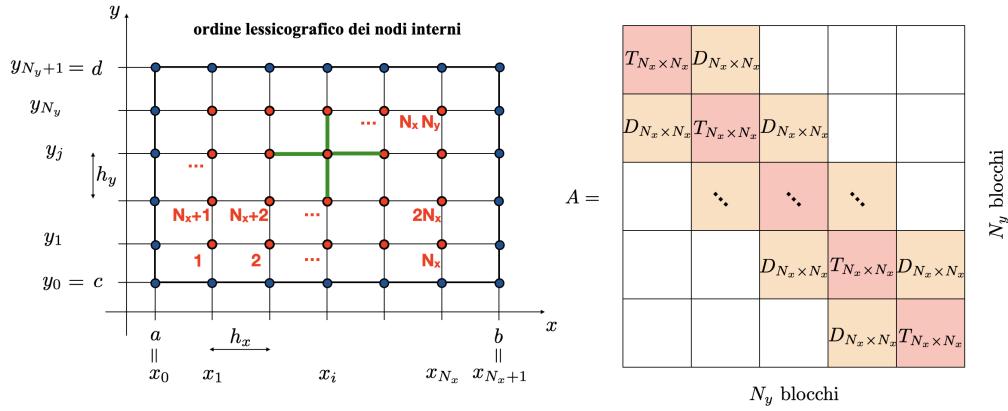


Figura 8.2: Riordinamento lessicografico (con raffigurazione dello schema a cinque punti) e matrice A

La matrice A è invece una matrice *tridiagonale a blocchi*, ovvero composta di N_y per N_y sottomatrici di dimensione N_x per N_x . In particolare,

$$A = \text{tridiag}_{N_y}(D_{N_x \times N_x}, T_{N_x \times N_x}, D_{N_x \times N_x}) \in \mathbb{R}^{N \times N},$$

dove

$$D_{N_x \times N_x} = -\frac{1}{h_y^2} I_{N_x \times N_x} \in \mathbb{R}^{N_x \times N_x},$$

e

$$T_{N_x \times N_x} = \text{tridiag}_{N_x} \left(-\frac{1}{h_x^2}, \frac{1}{h_x^2} + \frac{1}{h_y^2}, -\frac{1}{h_x^2} \right) \in \mathbb{R}^{N_x \times N_x}.$$

Osserviamo che la matrice A è sparsa, oltre che simmetrica e definita positiva. Si può inoltre mostrare che il suo numero di condizionamento è

$$K_2(A) = K(A) = \frac{C}{h^2},$$

dove $C > 0$ è una costante indipendente dalla dimensione caratteristica della griglia computazionale h .

In merito all'accuratezza dello schema a cinque punti, vale il seguente risultato.

Proposizione 8.3.1. Se $u \in C^4(\bar{\Omega})$, vale la seguente stima per l'errore e_h tra la soluzione esatta del problema di Poisson e la soluzione approssimata tramite lo schema cinque punti:

$$e_h = \max_{i=0, \dots, N_x+1, j=0, \dots, N_y+1} |u(x_i, y_j) - u_{i,j}| \leq Ch^2$$

per un'opportuna costante $C > 0$ indipendente da h , ovvero il metodo delle differenze finite converge con ordine 2 rispetto ad h .

8.4 Differenze Finite per l'Equazione del Calore (di Diffusione) 1D

Consideriamo infine l'approssimazione numerica di un problema ai limiti e ai valori iniziali per l'equazione del calore, per comprendere come, in casi come questi, una prima discretizzazione in spazio conduca a un problema semi-discretizzato che risulta un sistema di equazioni differenziali ordinarie, al quale applicare successivamente una tecnica di discretizzazione in tempo tra quelle già incontrate.

L'equazione del calore (o di diffusione) descrive la propagazione del calore in un mezzo omogeneo e isotropo (rispetto a tale fenomeno), come ad esempio un filo metallico. Tale equazione costituisce un caso particolare del modello riportato nell'equazione (8.9), in cui $u = u(x, t)$ assume il significato di temperatura e $V = \sigma = 0$, ovvero la propagazione del calore avviene solo per conduzione; effetti convettivi sarebbero inclusi considerando $V \neq 0$, mentre scambi termici ai bordi laterali del filo sarebbero considerati per $\sigma > 0$. Tale equazione si può ricavare in modo sostanzialmente analogo a quanto visto nella Sezione 8.1.1, ma a partire dalla legge di conservazione dell'energia anziché della massa.

In dimensione $d = 1$, il dominio spaziale su cui viene formulato il modello è un segmento, che può rappresentare un tratto di filo metallico (ovvero, una barra metallica di sezione cilindrica, di spessore trascurabile rispetto alla sua lunghezza). Assumiamo inoltre che il filo sia isolato sulla superficie laterale, ovvero che il calore possa fluire solo agli estremi $x = a$ e $x = b$ del filo. L'evoluzione della temperatura $u(x, t)$ in ogni punto $x \in (a, b)$ del filo, per ogni tempo $t > 0$, è descritta dall'equazione del calore, che modella il fenomeno del flusso di calore attraverso il filo, dalle condizioni al contorno (o condizioni al bordo) per descrivere la natura del problema agli estremi del filo, e da una condizione iniziale, che descrive il fenomeno all'inizio dell'esperimento.

Consideriamo per semplicità il seguente problema,

$$\begin{cases} \frac{\partial u}{\partial t} - \mu \frac{\partial^2 u}{\partial x^2} = f(x, t) & x \in (a, b), t > 0, \\ u(a, t) = u(b, t) = 0 & \text{per } t > 0 \\ u(x, 0) = u^0(x) & \text{per } x \in [a, b] \end{cases} \quad (8.36)$$

la cui soluzione $u : (a, b) \times (0, T) \rightarrow \mathbb{R}$, la temperatura, è funzione sia dello spazio che del tempo. Nel problema precedente, $f = f(x, t)$ è una sorgente di calore esterna, che può rappresentare una fonte di calore fisicamente collocata internamente al filo: ciò si verifica se ad esempio una corrente elettrica percorre una resistenza interna al filo. Il parametro $\mu > 0$ rappresenta il coefficiente di diffusione, ed è rappresentativo delle proprietà di conducibilità termica del materiale.

Il filo metallico dell'esempio ha dimensione finita e i suoi bordi $x = a$ e $x = b$ rappresentano il contorno del dominio del problema. Nel caso in esame, abbiamo assunto per semplicità che

$$u(a, t) = u(b, t) = 0, \quad t \in (0, T)$$

ovvero delle *condizioni al contorno* di Dirichlet omogenee. Infine, è necessario specificare la *condizione iniziale*, ovvero il problema evolutivo necessita della conoscenza di una configurazione di partenza della soluzione. Nel caso in esame, la condizione iniziale risulta:

$$u(x, 0) = u^0(x), \quad x \in (a, b).$$

Complessivamente, otteniamo quello che prende il nome di *problema di Cauchy-Dirichlet* per l'equazione del calore. Poiché x e t sono variabili indipendenti, procediamo con due discretizzazioni, partendo da quella in spazio.

8.4.1 Semi-discretizzazione in spazio

Introduciamo $N + 2$ nodi $\{x_j\}_{j=0}^{N+1}$ equispaziati nell'intervallo $[a, b]$, a distanza

$$h = \frac{b - a}{N + 1},$$

con $x_j = a + jh$, $j = 0, \dots, N + 1$, che individuano $N + 1$ sottointervalli. Osserviamo che $x_0 = a$ e che $x_{N+1} = b$, ovvero risultano N i nodi *interni* all'intervallo (a, b) , x_1, \dots, x_N . Indichiamo con $u_j(t)$ l'approssimazione di $u(x_j, t)$ per ogni $t > 0$, e per ogni $j = 0, \dots, N + 1$, con

$$u_j(0) = u^0(x_j).$$

Osserviamo come, a differenza dei problemi trattati in precedenza, in ogni nodo x_j non occorra determinare un valore, bensì una funzione del tempo, nel caso in esame. Una semi-discretizzazione spaziale dell'equazione del calore mediante differenze finite centrate può essere ottenuta approssimando la derivata seconda in spazio con una differenza finita centrale, analogamente a quanto fatto per l'equazione di Poisson, e risulta la seguente:

$$\begin{cases} \frac{d}{dt} u_j(t) - \frac{\mu}{h^2} (u_{j+1}(t) - 2u_j(t) + u_{j-1}(t)) = f(x_j, t) & j = 1, \dots, N, t > 0, \\ u_0(t) = u_{N+1}(t) = 0 & t > 0, \\ u_j(0) = u^0(x_j) & \text{per } j = 0, \dots, N + 1. \end{cases} \quad (8.37)$$

La semi-discretizzazione in spazio genera dunque un sistema di N equazioni differenziali ordinarie, di ordine 1, lineare, non omogeneo, a coefficienti costanti, che può essere riscritto in modo equivalente come

$$\boxed{\begin{cases} \frac{d\mathbf{u}_h}{dt}(t) + \frac{\mu}{h^2} A \mathbf{u}_h(t) = \mathbf{f}(t) & t > 0, \\ \mathbf{u}_h(0) = \mathbf{u}^0, \end{cases}} \quad (8.38)$$

dove $A \in \mathbb{R}^{N \times N}$, $A = \text{tridiag}_{N \times N}(-1, 2, -1)$ e

$$\mathbf{u}_h(t) = \begin{bmatrix} u_1(t) \\ u_2(t) \\ \vdots \\ u_N(t) \end{bmatrix}, \quad \mathbf{f}(t) = \begin{bmatrix} f(x_1, t) \\ f(x_2, t) \\ \vdots \\ f(x_N, t) \end{bmatrix}, \quad \mathbf{u}^0 = \begin{bmatrix} u^0(x_1) \\ u^0(x_2) \\ \vdots \\ u^0(x_N) \end{bmatrix}.$$

8.4.2 Discretizzazione in tempo

Possiamo sfruttare il θ -metodo (Sezione 7.5.3) per discretizzare in tempo il sistema di EDO (8.38). A tale scopo, introduciamo una suddivisione dell'intervallo temporale $[0, T]$ in N_t istanti $t^k = k\Delta t$, $k = 0, 1, \dots, N_t$ dove

$$\Delta t = \frac{T}{N_t}$$

è il passo di discretizzazione temporale, e indichiamo con $\mathbf{u}_h^k \in \mathbb{R}^N$ l'approssimazione di $\mathbf{u}_h(t^k)$. Usando il θ -metodo (si veda l'equazione (8.39)) otteniamo il seguente problema discreto: fissato $\theta \in [0, 1]$, trovare

$\{\mathbf{u}_h^k\}_{k=0}^{N_t}$ tale che

$$\begin{cases} \frac{\mathbf{u}_h^{k+1} - \mathbf{u}_h^k}{\Delta t} = -\frac{\mu}{h^2} A (\theta \mathbf{u}_h^{k+1} + (1-\theta)\mathbf{u}_h^k) + \theta \mathbf{f}^{k+1} + (1-\theta)\mathbf{f}^k & \text{per } k = 0, 1, \dots, N_t - 1, \\ \mathbf{u}_h^0 = \mathbf{u}^0. \end{cases} \quad (8.39)$$

dove abbiamo indicato con $\mathbf{f}^k = \mathbf{f}(t^k)$. In modo equivalente, si ottiene, a partire da $\mathbf{u}_h^0 = \mathbf{u}^0$, la seguente espressione ad ogni passo $k = 0, 1, \dots, N_t - 1$:

$$\left(I + \theta \frac{\mu}{h^2} \Delta t A \right) \mathbf{u}_h^{k+1} = \left(I - (1-\theta) \frac{\mu}{h^2} \Delta t A \right) \mathbf{u}_h^k + \mathbf{g}^{k+1} \quad \text{per } k = 0, \dots, N_t - 1 \quad (8.40)$$

dove

$$\mathbf{g}^{k+1} = \theta \mathbf{f}^{k+1} + (1-\theta)\mathbf{f}^k \quad \text{per } k = 0, \dots, N_t - 1.$$

In particolare, abbiamo che:

- se $\theta = 0$, il metodo di Eulero in avanti, esplicito, permette di ottenere, ad ogni passo di tempo

$$\mathbf{u}_h^{k+1} = \left(I - \frac{\mu}{h^2} \Delta t A \right) \mathbf{u}_h^k + \mathbf{g}^{k+1} \quad \text{per } k = 0, \dots, N_t - 1,$$

ovvero non occorre risolvere alcun sistema lineare per calcolare \mathbf{u}_h^{k+1} a partire da \mathbf{u}_h^k ;

- se $\theta > 0$, il metodo risulta implicito e richiede la soluzione di un sistema lineare della forma (8.40) ad ogni passo di tempo; si noti che la matrice del sistema dipende dalla discretizzazione temporale (e in particolare da Δt), ma non da k . Inoltre, poiché la matrice A è simmetrica e definita positiva, anche la matrice $\left(I + \theta \frac{\mu}{h^2} \Delta t A \right)$ risulta simmetrica e definita positiva; in particolare, tale matrice può essere fattorizzata una volta per tutte al tempo $t = 0$ – in dimensione $d = 1$, come abbiamo avuto già modo di osservare nel caso del problema di Poisson, tale fattorizzazione è basata sull'algoritmo di Thomas e richiede $O(8N)$ operazioni.

Casi particolari sono:

- se $\theta = 1$, il metodo di Eulero all'indietro, dove

$$\left(I + \frac{\mu}{h^2} \Delta t A \right) \mathbf{u}_h^{k+1} = \mathbf{u}_h^k + \mathbf{g}^{k+1} \quad \text{per } k = 0, \dots, N_t - 1;$$

- se $\theta = 1/2$, il metodo di Crank-Nicolson,

$$\left(I + \frac{\mu}{2h^2} \Delta t A \right) \mathbf{u}_h^{k+1} = \left(I - \frac{\mu}{2h^2} \Delta t A \right) \mathbf{u}_h^k + \mathbf{g}^{k+1} \quad \text{per } k = 0, \dots, N_t - 1.$$

Per quanto riguarda l'accuratezza di tale schema numerico, vale il seguente risultato:

Proposizione 8.4.1. Se $f \in C^2([a, b] \times [0, T])$, vale la seguente stima per l'errore e_h^k tra la soluzione esatta e la soluzione approssimata ottenuta risolvendo il sistema lineare (8.39) derivante dall'approssimazione con il θ -metodo e le differenze finite centrate:

$$e_h^k = \max_{j=0, \dots, N+1} |u(x_j, t^k) - u_j^k| \leq C \left(h^2 + \Delta t^{\eta(\theta)} \right)$$

per un'opportuna costante $C > 0$ indipendente da h , dove

$$\eta(\theta) = \begin{cases} 1 & \text{se } \theta \in [0, 1], \theta \neq 1/2, \\ 2 & \text{se } \theta = 1/2. \end{cases}$$

Notiamo come lo schema ottenuto con il θ -metodo e le differenze finite centrate abbia una *reminescenza* dell'accuratezza dell'approssimazione in spazio (le DF centrate sono accurate di ordine 2 rispetto ad h) e in tempo. In particolare, uno schema alle DF centrate che sfrutti il metodo di Crank-Nicolson per la discretizzazione in tempo converge con ordine 2 sia rispetto ad h che a Δt .

8.4.3 Stabilità (asintotica)

Nel caso in cui $f(x, t) = 0$, la soluzione esatta $u(x, t)$ dell'equazione del calore (8.36) tende a zero per ogni x , quando $t \rightarrow \infty$. Se anche la soluzione numerica mostra lo stesso tipo di comportamento asintotico, ovvero $u_j^k \rightarrow 0$ quando $k \rightarrow \infty$, per ogni $j = 0, \dots, N+1$, diciamo che lo schema numerico (8.40) è detto *asintoticamente stabile*, un concetto analogo a quello di assoluta stabilità introdotto nel caso delle equazioni differenziali ordinarie nel capitolo precedente. Allo stesso modo in cui l'assoluta stabilità richiede, eventualmente, una condizione sul passo di discretizzazione temporale nel caso dei metodi numerici per le EDO, anche nel caso della discretizzazione spazio-temporale dell'equazione del calore può sussistere una condizione tra h e Δt .

Se $f = 0$, allora $\mathbf{g}^{k+1} = \theta \mathbf{f}^{k+1} + (1 - \theta) \mathbf{f}^k = 0$ per ogni $k = 0, \dots, N_t - 1$, ovvero dal θ -metodo (8.40) ricaviamo che

$$\left(I + \theta \frac{\mu}{h^2} \Delta t A \right) \mathbf{u}_h^{k+1} = \left(I - (1 - \theta) \frac{\mu}{h^2} \Delta t A \right) \mathbf{u}_h^k \quad \text{per } k = 0, \dots, N_t - 1.$$

Nel caso del metodo di Eulero in avanti ($\theta = 0$) si ha che

$$\mathbf{u}_h^{k+1} = \left(I - \frac{\mu}{h^2} \Delta t A \right) \mathbf{u}_h^k \quad \text{per } k = 0, \dots, N_t - 1,$$

ovvero, in modo ricorsivo, otteniamo

$$\mathbf{u}_h^k = \left(I - \frac{\mu}{h^2} \Delta t A \right)^k \mathbf{u}_h^0 \quad \text{per } k = 0, \dots, N_t - 1,$$

e di conseguenza si ha che $\mathbf{u}^k \rightarrow \mathbf{0}$ per $k \rightarrow \infty$ a patto che

$$\rho \left(I - \frac{\mu}{h^2} \Delta t A \right) < 1. \tag{8.41}$$

In questo caso specifico, poiché gli autovalori della matrice A sono dati da

$$\lambda_j(A) = 4 \sin^2 \left(\frac{j\pi}{2(N+1)} \right) \quad \text{per } j = 1, \dots, N,$$

otteniamo che

$$\lambda_j \left(I - \frac{\mu}{h^2} \Delta t A \right) = 1 - \frac{\mu}{h^2} \Delta t \lambda_j(A) \quad \text{per } j = 1, \dots, N;$$

in particolare, tali autovalori sono tutti reali e positivi. Richiedere dunque che valga la condizione (8.41) significa richiedere che

$$-1 < 1 - \frac{\mu}{h^2} \Delta t \lambda_j(A) < 1 \quad \text{per } j = 1, \dots, N,$$

da cui si ottiene la seguente condizione di stabilità asintotica per il metodo delle differenze finite centrate in spazio, e di Eulero in avanti in tempo:

$$\Delta t < \frac{h^2}{2\mu}.$$

Come era logico aspettarsi, in maniera analoga a quanto visto nel caso delle EDO, il metodo di Eulero in avanti è asintoticamente stabile sotto la condizione che Δt decresca come il quadrato del parametro h di discretizzazione spaziale quando $h \rightarrow 0$.

Nel caso del metodo di Eulero all'indietro ($\theta = 1$), otteniamo che

$$\mathbf{u}_h^{k+1} = \left(I + \frac{\mu}{h^2} \Delta t A \right)^{-1} \mathbf{u}_h^k \quad \text{per } k = 0, \dots, N_t - 1$$

e si ha che $\mathbf{u}^k \rightarrow \mathbf{0}$ per $k \rightarrow \infty$ a patto che

$$\rho \left(I + \frac{\mu}{h^2} \Delta t A \right)^{-1} < 1. \tag{8.42}$$

Quest'ultima condizione è tuttavia sempre verificata, dal momento che il più piccolo autovalore di A è comunque positivo,

$$\lambda_{\min}(A) = \lambda_N(A) = 4 \sin^2 \left(\frac{N\pi}{2(N+1)} \right) > 0$$

e dunque, poiché $\left(I + \frac{\mu}{h^2} \Delta t A \right)$ ha autovalori tutti maggiori di 1 in modulo, la sua inversa ha autovalori tutti minori di 1 in modulo, e quindi la condizione (8.42) è sempre verificata. In conclusione, questo schema è incondizionatamente asintoticamente stabile, ovvero non occorre prescrivere alcuna restrizione sulla scelta di Δt rispetto ad h .

Più in generale, lo schema ottenuto con le differenze finite centrate in spazio e il θ -metodo in tempo è incondizionatamente asintoticamente stabile per tutti i valori $1/2 \leq \theta \leq 1$, e condizionatamente asintoticamente stabile se $0 \leq \theta < 1/2$.