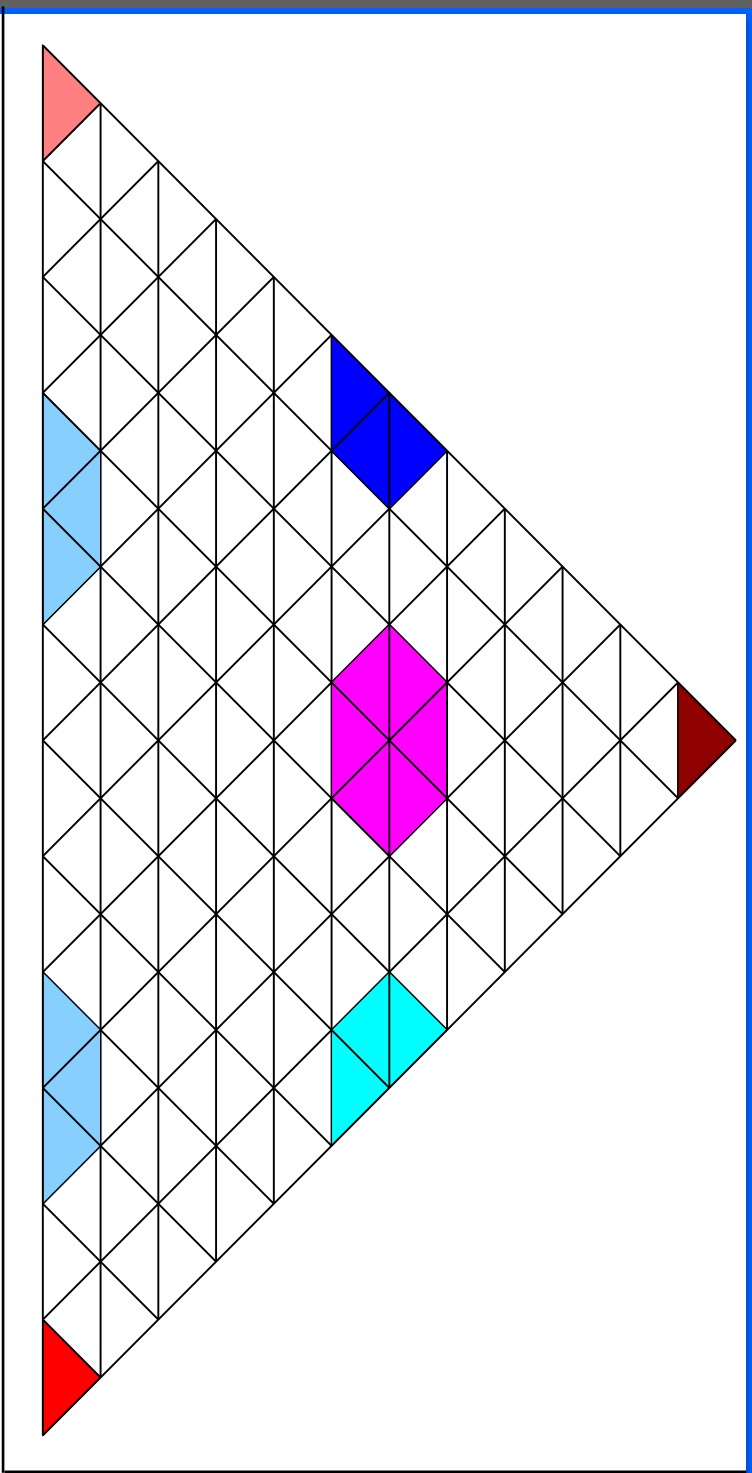
SIMAI E-Lecture notes Vol 13

ISSN 1970-4429

Sebastiano Seatzu
Cornelis Van der Mee
Pietro Contu



MATEMATICA APPLICATA

Un secondo corso

SIMAI
Società Italiana di Matematica Applicata e Industriale

Matematica Applicata

un secondo corso

Sebastiano Seatzu, Cornelis Van der Mee, Pietro Contu

Dipartimento di Matematica e Informatica
Università di Cagliari
viale Merello 92, 09123 Cagliari, Italia
seatzu@unica.it, cornelis@krein.unica.it

Rockley Photonics 234 E Colorado Blvd Suite 600,
Pasadena, CA 91101, USA
pietro.contu@rockleyphotonics.com

Vol. 13- 2017

ISBN-A: 10.978.88905708/10

Published by:
SIMAI - Società Italiana di Matematica Applicata e Industriale
Via dei Taurini, 19 c/o IAC/CNR
00185, ROMA (ITALY)

SIMAI e-Lecture Notes
ISSN: 1970-4429

Volume 13, 2013
ISBN-13: 978-88-905708-1-0
ISBN-A: 10.978.88905708/10

Dedicato ai nostri parenti

Indice

PREFAZIONE

Obiettivo primario del libro è la presentazione dei metodi basilari nella risoluzione analitica e numerica delle Equazioni Differenziali. Al fine di ridurre al minimo il ricorso ad altri testi, nel libro vengono richiamati i metodi e i risultati necessari alla loro reale comprensione.

Destinatari primi del libro sono gli studenti delle Lauree Magistrali in Ingegneria. Per questo motivo vengono presentati i risultati analitici e numerici che maggiormente interessano le applicazioni, rinviando a libri specialistici i risultati teorici e modellistici relativi a situazioni più complesse. È questa la ragione per la quale sono state omesse le dimostrazioni su diversi risultati analitici e numerici utilizzati. Il libro potrebbe essere adottato nella Laurea Magistrale in Matematica, a condizione che vengano apportate integrazioni sulle proprietà analitiche delle soluzioni delle equazioni differenziali e precisazioni varie sulla esistenza e unicità della soluzione in domini più generali. Considerazioni analoghe, in un certo senso complementari, valgono per la Laurea Magistrale in Fisica. In questo caso le integrazioni dovrebbero riguardare la parte modellistica. Sarebbe infatti opportuno integrare il libro con la presentazione di ulteriori problematiche fisiche, allo scopo di evidenziare maggiormente che le equazioni differenziali sono lo strumento matematico più utilizzato per descrivere le relazioni tra le diverse entità fisiche (massa, posizioni, forze, energie, momenti, ecc.). Le equazioni differenziali sono infatti utilizzate per rappresentare fenomeni stazionari ed evolutivi nella generalità dei settori dellIngegneria, della Fisica e di numerosissimi altri settori delle scienze applicate. Negli ultimi due decenni il loro utilizzo si è esteso a settori non tradizionali, come quello delle nanotecnologie, nel quale si presentano spesso con caratteristiche nuove, aprendo così nuovi fronti per la ricerca matematica. Due esempi relativi a questo settore sono illustrati nei Capitoli 6 e 8. Anche se buona parte degli argomenti trattati sono contenuti in [26], i due libri presentano importanti differenze. In primo luogo perché i cambiamenti, apportati nella presentazione degli argomenti, ne hanno accresciuto la leggibilità, in secondo luogo perché gli ampliamenti introdotti in alcuni capitoli lo hanno reso più completo, riducendo significativamente la necessità di ricorrere ad altri testi. Anche se la sua focalizzazione riguarda i metodi di risoluzione analitica e numerica delle equazioni

differenziali a derivate parziali, alcuni capitoli sono dedicati ai richiami e alle metodologie, le cui conoscenze sono essenziali alla loro effettiva comprensione. Dei suoi otto capitoli, il primo e il quarto riguardano la risoluzione analitica delle equazioni a derivate parziali. Più precisamente, mentre il primo capitolo contiene una presentazione di caratttere generale sulle equazioni alle derivate parziali con relativa classificazione, il quarto è dedicato alla presentazione dei due metodi di risoluzione analitica maggiormente utilizzati, con applicazioni a vari problemi di tipo ellittico, parabolico e iperbolico.

Il secondo richiama i risultati essenziali sulle equazioni differenziali ordinarie, con una certa enfasi sui metodi spettrali, dato che questi sono i più utilizzati nella risoluzione analitica di quelle alle derivate parziali. Il capitolo terzo è funzionale al quarto, in quanto la generalità delle soluzioni delle equazioni alle derivate parziali è rappresentabile mediante serie di Fourier o di autofunzioni relative a problemi di Sturm-Liouville. Per questo motivo, in aggiunta ai richiami essenziali, vengono messe in evidenza le proprietà di decadimento dei coefficienti delle serie di Fourier in funzione della regolarità della soluzione.

Nel quinto, dedicato allalgebra lineare, oltre ai richiami di carattere generale sugli spazi vettoriali, autovalori-autovettori e norme vettoriali e matriciali, vengono richiamate le proprietà spettrali basilari nella risoluzione dei sistemi lineari mediante i metodi iterativi. Vengono quindi presentati alcuni dei metodi iterativi più utilizzati nella risoluzione numerica dei sistemi lineari tipici delle equazioni a derivate parziali.

Il sesto è dedicato alla risoluzione delle equazioni differenziali ordinarie e alle derivate parziali mediante i metodi alle differenze finite. Nel primo caso viene considerata la risoluzione di problemi ai limiti per equazioni lineari e debolmente lineari del secondo ordine. Nel secondo caso la risoluzione numerica di problemi ellittici, parabolici e iperbolici lineari su domini rettangolari. Nel caso ellittico viene altresì considerata la risoluzione di problemi debolmente non lineari.

Il settimo capitolo è dedicato alla risoluzione numerica dei sistemi non lineari. Dopo un presentazione generale dei metodi di risoluzione e di alcuni metodi iterativi, vengono messe in evidenza le principali analogie e differenze con i metodi iterativi per i sistemi lineari. Il capitolo è corredato da esempi esplicativi e da applicazioni riguardanti la risoluzione numerica delle equazioni differenziali debolmente non lineari.

Il capitolo ottavo è il più vasto. Dopo aver richiamato alcune definizioni essenziali di calcolo vettoriale-differenziale, si procede alla formulazione variazionale dei problemi differenziali considerati, evidenziando le principali differenze tra le formulazioni classiche e quelle variazionali. Vengono quindi introdotti gli spazi di Sobolev, con la presentazione delle proprietà di specifico interesse per

il settore. Si passa quindi alla risoluzione dei problemi ai limiti per le equazioni differenziali lineari ordinarie e dei problemi con valori al bordo per quelle parziali di tipo ellittico. Per queste vengono considerati problemi con valori noti della soluzione al bordo, valori assegnati della derivata normale sulla frontiera e con condizioni miste.

L'ultima parte riguarda la risoluzione numerica dei problemi di tipo parabolico e iperbolico. Come nel capitolo sulle differenze finite, viene illustrata la risoluzione di un problema spettrale di Helmholtz, tipico di alcuni settori ingegneristici. Nell'Appendice A sono infine riportati i risultati dell'Analisi Funzionale, più frequentemente utillizzate nella Matematica Applicata.

Capitolo 1

INTRODUZIONE ALLE EQUAZIONI ALLE DERIVATE PARZIALI (PDEs)

1.1 Preliminari

Per equazione alle derivate parziali di ordine n in r variabili, si intende una relazione che stabilisca un legame funzionale tra una funzione incognita

$$u(x_1, x_2, \ldots, x_r)$$

e le derivate parziali prime $\frac{\partial u}{\partial x_i}$, seconde $\frac{\partial^2 u}{\partial x_i \partial x_j}$, ..., n-esime $\frac{\partial^n u}{\partial x_1^{k_1} \partial x_2^{k_2} \ldots \partial x_r^{k_r}}$, con $k_1 + k_2 + \ldots + k_r = n$, della stessa funzione.

Naturalmente nell'equazione possono non figurare varie derivate parziali. Perché l'ordine sia n, deve comparire almeno una delle sue derivate parziali di ordine n. Nel caso $r = 2$, un'equazione del primo ordine è del tipo

$$f\left(x, y; u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\right) = 0,$$

mentre una del secondo ordine ha la forma

$$f\left(x, y; u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial^2 u}{\partial x^2}, \frac{\partial^2 u}{\partial x \partial y}, \frac{\partial^2 u}{\partial y^2}\right) = 0.$$

Una funzione u che, assieme alle derivate parziali coinvolte, soddisfi l'equazione è detta soluzione o integrale dell'equazione. In accordo con la notazione internazionale, le equazioni differenziali alle derivate parziali verranno indicate come PDEs (Partial Differential Equations) e quelle ordinarie come ODEs (Ordinary Differential Equations).

Una fondamentale differenza fra le PDEs e le ODEs è che, mentre la soluzione generale delle ODEs, che rappresenta la totalità dei suoi integrali, dipende da n costanti arbitrarie (nel caso l'ordine sia n); nelle PDEs, di ordine n, l'integrale generale dipende da n funzioni arbitrarie.

Esempio 1.1 La ODE del primo ordine

$$y'(x) = f(x), \quad a \le x \le b, \quad f(x) \text{ continua in } [a, b],$$

ha come integrale generale la funzione

$$y(x) = \int_a^x f(t)dt + c,$$

dove c è una costante arbitraria.

La PDE del primo ordine $\frac{\partial u}{\partial x} = f(x, y)$, con $f(x, y)$ funzione continua nell'intervallo $a \le x \le b$, $c \le y \le d$, ha come integrale generale la funzione

$$u(x, y) = \int_a^x f(t, y)\, dt + \varphi(y),$$

dove $\varphi(y)$ è una funzione continua, peraltro arbitraria nell'intervallo $[c, d]$.

Esempio 1.2 La ODE del secondo ordine

$$y'' = f(x), \quad a \le x \le b, \text{ con } f(x) \text{ nota e continua in } [a, b],$$

ha come integrale generale la funzione

$$y(x) = \int_a^x \left(\int_a^t f(\tau)\, d\tau \right) dt + c_1 x + c_2,$$

essendo c_1 e c_2 costanti arbitrarie.

La PDE del secondo ordine

$$\frac{\partial^2 u}{\partial x \partial y} = f(x, y), \quad a \le x \le b, \quad c \le y \le d,$$

con $f(x, y)$ funzione nota e continua in $[a, b] \times [c, d]$, ha come soluzione generale un'integrale dipendente da due funzioni arbitrarie. Infatti, integrando l'equazione rispetto ad x, segue immediatamente che

$$\frac{\partial u}{\partial y} = \int_a^x f(t, y)\, dt + \varphi(y),$$

essendo φ una funzione continua e arbitraria nell'intervallo $[c, d]$. Da essa, integrando rispetto ad x, segue che

$$u(x, y) = \int_c^y \left(\int_a^x f(t, z)\, dt \right) dz + \int_a^y \varphi(z)\, dz + \psi(x),$$

essendo $\psi(x)$ una funzione continua, peraltro arbitraria, nell'intervallo $[a, b]$.

Il problema della soluzione generale di una PDE è, in generale, impossibile da risolvere in modo esplicito. Per fortuna, nelle applicazioni, esso non è generalmente di effettivo interesse. Interessa molto di più determinare l'unica soluzione che, unitamente ad altre condizioni prefissate, soddisfa il problema differenziale, ossia la PDE e le condizioni assegnate.

Nelle ODEs di ordine n, tipicamente, l'unicità della soluzione è ottenuta imponendo n condizioni iniziali (problema di Cauchy), oppure n condizioni lineari negli estremi di integrazione (problema con valori agli estremi). Nelle PDEs, il problema di Cauchy fondamentalmente riguarda le equazioni del primo ordine

$$f\left(x, y; u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\right) = 0. \tag{1.1}$$

Esso consiste nella determinazione di un integrale $u(x, y)$ che, nei punti di una curva regolare γ del piano xy, assuma dei valori prefissati. Supponendo che la curva γ sia assegnata mediante equazioni parametriche del tipo $x = \alpha(t)$, $y = \beta(t)$, $a \leq t \leq b$, i valori prefissati di $u(x, y)$ su γ risultano assegnati in funzione di t e, di conseguenza, si può dire che si cerca un integrale $u(x, y)$ della (1.1) che verifica la condizione iniziale

$$u[\alpha(t), \beta(t)] = \varphi(t), \quad a \leq t \leq b, \tag{1.2}$$

essendo $\varphi(t)$ una funzione continua, peraltro arbitraria, assegnata su $[a, b]$. Esso rappresenta la naturale estensione del problema di Cauchy per un'equazione $f(x, y, y') = 0$, con la condizione iniziale $y(x_0) = y_0$.

Geometricamente il problema (1.1)-(1.2) equivale alla determinazione della superficie integrale $u = u(x, y)$, soluzione della (1.1), che passa per la curva $x = \alpha(t)$, $y = \beta(t)$, $u = \varphi(t)$ dello spazio x, y, u con $a \leq t \leq b$.

Per maggiore chiarezza, consideriamo l'equazione

$$a(x, y)\frac{\partial u}{\partial x} + b(x, y)\frac{\partial u}{\partial y} = 0, \tag{1.3}$$

la quale, sotto l'ipotesi di continuità dei coefficienti, possiede infinite soluzioni dipendenti da una funzione arbitraria. Per la sua determinazione consideriamo una soluzione del tipo

$$\varphi(x, y; u(x, y)) = \text{costante}$$

con φ funzione da determinare. A tale scopo osserviamo che, per la derivazione delle funzioni composte, la φ soddisfa il sistema

$$\begin{cases} \dfrac{\partial \varphi}{\partial x} + \dfrac{\partial \varphi}{\partial u}\dfrac{\partial u}{\partial x} = 0, \\ \dfrac{\partial \varphi}{\partial y} + \dfrac{\partial \varphi}{\partial u}\dfrac{\partial u}{\partial y} = 0. \end{cases} \tag{1.4}$$

Da cui, moltiplicando la prima equazione per $\frac{\partial u}{\partial y}$ e la seconda per $\frac{\partial u}{\partial x}$ e sottraendo la prima dalla seconda. si ottiene l'equazione

$$\frac{\partial \varphi}{\partial y}\frac{\partial u}{\partial x} - \frac{\partial \varphi}{\partial x}\frac{\partial u}{\partial y} = 0.$$

Di conseguenza la $u(x, y)$ è soluzione della (1.3), nel caso che la φ soddisfi le condizioni:

$$\frac{\partial \varphi}{\partial y} = a(x, y) \quad e \quad \frac{\partial \varphi}{\partial x} = -b(x, y). \tag{1.5}$$

In questo caso, anche la $u = F[\varphi(x, y; u)]$, qualunque sia la F, purchè derivabile rispetto alla φ, è soluzione delle (1.3). Per verificarlo, basta osservare che per la (1.4),

$$\begin{aligned} &a(x, y)\frac{\partial F}{\partial x} + b(x, y)\frac{\partial F}{\partial y} \\ &\qquad = a(x, y)\frac{\partial F}{\partial \varphi}\left(\frac{\partial \varphi}{\partial x} + \frac{\partial \varphi}{\partial u}\frac{\partial u}{\partial x}\right) + b(x, y)\frac{\partial F}{\partial \varphi}\left(\frac{\partial \varphi}{\partial y} + \frac{\partial \varphi}{\partial u}\frac{\partial u}{\partial y}\right) \equiv 0. \end{aligned}$$

Esempio 1.3 Determinare la soluzione generale dell'equazione

$$3\frac{\partial u}{\partial x} + 2\frac{\partial u}{\partial y} = 0. \tag{1.6}$$

In virtù della (1.5)

$$\frac{\partial \varphi}{\partial y} = 3 \quad \text{e} \quad \frac{\partial \varphi}{\partial x} = -2,$$

da cui $\varphi(x, y) = 3y + f(x)$ e $\varphi(x, y) = -2x + g(y)$, con $f(x)$ e $g(y)$ funzioni continue ed arbitrarie. Di conseguenza, prendendo $f(x) = -2x$ e $g(y) = 3y$, una soluzione particolare è

$$\varphi(x, y) = -2x + 3y,$$

mentre la soluzione generale è

$$u(x, y) = F(-2x + 3y),$$

con F arbitraria, purchè derivabile rispetto a x e y. Per la verifica basta osservare che

$$\frac{\partial u}{\partial x} = -2F'(-2x + 3y) \quad \text{e} \quad \frac{\partial u}{\partial y} = 3F'(-2x + 3y),$$

da cui segue la (2.6).

Determinare ora la soluzione del problema di Cauchy

$$\begin{cases} 3\dfrac{\partial u}{\partial x} + 2\dfrac{\partial u}{\partial y} = 0, \\ u(x,0) = \sin x. \end{cases} \tag{1.7}$$

Essendo $F(-2x+3y)$ la soluzione generale, dobbiamo imporre che

$$F(-2x) = \sin x,$$

ossia che risulti

$$F(t) = \sin\left(-\frac{t}{2}\right), \quad t = -2x.$$

Da cui, posto $t = -2x + 3y$, la soluzione del problema (1.7) è

$$u(x,y) = \sin\left(-\tfrac{1}{2}(-2x+3y)\right) = \sin\left(x - \tfrac{3}{2}y\right),$$

come è immediato verificare per sostituzione diretta.

L'estensione di questo problema alle PDEs di ordine $n \geq 2$, oltre che di difficilissima soluzione, ha scarso interesse applicativo. Per questo motivo, nel seguito, per tali equazioni faremo riferimento soltanto ad altri tipi di condizioni.

Passiamo ora alla descrizione dei metodi analitici più comunemente usati nella risoluzione esatta delle PDEs lineari del secondo ordine, che rappresenta l'argomento fondamentale del capitolo. Per la loro descrizione è indispensabile fare ricorso alle serie di Fourier e alle proprietà spettrali tipiche dei problemi di Sturm-Liouville. Le definizioni, proprietà e risultati sulle serie di Fourier e sui problemi di Sturm-Liouville , funzionali alla illustrazione dei metodi sulle PDEs, sono contenuti nel Capitolo 3.

1.2 PDEs del secondo ordine

Una PDE del secondo ordine è definita lineare quando in essa la u e le sue derivate parziali del primo e del secondo ordine compaiono linearmente. Nel caso le variabili indipendenti siano due, essa è rappresentabile nel modo seguente:

$$\begin{aligned} &a_{11}(x,y)u_{xx} + a_{12}(x,y)u_{xy} + a_{22}(x,y)u_{yy} \\ &\qquad + b_1(x,y)u_x + b_2(x,y)u_y + c(x,y)u = f(x,y), \end{aligned} \tag{1.8}$$

dove i coefficienti $a_{11}(x,y), \ldots, c(x,y)$ e il termine noto $f(x,y)$ rappresentano funzioni che, per semplicità, supponiamo continue in un assegnato dominio Ω del piano xy. In essa, come usuale, $u_x = \frac{\partial u}{\partial x}$, $u_y = \frac{\partial u}{\partial y}$, $u_{xx} = \frac{\partial^2 u}{\partial x^2}$, $u_{xy} = \frac{\partial^2 u}{\partial x \partial y}$ e $u_{yy} = \frac{\partial^2 u}{\partial y^2}$. L'equazione è detta omogenea se $f(x,y) = 0$ in Ω.

Per le equazioni del tipo (1.8) il problema di Cauchy consiste nel richiedere che la u e la sua derivata normale $\frac{\partial u}{\partial n}$ assumano valori prefissati lungo i punti di una curva γ definita nel dominio Ω dell'equazione differenziale. L'imposizione di questa condizione induce tre situazioni, sostanzialmente diverse, a seconda che il discriminante

$$\Delta(x,y) = a_{12}(x,y)^2 - 4a_{11}(x,y)a_{22}(x,y), \quad (x,y) \in \Omega, \tag{1.9}$$

sia negativo, uguale a zero o positivo. Nel primo caso ($\Delta < 0$) l'equazione è detta di tipo *ellittico*, nel secondo caso ($\Delta = 0$) di tipo *parabolico* e, nel terzo caso ($\Delta > 0$), di tipo *iperbolico*. Essa è detta di tipo misto se, in un subdominio di Ω è di un tipo e in un altro di tipo diverso.

Se i coefficienti delle derivate di secondo ordine sono costanti, la classificazione è la stessa in tutti i punti del dominio Ω. Diversamente essa può dipendere dal punto del dominio, come nel seguente esempio di Tricomi:

$$y\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0.$$

In questo caso, essendo $a_{11}(x,y) = y$, $a_{12}(x,y) = 0$, $a_{22}(x,y) = 1$ e

$$\Delta(x,y) = A_{12}(x,y)^2 - 4\, a_{11}(x,h)b_{21}(x,y) = -4y.$$

Di conseguenza l'equazione è ellitiica per $y > 0$, parabolica per $y = 0$ e iperbolica per $y < 0$. Se il dominio contiene valori di $y > 0$ e $y \neq 0$, l'equazione è di tipo misto. Se ovunque in Ω, è $y > 0$, l'equazione è di tipo ellittico in Ω, se è $y = 0$, è di tipo parabolico ed è di tipo iperbolico se $y < 0$.

Per le equazioni di tipo ellittico, il problema di Cauchy ha pochissima importanza, mentre sono di grande importanza i problemi detti di Dirichlet e di Neumann, il cui significato verrà precisato nel seguito.

Per quelle di tipo parabolico, interessa pochissimo il problema di Cauchy e interessa molto, come vedremo nel seguito, il problema di *propagazione non vibratoria*.

In quelle di tipo iperbolico, il problema di Cauchy è di qualche interesse applicativo, anche se quello che maggiormente interessa è quello detto di *propagazione vibratoria*, il cui significato verrà chiarito nel seguito.

Nelle applicazioni, le equazioni ellittiche rappresentano modelli di tipo stazionario a differenza di quelli parabolici, che rappresentano modelli evolutivi di prima specie, e di quelli iperbolici, che rappresentano problemi evolutivi di seconda specie.

Una PDE del secondo ordine, in due variabili, è definita debolmente nonlineare quando in essa la u non compare linearmente, a differenza delle sue

derivate parziali che compaiono tutte linearmente. Essa è dunque del tipo

$$a_{11}(x,y)u_{xx} + a_{12}(x,y)u_{xy} + a_{22}(x,y)u_{yy} \\ + b_1(x,y)u_x + b_2(x,y)u_y + c(x,y;u) = f(x,y)$$

con $c(x,y;u)$ nonlineare in u.

Esempio 1.4 L'equazione seguente

$$\begin{cases} u_{xx} + u_{yy} + (e^x \sin y)u_x + (e^y \sin x)u_y - (x+y)^2 u^5 = xy \cos xy, \\ -\pi \le x \le \pi, \quad -\frac{\pi}{2} \le y \le \frac{\pi}{2}, \end{cases}$$

è pertanto un esempio di PDE debolmente non lineare. L'equazione è detta fortemente nonlineare, quando la nonlinearità riguarda almeno una delle derivate parziali della u.

La trattazione proposta, anche se fondamentalmente riferita alle PDEs del secondo ordine in due variabili, può estendersi (in modo del tutto naturale) ai modelli in più variabili. Allo scopo di rendere concreta questa affermazione, nel seguito verranno considerati alcuni esempi di PDE in tre variabili. La generalizzazione formale a $\mathbb{R}^n$ risulta sostanzialmente immediata, una volta precisata l'estensione dal caso 2D a quello 3D.

I problemi applicativi governati da PDEs sono innumerevoli. Essi riguardono soprattutto la Fisica, la Chimica, la Biologia e la generalità dei settori dell'Ingegneria, Relativamento parlando, i modelli differenziali di tipo ordinario e alle derivate parziali rappresentano la maggior parte dei modelli matematici di reale interesse applicativo.

Tra di essi, una prima distinzione riguarda i problemi di tipo stazionario da quelli di tipo evolutivo, tra i quali una ulteriore distinzione riguarda quelli di primo ordine e di seconda ordine.

A titolo esemplicativo, ci soffermiamo ora su tre classici problemi fisici la cui modellizzazione matematica conduce a tre PDEs, rispettivamente di tipo ellittico, parabolico e iperbolico,

Esempi:

(1′) Caso ellittico (*equazione di Laplace*):

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \quad (x,y) \in \Omega.$$

Tale equazione può rappresentare la conservazione di un flusso in assenza di sorgenti e di scarichi, come la temperatura in un dominio Ω in condizioni stazionarie, dunque indipendenti dal tempo, oppure un potenziale gravitazionale o elettrostatico in Ω, in condizioni stazionarie.

(1″) Caso ellittico (*equazione di Poisson*):

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = -\frac{1}{4\pi}\rho(x,y), \quad (x,y) \in \Omega.$$

Tale equazione può rappresentare la conservazione di un flusso in presenza di una sorgente, come in elettrostatica, fluidodinamica e gravitazione Newtoniana.

(2) Caso parabolico (*equazione del calore*):

$$\frac{\partial u}{\partial t} = k\frac{\partial^2 u}{\partial x^2}, \qquad k = \frac{K}{\sigma\mu} = \frac{\text{conducibilità termica}}{\text{calore specifico} \times \text{densità}}.$$

Tipicamente la $u(x,t)$ rappresenta la temperatura in un punto x di un intervallo $[a,b]$ all'istante t, con $0 \le t \le T$.

(3) Caso iperbolico (*equazione delle onde*):

(3′) *Corda vibrante*

$$\frac{\partial^2 u}{\partial t^2} = a^2\frac{\partial^2 u}{\partial x^2}, \qquad a^2 = \frac{\tau}{\mu} = \frac{\text{tensione}}{\text{densità}}.$$

In genere la $u(x,t)$ rappresenta la tensione all'istante t $(0 \le t \le T)$ di una corda in un punto x di un intervallo $[a,b]$.

(3″) *Vibrazioni trasversali di una trave*

$$\frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial u}{\partial x}, \qquad c^2 = \frac{e}{\mu} = \frac{\text{elasticità}}{\text{densità}}.$$

In questo caso la $u(x,t)$ indica la vibrazione della sezione trasversale di una trave in un punto x di un intervallo $[a,b]$ all'istante t (deflessione rispetto ad una posizione iniziale).

Nel caso di una PDE lineare in n variabili ($n \ge 2$) indicata, come usuale, con $\partial\Omega$ la frontiera di riferimento e con $[0,T]$ l'intervallo temporale di riferimento nei modelli evolutivi, i modelli differenziali di riferimento, per i quali è facile dare le condizioni che assicurano l'esistenza e l'unicità della soluzione sono i seguenti:

(a) *Modello ellittico*

$$\begin{cases} -\displaystyle\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right) + a_0 u = f, & x \in \Omega \subset \mathbb{R}^n, \\ u = \varphi, & \text{su } \partial\Omega. \end{cases}$$

(b) *Modello parabolico*

$$\begin{cases} \dfrac{\partial u}{\partial t} - \displaystyle\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right) + a_0 u = f, & (x,t) \in \Omega \times [0,T], \\ u = \varphi, & \text{su } \partial\Omega \times [0,T], \\ u(x,0) = \psi, & x \in \Omega \subset \mathbb{R}^n. \end{cases}$$

(c) *Modello iperbolico*

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} - \displaystyle\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right) + a_0 u = f, & (x,t) \in \Omega \times [0,T], \\ u = \varphi, & \text{su } \partial\Omega \times [0,T], \\ u(x,0) = \psi_0, & x \in \Omega \subset \mathbb{R}^n, \\ u_t(x,0) = \psi_1, & x \in \Omega \subset \mathbb{R}^n. \end{cases}$$

Le definizioni (a), (b) e (c) introdotte sono valide nell'ipotesi che i coefficienti $\{a_{ij}\}$ e $\{a_0\}$ soddisfino le condizioni di ellitticità, ossia che esistano:

(a) un numero α_0 tale che $a_0(x) \geq \alpha_0$, qualunque sia $x \in \Omega$;

(b) esista una costante positiva α tale che

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geq \alpha\|\xi\|^2, \quad \text{per ogni } x \in \Omega \text{ e } \xi \in \mathbb{R}^n,$$

essendo $\|\xi\| = \sqrt{\xi_1^2 + \ldots + \xi_n^2}$ la norma Euclidea del vettore ξ.

Da notare che α_0 può non essere positivo, a condizione che $\alpha + \alpha_0 > 0$.

Nel caso dell'equazione di Laplace, la condizione di ellitticità è verificata in ogni punto $(x_1, x_2) \in \mathbb{R}^2$. Basta infatti osservare che, essendo $a_{11} = a_{22} = 1$ e $a_{12} = a_{21} = 0$, la condizione di ellitticità è $\xi_1^2 + \xi_2^2 \geq \alpha||\xi||^2$, ovviamente verificata per ogni $\xi = (\xi_1, \xi_2) \in \mathbb{R}^2$ e $0 < \alpha < 1$. Nel caso la relazione sia valida per $\alpha = 0$ e non $\alpha > 0$, la condizione è detta di debole ellitticità.

Ciascuno dei tre problemi indicati possiede una e una sola soluzione, sotto ipotesi molto generali sui coefficienti $\{a_{ij}, a_0\}$, sulla funzione f in Ω, come anche su quelle assegnate su $\partial\Omega$ e su $\partial\Omega \times [0,T]$. Questo è vero, in particolare, quando l'operatore ellittico

$$A = -\sum_{i,.=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial}{\partial x_j}\right) + a_0(x)$$

soddisfa le condizioni di ellitticità in Ω e la frontiera $\partial\Omega$ è sufficientemente regolare (C^1 a tratti). Ipotesi che, nel seguito, supporremo sempre soddisfatta.

I problemi modello indicati, in ciascuno dei quali è assegnata la soluzione su $\partial\Omega$, sono definiti *problemi di Dirichlet*. Se invece, in luogo della u su $\partial\Omega$, si assegna la derivata normale $\frac{\partial u}{\partial n}$, essi sono noti come *problemi di Neumann*.

Infine sono definiti di *tipo misto* quando su una parte di $\partial\Omega$ è assegnata la u e sulla parte complementare la $\frac{\partial u}{\partial n}$, oppure se su $\partial\Omega$ è assegnata un'opportuna combinazione lineare di u e $\frac{\partial u}{\partial n}$.

Esempi di *problemi di Dirichlet, di Neumann e di tipo misto* per l'equazione di Laplace:

$$\begin{cases} \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = 0, \quad 0 \le x \le 5, \quad 0 \le y \le 3, \\ u(x,0) = 0, \quad u(x,3) = 0, \\ u(0,y) = 0, \quad u(5,y) = f(y), \end{cases}$$

$$\begin{cases} \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = 0, \quad 0 \le x \le 5, \quad 0 \le y \le 3, \\ u_y(x,0) = 0, \quad u_y(x,3) = 0, \\ u_x(0,y) = g(y), \quad u_x(5,y) = 0, \end{cases}$$

$$\begin{cases} \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = 0, \quad 0 \le x \le 5, \quad 0 \le y \le 3, \\ u_y(x,0) = 0, \quad u(x,3) = 0, \\ u_x(0,y) = 0, \quad u(5,y) = f(y), \end{cases}$$

essendo $g(x)$ una funzione continua sull'intervallo $[0,5]$ e $f(y)$ una funzione continua sull'intervallo $[0,3]$.

Esempi di PDEs per le quali è facile determinare la soluzione generale:

- Consideriamo la PDE
$$\frac{\partial^2 u}{\partial x \partial y} = 2x - y.$$
Integrando ordinatamente rispetto alla x e alla y si ottiene la soluzione generale
$$u(x,y) = x^2 y - \tfrac{1}{2} x y^2 + F(y) + G(x),$$
con F e G funzioni arbitrarie di classe C^1. Le funzioni
$$u_1(x,y) = x^2 y - \tfrac{1}{2} x y^2 + 3y^4 + \sin x - 5$$
e
$$u_2(x,y) = x^2 y - \tfrac{1}{2} x y^2 + y + \cos^2 x$$
sono invece due soluzioni particolari.

- Consideriamo ora l'equazione di Clairaut

$$u = x\frac{\partial u}{\partial x} - \left(\frac{\partial u}{\partial x}\right)^2.$$

La soluzione generale è

$$u(x,y) = xF(y) - F^2(y), \qquad F \text{ arbitraria di classe } C^1.$$

Le funzioni

$$u_1(x,y) = \frac{1}{4}x^2$$

e

$$u_2(x,y) = x\sin y - \sin^2 y$$

sono due soluzioni particolari.

- Consideriamo ora la PDE

$$u_{xy} = \frac{\partial^2 u}{\partial x \partial y} = x\cos y.$$

La soluzione generale è

$$u(x,y) = \tfrac{1}{2}x^2\sin y + F(y) + G(x),$$

con F e G di classe C^1. Le funzioni

$$u_1(x,y) = \frac{1}{2}x^2\sin y - \tfrac{1}{2}y^3 + 3x^4 + \sinh x$$

e

$$u_2(x,y) = \tfrac{1}{2}x^2\sin y - \cosh y + y - \cos^2 x$$

sono invece due soluzioni particolari, in quanto ottenute particolarizzando $F(y)$ e $G(x)$.

La risoluzione analitica delle PDEs è ottenibile soltanto in casi particolari. Per questo motivo, in generale, si ricorre alla sua risoluzione numerica. Quando possibile, la risoluzione analitica è preferibile sia per l'esattezza dei risultati, sia per la quantità e qualità di informazioni matematiche e fisiche da essa deducibili. I metodi analitici più utilizzati sono il metodo degli integrali generali e il metodo della separazione delle variabili. Per la loro illustrazione è necessario utilizzare i risultati essenziali sulle ODEs (presentati nel Cap. 2) e quelli sulle serie di Fourier e sulle proprietà spettrali dei problemi di Sturm-Liouville (introdotti nel Cap. 3). Per questi motivi la illustrazione dei metodi analitici delle PDEs è rinviata al Cap. 4.

Capitolo 2

RISULTATI ESSENZIALI SULLE EQUAZIONI DIFFERENZIALI ORDINARIE (ODEs)

2.1 Preliminari

In questa sezione vengono illustrati alcuni metodi basilari per la risoluzione analitica delle ODEs e richiamati, molto sinteticamente, alcuni metodi per la loro risoluzione numerica. I metodi analitici richiamati sono utilizzati nella risoluzione analitica delle equazioni a derivate parziali (PDEs) ed è questo il motivo per cui sono più dettagliati rispetto a quelli numerici. Per ODE di ordine n intendiamo una relazione funzionale tra una variabile indipendente, la funzione incognita e le sue derivate fino all'ordine n. La sua forma generale è del tipo

$$f(x; y, y', \ldots, y^{(n)}) = 0, \tag{2.1}$$

dove x è la variabile indipendente, y la funzione incognita, $y', \ldots, y^{(n)}$ le sue derivate fino all'ordine n e f una funzione nota. L'equazione

$$y' + a(x)y = f(x), \qquad a \leq x \leq b, \tag{2.2}$$

è evidentemente lineare e del primo ordine, mentre l'equazione

$$y'' + a(x)y' + b(x)y = f(x), \qquad a \leq x \leq b, \tag{2.3}$$

è lineare del secondo ordine. Più in generale, l'equazione

$$y^{(n)} + a_1(x)y^{(n-1)} + \ldots + a_n(x)y = f(x), \qquad a \leq x \leq b, \tag{2.4}$$

è lineare di ordine n.

È ben noto che la soluzione generale di una ODE di ordine n dipende da n costanti arbitrarie. Conseguentemente per la sua unicità è necessario assegnare n condizioni linearmente indipendenti: tipicamente n condizioni in a (oppure m in a e $n-m$ in b). Nel caso sia $a = -\infty$ e/o $b = +\infty$ le condizioni assegnate sono di tipo asintotico. Esempi:

(1) $y' = f(x)$ per $a \le x \le b$, $y(x) = \int_a^x f(t)\,dt + c$.

(2) $y'' + \omega^2 y = 0$ per $a \le x \le b$, $y(x) = c_1 \cos \omega x + c_2 \sin \omega x$.

dove c, c_1 e c_2 sono costanti arbitrarie. Nell'ipotesi che sia $\sin \omega(b-a) \neq 0$, esse possono essere univocamente determinate fissando il valore della y in a per la (2.2) e i valori della y e della y' in a oppure, in alternativa, i valori di y in a e b. Nel caso della (2.4) le costanti arbitrarie sono n e si parla di problema di Cauchy, nel caso vengono assegnati in a i valori delle $y, y', \ldots, y^{(n-1)}$; di problema agli estremi nel caso vengono (complessivamente) assegnati n valori di y e della sua derivata prima in a e b. È bene ricordare che la (2.2) è facilmente risolubile analiticamente. Infatti, indicata con $\alpha(x)$ una primitiva di $a(x)$, molteplicando primo e secondo membro per $e^{\alpha(x)}$ si ottiene l'equazione

$$\frac{d}{dx}\left(e^{\alpha(x)} y\right) = e^{\alpha(x)} f(x), \qquad \alpha(x) = \int_a^x a(t)\,dt,$$

dalla quale segue immediatamente che

$$y(x) = e^{-\alpha(x)} \int_a^x e^{\alpha(t)} f(t)\,dt + c\, e^{-\alpha(x)}. \tag{2.5}$$

Esempio 2.1

$$\begin{cases} y' - y = \cos t, \\ y(0) = 1. \end{cases}$$

Molteplicando primo e secondo membro per e^{-t} l'equazione diventa

$$(e^{-t} y)' = e^{-t} \cos t \Longrightarrow y(t) = e^t \left[\int_0^t e^{-\tau} \cos \tau\, d\tau + c \right],$$

dalla quale, osservato che

$$\int_0^t e^{-\tau} \cos \tau\, d\tau = \frac{1}{2} \left[e^{-t} (\sin t - \cos t) + 1 \right],$$

deriva che

$$y(t) = \tfrac{1}{2}(\sin t - \cos t + e^t) + c\, e^t,$$

da cui, richiedendo che $y(0) = 1$, risulta che

$$y(t) = \tfrac{1}{2}(\sin t - \cos t) + \tfrac{3}{2} e^t.$$

Esempio 2.2

$$\begin{cases} y'' + \omega^2 y = 0, & a \le x \le b, \\ y(a) = \alpha, \quad y(b) = \beta, & \text{con } 0 < \omega(b-a) < \pi. \end{cases}$$

Essendo $\cos \omega x$ e $\sin \omega x$ soluzioni particolari, la soluzione generale è

$$y(x) = c_1 \cos \omega x + c_2 \sin \omega x.$$

La soluzione del problema agli estremi viene quindi ottenuta imponendo che c_1 e c_2 soddisfino il sistema

$$\begin{cases} c_1 \cos \omega a + c_2 \sin \omega a = a, \\ c_1 \cos \omega b + c_2 \sin \omega b = b, \end{cases} \quad \Longrightarrow \begin{cases} c_1 = \dfrac{\alpha \sin \omega b - \beta \sin \omega a}{\sin \omega(b-a)}, \\ c_2 = \dfrac{\beta \cos \omega a - \alpha \cos \omega b}{\sin \omega(b-a)}. \end{cases}$$

Nel caso particolare $a = 0$, $b = 1$, $\alpha = 1$ e $\beta = 0$

$$y(x) = \cos \omega x - \frac{\cos \omega}{\sin \omega} \sin \omega x.$$

2.2 Trasformazione di una ODE di ordine superiore al primo in un sistema

Iniziamo con la trasformazione di una ODE lineare del secondo ordine in un sistema lineare di due equazioni del primo ordine. Sia

$$a\, y'' + b\, y' + c\, y = f(x), \qquad a \neq 0, \tag{2.6}$$

una ODE nella quale a, b e c sono numeri reali noti e $f(x)$ una funzione nota. Per la sua trasformazione, posto

$$y = y_1 \quad \text{e} \quad y' = y_1' = y_2,$$

la (2.6) diventa

$$\begin{cases} y_1' = y_2, \\ a y_2' + b y_1' + c y_1 = f(x), \end{cases} \quad \Longleftrightarrow \begin{cases} y_1' = y_2, \\ y_2' = -\dfrac{c}{a} y_1 - \dfrac{b}{a} y_2 + \dfrac{f(x)}{a}, \end{cases}$$

che, con notazione matriciale, possiamo scrivere nella forma

$$\frac{d\boldsymbol{y}}{dx} = A\, \boldsymbol{y} + \boldsymbol{f}, \qquad = \begin{pmatrix} y_1(x) \\ y_2(x) \end{pmatrix}, \tag{2.7}$$

dove $A = \left(\begin{smallmatrix} 0 & 1 \\ -c/a & -b/a \end{smallmatrix} \right)$, $\boldsymbol{f}(x) = \left(\begin{smallmatrix} 0 \\ f(x)/a \end{smallmatrix} \right)$.

Esempio 2.3

$$y'' + \omega^2 y = \cos x.$$

Ponendo $y_1 = y$, $y_2 = y_1' = y'$, l'equazione precedente si trasforma nel sistema

$$\begin{cases} y_1' = y_2, \\ y_2' = -\omega^2 y_1 + \cos x, \end{cases} \iff \frac{d\boldsymbol{y}}{dx} = \begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix} \boldsymbol{y}(x) + \begin{pmatrix} 0 \\ \cos x \end{pmatrix},$$

essendo $\boldsymbol{y}(x) = \begin{pmatrix} y_1(x) \\ y_2(x) \end{pmatrix}$.

Esempio 2.4

$$y^{IV} + 2y'' + y' + y = \sin x.$$

Ponendo $y_1 = y$, $y_2 = y_1' = y'$, $y_3 = y_2' = y''$ e $y_3' = y^{IV}$, l'equazione può essere scritta nella forma

$$\begin{cases} y_1' = y_2, \\ y_2' = y_3, \\ y_3' = y_4, \\ y_4' = -2y_3 - y_2 - y_1 + \sin x, \end{cases} \iff \frac{d\boldsymbol{y}}{dx} = A\,\boldsymbol{y} + \boldsymbol{f},$$

dove

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & -1 & -2 & 0 \end{pmatrix}, \quad \boldsymbol{f}(x) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \sin x \end{pmatrix}.$$

Lo stesso tipo di tecnica permette di trasformare una ODE lineare di ordine n in un sistema lineare di ODE del primo ordine.

La tecnica è ugualmente applicabile, senza modifiche significative, al caso di ODE non lineari. Un'equazione non lineare del secondo ordine

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

può essere trasformata in un sistema non lineare di due equazioni del primo ordine. Ponendo $y_1 = y$ e $y_2 = y_1'$, la (2.7) si trasforma infatti nel sistema

$$\begin{cases} y_1' = y_2, \\ y_2' = f(x, y_1, y_2). \end{cases}$$

2.3 Risoluzione di semplici ODEs lineari del secondo ordine

Iniziamo con una ODE lineare *omogenea* del secondo ordine a coefficienti costanti

$$a\,y'' + b\,y' + c\,y = 0, \tag{2.8}$$

dove a, b e c sono costanti reali con $a \neq 0$. Il metodo più utilizzato si basa sull'osservazione che la soluzione generale è esprimibile come combinazione lineare di due soluzioni particolari, linearmente indipendenti. Per determinare le quali, seguendo il metodo di D'Alembert, poniamo

$$y(x) = e^{\alpha x}$$

con α costante complessa da determinare. Sostituendo la sua espressione nella (2.8) otteniamo

$$e^{\alpha x}\left(a\,\alpha^2 + b\,\alpha + c\right) = 0,$$

dalla quale, dato che $e^{\alpha x} \neq 0$, si deduce che α deve essere soluzione dell'equazione *caratteristica*

$$a\,\alpha^2 + b\,\alpha + c = 0$$

le cui radici

$$\alpha_{1,2} = \frac{-b \mp \sqrt{b^2 - 4ac}}{2a}$$

identificano le soluzioni caratteristiche della (2.8).

(a) Le radici caratteristiche sono reali e distinte ($b^2 - 4ac > 0$). In questo caso

$$y_1(x) = e^{\alpha_1 x} \quad \text{e} \quad y_2(x) = e^{\alpha_2 x}$$

sono entrambe soluzioni e la loro combinazione lineare

$$y(x) = c_1 e^{\alpha_1 x} + c_2 e^{\alpha_2 x}, \quad \text{con } c_1, c_2 \in \mathbb{R},$$

è la soluzione generale della (2.8). L'arbitrarietà di c_1 e c_2 consente di soddisfare due condizioni iniziali (problema di Cauchy) oppure due condizioni agli estremi (problema con valori agli estremi).

(b) Le due radici sono reali e coincidenti ($b^2 - 4ac = 0$): $\alpha_1 = \alpha_2 = \alpha = -\frac{b}{2a}$. Alla radice α corrispondono le soluzioni:

$$y_1(x) = e^{\alpha x} \quad \text{e} \quad y_2(x) = x\,e^{\alpha x}.$$

Essendo ovvio che y_1 sia una soluzione, basta osservare che lo è anche la seconda. Derivando due volte y_2 e sostituendo nella (2.8) si ha infatti che

$$a(2\alpha+\alpha^2 x)e^{\alpha x} + b(1+\alpha x)e^{\alpha x} + cxe^{\alpha x} = x\,e^{\alpha x}(a\alpha^2 + b\alpha + c) + e^{\alpha x}(2a\alpha + b) = 0,$$

dato che $a\alpha^2 + b\alpha + c = 0$ (essendo α radice dell'equazione caratteristica) e $2a\alpha + b = 0$, in quanto $\alpha = -\frac{b}{2a}$. La soluzione generale è dunque

$$y(x) = (c_0 + c_1\, x)e^{\alpha x}, \qquad c_0, c_1 \in \mathbb{R}.$$

(c) Le radici caratteristiche sono una complessa coniugata dell'altra ($b^2 - 4ac < 0$). Anche in questo caso $e^{\alpha_1 x}$ e $e^{\alpha_2 x}$ sono soluzioni. La loro forma non è però (in generale) utilizzata in quanto si preferisce evitare il ricorso a funzioni complesse, dato che la ODE è a coefficienti reali. Per avere una rappresentazione reale della soluzione generale, basta utilizzare opportunamente la formula di Eulero. Più precisamente, posto $\alpha_1 = -\frac{b}{2a} + i\beta$ e $\alpha_2 = -\frac{b}{2a} - i\beta$, $\beta = \frac{\sqrt{b^2-4ac}}{2a} > 0$, per la formula di Eulero

$$e^{\alpha_1 x} = e^{-\frac{b}{2a}x}\left(\cos\beta x + i\sin\beta x\right),$$
$$e^{\alpha_2 x} = e^{-\frac{b}{2a}x}\left(\cos\beta x - i\sin\beta x\right),$$

da cui deriva immediatamente che

$$\frac{1}{2}\left(e^{\alpha_1 x} + e^{\alpha_2 x}\right) = e^{-\frac{b}{2a}x}\cos\beta x,$$
$$\frac{1}{2i}\left(e^{\alpha_1 x} - e^{\alpha_2 x}\right) = e^{-\frac{b}{2a}x}\sin\beta x.$$

Questo implica (essendo l'equazione lineare) che

$$e^{-\frac{b}{2a}x}\cos\beta x \quad \text{e} \quad e^{-\frac{b}{2a}x}\sin\beta x$$

sono soluzioni (linearmente indipendenti) della (2.8) (come si può verificare per sostituzione). Conseguentemente, in questo caso, la soluzione generale è

$$y(x) = e^{-\frac{b}{2a}x}\left(c_1\,\cos\beta x + c_2\,\sin\beta x\right).$$

Esempio 2.5 Determinare la soluzione dei seguenti problemi di Cauchy:

(a)

$$\begin{cases} y'' + 2y - y = 0, \\ y(0) = 0, \quad y'(0) = 1. \end{cases}$$

Posto $y = e^{\alpha x}$, l'equazione caratteristica è

$$\alpha^2 + 2\alpha - 1 = 0 \Longrightarrow \alpha_{1,2} = -1 \mp \sqrt{2}.$$

Di conseguenza la soluzione generale è

$$y(x) = e^{-x}\left(c_1 e^{-x\sqrt{2}} + c_2 e^{x\sqrt{2}}\right),$$

con (c_1, c_2) soluzione ca) (unica) del sistema

$$\begin{cases} c_1 + c_2 = 0, \\ (1+\sqrt{2})c_1 + (1-\sqrt{2})c_2 = -1. \end{cases}$$

(b)

$$\begin{cases} y'' + \omega^2 y = 0, \\ y(0) = 1, \quad y'(0) = 0, \end{cases}$$

dove $\omega > 0$. Essendo $\alpha_{1,2} = \pm i\omega$, la soluzione generale è

$$y(x) = c_1 \cos\omega x + c_2 \sin\omega x.$$

La soluzione particolare si ottiene imponendo che la y soddisfi le condizioni iniziali, ossia che

$$\begin{cases} c_1 = 1, \\ c_2\,\omega = 0, \end{cases} \implies y(x) = \cos\omega x.$$

(c)

$$\begin{cases} y'' - 2y' + y = 0, \\ y(0) = 1, \quad y'(0) = 0. \end{cases}$$

L'equazione caratteristica associata è

$$\alpha^2 - 2\alpha + 1 = 0 \implies \alpha_1 = \alpha_2 = 1.$$

Di conseguenza la soluzione generale è

$$y(x) = e^x\,(a + bx)$$

con (a, b) soluzione unica del sistema

$$\begin{cases} y(0) = a = 1, \\ y'(0) = a + b = 0, \end{cases} \implies y(x) = e^x\,(1 - x).$$

2.4 Esempi di ODEs di tipo inomogeneo facilmente risolubili

Le ODEs più ricorrenti nelle applicazioni sono di secondo ordine che, in vari casi, è possibile risolvere facilmente. Nel caso i coefficienti siano costanti, abbiamo già visto come si calcola la soluzione generale dell'equazione omogenea

associata. Nel caso inomogeneo a questa deve essere aggiunta una particolare della inomogenea. Supponendo che l'equazione sia del tipo

$$a\,y'' + b\,y' + c\,y = f(x) \tag{2.9}$$

con a, b e c costanti, la sua determinazione è facile se: **(a)** $f(x)$ è esponenziale; **(b)** un polinomio algebrico; **(c)** un polinomio trigonometrico.

(a) $f(x) = e^{\alpha x}$, $\alpha \in \mathbb{R}$. In questo caso si cerca una soluzione del tipo $y = \hat{c}\,e^{\alpha x}$ con $\hat{c}$ costante da determinare. Sostituendo la y e le sue prime derivate nella (2.9) si ottiene la relazione

$$\hat{c}\,e^{\alpha x}\left(a\,\alpha^2 + b\,\alpha + c\right) = e^{\alpha x},$$

dalla quale si ricava immediatamente $\hat{c}$, nell'ipotesi che $a\,\alpha^2 + b\,\alpha + c \neq 0$. Nel caso non lo sia, si pone $y = x\,e^{\alpha x}$, da cui deriva che $\hat{c} = \frac{1}{2a\,\alpha + b}$ se $2a\,\alpha + b \neq 0$. Nel caso $a\,\alpha^2 + b\,\alpha + c = 2a\,\alpha + b = 0$, si pone $y = x^2\,e^{\alpha x}$, da cui deriva che $\hat{c} = \frac{1}{2a}$, essendo $a \neq 0$.

(b) Supponiamo, per ipotesi, che $f(x)$ sia un polinomio di secondo grado

$$f(x) = a_0x^2 + a_1x + a_2, \qquad a_0 \neq 0.$$

La previsione è che anche $y(x)$ sia un polinomio di secondo grado

$$y(x) = y_0x^2 + y_1x + y_2.$$

Derivando e sostituendo si trova la relazione

$$a(2y_0) + b(2y_0x + y_1) + c(y_0x^2 + y_1x + y_2) = a_0x^2 + a_1x + a_2,$$

dalla quale segue che y_0, y_1 e y_2 debbono soddisfare il sistema

$$\begin{cases} cy_0 = a_0, \\ (2b)y_0 + cy_1 = a_1, \\ (2a)y_0 + by_1 + cy_2 = a_2, \end{cases}$$

che, nell'ipotesi che $c \neq 0$, ammette un'unica soluzione. Nel caso $c = 0$, si ripete il procedimento ponendo

$$y(x) = y_0x^3 + y_1x^2 + y_2x + y_3.$$

Questo implica che a_0, a_1 e a_2 sono la soluzione unica del sistema

$$\begin{cases} 3by_0 = a_0, \\ 6ay_0 + 2by_1 = a_1, \\ 2ay_1 + by_2 = a_2, \end{cases}$$

nell'ipotesi che $b \neq 0$. Nel caso $b = c = 0$ e $a \neq 0$ la ODE si riduce all'equazione

$$y'' = \frac{a_0 x^2 + a_1 x + a_2}{a}$$

con soluzione generale

$$y(x) = \frac{a_0}{12a} x^4 + \frac{a_1}{6a} x^3 + \frac{a_2}{2a} x^2 + c_0 + c_1 x,$$

dove le costanti c_0 e c_1 sono arbitrarie.

(c) Nel caso

$$f(x) = f_0 \cos \omega x + f_1 \sin \omega x, \qquad 0 \neq \omega \in \mathbb{R},$$

si procede nell'ipotesi che anche la y sia dello stesso tipo. Ponendo

$$y(x) = y_0 \cos \omega x + y_1 \sin \omega x,$$

derivando e sostituendo nella (2.9) si trova che la coppia (y_0, y_1) è la soluzione unica del sistema

$$\begin{cases} (c - a\,\omega^2) y_0 + b\,\omega y_1 = f_0, \\ (c - a\,\omega^2) y_1 - b\,\omega y_0 = f_1, \end{cases}$$

nell'ipotesi che $(c - a\,\omega^2)^2 + (b\,\omega)^2 > 0$. Nel caso $b = c - a\,\omega^2 = 0$ e $a \neq 0$ la ODE si riduce all'equazione $y'' + \omega^2 y = 0$. In tal caso, ponendo

$$y(x) = f_0 x \cos \omega x + f_1 x \sin \omega x,$$

si ottiene il sistema $-2f_0\omega = y_1$ e $2f_1\omega = y_0$ di soluzione $f_0 = -\frac{y_1}{2\omega}$ e $f_1 = \frac{y_0}{2\omega}$.

Essendo l'equazione lineare, per il principio di sovrapposizione, il problema è facilmente risolubile anche nel caso $f(x)$ sia di tipo misto. Nel caso, ad esempio, sia

$$f(x) = f_1(x) + f_2(x)$$

con f_1 polinomio algebrico e f_2 polinomio trigonometrico, si cercherà una soluzione del tipo $y(x) = y_1(x) + y_2(x)$ con y_1 (polinomio algebrico) soluzione del sistema

$$a\,y'' + b\,y' + c\,y = f_1(x)$$

e y_2 (polinomio trigonometrico) soluzione del sistema

$$a\,y'' + b\,y' + c\,y = f_2(x).$$

Esempio 2.6

$$\begin{cases} y'' + 2y' + 4y = x^2 + 1, \\ y(0) = 1, \quad y(1) = 0. \end{cases}$$

Poiché le radici dell'equazione caratteristica associata sono $\alpha_{1,2} = -1 \mp i\sqrt{3}$, la soluzione dell'equazione omogenea associata è

$$\hat{y}(x) = e^{-x}\left(a\,\cos(x\sqrt{3}) + b\,\sin(x\sqrt{3})\right).$$

Essendo il secondo membro dell'equazione un polinomio di secondo grado, è facile dimostrare che una sua soluzione particolare è

$$y_p(x) = \tfrac{1}{4}(x^2 - x + 1).$$

Di conseguenza la soluzione è $y(x) = \hat{y}(x) + y_p(x)$, essendo a e b la soluzione unica del sistema

$$\begin{cases} \frac{1}{4} + a = 1, \\ \frac{1}{4} + \frac{1}{e}\left(a\cos(\sqrt{3}) + b\sin(\sqrt{3})\right) = 0, \end{cases} \quad \Longrightarrow a = \tfrac{3}{4}, \quad b = -\frac{e + 3\cos(\sqrt{3})}{4\sin(\sqrt{3})}.$$

2.5 Risoluzione di sistemi di ODEs lineari omogeni

Consideriamo ora la risoluzione analitica di un sistema lineare omogeneo di ODE a coefficienti costanti, pertanto esprimibile nella forma

$$\frac{d\boldsymbol{y}}{dx} = A\,\boldsymbol{y}, \tag{2.10}$$

dove A è una matrice a termini costanti di ordine n e $\boldsymbol{y}$ è un vettore di n funzioni incognite della x. Anche se esistono metodi a carattere più generale, procediamo nell'ipotesi che la matrice A ammetta un sistema di autovettori linearmenti indipendenti. Sotto tale ipotesi la soluzione generale del sistema può essere ottenuta con un *metodo spettrale* che consente di ricondurre il calcolo alla risoluzione di n ODE disgiunte del primo ordine. Come sarà evidente nel seguito, la denominazione spettrale deriva dall'utilizzo nel metodo dello spettro della matrice A. Per la sua applicazione, si cercano soluzioni della (2.10) della forma [8]

$$\boldsymbol{y}(x) = \alpha(x)\,\boldsymbol{u},$$

dove $\alpha(x)$ è una funzione scalare e $\boldsymbol{u}$ è un vettore non nullo in $\mathbb{R}^n$. Sostituendo la sua espressione nella (2.10) abbiamo che

$$\frac{d\boldsymbol{y}}{dx} = \frac{d\alpha}{dx}\,\boldsymbol{u} = \alpha(x)\,A\,\boldsymbol{u},$$

da cui deriva che i due vettori $\boldsymbol{u}$ e $A\,\boldsymbol{u}$ debbono essere proporzionali, ossia che $\boldsymbol{u}$ deve essere un autovettore di A. In altre parole perché valga l'ultima relazione deve essere

$$A\,\boldsymbol{u} = \lambda\,\boldsymbol{u},$$

dalla quale segue che

$$\frac{d\alpha}{dx} = \lambda\,\alpha(x) \Longrightarrow \alpha(x) = c\,e^{\lambda x},$$

dove c è una costante. Di conseguenza, ad ogni coppia autovalore-autovettore $(\lambda, \boldsymbol{u})$ di A si può associare una soluzione della (2.10) del tipo

$$\boldsymbol{y}(x) = c\,e^{\lambda x}\,\boldsymbol{u}$$

determinata a meno di una costante arbitraria. Se lo spettro di A è $\{\lambda_i, \boldsymbol{u}_i\}_{i=1}^n$ e i suoi autovettori sono linearmente indipendenti, la soluzione generale della (2.10) assume la forma

$$\boldsymbol{y}(x) = \sum_{i=1}^{n} c_i\,e^{\lambda_i x}\,\boldsymbol{u}_i \tag{2.11}$$

con $c_1, c_2, \ldots, c_n$ costanti arbitrarie, pertanto utilizzabili per soddisfare condizioni iniziali o agli estremi.

Esempio 2.7

$$\begin{cases} y_1' = 5y_1 - 4y_2 - y_3, \\ y_2' = -4y_1 + 8y_2 - 4y_3, \\ y_3' = -y_1 - 4y_2 + 5y_3, \\ y_1(0) = 1, \quad y_2(0) = 1, \quad y_3(0) = -1. \end{cases}$$

La matrice A del sistema è pertanto

$$A = \begin{pmatrix} 5 & -4 & -1 \\ -4 & 8 & -4 \\ -1 & -4 & 5 \end{pmatrix},$$

il cui sistema spettrale è

$$\lambda_1 = 0,\ \boldsymbol{u}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}; \quad \lambda_2 = 6,\ \boldsymbol{u}_2 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}; \quad \lambda_3 = 12,\ \boldsymbol{u}_3 = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}.$$

La soluzione generale del sistema differenziale è pertanto

$$\boldsymbol{y}(x) = \sum_{i=1}^{3} c_i\, e^{\lambda_i x}\, \boldsymbol{u}_i = \begin{cases} c_1 + c_2 e^{6x} + c_3 e^{12x}, \\ c_1 - 2c_3 e^{12x}, \\ c_1 - c_2 e^{6x} + c_3 e^{12x}, \end{cases}$$

con c_1, c_2 e c_3 da determinare imponendo che siano soddisfatte le condizioni iniziali, ossia che soddisfino il sistema

$$\begin{cases} c_1 + c_2 + c_3 = 1, \\ c_1 - 2c_3 = 1, \\ c_1 - c_2 + c_3 = -1, \end{cases} \qquad \Longrightarrow c_1 = \tfrac{1}{3}, \quad c_2 = 1, \quad c_3 = -\tfrac{1}{3}.$$

Di conseguenza

$$y_1(x) = \tfrac{1}{3}\left(1 + 3\, e^{6x} - e^{12x}\right),\ y_2(x) = \tfrac{1}{3}\left(1 + 2\, e^{12x}\right),\ y_3(x) = \tfrac{1}{3}\left(1 - 3\, e^{6x} - e^{12x}\right).$$

2.6 Risoluzione di sistemi di ODEs lineari inomogenei

Consideriamo ora la risoluzione di sistemi di ODE lineari di tipo

$$\frac{d\boldsymbol{y}}{dx} = A\,\boldsymbol{y} + \boldsymbol{f}(x), \tag{2.12}$$

dove A, come nel caso omogeneo, è una matrice costante di ordine n e $\boldsymbol{f}(x)$ è un vettore di n funzioni della x. Come verrà evidenziato nel seguito, esistono situazioni particolari nelle quali, nota la soluzione generale del sistema omogeneo, è molto facile trovare una soluzione particolare di quello inomogeneo e conseguentemente determinare la soluzione generale di quello non omogeneo, dato che è la somma della soluzione generale del sistema omogeneo associato (sistema (2.10)) e di una particolare di quello inomogeneo (sistema (2.12)). Il metodo a carattere generale più utilizzato è il *metodo della variazione dei parametri* che ci limitiamo ad illustrare nel caso la matrice A abbia n autovettori linearmente indipendenti (stessa ipotesi del caso omogeneo). Di conseguenza anche in questo caso possiamo parlare di metodo spettrale. Come nel caso omogeneo, indichiamo con

$$\{\lambda_i . \boldsymbol{u}_i\}_{i=1}^{n}$$

lo spettro (autovalori-autofunzioni) di A. Essendo (per ipotesi) linearmente indipendenti, ogni vettore di $\mathbb{R}^n$ è esprimibile come loro combinazione lineare. Di

conseguenza, per ogni x, possiamo affermare che $\boldsymbol{f}(x)$ può essere rappresentato nel modo seguente:

$$\boldsymbol{f}(x) = c_1(x)\boldsymbol{u} + c_2(x)\boldsymbol{u}_2 + \ldots + c_n(x)\boldsymbol{u}_n, \tag{2.13}$$

dove i coefficienti sono, ovviamente, funzioni di x. Essendo $\boldsymbol{f}(x)$ nota, i coefficienti $\{c_i(x)\}_{i=1}^n$ possono essere calcolati risolvendo un sistema lineare. Nel caso A sia simmetrica e, conseguentemente, gli autovalori siano ortogonali, i coefficienti sono le proiezioni ortogonali di $\boldsymbol{f}(x)$ sugli autovettori, ossia

$$c_i(x) = \boldsymbol{f}(x)^T\boldsymbol{u}_i, \qquad i = 1, 2, \ldots, n,$$

con

$$\boldsymbol{f}(x) = (f_1(x), \ldots, f_n(x))^T,\ \boldsymbol{u}_i = (u_{1i}, u_{2i}, \ldots, u_{ni})^T,\ \boldsymbol{f}(x)^T\boldsymbol{u}_i = \sum_{j=1}^{n} f_j(x)u_{ji}.$$

Per lo stesso motivo anche la soluzione $\boldsymbol{y}(x)$ della (2.12), per ogni x, può essere espressa in modo analogo, ossia ponendo

$$\boldsymbol{y}(x) = a_1(x)\boldsymbol{u}_1 + a_2(x)\boldsymbol{u}_2 + \ldots + a_n(x)\boldsymbol{u}_n, \tag{2.14}$$

i cui coefficienti $\{a_i(x)\}_{i=1}^n$ non possono essere calcolati come nel caso della $\boldsymbol{f}(x)$, dato che $\boldsymbol{y}(x)$ è un vettore incognito. Tuttavia il metodo permette di calcolarli mediante la risoluzione di n ODE disaccopiate (una per coefficiente). Tenuto conto della (2.14) e del significato degli $\{\boldsymbol{u}_i\}_{i=1}^n$, possiamo affermare che

$$\begin{aligned}
\frac{d\boldsymbol{y}}{dx} - A\,\boldsymbol{y} &= \frac{d}{dx}\left(\sum_{i=1}^{n} a_i(x)\boldsymbol{u}_i\right) - A\left(\sum_{i=1}^{n} a_i(x)\boldsymbol{u}_i\right) \\
&= \sum_{i=1}^{n} \frac{da_i}{dx}\boldsymbol{u}_i - \sum_{i=1}^{n} a_i(x)A\,\boldsymbol{u}_i = \sum_{i=1}^{n} \frac{da_i}{dx}\boldsymbol{u}_i - \sum_{i=1}^{n} \lambda_i a_i(x)\boldsymbol{u}_i \\
&= \sum_{i=1}^{n} \left\{\frac{da_i}{dx} - \lambda_i a_i(x)\right\}\boldsymbol{u}_i.
\end{aligned}$$

Di conseguenza, tenendo conto della (2.14), il sistema (2.12) può essere scritto nella forma seguente

$$\sum_{i=1}^{n} \left\{\frac{da_i}{dx} - \lambda_i a_i(x)\right\}\boldsymbol{u}_i = \sum_{i=1}^{n} c_i(x)\boldsymbol{u}_i,$$

dalla quale segue che

$$\frac{da_i}{dx} - \lambda_i a_i = c_i(x), \qquad i = 1, 2, \ldots, n. \tag{2.15}$$

La (2.15) evidenzia che, in questo modo, la risoluzione del sistema (2.12) viene ricondotto a quella di n ODE del primo ordine, ciascuna indipendente dalle altre. La loro risoluzione consente di scrivere immediatamente la soluzione generale della (2.12), la quale dipende da n costanti arbitrarie derivanti dalla risoluzione delle n ODEs (2.15) (una costante per equazione). I loro valori vengono determinati uno per volta, nel caso siano assegnati i valori iniziali per le componenti del vettore $\boldsymbol{y}$ ($y_i(a) = y_{i0}$, $i = 1, \ldots, n$).

Esempio 2.8

$$\begin{cases} y_1' = y_1 + y_2, \\ y_2' = y_1 - y_2, \\ y_1(0) = 1, \quad y_2(0) = -1. \end{cases}$$

Lo spettro della matrice $A = \left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$ è $\lambda_1 = \sqrt{2}$, $\boldsymbol{u}_1 = \left(\begin{smallmatrix} 1 \\ \sqrt{2}-1 \end{smallmatrix}\right)$; $\lambda_2 = -\sqrt{2}$, $\boldsymbol{u}_2 = \left(\begin{smallmatrix} 1-\sqrt{2} \\ 1 \end{smallmatrix}\right)$. Di conseguenza la soluzione generale è

$$\boldsymbol{y}(x) = c_1\, e^{x\sqrt{2}}\, \boldsymbol{u}_1 + c_2\, e^{-x\sqrt{2}}\, \boldsymbol{u}_2$$

con c_1 e c_2 determinati imponendo che siano soddisfatte le condizioni iniziali, ossia come soluzione del sistema

$$\begin{cases} c_1 + (1-\sqrt{2})c_2 = 1, \\ (\sqrt{2}-1)c_1 + c_2 = -1, \end{cases} \quad \Longrightarrow c_1 = \tfrac{1}{2},\ c_2 = -\tfrac{1}{2}(1+\sqrt{2}).$$

Questo implica che

$$y_1(x) = \tfrac{1}{2}\left[e^{x\sqrt{2}} + e^{-x\sqrt{2}}\right], \quad y_2(x) = \tfrac{1}{2}\left[(\sqrt{2}-1)e^{x\sqrt{2}} - (\sqrt{2}+1)e^{-x\sqrt{2}}\right].$$

Esempio 2.9

$$\begin{cases} y''(x) + \theta^2 y = f(x), & a \le x \le b, \\ y(a) = \alpha, \quad y(b) = \beta, & \theta > 0,\ 0 < b-a < \pi, \end{cases} \tag{2.16}$$

dove $0 \neq \theta \in \mathbb{R}$. Essendo il problema lineare, per il principio della sovrapposizione, possiamo porre $y = v + w$, con v e w soluzioni dei seguenti due problemi:

$$\begin{cases} v'' + \theta^2 v = f(x), \\ v(a) = v(b) = 0, \end{cases} \tag{2.17a}$$

$$\begin{cases} w'' + \theta^2 w = 0, \\ w(a) = \alpha, \quad w(b) = \beta. \end{cases} \tag{2.17b}$$

La ODE che caratterizza il primo problema è del tipo inomogeneo e quella che caratterizza il secondo è di tipo omogeneo. L'inverso si verifica per le condizioni agli estremi, essendo di tipo omogeneo il primo e di tipo inomogeneo il secondo. La (2.17b) è facilmente risolubile in quanto, essendo $\cos\theta x$ e $\sin\theta x$ due soluzioni linearmente indipendenti, la soluzione generale è

$$w(x) = c_1 \cos\theta x + c_2 \sin\theta x,$$

con (c_1, c_2) soluzione unica del sistema

$$\begin{cases} (\cos\theta a)c_1 + (\sin\theta a)c_2 = \alpha, \\ (\cos\theta b)c_1 + (\sin\theta b)c_2 = \beta, \end{cases}$$

chiaramente non singolare, dato che il determinante del sistema è $\sin\theta(b-a)$. Per la soluzione del secondo problema, posto $v_1 = v$ e $v_1' = v_2$, trasformiamo l'equazione nel sistema lineare

$$\boldsymbol{v}'(x) = \begin{pmatrix} v_1'(x) \\ v_2'(x) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -\theta^2 & 0 \end{pmatrix} \begin{pmatrix} v_1(x) \\ v_2(x) \end{pmatrix} + \begin{pmatrix} 0 \\ f(x) \end{pmatrix} = A\,\boldsymbol{v}(x) + \boldsymbol{f}(x). \quad (2.18)$$

Lo spettro di A è: $\lambda_1 = i\theta$, $\boldsymbol{u}_1 = \left(\begin{smallmatrix} 1 \\ i\theta \end{smallmatrix}\right)$; $\lambda_2 = -i\theta$, $\boldsymbol{u}_2 = \left(\begin{smallmatrix} 1 \\ -i\theta \end{smallmatrix}\right)$. Di conseguenza, sulla base del metodo della variazione dei parametri, possiamo scrivere

$$\boldsymbol{v}(x) = \alpha_1(x)\,\boldsymbol{u}_1 + \alpha_2(x)\,\boldsymbol{u}_2 \quad \text{e} \quad \boldsymbol{f}(x) = c_1(x)\,\boldsymbol{u}_1 + c_2(x)\,\boldsymbol{u}_2$$

con $\alpha_1(x)$ e $\alpha_2(x)$ incognite e $c_1(x)$ e $c_2(x)$ soluzioni del sistema

$$\begin{cases} c_1(x) + c_2(x) = 0, \\ i\theta[c_1(x) - c_2(x)] = f(x), \end{cases} \quad \Longrightarrow c_1(x) = \frac{f(x)}{2i\theta},\ c_2(x) = -\frac{f(x)}{2i\theta}.$$

Sostituendo $\boldsymbol{v}(x)$ nella (2.18) e tenendo conto dello spettro della A e della rappresentazione di $\boldsymbol{f}(x)$, possiamo affermare che

$$\begin{aligned} \boldsymbol{v}'(x) &= \alpha_1'(x)\,\boldsymbol{u}_1 + \alpha_2'(x)\,\boldsymbol{u}_2 = A\,\boldsymbol{v}(x) + \boldsymbol{f}(x) \\ &= \alpha_1(x)\lambda_1\boldsymbol{u}_1 + \alpha_2(x)\lambda_2\boldsymbol{u}_2 + c_1(x)\,\boldsymbol{u}_1 + c_2(x)\,\boldsymbol{u}_2, \end{aligned}$$

dalla quale segue che $\alpha_1(x)$ e $\alpha_2(x)$ sono soluzione delle due equazioni disaccoppiate

$$\begin{cases} \alpha_1'(x) = \lambda_1\alpha_1(x) + c_1(x), \\ \alpha_2'(x) = \lambda_2\alpha_2(x) + c_2(x), \end{cases} \Longrightarrow \begin{cases} \alpha_1(x) = e^{\lambda_1 x}\left[\displaystyle\int_0^x e^{-\lambda_1 t}c_1(t)\,dt + c\right], \\ \alpha_2(x) = e^{\lambda_2 x}\left[\displaystyle\int_0^x e^{-\lambda_2 t}c_2(t)\,dt + \hat{c}\right], \end{cases}$$

con c e $\hat{c}$ costanti arbitrarie. La soluzione generale dell'equazione (2.17a) è pertanto

$$\begin{aligned} v(x) &= \alpha_1(x) + \alpha_2(x) \\ &= \frac{1}{2i\theta}\left\{e^{i\theta x}\left[\int_0^x e^{-i\theta t} f(t)\,dt + c\right] - e^{-i\theta x}\left[\int_0^x e^{i\theta t} f(t)\,dt + \hat{c}\right]\right\} \\ &= \frac{1}{2i\theta}\left\{\int_0^x \left[e^{i\theta(x-t)} - e^{-i\theta(x-t)}\right] f(t)\,dt + c\,e^{i\theta x} + \hat{c}\,e^{-i\theta x}\right\} \\ &= \int_0^x \frac{\sin\theta(x-t)}{\theta} f(t)\,dt + \frac{c+\hat{c}}{2i}\,\frac{\cos\theta x}{\theta} + \frac{c-\hat{c}}{2}\,\frac{\sin\theta x}{\theta}, \end{aligned}$$

con c e $\hat{c}$ determinabili imponendo le condizioni $v(a) = \alpha$ e $v(b) = \beta$.

2.7 Risoluzione numerica dei sistemi di ODEs

Sull'argomento esistono numerosi libri a carattere introduttivo e specialistico. Nel primo settore rientrono i libri di Analisi Numerica (per es. [17, 28, 24]) e nel secondo quelli specifici (per es. [3, 14]). In ciascuno di essi ampio spazio viene riservato alla famiglia di metodi di tipo Runge-Kutta, con particolare riguardo a quelli a passo variabile di ordine 4 e 5 e loro combinazioni. Il metodo, probabilmente più diffuso, è il cosiddetto ODE45 (presente nel MATLAB) che utilizza una combinazione di metodi di Runge-Kutta (a passo variabile) del quarto e del quinto ordine. Esso consente di risolvere con successo la generalità dei sistemi di ODE, con l'importante eccezione di quelli definiti STIFF, ossia di quelli nei quali è particolarmente elevato il rapporto tra il massimo e minimo valore assoluto degli autovalori della matrice A del sistema. Vediamo di illustrare il problema con un interessante esempio riportato in [8, pag. 115].

Consideriamo

$$\begin{cases} \dfrac{d\boldsymbol{x}}{dt} = A_1\boldsymbol{x}, \\ \boldsymbol{x}(0) = \boldsymbol{\alpha}, \end{cases} \qquad \begin{cases} \dfrac{d\boldsymbol{y}}{dt} = A_2\boldsymbol{y}, \\ \boldsymbol{y}(0) = \boldsymbol{\alpha}, \end{cases} \tag{2.19}$$

due ODE di tipo omogeneo con le stesse condizioni iniziali. Per semplificare l'analisi, assegniamo A_1 e A_2 mediante le loro decomposizioni spettrali, supponendo che siano caratterizzate da autovalori diversi e autovettori uguali. Più precisamente, supponiamo che

$$A_1 = UD_1U^T \quad \text{e} \quad A_2 = UD_2U^T, \text{ con } U = \begin{pmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \boldsymbol{u}_3 \end{pmatrix},$$

essendo

$$\boldsymbol{u}_1 = \frac{1}{\sqrt{3}}\begin{pmatrix}1\\1\\1\end{pmatrix}, \quad \boldsymbol{u}_2 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\0\\-1\end{pmatrix}, \quad \boldsymbol{u}_3 = \frac{1}{\sqrt{6}}\begin{pmatrix}1\\-2\\1\end{pmatrix};$$

$$D_1 = \mathrm{diag}(-1,-2,-3) \quad \text{e} \quad D_2 = \mathrm{diag}(-1,-10,-100) \quad \text{e} \quad \boldsymbol{\alpha} = \sqrt{3}\,\boldsymbol{u}_1.$$

Avendo i due problemi lo stesso vettore iniziale e gli stessi autovettori, la differenza riguarda gli autovalori delle due matrici e soprattutto la differenza tra il massimo e minimo valore assoluto degli autovalori. Tale rapporto è infatti 3 per A_1 e 100 per A_2. Il metodo spettrale precedentemente illustrato permette di calcolare facilmente le rispettive soluzioni. Ponendo infatti

$$\boldsymbol{x}(t) = \sum_{j=1}^{3} a_j(t)\,\boldsymbol{u}_j,$$

si ottiene che

$$\sum_{j=1}^{3} a_j'(t)\,\boldsymbol{u}_j = \sum_{j=1}^{3} \lambda_j a_j(t)\,\boldsymbol{u}_j,$$

dalla quale segue che $a_j'(t) = \lambda_j a_j(t)$, ossia $a_j(t) = e^{\lambda_j t} a_j(0)$, $j = 1,2,3$. Di conseguenza,

$$\boldsymbol{x}(t) = a_1(0)e^{-t}\,\boldsymbol{u}_1 + a_2(0)e^{-2t}\,\boldsymbol{u}_2 + a_3(0)e^{-3t}\,\boldsymbol{u}_3$$

con $a_1(0)$, $a_2(0)$ e $a_3(0)$ determinati imponendo che

$$\boldsymbol{x}(0) = \boldsymbol{\alpha} = \sqrt{3}\,\boldsymbol{u}_1 \Longrightarrow a_1(0) = \sqrt{3},\ a_2(0) = a_3(0) = 0,$$

per cui

$$\boldsymbol{x}(t) = \sqrt{3}\,e^{-t}\,\boldsymbol{u}_1 = e^{-t}\boldsymbol{\alpha}.$$

Utilizzando lo stesso procedimento si trova che

$$\boldsymbol{y}(t) = b_1(0)e^{-t}\,\boldsymbol{u}_1 + b_2(0)e^{-10t}\,\boldsymbol{u}_2 + b_3(0)e^{-100t}\,\boldsymbol{u}_3,$$

da cui, imponendo che $\boldsymbol{y}(0) = \boldsymbol{\alpha} = \sqrt{3}\,\boldsymbol{u}_1$, segue che $b_1(0) = \sqrt{3}$, $b_2(0) = b_3(0) = 0$. Di conseguenza

$$\boldsymbol{y}(t) = \boldsymbol{x}(t) = e^{-t}\,\boldsymbol{\alpha},$$

ossia $e^{-t}\,\boldsymbol{\alpha}$ è la soluzione di ciascuno dei due problemi.

Nonostante ciò, il numero dei passi occorrente per ottenere una precisione dell'ordine di $(\Delta t)^4$ in $t = 0, 1$ è, nel secondo esempio, circa 200 volte quello richiesto nel primo. Risultato che dipende del diverso rapporto $(|\lambda_{\max}|/|\lambda_{\min}|) = 3$ nel primo caso e 100 nel secondo. Nel secondo esempio, come in tutte le situazioni simili è importante, e talvolta necessario, ricorrere a metodi di tipo implicito, in luogo di metodi espliciti, come lo sono i metodi di tipo Runge-Kutta in generale e il metodo ODE45 in particolare.

Esempio 2.10 (di problemi stiff)

$$\frac{d\boldsymbol{y}}{dx} = f(x, \boldsymbol{y}) = \begin{pmatrix} -10^6\, y_1 \\ -y_2 \end{pmatrix}, \qquad \boldsymbol{y}(0) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

La soluzione, come è immediato verificare, è

$$\boldsymbol{y}(t) = \begin{pmatrix} e^{-10^6 t} \\ -e^{-t} \end{pmatrix}.$$

L'applicazione del metodo di Eulero esplicito, indicato con Δt il passo temporale unitario, richiede che

$$\boldsymbol{y}^{(n+1)} = \boldsymbol{y}^{(n)} + f(x_n, \boldsymbol{y}^{(n)})\Delta t, \qquad n = 0, 1, 2, \ldots,$$

ossia che

$$\begin{cases} y_1^{(n+1)} = y_1^{(n)} - 10^6 \Delta t\, y_1^{(n)} = (1 - 10^6 \Delta t)^2 y_1^{(n-1)} = \ldots = (1 - 10^6 \Delta t)^{n+1}, \\ y_2^{(n+1)} = y_2^{(n)} - (\Delta t) y_2^{(n)} = (1 - \Delta t) y_2^{(n)} = (1 - \Delta t)^2 y_2^{(n-1)} = \ldots = (1 - \Delta t)^{n+1}. \end{cases}$$

Da tali formule di ricorrenza segue che, per ragioni di stabilità,

$$|1 - 10^6 \Delta t| < 1 \quad \text{e} \quad |1 - \Delta t| < 1 \Longrightarrow 0 < \Delta t < 2\, 10^{-6}.$$

In altri termini, la stabilità numerica richiede che il passo temporale sia molto piccolo. Se Δt è esterno alla regione di stabilità, si verifica il cosiddetto "blow up", ossia una abnorme crescita dei valori assoluti degli iterati. Se, per esempio, $\Delta t = 10^{-5}$, si ottengono gli iterati

$$y_1^{(1)} = (1 - 10) y_1^{(0)} = -9, \; y_1^{(2)} = (-9)^2, \ldots, y_1^{(n)} = (-9)^n,$$

per cui $|y_1^{(n)}| \to +\infty$ per $n \to \infty$. Inoltre

$$\begin{aligned} y_2^{(1)} &= (1 - 10^{-5}) y_2^{(0)} = 1 - 10^{-5}, \\ y_2^{(2)} &= (1 - 10^{-5})^2, \ldots, \\ y_2^{(n)} &= (1 - 10^{-5})^n \simeq 1 - 10^{-5}\, n. \end{aligned}$$

Situazioni non accettabili, anche se non "esplosive", si presentano quando ci si avvicina alla frontiera sella regione di stabilità. Per $\Delta t = 10^{-6}$, ad esempio, si ottengono gli iterati

$$\begin{cases} y_1^{(n)} = 0 \text{ per } n = 0, 1, 2, \dots, \\ y_2^{(n)} = (1 - 10^{-6})^n \text{ per } n = 0, 1, 2, \dots. \end{cases}$$

Per evitare questo tipo di problemi è necessario ricorrere a metodi impliciti, ossia a metodi che, ad ogni passo, richiedono la risoluzione di un sistema di equazioni. Situazione che, per quanto possibile, si cerca di evitare soprattutto quando il sistema da risolvere è non lineare. Utilizzando (nel caso specifico) il metodo di Eulero implicito, ad ogni iterazione si deve risolvere il sistema

$$\boldsymbol{y}^{(n+1)} = \boldsymbol{y}^{(n)} + \boldsymbol{f}(x_n, \boldsymbol{y}^{(n+1)}),$$

che, in forma scalare, diventa

$$y_1^{(n+1)} = y_1^{(n)} - 10^6 \Delta t \, y_1^{(n+1)} \Rightarrow y_1^{(n+1)} = \frac{y_1^{(n)}}{1 + 10^6 \Delta t} = \dots = \frac{1}{(1 + 10^6 \Delta t)^{n+1}},$$

$$y_2^{(n+1)} = y_2^{(n)} - \Delta t \, y_2^{(n+1)} \Rightarrow y_2^{(n+1)} = \frac{y_2^{(n)}}{1 + \Delta t} = \frac{y_2^{(n-1)}}{(1 + \Delta t)^2} = \dots = \frac{1}{(1 + \Delta t)^{n+1}}.$$

Ricorrenze che evidenziano la stabilità del metodo, indipendentemente da Δt, in quanto

$$\left|(1 + 10^6 \Delta t)^{-1}\right| < 1 \quad \text{e} \quad \left|(1 + \Delta t)^{-1}\right| < 1,$$

per ogni $\Delta t > 0$. Per $\Delta t < 10^{-5}$, ad esempio, si ottengono gli iterati

$$y_1^{(n)} = \left(\frac{1}{1 + 10}\right)^n, \quad y_2^{(n)} = \left(\frac{1}{1 + 10^{-5}}\right)^n, \qquad n = 0, 1, 2, \dots.$$

Valori sostanzialmente accettabili, in quanto

$$y_1(10^{-5} n) = e^{-10n} = e^{-10n}, \qquad y_2(10^{-5} n) = e^{-10^{-5} n} = e^{-10^{-5} n}.$$

Considerazioni analoghe valgono se si considera l'esempio (2.19) caratterizzato dalla matrice A_2. In questo caso la regione di stabilità, relativa al metodo di Eulero diretto, è caratterizzata da un passo temporale

$$0 < \Delta t < 0.02.$$

Tale limitazione segue facilmente dalla applicazione del metodo di Eulero e dell'osservazione che

$$A_2 = V D_2 V^T \quad \text{con } V \text{ matrice ortogonale } (VV^T = I).$$

Applicando il metodo, per $n = 0, 1, 2, \ldots$ abbiamo

$$\begin{aligned}\boldsymbol{y}_2^{(n+1)} &= \boldsymbol{y}_2^{(n)} + \Delta t\, V D_2 V^T \boldsymbol{y}_2^{(n)} = V V^T \boldsymbol{y}_2^{(n)} + \Delta t\, V D_2 V^T \boldsymbol{y}_2^{(n)} \\ &= V(I + \Delta t\, D_2) V^T \boldsymbol{y}_2^{(n)},\end{aligned}$$

da cui, posto $V^T \boldsymbol{y}_2^{(n)} = \boldsymbol{z}^{(n)}$, segue

$$\boldsymbol{z}^{(n+1)} = (I + \Delta t D_2)\boldsymbol{z}^{(n)} \Longleftrightarrow \begin{cases} z_1^{(n+1)} = (1 - \Delta t) z_1^{(n)}, \\ z_2^{(n+1)} = (1 - 10\Delta t) z_2^{(n)}, \\ z_3^{(n+1)} = (1 - 100\Delta t) z_3^{(n)}. \end{cases}$$

Il metodo è pertanto stabile solo se

$$|1 - \Delta t| < 1, \qquad |1 - 10\Delta t| < 1, \qquad |1 - 100\Delta t| < 1,$$

ossia se $0 < \Delta t < 0.02$. Di conseguenza, se si considera un passo temporale maggiore, per esempio $\Delta t = 0.03$, si presenta il blow-up ossia una crescita abnorme dei valori degli iterati. Se invece si applica il metodo di Eulero implicito, tale fenomeno non si presenta e si ottengono ragionevoli valori della relazione anche con un passo temporale $\Delta t = 0.03$. Il metodo implicito non richiede invece alcuna limitazione sul passo Δt. Per rendersene conto basta osservare che la sua applicazione richide che

$$\boldsymbol{y}_2^{(n+1)} = \boldsymbol{y}_2^{(n)} + \Delta t\, V D_2 V^T \boldsymbol{y}_2^{(n+1)},$$

dalla quale, usando le stesse precedenti notazioni, deriva la successione

$$(1 - \Delta t\, D_2)\boldsymbol{z}^{(n+1)} = \boldsymbol{z}^{(n)} \Longleftrightarrow \boldsymbol{z}^{(n+1)} = (I - \Delta t\, D_2)^{-1} \boldsymbol{z}^{(n)}.$$

Successione che, in forma scalare, è così esprimibile

$$\begin{cases} z_1^{(n+1)} = (1 + \Delta t)^{-1} z_1^{(n)}, \\ z_2^{(n+1)} = (1 + 10\Delta t)^{-1} z_2^{(n)}, \\ z_3^{(n+1)} = (1 + 100\Delta t)^{-1} z_3^{(n)}. \end{cases}$$

Le precedenti considerazioni valgono unicamente per sistemi lineari. Nel caso non lineare il comportamento è molto più complicato, in quanto possono presentarsi singolarità anche al finito. In questo caso il "blow-up" può risultare intrinseco al problema come si può osservare analizzando, ad esempio, l'equazione [20]

$$y' = ty(y - 2), \qquad y(0) = y_0.$$

La sua esatta soluzione è

$$y(t) = \frac{2y_0}{y_0 + (2 - y_0)e^{t^2}}.$$

Dalla sua espressione segue immediatamente che:

(a) $y(t) \equiv 2$ per $y_0 = 2$;

(b) $y(t)$ converge asintoticamente a zero per $0 < y_0 < 2$;

(c) Per $y_0 > 2$, $y(t)$ ha una singolarità al finito a $t = \tau$, con

$$\tau(y_0) = \sqrt{\ln\left(\frac{y_0}{y_0 - 2}\right)}.$$

Inoltre, $y(t) \to +\infty$ se $t \to \tau(y_0)^-$.

La soluzione dell'equazione differenziale è pertanto instabile rispetto alla sua condizione iniziale in quanto per $y_0 = 2$ è costante, per $0 < y_0 < 2$ tende asintoticamente a zero per $t \to +\infty$ e per $y_0 > 2$ presenta un asintoto verticale in $t = \tau(y_0)$.

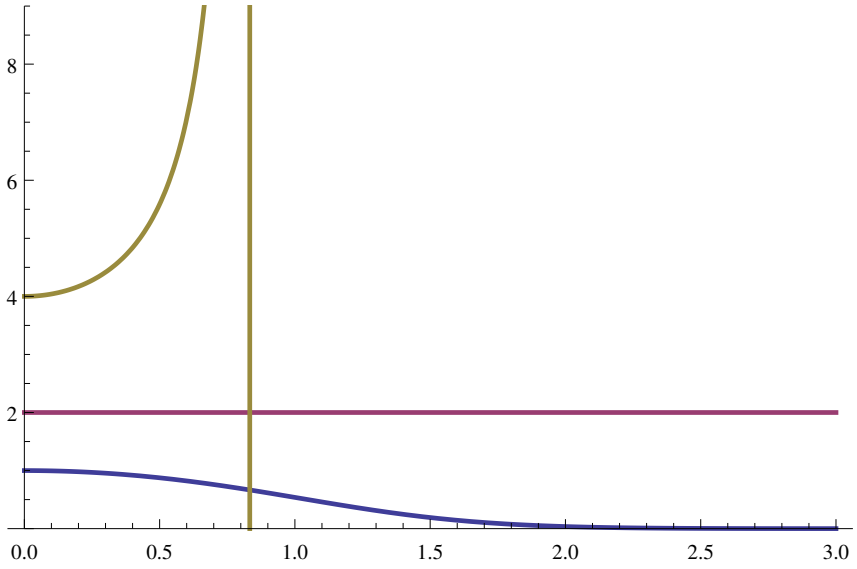


Figura 2.1: La figura mostra i grafici della funzione $y(t) = \frac{2y_0}{y_0+(2-y_0)e^{t^2}}$ per $y_0 = 4$ e per $y_0 = 1$.

Capitolo 3

SERIE DI FOURIER E PROBLEMI SPETTRALI DI STURM-LIOUVILLE

3.1 Funzioni periodiche e polinomi trigonometrici

Definizione. Una funzione $f(x)$ definita in un dominio D è detta periodica di periodo T se, qualunque sia $x \in D$, $x+T$ appartiene a D ed inoltre $f(x) = f(x+T)$.

In tal caso T è il periodo e $1/T$ è la frequenza della f. Se $f(x)$ è periodica con periodo T, lo è anche di periodo $2T, 3T, \ldots$. Per questo motivo, talvolta si preferisce precisare che T è il *periodo fondamentale*.

Per esempio le funzioni $\sin x$ e $\cos x$ sono periodiche di periodo 2π, mentre $\sin \omega x$ e $\cos \omega x$ ($\omega \neq 0$), sono periodiche di periodo $\frac{2\pi}{\omega}$. Se k è un intero positivo, anche $\sin k\omega x$ e $\cos k\omega x$ sono periodiche con periodo fondamentale $(2\pi/k\omega)$.

Spesso, in vari contesti applicativi, una funzione definita in un intervallo di ampiezza limitata viene estesa per periodicità a tutto $\mathbb{R}$. Ad esempio, la funzione

$$f(x) = \begin{cases} x, & 0 \leq x < 1, \\ 2 - x, & 1 \leq x < 2, \end{cases}$$

può essere estesa per periodicità su $\mathbb{R}$, ponendo $f(x) = f(x+2)$.

Armoniche elementari. Le funzioni

1. $f_k(x) = a_k \cos k\omega x + b_k \sin k\omega x$, k intero positivo, $\omega \neq 0$,

2. $g_k(x) = \rho_k \cos(k\omega x + \theta_k)$, k intero positivo, $\rho > 0$, $\omega \neq 0$, $\theta_k \in \mathbb{R}$,

dette *armoniche elementari*, sono periodiche di periodo $\frac{2\pi}{\omega}$. Esse sono inoltre equivalenti, nel senso che, assegnate le costanti che caratterizzano una forma, è sempre possibile passare dalla (1) alla (2) e viceversa. Per esempio, assegnati ρ_k e θ_k, è immediato calcolare a_k e b_k. Infatti, essendo

$$\begin{aligned} g_k(x) &= \rho_k \cos(k\omega x + \theta_k) = \rho_k \{\cos(k\omega x)\cos(\theta_k) - \sin(k\omega x)\sin(\theta_k\} \\ &= \rho_k \cos(\theta_k)\cos(k\omega x) - \rho_k \sin(\theta_k)\sin(k\omega x) \end{aligned}$$

basta porre $a_k = \rho_k \cos(\theta_k)$ e $b_k = -\rho_k \sin(\theta_k)$.

Viceversa, noti a_k e b_k, è immediato determinare ρ_k e θ_k. Infatti, essendo

$$\begin{cases} \rho_k \cos\theta_k = a_k, \\ \rho_k \sin\theta_k = -b_k, \end{cases}$$

si avrà $\rho_k = \sqrt{a_k^2 + b_k^2}$ e $\theta_k = \operatorname{arctg}\left(-\frac{b_k}{a_k}\right)$ se $a_k \neq 0$. In particolare, se $b_k = 0$ si ha $\rho_k = |a_k|$ e $\theta = 0$ oppure $\theta = \pi$ a seconda che sia $a_k > 0$ o $a_k < 0$.

Polinomio trigonometrico. Per polinomio trigonometrico di ordine n, intendiamo una funzione del tipo

$$T_n(x) = a_0 + \sum_{k=1}^{n} (a_k \cos k\omega x + b_k \sin k\omega x),$$

dove a_0, a_k, b_k e $\omega \neq 0$ sono numeri reali.

T_n è una funzione periodica, in quanto combinazione lineare delle $2n+1$ funzioni elementari

$$1,\ \cos\omega x,\ \sin\omega x, \ldots,\ \cos n\omega x,\ \sin n\omega x, \tag{3.1}$$

tutte periodiche con periodo fondamentale $T = \frac{2\pi}{\omega}$.

Teorema 3.1 *Le funzioni trigonometriche* (3.1)*, qualunque sia n, sono mutuamente ortogonali in* $[0, T]$.

Dimostrazione. La funzione costante $f(x) = 1$ è ortogonale a tutte le altre, in quanto, per $k = 1, 2, \ldots, n$,

$$\begin{aligned} \int_0^T 1 \cdot \cos k\omega x\, dx &= \left[\frac{\sin k\omega x}{k\omega}\right]_0^T = \frac{\sin k\omega T}{k\omega} = \frac{\sin 2ik\pi}{k\omega} = 0, \\ \int_0^T 1 \cdot \sin k\omega x\, dx &= -\left[\frac{\cos k\omega x}{k\omega}\right]_0^T = \frac{1 - \cos 2k\pi}{k\omega} = 0. \end{aligned}$$

Inoltre se $h \neq k$,

$$\int_0^T \cos k\omega x \cos h\omega x \, dx = \int_0^T \tfrac{1}{2} \{\cos(k-h)\omega x + \cos(k+h)\omega x\} \, dx$$
$$= \frac{1}{2} \left[\frac{\sin(k-h)\omega x}{(k-h)\omega} + \frac{\sin(k+h)\omega x}{(k+h)\omega} \right]_0^T = 0,$$

in quanto $\omega T = 2\pi$. Da notare che se $h = k$,

$$\int_0^T \cos^2 k\omega x \, dx = \int_0^T \tfrac{1}{2} \{1 + \cos 2k\omega x\} \, dx$$
$$= \frac{1}{2} \left[x + \frac{\sin 2k\omega x}{2k\omega} \right]_0^T = \frac{T}{2}.$$

Pertanto i due risultati possono essere espressi nella forma

$$\int_0^T \cos k\omega x \, \cos h\omega x \, dx = \frac{T}{2} \delta_{hk},$$

essendo δ_{hk} il simbolo di Kronecker, così definito:

$$\delta_{hk} = \begin{cases} 1, & h = k, \\ 0, & h \neq k. \end{cases}$$

Analogamente, se $h \neq k$,

$$\int_0^T \sin k\omega x \, \sin h\omega x \, dx = \int_0^T \tfrac{1}{2} \{\cos(k-h)\omega x - \cos(k+h)\omega x\} \, dx$$
$$= \frac{1}{2} \left[\frac{\sin(k-h)\omega x}{(k-h)\omega} - \frac{\sin(k+h)\omega x}{(k+h)\omega} \right]_0^T = 0,$$

e, se $h = k$,

$$\int_0^T \sin^2 k\omega x \, dx = \int_0^T \tfrac{1}{2} \{1 - \cos 2k\omega x\} \, dx$$
$$= \frac{1}{2} \left[x - \frac{\sin 2k\omega x}{2k\omega} \right]_0^T = \frac{T}{2}.$$

Pertanto, indipendentemente dall'essere h e k uguali o diversi,

$$\int_0^T \sin k\omega x \, \sin h\omega x \, dx = \frac{T}{2} \delta_{hk}.$$

Infine è immediato osservare che $\sin h\omega x$ e $\cos k\omega x$ sono ortogonali in $[0, T]$, qualunque sia la coppia di valori h, k. Infatti, se $h \neq k$,

$$\begin{aligned}\int_0^T \sin k\omega x \, \cos h\omega x \, dx &= \int_0^T \tfrac{1}{2} \left\{\sin(k-h)\omega x + \sin(k+h)\omega x\right\} dx \\ &= \frac{1}{2}\left[-\frac{\cos(k-h)\omega x}{(k-h)\omega} - \frac{\cos(k+h)\omega x}{(k+h)\omega}\right]_0^T = 0,\end{aligned}$$

e inoltre, se $k = h$,

$$\int_0^T \sin k\omega x \, \cos h\omega x \, dx = \int_0^T \tfrac{1}{2} \sin 2h\omega x \, dx = \frac{1}{2}\left[-\frac{\cos 2h\omega x}{2h\omega}\right]_0^T = 0.$$

□

Dimostriamo ora un'importante proprietà sull'integrazione delle funzioni periodiche.

Teorema 3.2 *Indicato con T il periodo di una funzione periodica e integrabile e con a un qualunque numero reale, vale la seguente proprietà:*

$$\int_a^{a+T} f(x)\, dx = \int_0^T f(x)\, dx.$$

Dimostrazione. Qualunque sia $a \in \mathbb{R}$, esiste un intero n tale che $nT \leq a < (n+1)T$. Pertanto, essendo $(n+1)T \leq a + T$,

$$\begin{aligned}\int_a^{a+T} f(x)\, dx &= \int_a^{(n+1)T} f(x)\, dx + \int_{(n+1)T}^{a+T} f(x)\, dx \\ &= \int_a^{(n+1)T} f(x)\, dx + \int_{nT}^{a} f(y+T)\, dy \\ &= \int_a^{(n+1)T} f(x)\, dx + \int_{nT}^{a} f(x)\, dx \\ &= \int_{nT}^{(n+1)T} f(x)\, dx \\ &= \int_0^T f(z+nT)\, dz \\ &= \int_0^T f(x)\, dx.\end{aligned}$$

Dalla proprietà segue, in particolare, che

$$\int_{-\frac{T}{2}}^{\frac{T}{2}} f(x)\, dx = \int_0^T f(x)\, dx,$$

qualunque sia il valore di T. □

Coefficienti di Fourier. Supponiamo che una funzione $f(x)$ sia *bene approssimata* in $[0,T]$ mediante un polinomio trigonometrico di ordine n

$$f(x) \sim a_0 + \sum_{k=1}^{n}(a_k \cos k\omega x + b_k \sin k\omega x), \qquad T = \frac{2\pi}{\omega}.$$

È allora naturale richiedere che l'integrazione in $[0,T]$ della $f(x)$ per le funzioni di base

$$\{1, \cos\omega x, \sin\omega x, \ldots, \cos k\omega x, \sin k\omega x\}$$

risulti sostanzialmente uguale all'integrazione della funzione approssimante per le stesse funzioni di base. Questo fatto, in conseguenza della ortogonalità delle funzioni di base, implica che

$$\begin{aligned}
\int_0^T f(x)dx &\simeq a_0 \int_0^T 1\,dx = a_0 T, \\
\int_0^T f(x)\cos k\omega x\,dx &\simeq a_k \int_0^T \cos^2 k\omega x\,dx = a_k \tfrac{T}{2}, \qquad k = 1,2,\ldots,n, \\
\int_0^T f(x)\sin k\omega x\,dx &\simeq b_k \int_0^T \sin^2 k\omega x\,dx = b_k \tfrac{T}{2}, \qquad k = 1,2,\ldots,n.
\end{aligned}$$

Poiché questa osservazione è dovuta a Fourier, i coefficienti

$$\begin{aligned}
a_0 &= \frac{1}{T}\int_0^T f(x)\,dx, \\
a_k &= \frac{2}{T}\int_0^T f(x)\cos k\omega x\,dx, \qquad k = 1,2,\ldots,n, \\
b_k &= \frac{2}{T}\int_0^T f(x)\sin k\omega x\,dx, \qquad k = 1,2,\ldots,n,
\end{aligned}$$

vengono definiti *coefficienti di Fourier* della f.

Energia di un segnale. Supponendo che una funzione $f(x)$, al quadrato integrabile in $[0,T]$, rappresenti un segnale a valori reali o complessi, si definisce energia del segnale la norma della f così definita

$$\|f\| = \left(\int_0^T |f(x)|^2\,dx\right)^{\frac{1}{2}}, \qquad |f(x)|^2 = f(x)\overline{f(x)},$$

dove $\overline{f(x)}$ indica il complesso coniugato di $f(x)$ se $f(x)$ è complessa, $f(x)$ stessa se $f(x)$ è reale.

Indicato con T_n un polinomio trigonometrico, il quadrato della sua energia può essere agevolmente calcolato nel modo seguente:

$$\begin{aligned}
\|T_n\|^2 &= \int_0^T \left[a_0 + \sum_{k=1}^{n} (a_k \cos k\omega x + b_k \sin k\omega x) \right]^2 dx \\
&= \int_0^T \left[a_0^2 + 2a_0 \sum_{k=1}^{n} (a_k \cos k\omega x + b_k \sin k\omega x) \right. \\
&\qquad \left. + \sum_{k,l=1}^{n} (a_k \cos k\omega x + b_k \sin k\omega x)(a_l \cos l\,\omega x + b_l \sin l\,\omega x) \right] dx \\
&= \int_0^T \left[a_0^2 + \sum_{k=1}^{n} (a_k^2 \cos^2 k\omega x + b_k^2 \sin^2 k\omega x) \right] dx \\
&= T \left[a_0^2 + \tfrac{1}{2} \sum_{k=1}^{n} \left(a_k^2 + b_k^2 \right) \right],
\end{aligned}$$

essendo $\omega = (2\pi/T)$ e $\int_0^T \cos^2 k\omega x\, dx = \int_0^T \sin^2 k\omega x\, dx = \frac{T}{2}$.

È interessante capire quale sia la distanza minima, in termini di energia, tra una $f(x)$ in $[0, T]$ ed una sua approssimante trigonometrica. Interessa dunque minimizzare la funzione

$$I(a_0, a_1, b_1, \ldots, a_n, b_n) = \int_0^T [f(x) - T_n(x)]^2\, dx$$

al variare dei coefficienti $a_0, a_1, b_1, \ldots, a_n, b_n$ di $T_n(x)$. È possibile dimostrare che, grazie alla ortogonalità delle funzioni di base, tale minimo è determinato dai coefficienti di Fourier.

Dimostrazione. Per brevità, limitiamoci a dimostrare il risultato per $n = 1$. A tale scopo consideriamo la funzione

$$I(a_0, a_1, b_1) = \int_0^T [f(x) - (a_0 + a_1 \cos \omega x + b_1 \sin \omega x)]^2\, dx$$

e indichiamo con

$$\begin{aligned}
\hat{a}_0 &= \frac{1}{T} \int_0^T f(x)\, dx, \\
\hat{a}_1 &= \frac{2}{T} \int_0^T f(x) \cos \omega x\, dx, \\
\hat{b}_1 &= \frac{2}{T} \int_0^T f(x) \sin \omega x\, dx,
\end{aligned}$$

i coefficienti di Fourier della $f(x)$ in $[0,T]$. Per l'ortogonalità delle funzioni di base e la definizione di $\hat{a}_0$, $\hat{a}_1$ e $\hat{b}_1$

$$\begin{aligned} I(a_0,a_1,b_1) &= \int_0^T f^2(x)\,dx + T\,a_0^2 + \tfrac{T}{2}(a_1^2+b_1^2) \\ &\quad - 2\left[a_0\int_0^T f(x)\,dx + a_1\int_0^T f(x)\cos\omega x\,dx + b_1\int_0^T f(x)\sin\omega x\,dx\right] \\ &= \int_0^T f^2(x)dx + T\,a_0^2 + \tfrac{T}{2}(a_1^2+b_1^2) - 2\left[T\,a_0\hat{a}_0 + \tfrac{T}{2}(a_1\hat{a}_1 + b_1\hat{b}_1)\right], \end{aligned}$$

da cui, aggiungendo e togliendo $T\hat{a}_0^2$, $\frac{T}{2}\hat{a}_1^2$ e $\frac{T}{2}\hat{b}_1^2$,

$$\begin{aligned} I(a_0,a_1,b_1) = \int_0^T f^2(x)\,dx - T\,\hat{a}_0^2 + T(a_0-\hat{a}_0)^2 - \tfrac{T}{2}\hat{a}_1^2 \\ + \tfrac{T}{2}(a_1-\hat{a}_1)^2 - \tfrac{T}{2}\hat{b}_1^2 + \tfrac{T}{2}(b_1-\hat{b}_1)^2, \end{aligned}$$

il cui minimo è ovviamente ottenuto per $a_0=\hat{a}_0$, $a_1=\hat{a}_1$ e $b_1=\hat{b}_1$.

Il quadrato dell'energia corrispondente alla differenza tra la $f(x)$ e la sua approssimante ottimale $\hat{T}_1(x) = \hat{a}_0 + \hat{a}_1\cos\omega x + \hat{b}_1\sin\omega x$ è dunque

$$I(\hat{a}_0,\hat{a}_1,\hat{b}_1) = \int_0^T f^2(x)dx - T\,\hat{a}_0^2 - \tfrac{T}{2}(\hat{a}_1^2+\hat{b}_1^2).$$

Estendendo tali considerazioni al caso generale si ottiene che il minimo di $I(a_0,a_1,b_1,\ldots,a_n,b_n)$ è ottenuto per $a_0=\hat{a}_0$, $a_k=\hat{a}_k$ e $b_k=\hat{b}_k$, $k=1,2,\ldots,n$. In altri termini, prefissato n, i coefficienti di Fourier identificano il polinomio trigonometrico che, in termini di energia, meglio approssima un segnale, esprimibile con una funzione al quadrato integrabile in $[0,T]$.

L'estensione delle precedenti considerazioni permette inoltre di affermare che

$$I(\hat{a}_0,\hat{a}_1,\hat{b}_1,\ldots,\hat{a}_n,\hat{b}_n) = \int_0^T f^2(x)\,dx - T\,\hat{a}_0^2 - \frac{T}{2}\sum_{k=1}^n\left(\hat{a}_k^2+\hat{b}_k^2\right)$$

da cui, essendo $I(\hat{a}_0,\hat{a}_1,\hat{b}_1,\ldots,\hat{a}_n,\hat{b}_n)\geq 0$, segue la *diseguaglianza di Bessel*

$$\hat{a}_0^2 + \frac{1}{2}\sum_{k=1}^n\left(\hat{a}_k^2+\hat{b}_k^2\right) \leq \frac{1}{T}\int_0^T f^2(x)\,dx.$$

Tale diseguaglianza esprime il fatto che l'energia associata al polinomio trigonometrico che meglio approssima un segnale è sempre inferiore a quella del segnale stesso.

Un'altra interessante osservazione è che

$$\|T_{n+1}\|^2 = \|T_n\|^2 + \frac{T}{2}\left(\hat{a}_{n+1}^2 + \hat{b}_{n+1}^2\right),$$

ossia: l'energia dell'approssimante T_n è una funzione *monotonamente* crescente rispetto a n e limitata dall'energia del segnale.

3.2 Serie di Fourier

Definizione. Sia f una funzione integrabile in $[-L, L]$. Allora ad essa è formalmente associabile la serie di Fourier

$$\mathcal{F}(x) = a_0 + \sum_{n=1}^{\infty}\left(a_n \cos\frac{n\pi x}{L} + b_n \sin\frac{n\pi x}{L}\right), \tag{3.2}$$

dove, a_0 e gli $\{a_n, b_n\}$ sono i coefficienti di Fourier, così definiti:

$$\begin{aligned} a_0 &= \frac{1}{2L}\int_{-L}^{L} f(x)\,dx \\ a_n &= \frac{1}{L}\int_{-L}^{L} f(x)\cos\frac{n\pi x}{L}\,dx \\ b_n &= \frac{1}{L}\int_{-L}^{L} f(x)\sin\frac{n\pi x}{L}\,dx. \end{aligned}$$

Esempio 3.3 Scrivere la serie di Fourier di $f(x) = x$, $-\pi \le x \le \pi$.

Dalla definizione della sua serie di Fourier segue immediatamente che:

$$\begin{aligned} a_0 &= \frac{1}{2\pi}\int_{-\pi}^{\pi} x\,dx = \frac{1}{4\pi}[x^2]_{-\pi}^{\pi} = 0; \\ a_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} x\,\cos kx\,dx = \frac{1}{\pi}\int_{-\pi}^{\pi} x\,d\frac{\sin kx}{k} = \\ &= \frac{1}{\pi}\left[x\frac{\sin kx}{k}\right]_{-\pi}^{\pi} - \frac{1}{\pi}\int_{-\pi}^{\pi}\frac{\sin kx}{k}\,dx = \frac{1}{\pi}\left[\frac{\cos kx}{k}\right]_{-\pi}^{\pi} = 0; \\ b_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} x\,\sin kx\,dx = -\frac{1}{\pi}\int_{-\pi}^{\pi} x\,d\frac{\cos kx}{k} = \\ &= -\frac{1}{\pi}\left[x\frac{\cos kx}{k}\right]_{-\pi}^{\pi} + \frac{1}{\pi}\int_{-\pi}^{\pi}\frac{\cos kx}{k}\,dx = \frac{2}{k}(-1)^{k+1}, \end{aligned}$$

in quanto $\cos k\pi = \cos(-k\pi) = (-1)^k$.

La serie di Fourier di x su $[-\pi, \pi]$ è dunque la

$$\mathcal{F}(x) = \sum_{k=1}^{\infty} \frac{2}{k}(-1)^{k+1} \sin kx.$$

Esempio 3.4 Scrivere la serie di Fourier di $f(x) = x^2$, $-\pi \leq x \leq \pi$.

$$\begin{aligned}
a_0 &= \frac{1}{2\pi}\int_{-\pi}^{\pi} x^2\,dx = \frac{1}{2\pi}\left[\frac{x^3}{3}\right]_{-\pi}^{\pi} = \frac{\pi^2}{3};\\
a_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} x^2\,\cos kx\,dx = \frac{1}{\pi}\int_{-\pi}^{\pi} x^2\,d\frac{\sin kx}{k} =\\
&= \frac{1}{\pi}\left[x^2\,\frac{\sin kx}{k}\right]_{-\pi}^{\pi} + \frac{1}{\pi}\int_{-\pi}^{\pi} 2x\,d\frac{\cos kx}{k^2} =\\
&= \frac{1}{\pi}\left[2x\,\frac{\cos kx}{k^2}\right]_{-\pi}^{\pi} - \frac{2}{\pi}\int_{-\pi}^{\pi}\frac{\cos kx}{k^2}\,dx = \frac{4}{k^2}(-1)^k;\\
b_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} x^2\,\sin kx\,dx = \frac{1}{\pi}\int_{-\pi}^{\pi} x^2\,d\frac{\cos kx}{k} =\\
&= \frac{1}{\pi}\left[x^2\,\frac{\cos kx}{k}\right]_{-\pi}^{\pi} + \frac{1}{\pi}\int_{-\pi}^{\pi} 2x\,d\frac{\sin kx}{k^2} =\\
&= \frac{1}{\pi}\left[2x\,\frac{\sin kx}{k^2}\right]_{-\pi}^{\pi} - \frac{2}{\pi}\int_{-\pi}^{\pi}\frac{\sin kx}{k^2}\,dx = \frac{2}{k}\left[\frac{\cos kx}{k^2}\right]_{-\pi}^{\pi} = 0.
\end{aligned}$$

Pertanto la serie di Fourier di x^2 su $[-\pi, \pi]$ è

$$\mathcal{F}(x) = \frac{\pi^2}{3} + \sum_{k=1}^{\infty}\frac{4}{k^2}(-1)^k \cos kx.$$

Esempio 3.5 Scrivere la serie di Fourier di

$$f(x) = \begin{cases} 0, & -\pi \leq x < 1,\\ 1, & 1 \leq x < 2,\\ 2, & 2 \leq x < \pi. \end{cases}$$

In conseguenza della definizione della sua serie di Fourier,

$$\begin{aligned}
a_0 &= \frac{1}{2\pi}\int_{-\pi}^{\pi} x\,dx = \frac{1}{2\pi}\left[\int_1^2 dx + \int_2^{\pi} 2\,dx\right] =\\
&= \frac{1}{2\pi}[1 + 2(\pi - 2)] = 1 - \frac{3}{2\pi};
\end{aligned}$$

$$
\begin{aligned}
a_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} x\, f(x)\cos kx\, dx = \frac{1}{\pi}\left[\int_1^2 \cos kx\, dx + 2\int_2^{\pi}\cos kx\, dx\right] = \\
&= \frac{1}{k\pi}[(\sin 2k - \sin k) + 2(\sin k\pi - 2\sin 2k)] = -\frac{1}{k\pi}(\sin k)(2\cos k + 1); \\
b_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin kx\, dx = \left[\int_1^2 \sin kx\, dx + 2\int_2^{\pi}\sin kx\, dx\right] = \\
&= -\frac{1}{k\pi}[(\cos 2k - \cos k) + 2(\cos k\pi - 2\cos 2k)] = \\
&= -\frac{1}{k\pi}\left[-\cos 2k - \cos k + 2(-1)^k\right].
\end{aligned}
$$

Di conseguenza, alla $f(x)$ possiamo associare la serie di Fourier

$$
\begin{aligned}
\mathcal{F}(x) = 1 - \frac{3}{2\pi} - \frac{1}{\pi}\sum_{k=1}^{\infty}\frac{1}{k}\{&[\sin k(\cos 2k + 1)]\cos kx \\
&+ \left[2(-1)^k - \cos k\,\cos 2k\right]\sin kx\Big\}.
\end{aligned}
$$

Esempio 3.6 Scrivere la serie di Fourier di

$$
f(x) = \begin{cases} 0, & -3 \le x < 0, \\ x, & 0 \le x < 3. \end{cases}
$$

Per la definizione della sua serie di Fourier,

$$
\begin{aligned}
a_0 &= \tfrac{1}{6}\int_{-3}^{3} f(x)\, dx = \tfrac{1}{6}\int_0^3 x\, dx = \tfrac{3}{4}; \\
a_k &= \tfrac{1}{3}\int_3^3 f(x)\cos\frac{k\pi x}{3}\, dx = \tfrac{1}{3}\int_3^3 x\cos\frac{k\pi x}{3}\, dx = \frac{3}{k^2\pi^2}\left[(-1)^k - 1\right]; \\
b_k &= \frac{1}{3}\int_3^3 f(x)\sin\frac{k\pi x}{3}\, dx = \frac{1}{3}\int_3^3 x\sin\frac{k\pi x}{3} dx = -\frac{3}{k\pi}(-1)^k.
\end{aligned}
$$

Pertanto, la serie di Fourier di $f(x)$ in $[-3,3]$ è

$$
\mathcal{F}(x) = \tfrac{3}{4} + \sum_{k=1}^{\infty}\left\{\frac{3}{k^2\pi^2}\left[(-1)^k - 1\right]\cos\frac{k\pi x}{3} - \frac{3}{k\pi}(-1)^k\sin\frac{k\pi x}{3}\right\}.
$$

Funzioni pari e dispari. Sia f una funzione definita in un intervallo $[-L, L]$ for every $L \in \mathbb{R}^+$. Se per ogni $x \in [-L, L]$, risulta:

$$
\begin{aligned}
&f(x) = f(-x), && \text{la funzione si dice pari;} \\
&f(x) = -f(-x), && \text{la funzione si dice dispari.}
\end{aligned}
$$

Se per almeno un valore di x non vale nessuna delle due precedenti condizioni, la f non è né pari né dispari. Di conseguenza, la funzione dell'Es. 3.3 è dispari, quella dell'Es. 3.4 è pari, mentre quelle degli esempi 3.5 e 3.6 non sono né pari né dispari.

È importante notare che se la funzione $f(x)$ è pari, allora $b_n = 0$ per $n = 1, 2, \ldots$ e la sua serie di Fourier si riduce ad una serie di soli coseni. Più precisamente risulta

Teorema 3.7 *Se la funzione f è pari, la sua serie di Fourier è*

$$\mathcal{F}(x) = a_0 + \sum_{n=1}^{\infty} a_n \cos \frac{n\pi x}{L}$$

dove

$$a_0 = \frac{1}{L}\int_0^L f(x)\,dx \quad e \quad a_n = \frac{2}{L}\int_0^L f(x)\cos\frac{n\pi x}{L}\,dx.$$

Dimostrazione. Tenuto conto dell'ipotesi di parità

$$\begin{aligned} a_0 &= \frac{1}{2L}\int_{-L}^{L} f(x)\,dx = \frac{1}{2L}\left[\int_{-L}^{0} f(x)\,dx + \int_0^L f(x)\,dx\right] \\ &= \frac{1}{2L}\left[-\int_{L}^{0} f(x)\,dx + \int_0^L f(x)\,dx\right] = \frac{1}{L}\int_0^L f(x)\,dx, \end{aligned}$$

dove nel secondo passaggio si è posto $t = -x$ nel primo integrale e si è fatto uso della parità della $f(x)$.

Analogamente

$$\begin{aligned} a_k &= \frac{1}{L}\int_{-L}^{L} f(x)\cos\frac{k\pi x}{L}\,dx \\ &= \frac{1}{L}\left[\int_{-L}^{0} f(x)\cos\frac{k\pi x}{L}\,dx + \int_0^L f(x)\cos\frac{k\pi x}{L}\,dx\right] \\ &= \frac{2}{L}\int_0^L f(x)\cos\frac{k\pi x}{L}\,dx, \end{aligned}$$

e infine

$$\begin{aligned} b_k &= \frac{1}{L}\int_{-L}^{L} f(x)\sin\frac{k\pi x}{L}\,dx \\ &= \frac{1}{L}\left[\int_{-L}^{0} f(x)\sin\frac{k\pi x}{L}\,dx + \int_0^L f(x)\sin\frac{k\pi x}{L}\,dx\right] = 0. \end{aligned}$$

□

Se la funzione $f(x)$ è dispari, allora $a_0 = 0$ e $a_n = 0$ per $n = 1, 2, \ldots$ e la sua serie di Fourier si riduce ad una serie di soli seni. Più precisamente risulta

Teorema 3.8 *Se la funzione f è dispari, la sua serie di Fourier è*

$$\mathcal{F}(x) = \sum_{n=1}^{\infty} b_n \sin \frac{n\pi x}{L},$$

dove

$$b_n = \frac{2}{L} \int_0^L f(x) \sin \frac{n\pi x}{L}\, dx.$$

Dimostrazione. Si dimostra in modo del tutto analogo, ossia decomponendo l'integrale tra $[-L, L]$ in uno tra $[-L, 0]$ e uno tra $[0, L]$, ponendo $t = -x$ nel primo integrale e ricordando che $f(-t) = -f(t)$. □

Osservazione. Non è detto che il valore della f in un generico $x \in [-L, L]$ coincida con il valore della sua serie di Fourier in x. In altri termini, non necessariamente la serie di Fourier di una funzione risulta ad essa convergente in ogni punto.

Ad esempio, la serie di Fourier di $f(x) = x$ in $[-\pi, \pi]$ è

$$\mathcal{F}(x) = \sum_{k=1}^{\infty} \frac{2}{k}(-1)^{k+1} \sin kx,$$

la quale assume il valore 0 in $x = \pm\pi$, mentre in tali punti $f(x) = \pm\pi$. Inoltre, mentre $f(\frac{\pi}{2}) = \frac{\pi}{2}$, non è affatto evidente che $\frac{\pi}{2}$ sia il valore della serie in tale punto.

Allo scopo di discutere il problema della convergenza è bene premettere le seguenti definizioni di derivata destra e derivata sinistra.

Definizione. Se la $f(x)$ ammette limite destro in x_0 ($f(x_0^+)$), si definisce *derivata destra* della f in x_0 il limite

$$f'(x_0^+) = \lim_{h \to 0^+} \frac{f(x_0 + h) - f(x_0^+)}{h},$$

nell'ipotesi che esso esista e sia finito.

Analogamente, se la f ammette limite sinistro in x_0 ($f(x_0^-)$), si definisce *derivata sinistra* della f in x_0 il limite

$$f'(x_0^-) = \lim_{h \to 0^+} \frac{f(x_0 - h) - f(x_0^-)}{-h},$$

nell'ipotesi che esso esista e sia finito.

Esempio 3.9 Sia

$$f(x) = \begin{cases} 1+x, & -\pi \le x < 0, \\ x^2, & 0 \le x < \pi. \end{cases}$$

In questo caso esistono derivata destra e sinistra in tutti i punti interni. Per $x \in (-\pi, \pi) \setminus \{0\}$ è evidente, mentre in $x = 0$, essendo $f(0^+) = 0$ e $f(0^-) = 1$, risulta

$$\lim_{h\to 0^+} \frac{h^2 - 0}{h} = 0, \quad \lim_{h\to 0^+} \frac{1 - h - 1}{-h} = 1.$$

Da notare che in $-\pi$ esiste la derivata destra e in π quella sinistra con $f'(-\pi^+) = 1$ e $f'(\pi^-) = 2\pi$.

Definizione. La funzione f è regolare a tratti in $[a, b]$ se valgono le seguenti proprietà:

1) esiste un numero finito di punti $x_1, \ldots, x_n$ con $a < x_1 < \ldots < x_n < b$, tale che f sia di classe C^1 negli intervalli (a, x_1), (x_j, x_{j+1}) (per $j = 1, 2, \ldots, n-1$) e (x_n, b);

2) nei punti $x_1, \ldots, x_n$, esistono finite le derivate destra e sinistra;

3) nei punti $x_1, \ldots, x_n$, esistono finiti i limiti destro e sinistro.

Teorema 3.10 (*Convergenza della serie di Fourier*) *Sia f regolare a tratti su $[-L, L]$. Allora:*

1) se la f è continua in $x_0 \in (-L, L)$, la serie di Fourier assume in x_0 il valore $f(x_0)$;

2) se $x_0 \in (-L, L)$ e la f è discontinua in x_0, la serie di Fourier in x_0 converge a

$$\tfrac{1}{2}[f(x_0^+) + f(x_0^-)];$$

3) la serie di Fourier converge a

$$\tfrac{1}{2}[f(-L^+) + f(L^-)]$$

sia in $-L$ che in L.

La 1) consente, ad esempio, di affermare che $\sum_{k=1}^{\infty} (-1)^{k+1} \frac{2}{k} \sin k\frac{\pi}{2} = \frac{\pi}{2}$, dalla quale segue che

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}.$$

Per convincersene è sufficiente osservare che, per l'esempio 3.4,

$$\pi^2 = \frac{\pi^2}{3} + \sum_{k=1}^{\infty}(-1)^k \frac{4}{k^2}\cos k\pi = \frac{\pi^2}{3} + \sum_{k=1}^{\infty}\frac{4}{k^2}.$$

Teorema 3.11 (*Lemma di Riemann-Lebesgue*) *Sia f sommabile in $[-L, L]$ (nel senso che esiste finito l'integrale $\int_{-L}^{L} |f(x)|\,dx$) e ivi sviluppabile in serie di Fourier, allora*

$$\lim_{n\to\infty}\int_{-L}^{L} f(x)\cos\frac{n\pi x}{L}\,dx = 0, \quad \lim_{n\to\infty}\int_{-L}^{L} f(x)\sin\frac{n\pi x}{L}\,dx = 0,$$

limiti che stabiliscono la convergenza a zero dei coefficienti della serie di Fourier.

Questioni rilevanti sulle serie di Fourier riguardano la sua integrazione e derivazione termine a termine. Come si evince dai due teoremi che seguono, la differenziabilità termine a termine è molto più problematica rispetto alla integrabilità.

Teorema 3.12 (*Integrazione termine a termine*) *Se la f è regolare a tratti in $[-L, L]$, essa è integrabile termine a termine, ossia, per ogni $x \in [-L, L]$,*

$$\int_{-L}^{x} f(t)\,dt = a_0(x+L) + \frac{L}{\pi}\sum_{n=1}^{\infty}\frac{1}{n}\left\{-a_n \sin\frac{n\pi x}{L} + b_n\left[\cos\frac{n\pi x}{L} - \cos n\pi\right]\right\}.$$

Esempio 3.13 Dall'esempio 3.3, la serie di Fourier di $f(x) = 2x$ in $[-\pi, \pi]$ è

$$\mathcal{F}(x) = \sum_{k=1}^{\infty}\frac{4}{k}\sin kx.$$

Poiché f è continua in $[-\pi, \pi]$, in virtù del Teorema 3.12, si può scrivere

$$\begin{aligned}\int_{-\pi}^{x} 2t\,dt = x^2 - \pi^2 &= \sum_{k=1}^{\infty}\frac{4}{k}(-1)^{k+1}\int_{-\pi}^{x}\sin kt\,dt \\ &= \sum_{k=1}^{\infty}\frac{4}{k}(-1)^{k+1}\left[-\frac{1}{k}(\cos kx - \cos k\pi)\right] \\ &= \sum_{k=1}^{\infty}\frac{4}{k^2}(-1)^k\left[\cos kx - (-1)^k\right],\end{aligned}$$

da cui, ricordando che $\sum_{k=1}^{\infty}\frac{1}{k^2} = \frac{\pi^2}{6}$, si ottiene esattamente, come nell'esempio 3.4, che

$$x^2 = \frac{\pi^2}{3} + \sum_{k=1}^{\infty}\frac{4}{k}(-1)^k\cos kx.$$

La differenziazione di una serie di Fourier si presenta in modo molto diverso. Infatti, differenziando termine a termine la serie di Fourier di x in $[-\pi, \pi]$ si ottiene

$$\sum_{k=1}^{\infty} \frac{d}{dx}\left(\frac{2}{k}(-1)^{k+1}\sin kx\right) = \sum_{k=1}^{\infty} 2(-1)^{k+1}\cos kx,$$

la quale non converge per alcun valore di x in $(-\pi, \pi)$ e tanto meno risulta $f'(x) = 1$.

Teorema 3.14 (*Differenziazione termine a termine*) *Se la f è continua in $[-L, L]$ con $f(-L) = f(L)$ e la f' è regolare a tratti in $[-L, L]$, allora la serie di Fourier è derivabile termine a termine. Ossia, in ogni punto $x \in (-L, L)$ in cui la $f'(x)$ è continua,*

$$\mathcal{F}'(x) = \frac{\pi}{L}\sum_{n=1}^{\infty}\left[-n\,a_n \sin\frac{n\pi x}{L} + n\,b_n\cos\frac{n\pi x}{L}\right].$$

La serie di Fourier di $f(x) = x$ non è dunque derivabile termine a termine in quanto $f(-\pi) \neq f(\pi)$.

Esercizi.

1. Risolvere l'equazione differenziale

$$y'' + 8y = f(t), \qquad f(t) = \begin{cases} \dfrac{\pi + 2t}{\pi}, & -\pi \leq t < 0, \\ \dfrac{\pi - 2t}{\pi}, & 0 \leq t < \pi, \\ f(t + 2\pi), & t \in \mathbb{R}. \end{cases}$$

Sviluppando la $f(t)$ in serie di Fourier si ha che:

$$\mathcal{F}(t) = \frac{4}{\pi^2}\sum_{k=1}^{\infty}\frac{1-(-1)^k}{k^2}\cos kt.$$

Posto $y(t) = a_0 + \sum_{k=1}^{\infty} a_k \cos kt$, derivando e sostituendo nell'equazione differenziale, si ottiene

$$8\,a_0 + \sum_{k=1}^{\infty}(-k^2+8)a_k\cos kt = \frac{4}{\pi^2}\sum_{k=1}^{\infty}\frac{1-(-1)^k}{k^2}\cos kt$$

da cui

$$a_0 = 0, \quad (-k^2+8)a_k = \frac{4}{\pi^2}\sum_{k=1}^{\infty}\frac{1-(-1)^k}{k^2}\cos kt, \quad k = 1.2, \ldots$$

e infine

$$y(t) = \frac{8}{\pi^2} \sum_{k=1}^{\infty} \frac{1}{(2\,k-1)^2[8-(2k-1)^2]} \cos(2k-1)t.$$

2. Calcolare la serie di Fourier di $f(x) = |x|$, $-\pi \leq x \leq \pi$.

   Soluzione:

$$\mathcal{F}(x) = \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{1}{(2\,n-1)^2} \cos(2\,n-1)x.$$

3. Calcolare la serie di Fourier di $f(x) = \begin{cases} -1, & -\pi \leq x < 0, \\ 1, & 0 \leq x < \pi. \end{cases}$

   Soluzione:

$$\mathcal{F}(x) = \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{1}{2\,n-1} \sin(2\,n-1)x.$$

Decadimento dei coefficienti. La rapidità con cui i coefficienti di Fourier $\{a_k; b_k\}$ convergono a zero per $k \to \infty$, dipende dalla regolarità della funzione $f(x)$ cui la serie si riferisce. In generale, si può affermare che esso è tanto più rapido quanto più la f è regolare. Più precisamente se la $f(x)$ presenta dei punti di discontinuità $a_k = O\left(\frac{1}{k}\right)$ e $b_k = O\left(\frac{1}{k}\right)$, ossia i coefficienti decadono a zero come $\frac{1}{k}$ per $k \to \infty$. Se la funzione è continua ed esiste la derivata prima, come funzione continua a tratti, $a_k = O\left(\frac{1}{k^2}\right)$ e $b_k = O\left(\frac{1}{k^2}\right)$, ossia i coefficienti decadono a zero, per $k \to \infty$, come $\frac{1}{k^2}$. Se, infine, la f è derivabile r volte ($r \geq 1$) ed esiste la derivata $(r+1)$-esima, come funzione continua a tratti, $a_k = O\left(\frac{1}{k^{r+1}}\right)$ e $b_k = O\left(\frac{1}{k^{r+1}}\right)$, ossia i coefficienti decadono a zero, per $k \to \infty$, come $\frac{1}{k^{r+2}}$.

3.3 Serie di Fourier in due variabili

La sviluppabilità in serie di Fourier non è limitata alle funzioni in una variabile, ma riguarda anche le funzioni in più variabili. Supponiamo, ad esempio, di avere una funzione

$$f(x,y) \text{ con } -L \leq x \leq L \text{ e } -M \leq y \leq M,$$

integrabile nel rettangolo $[-L, L] \times [-M, M]$. Supponiamo inoltre che, prefissato y, la funzione $g_{[y]}(x) = f(x,y)$ sia dispari nell'intervallo $[-L, L]$. La sua

espansione in serie di Fourier è allora

$$g_{[y]}(x) = \sum_{n=1}^{\infty} b_n(y) \sin \frac{n\pi x}{L},$$

dove

$$b_n(y) = \frac{2}{L} \int_0^L g_{[y]}(x) \sin\left(\frac{n\pi x}{L}\right) dx = \frac{2}{L} \int_0^L f(x,y) \sin\left(\frac{n\pi x}{L}\right) dx,$$

per $n = 1, 2, \ldots$.

Supponiamo ora che la $b_n(y)$ sia, a sua volta, sviluppabile in una serie di funzioni seno (ipotesi verificata nel caso sia dispari in $[-M, M]$). In tal caso si può scrivere

$$b_n(y) = \sum_{m=1}^{\infty} b_{n,m} \sin\left(\frac{m\pi y}{M}\right),$$

essendo

$$b_{n,m} = \frac{2}{M} \int_0^M b_n(y) \sin\left(\frac{m\pi y}{M}\right) dy.$$

In definitiva, la $f(x,y)$ può essere sviluppata come una doppia serie di Fourier, nel modo seguente:

$$\begin{aligned} f(x,y) &= \sum_{n=1}^{\infty} \left[\sum_{m=1}^{\infty} b_{n,m} \sin \frac{m\pi y}{M} \right] \sin \frac{n\pi x}{L} \\ &= \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} b_{n,m} \sin \frac{m\pi y}{M} \sin \frac{n\pi x}{L}, \end{aligned}$$

essendo

$$\begin{aligned} b_{n,m} &= \frac{2}{M} \int_0^M \left(\frac{2}{L} \int_0^L f(x,y) \sin \frac{n\pi x}{L}\, dx \right) \sin \frac{m\pi y}{M} dy \\ &= \frac{4}{LM} \int_0^M \int_0^L f(x,y) \sin \frac{n\pi x}{L} \sin \frac{m\pi y}{M}\, dx dy. \end{aligned}$$

Esempio 3.15 Sia $f(x,y) = xy$, $-\frac{\pi}{2} \le x \le \frac{\pi}{2}$, $-\pi \le y \le \pi$.

Essendo la f dispari rispetto ad ambedue le variabili, sulla base delle precedenti considerazioni, possiamo scrivere

$$f(x,y) = \sum_{i=1}^{n} \sum_{j=1}^{m} b_{nm} \sin my \sin 2nx,$$

dove

$$b_{nm} = \frac{8}{\pi^2}\int_0^{\pi}\int_0^{\pi/2} xy\,\sin my \sin 2nx\,dxdy$$
$$= \frac{8}{\pi^2}\int_0^{\pi} y\,\sin my\,dy \int_0^{\pi/2} x\,\sin 2nx\,dx = \frac{2}{nm}(-1)^{n+m}.$$

Nel caso generale, la rappresentazione in serie di Fourier di una $f(x,y)$, sull'intervallo $-L \le x \le L$ e $-M \le y \le M$, è la seguente

$$\begin{aligned} f(x,y) = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} \Big(& a_{nm}\cos\frac{n\pi x}{L}\cos\frac{m\pi y}{M} + b_{nm}\cos\frac{n\pi x}{L}\sin\frac{m\pi y}{M} \\ & + c_{nm}\sin\frac{n\pi x}{L}\cos\frac{m\pi y}{M} + d_{nm}\sin\frac{n\pi x}{L}\sin\frac{m\pi y}{M}\Big) \\ = a_{00} + \sum_{n=1}^{\infty}\Big(& a_{n0}\cos\frac{n\pi x}{L} + c_{n0}\sin\frac{n\pi x}{L}\Big) \\ + \sum_{m=1}^{\infty}\Big(& a_{0m}\cos\frac{m\pi y}{M} + b_{0m}\sin\frac{m\pi y}{M}\Big) \\ + \sum_{n=1}^{\infty}\sum_{m=1}^{\infty}\Big(& a_{nm}\cos\frac{n\pi x}{L}\cos\frac{m\pi y}{M} + b_{nm}\cos\frac{n\pi x}{L}\sin\frac{m\pi y}{M} \\ & + c_{nm}\sin\frac{n\pi x}{L}\cos\frac{m\pi y}{M} + d_{nm}\sin\frac{n\pi x}{L}\sin\frac{m\pi y}{M}\Big), \end{aligned}$$

con i coefficienti così determinati:

$$a_{00} = \frac{1}{4LM}\int_{-L}^{L}\int_{-M}^{M} f(x,y)\,dydx,$$
$$a_{0m} = \frac{1}{2LM}\int_{-M}^{M}\int_{-L}^{L} f(x,y)\cos\frac{m\pi x}{L}\,dxdy, \quad m = 1,2,\ldots;$$
$$a_{n0} = \frac{1}{2LM}\int_{-L}^{L}\int_{-M}^{M} f(x,y)\cos\frac{n\pi y}{M}\,dydx, \quad n = 1,2,\ldots;$$
$$a_{nm} = \frac{1}{LM}\int_{-M}^{M}\cos\frac{m\pi y}{M}\int_{-L}^{L} f(x,y)\cos\frac{n\pi x}{L}\,dxdy, \quad m,n = 1,2,\ldots;$$
$$b_{nm} = \frac{1}{LM}\int_{-M}^{M}\sin\frac{m\pi y}{M}\int_{-L}^{L} f(x,y)\cos\frac{n\pi x}{L}\,dxdy, \quad m,n = 1,2,\ldots;$$
$$c_{nm} = \frac{1}{LM}\int_{-M}^{M}\cos\frac{m\pi y}{M}\int_{-L}^{L} f(x,y)\sin\frac{n\pi x}{L}\,dxdy, \quad m,n = 1,2,\ldots;$$
$$d_{nm} = \frac{1}{LM}\int_{-M}^{M}\sin\frac{m\pi y}{M}\int_{-L}^{L} f(x,y)\sin\frac{n\pi x}{L}\,dxdy, \quad m,n = 1,2,\ldots.$$

Da notare che b_{0n}, c_{m0}, d_{0n} e d_{m0} non compaiono in quanto coefficienti di funzioni che si annullano rispettivamente per $m = 0$, $n = 0$, $m = 0$ e $n = 0$ come è immediato notare, guardando lo sviluppo della $f(x, y)$. Naturalmente, come abbiamo già visto nel caso di una funzione $f(x, y)$ dispari rispetto ad entrambe le variabili, lo sviluppo si semplifica notevolmente nel caso di importanti simmetrie. Nel caso la $f(x, y)$ sia pari rispetto ad ambedue le variabili, essa assume la forma seguente:

$$\begin{aligned} f(x,y) =& a_{00} + \sum_{n=1}^{\infty} a_{0n} \cos\frac{n\pi x}{L} + \sum_{m=1}^{\infty} a_{m0} \cos\frac{m\pi y}{M} \\ &+ \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} a_{mn} \cos\frac{n\pi x}{L} \cos\frac{m\pi y}{M}, \end{aligned}$$

dove a_{00}, a_{0n} e a_{m0} sono determinati come già specificato nel caso generale. Qualora la funzione sia pari rispetto alla x e dispari rispetto alla y, si ha il seguente sviluppo:

$$f(x,y) = \sum_{m=1}^{\infty} b_{0m} \sin\frac{m\pi y}{M} + \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} b_{mn} \cos\frac{n\pi x}{L} \sin\frac{m\pi y}{M}.$$

Se, infine la f è dispari rispetto alla x e pari rispetto alla y, si ha il seguente sviluppo:

$$f(x,y) = \sum_{n=1}^{\infty} c_{0m} \sin\frac{n\pi x}{L} + \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} c_{mn} \sin\frac{n\pi x}{L} \cos\frac{m\pi y}{M}.$$

Esercizio. Dimostrare che lo sviluppo di $f(x, y) = x^3 y$, per $-\pi \le x \le \pi$ e $-\pi \le y \le \pi$ è la serie

$$f(x,y) = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} (-1)^{n+m} \frac{4}{m^3 n}(m^2\pi^2 - 6) \sin nx \sin mx.$$

3.4 Problema di Sturm-Liouville: Forma canonica

Si tratta di un problema spettrale che, nella sua forma canonica, viene rappresentato nel modo seguente:

$$\begin{cases} -\dfrac{d}{dx}\left[p(x)\dfrac{dy}{dx}\right] + q(x)y = \lambda r(x) y, \quad a \le x \le b, \\ \alpha_1\, y(a) + \alpha_2\, y'(a) = 0, \quad \beta_1\, y(b) + \beta_2\, y'(b) = 0, \end{cases} \tag{3.3}$$

dove α_1, α_2, β_1, β_2 sono costanti assegnate tali che $\alpha_1^2 + \alpha_2^2 > 0$ e $\beta_1^2 + \beta_2^2 > 0$; $p(x) > 0$, $q(x)$ e $r(x) \geq 0$ sono funzioni note e λ è un parametro indipendente da x.

Ogni equazione del tipo

$$y'' + a(x)y' + [b(x) + \lambda c(x)]y = 0, \quad a \leq x \leq b,$$

può essere scritta nella forma (3.3), a condizione che $c(x) \geq 0$. A tale scopo, indicata con $\alpha(x)$ una primitiva di $a(x)$ $\left(\alpha(x) = \int a(x)\,dx\right)$, basta moltiplicare l'equazione per $e^{\alpha(x)}$. Così operando si ottiene infatti

$$\begin{aligned} e^{\alpha(x)}y'' + e^{\alpha(x)}a(x)y' + \left[e^{\alpha(x)}b(x) + \lambda e^{\alpha(x)}c(x)\right] y = \\ = \frac{d}{dx}\left[e^{\alpha(x)}y'\right] + [-q(x) + \lambda r(x)]y = 0, \end{aligned}$$

con $p(x) = e^{\alpha(x)} > 0$, $q(x) = -e^{\alpha(x)}b(x)$ e $r(x) = e^{\alpha(x)}c(x) \geq 0$, essendo $c(x) \geq 0$ per ipotesi.

Per il suo ruolo particolare la funzione $r(x)$ è definita funzione peso. I valori $\{\lambda_k\}_{k=1}^{\infty}$ del parametro spettrale λ ai quali corrispondono soluzioni $\{y_k\}_{k=1}^{\infty}$ non identicamente nulle, sono detti autovalori e le corrispondenti y_k sono definite autofunzioni. L'insieme $\{\lambda_k, y_k\}_{k=1}^{\infty}$ rappresenta lo spettro del problema.

Lo spettro è caratterizzato dalle seguenti due proprietà:

- gli autovalori sono reali e formano una infinità numerabile non limitata, con $\lim_{k\to\infty} |\lambda_k| = +\infty$;

- le autofunzioni sono mutuamente ortogonali, rispetto alla funzione peso $r(x)$, nell'intervallo $[a, b]$, ossia $\int_a^b r(x)y_n(x)y_m(x)\,dx = 0$ per ogni coppia (n, m) con $n \neq m$.

Da notare che ambedue le proprietà valgono anche nell'ipotesi che $r(x) \leq 0$, nel qual caso come peso si assume la funzione $|r(x)|$.

Alcune dimostrazioni.

1) Gli autovalori sono reali.

   Dimostrazione: Supponendo α_1, α_2, β_1, β_2, p, q ed r reali, siano λ un autovalore e y la corrispondente autofunzione. In tal caso λ e y soddisfano il sistema

$$\begin{cases} -\dfrac{d}{dx}\left[p(x)\,\dfrac{dy}{dx}\right] + q(x)y = \lambda\, r(x)y, \quad a \leq x \leq b, \\ \alpha_1\, y(a) + \alpha_2\, y'(a) = 0, \quad \beta_1\, y(b) + \beta_2\, y'(b) = 0, \end{cases} \tag{3.4}$$

da cui, considerando i complessi coniugati, si ha

$$\begin{cases} -\dfrac{d}{dx}\left[p(x)\,\dfrac{d\overline{y}}{dx}\right] + q(x)\overline{y} = \overline{\lambda}\, r(x)\overline{y}, \quad a \le x \le b, \\ \alpha_1\,\overline{y}(a) + \alpha_2\,\overline{y}'(a) = 0, \quad \beta_1\,\overline{y}(b) + \beta_2\,\overline{y}'(b) = 0. \end{cases} \tag{3.5}$$

Moltiplicando l'equazione (3.4) per $\overline{y}$ e la (3.5) per y e sottraendo membro a membro si ottiene

$$\frac{d}{dx}\left[p(x)\,(y\overline{y}' - \overline{y}\,y')\right] = (\lambda - \overline{\lambda})\, r(x)\,|y|^2,$$

da cui, integrando tra a e b,

$$[p(x)(y\overline{y}' - \overline{y}\,y')|_a^b = (\lambda - \overline{\lambda}) \int_a^b r(x)|y|^2\,dx = 0,$$

in quanto $y(a)\overline{y}'(a) - \overline{y}(a)y'(a) = 0$ e $y(b)\overline{y}'(b) - \overline{y}(b)y'(b) = 0$, dato che α_1 e α_2 non si annullano simultaneamente[1] e neanche β_1 e β_2. Di conseguenza, $\lambda = \overline{\lambda}$, dato che $\displaystyle\int_a^b r(x)|y|^2\,dx > 0$.

2) Le autofunzioni corrispondenti ad autovalori diversi sono ortogonali in $[a, b]$ rispetto alla funzione peso $r(x)$.

Dimostrazione: Se y_1 e y_2 sono le autofunzioni corrispondenti agli autovalori $\lambda_1 \neq \lambda_2$, sono validi i sistemi

$$\begin{cases} -\dfrac{d}{dx}\left[p(x)\,\dfrac{dy_1}{dx}\right] + q(x)y_1 = \lambda_1\, r(x)y_1, \quad a \le x \le b, \\ \alpha_1\,y_1(a) + \alpha_2\,y_1'(a) = 0, \quad \beta_1\,y_1(b) + \beta_2\,y_1'(b) = 0, \end{cases} \tag{3.6}$$

$$\begin{cases} -\dfrac{d}{dx}\left[p(x)\,\dfrac{dy_2}{dx}\right] + q(x)y_2 = \lambda_2\, r(x)y_2, \quad a \le x \le b \\ \alpha_1\,y_2(a) + \alpha_2\,y_2'(a) = 0, \quad \beta_1\,y_2(b) + \beta_2\,y_2'(b) = 0, \end{cases} \tag{3.7}$$

Moltiplicando l'equazione (3.6) per y_2 e la (3.7) per y_1 e sottraendo membro a membro si ottiene

$$\frac{d}{dx}\left[p(x)(y_1y_2' - y_2y_1')\right] = (\lambda_1 - \lambda_2)r(x)y_1y_2.$$

[1] La relazione $y(a)\overline{y}'(a) - \overline{y}(a)y'(a) = 0$ è vera in quanto il primo membro è il determinante del sistema omogeneo

$$\begin{cases} y(a)\alpha_1 - y'(a)\alpha_2 = 0, \\ \overline{y}(a)\alpha_1 + \overline{y}'(a)\alpha_2 = 0, \end{cases}$$

nel quale α_1 e α_2 non sono entrambi nulli. Per lo stesso tipo di considerazioni è valida anche la seconda relazione in b.

Integrando tra a e b e tenendo conto delle condizioni agli estremi si ha

$$(\lambda_1 - \lambda_2) \int_a^b r(x) y_1 y_2 \, dx = [p(x)(y_1 y_2' - y_2 \, y_1')]_a^b = 0.$$

Conseguentemente, essendo, $\lambda_1 \neq \lambda_2$, y_1 e y_2 sono ortogonali rispetto a $r(x)$ in $[a, b]$.

I sistemi di Sturm-Liouville sono spesso generati dalla risoluzione dei problemi alle derivate parziali mediante la tecnica di separazione delle variabili.

Esercizio 3.16 Sviluppare la funzione $f(x) = e^{-x\sqrt{2}}$ mediante le autofunzioni del seguente problema di Sturm-Liouville:

$$\begin{cases} \varphi'' + 2\varphi' + (2 + \lambda)\, \varphi = 0, \quad \text{con } 0 \leq x \leq \frac{\pi}{2}, \\ \varphi'(0) + 2\varphi(0) = 0, \\ \varphi(\frac{\pi}{2}) = 0. \end{cases}$$

Si cercano soluzioni del tipo $\varphi(x) \simeq e^{\alpha\, x}$ che conducono all'equazione caratteristica

$$\alpha^2 + 2\alpha + (2 + \lambda) = 0.$$

Questa equazione ha come soluzioni $\alpha_{1,2} = -1 \pm \sqrt{-(\lambda + 1)}$. Occorre distinguere tre casi a seconda del segno del discriminante.

• $\lambda < -1$. In tal caso abbiamo due radici reali e distinte $\alpha_{1,2} = -1 \pm \sqrt{-(\lambda + 1)}$, per cui la soluzione generale è

$$\varphi(x) = e^{-x} \left(c_1 \, e^{-x\sqrt{-(\lambda+1)}} + c_2 \, e^{x\sqrt{-(\lambda+1)}} \right).$$

Poiché

$$\begin{aligned} \varphi(0) &= c_1 + c_2, \\ \varphi'(0) &= -\left(c_1 + c_2\right) + \left(-c_1 \sqrt{-(\lambda + 1)} + c_2 \sqrt{-(\lambda + 1)} \right), \\ \varphi(\tfrac{\pi}{2}) &= e^{-\frac{\pi}{2}} \left[c_1 e^{-\frac{\pi}{2}\sqrt{-(\lambda+1)}} + c_2 e^{\frac{\pi}{2}\sqrt{-(\lambda+1)}} \right], \end{aligned}$$

le costanti c_1 e c_2 devono soddisfare il sistema

$$\begin{cases} \left(1 - \sqrt{-(\lambda + 1)}\right) c_1 + \left(1 + \sqrt{-(\lambda + 1}\right) c_2 = 0, \\ e^{-\frac{\pi}{2}\sqrt{-(\lambda+1)}} c_1 + e^{\frac{\pi}{2}\sqrt{-(\lambda+1)}} c_2 = 0. \end{cases}$$

Affinché c_1 e c_2 non siano simultaneamente nulli il determinante $\Delta(\lambda)$ di questo sistema, ossia

$$\Delta(\lambda) = \left(1 - \sqrt{-(\lambda + 1)}\right) \, e^{\frac{\pi}{2}\sqrt{-(\lambda+1)}} - \left(1 + \sqrt{-(\lambda + 1)}\right) \, e^{-\frac{\pi}{2}\sqrt{-(\lambda+1)}}$$

deve risultare nullo. Posto $b = \sqrt{-(\lambda+1)} > 0$, si può affermare che il determinante si annulla quando

$$(1-b)\,e^{b\frac{\pi}{2}} - (1+b)\,e^{-b\frac{\pi}{2}} = 0.$$

Riscritta l'equazione come $\frac{1-b}{1+b} = e^{-\pi\,b}$, le sue soluzioni sono date dall'intersezione in $(0,1)$ fra la funzione $y(b) = \frac{1-b}{1+b}$ e la funzione $y(b) = e^{-\pi\,b}$ per $0 \leq b \leq 1$. Dalla rappresentazione grafica delle due funzioni [Fig. 3.1], si vede che esiste un singolo valore di $b \in (0,1)$, $\tilde{b}$, dove i due grafici si intersecano. Ad esso corrispondono l'autovalore $\tilde{\lambda} = -1 - \tilde{b}^2$ appartenente all'intervallo $(-2,-1)$ e l'autofunzione

$$\tilde{\varphi}(x) = e^{-x}\left(e^{-x\sqrt{-(\tilde{\lambda}+1)}} - \frac{1-\sqrt{-(\tilde{\lambda}+1)}}{1+\sqrt{-(\tilde{\lambda}+1)}}e^{x\sqrt{-(\tilde{\lambda}+1)}}\right).$$

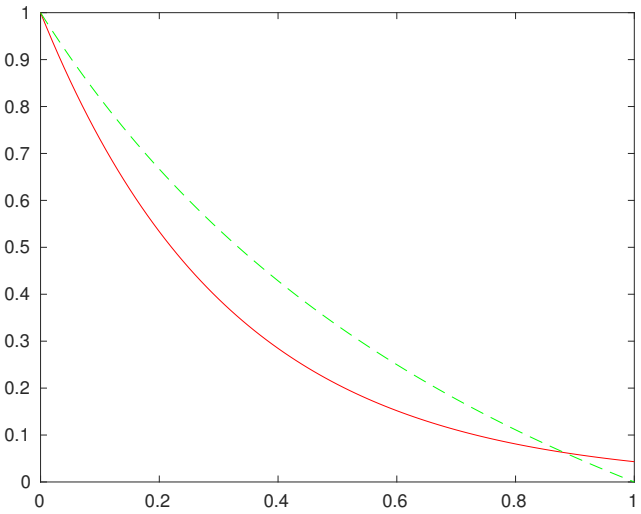


Figura 3.1: La figura rappresenta i grafici delle funzioni $y = \frac{1-x}{1+x}$ e $y = e^{-\pi x}$.

• $\lambda = -1$. In tal caso abbiamo due radici reali e coincidenti $\alpha_1 = \alpha_2 = -1$. La soluzione generale è $\varphi(x) = e^{-x}\,(c_1 + c_2\,x)$, con c_1 e c_2 soluzioni del sistema omogeneo

$$\begin{cases} c_1 + c_2 = 0, \\ c_1 + \dfrac{\pi}{2}\,c_2 = 0. \end{cases}$$

Essendo il suo determinante diverso da zero, l'unica soluzione è quella banale $(c_1 = c_2 = 0)$.

• $\lambda > -1$. In questo caso abbiamo due radici complesse coniugate $\alpha_{1,2} = -1 \pm i\sqrt{\lambda+1}$, per cui la soluzione generale è

$$\varphi(x) = e^{-x}\left(c_1 \cos(x\sqrt{\lambda+1}\,) + c_2\, \sin(x\sqrt{\lambda+1}\,)\right),$$

le cui costanti c_1 e c_2 sono da determinare mediante il sistema di equazioni lineari e omogenee

$$\begin{cases} c_1 + c_2\sqrt{\lambda+1} = 0, \\ c_1\, \cos(\frac{\pi}{2}\sqrt{\lambda+1}\,) + c_2\, \sin(\frac{\pi}{2}\sqrt{\lambda+1}\,) = 0. \end{cases}$$

Dalla prima equazione del sistema si ricava $c_1 = -c_2\sqrt{\lambda+1}$, che sostituita nella seconda equazione fornisce

$$c_2\left[-\sqrt{\lambda+1}\, \cos(\tfrac{\pi}{2}\sqrt{\lambda+1}\,) + \sin(\tfrac{\pi}{2}\sqrt{\lambda+1}\,)\right] = 0.$$

Non consideriamo $c_2 = 0$, in quanto questa scelta comporta $c_1 = 0$. L'equazione è quindi non banalmente soddisfatta quando

$$\mathrm{tg}\,(\tfrac{\pi}{2}\sqrt{\lambda+1}\,) = \sqrt{\lambda+1}\,.$$

Per risolvere tale equazione poniamo $\frac{\pi}{2}\sqrt{\lambda+1} = z > 0$, da cui segue che

$$\mathrm{tg}z = \frac{2}{\pi}\, z.$$

Come si vede nella Fig. 3.2, i due grafici si intersecano in un insieme nu-

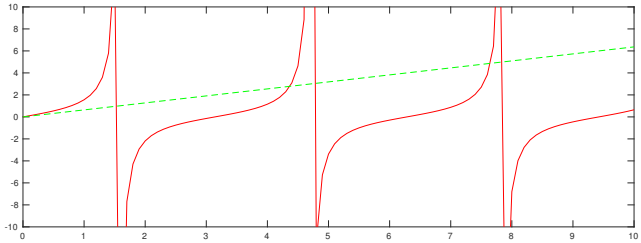


Figura 3.2: La figura mostra i grafici delle funzioni $y = \frac{2}{\pi}\, z$ e $y = \mathrm{tg}z$.

merabile di punti $\{z_k : k = 1, 2, \ldots\}$, dove $k\pi < z_k < \frac{2k+1}{2}\,\frac{\pi}{2}$. Gli autovalori corrispondenti sono i $\lambda_k = -1 + \left(\frac{2z_k}{\pi}\right)^2$, cui sono associate le autofunzioni

$$\varphi_k(x) = e^{-x}\left(-\frac{2z_k}{\pi}\cos\frac{2z_k x}{\pi} + \sin\frac{2z_k x}{\pi}\right).$$

La soluzione generale è pertanto la serie

$$\varphi_(x) = \tilde{a}\tilde{\varphi}(x) + \sum_{k=1}^{\infty} a_k \varphi_k(x),$$

essendo $\tilde{\varphi}(x)$ l'autofunzione relativa all'autovalore $-2 < \tilde{\lambda} < -1$. Per determinare i coefficienti si deve imporre la condizione

$$e^{-\sqrt{2}x} = \tilde{a}\tilde{\varphi}(x) + \sum_{k=1}^{\infty} a_k \varphi_k(x).$$

Per fare questo osserviamo che l'equazione $\varphi'' + 2\varphi' + (2+\lambda)\,\varphi = 0$ può essere riscritta nella forma canonica

$$\frac{d}{dx}(e^{2x}\varphi') + e^{2x}(2+\lambda)\varphi = 0,$$

per cui la funzione peso è $r(x) = e^{2x}$. Essendo le autofunzioni ortogonali in $[0, \frac{\pi}{2}]$ rispetto al peso $r(x)$, i coefficienti della serie sono ottenibili nel modo seguente:

$$\tilde{a} = \frac{\int_0^{\frac{\pi}{2}} e^{2x} e^{-\sqrt{2}x} \tilde{\varphi}(x) dx}{\int_0^{\frac{\pi}{2}} e^{2x} \tilde{\varphi}(x)^2 dx},$$

$$a_k = \frac{\int_0^{\frac{\pi}{2}} e^{2x} e^{-\sqrt{2}x} \varphi_k(x)\, dx}{\int_0^{\frac{\pi}{2}} e^{2x} \varphi_k(x)^2\, dx}, \text{ con } k = 1, 2, \ldots.$$

Esercizio: Risolvere il seguente problema di Sturm-Liouville:

$$\begin{cases} 2y'' + 3y' + (\lambda - 2)y = 0, & \text{con } 0 < x < 5, \\ 2y(0) + y'(0) = 0, \\ y(5) = 5. \end{cases}$$

3.5 Problema di Sturm-Liouville con condizioni di periodicità

Consideriamo preliminarmente il classico problema spettrale

$$\begin{cases} -y'' = \lambda y, & a \le x \le b, \\ y(a) = y(b), \\ y'(a) = y'(b). \end{cases} \tag{3.8}$$

È facile osservare che non esistono autovalori negativi. Infatti se $\lambda = -\beta^2$, $\beta > 0$, la soluzione risulta del tipo $y(x) = c_1 e^{\beta x} + c_2 e^{-\beta x}$ con (c_1, c_2) soluzione del sistema lineare omogeneo

$$\begin{cases} (e^{\beta a} - e^{\beta b})c_1 + (e^{-\beta a} - e^{-\beta b})c_2 = 0, \\ (e^{\beta a} - e^{\beta b})c_1 - (e^{-\beta a} - e^{-\beta b})c_2 = 0, \end{cases}$$

il cui determinante

$$(e^{\beta a} - e^{\beta b})(e^{-\beta a} - e^{-\beta b}) + (e^{-\beta a} - e^{-\beta b})(e^{\beta a} - e^{\beta b}) \neq 0$$

per $\beta > 0$. Per $\lambda = 0$, si ottiene un'autofunzione $y_0(x) = c_0 =$ costante. Per $\lambda > 0$, la soluzione generale dell'equazione è

$$y(x) = c_1 \cos(x\sqrt{\lambda}) + c_2 \sin(x\sqrt{\lambda}),$$

con c_1 e c_2 costanti arbitrarie.

Imponendo le due condizioni di periodicità, si trova che (c_1, c_2) debbono soddisfare il sistema

$$\begin{cases} (\cos(a\sqrt{\lambda}) - \cos(b\sqrt{\lambda})c_1 + (\sin(a\sqrt{\lambda}) - \sin(b\sqrt{\lambda})c_2 = 0, \\ -(\sin(a\sqrt{\lambda}) - \sin(b\sqrt{\lambda})c_1 + (\cos(a\sqrt{\lambda}) - \cos(b\sqrt{\lambda})c_2 = 0, \end{cases}$$

il quale possiede una soluzione non banale (c_1 e c_2 non contemporaneamente nulli) se e solo se il suo determinante

$$(\cos(a\sqrt{\lambda}) - \cos(b\sqrt{\lambda})^2 + (\sin(a\sqrt{\lambda}) - \sin(b\sqrt{\lambda})^2 = 0$$

ossia se e solo se $\cos(a\sqrt{\lambda}) = \cos(b\sqrt{\lambda})$ e $\sin(a\sqrt{\lambda}) = \sin(b\sqrt{\lambda})$. Quest'ultima condizione, tenendo conto dell'ipotesi di nonnegatività di λ, è soddisfatta se e solo se $(b-a)\sqrt{\lambda_k} = 2k\pi$, $k = 1, 2, \ldots$, ossia

$$\lambda_k = \left(\frac{2k\pi}{b-a}\right)^2, \qquad k = 1, 2, \ldots.$$

Lo spettro del problema (3.8), tenuto conto del fatto che le autofunzioni sono definite a meno di un fattore moltiplicativo e che non esiste alcuna relazione tra c_1 e c_2, è pertanto costituito dalla seguente infinità numerabile di autovalori e autofunzioni:

$$S(\lambda_k, y_k) = \begin{cases} \{\lambda_0 = 0; 1\}, \quad k = 0, \\ \left\{\lambda_k = \left(\dfrac{2k\pi}{b-a}\right)^2; \cos\left(\dfrac{2k\pi}{b-a}\right), \sin\left(\dfrac{2k\pi}{b-a}\right)\right\}, \end{cases}$$

dove $k = 1, 2, \ldots$. Di conseguenza, il problema (3.8) possiede una infinità di autovalori λ_k, con $\lambda_k \to +\infty$ per $k \to \infty$, e un'infinità di autofunzioni mutuamente ortogonali in $[a, b]$, come è immediato verificare.

Consideriamo ora la seguente estensione bidimensionale del problema (3.8):

$$\begin{cases} -(u_{xx} + u_{yy}) = \lambda u, \quad a \le x \le b,\ c \le y \le d, \\ u(a, y) = u(b, y), \quad u_x(a, y) = u_x(b, y), \\ u(x, c) = u(x, d), \quad u_y(x, c) = u_y(x, d). \end{cases} \tag{3.9}$$

Per la sua risoluzione, utilizzando la separazione delle variabili, poniamo

$$U(x, y) = X(x)Y(y).$$

Procedendo come al solito, indicato con α il parametro che caratterizza la separazione nella FDE, otteniamo

$$-X'' = \alpha X \text{ e } -Y'' = (\lambda - \alpha)Y.$$

Tenuto conto della periodicità, si perviene ai due sistemi

$$\begin{cases} -X'' = \alpha X, \\ X(a) = X(b), \\ X'(a) = X'(b), \end{cases} \tag{3.10}$$

$$\begin{cases} -Y'' = (\lambda - \alpha)Y, \\ Y(c) = Y(d), \\ Y'(c) = Y'(d). \end{cases} \tag{3.11}$$

Procedendo come nel caso (3.8), per il sistema (3.10) si ottiene lo spettro

$$S(\alpha_n, X_n) = \left\{ \alpha_n = \left(\frac{2n\pi}{b-a} \right)^2, X_n(x) = \cos\left(\frac{2n\pi x}{b-a} \right), \sin\left(\frac{2n\pi x}{b-a} \right) \right\}_{n=0}^{\infty}.$$

Per quanto riguarda lo spettro del sistema (3.11), nel quale α assume i valori α_n già identificati, si procede come nel caso (3.8). Così procedendo, per ogni $n = 0, 1, \ldots$, si ottiene lo spettro

$$S_n(\lambda_{n,m}, Y_{n,m}) = \left\{ \lambda_{n,m} = \alpha_n + \left(\frac{2m\pi}{d-c} \right)^2,\ Y_{n,m}(y) = \cos\left(\frac{2m\pi y}{d-c} \right), \right.$$
$$\left. \text{oppure } Y_{n,m}(y) = \sin\left(\frac{2m\pi y}{d-c} \right); m = 0, 1, 2, \ldots \right\},$$

dove $\lambda_{n,m} \to +\infty$ per $m \to \infty$ e le $Y_{n,m}$ sono mutuamente ortogonali nell'intervallo $[c, d]$.

Come conseguenza, tenuto conto della separazione delle variabili, lo spettro per il problema (3.9) è il seguente insieme bi-infinito

$$\begin{aligned} S(\lambda_{n,m}; U_{n,m}) = \Big\{ & \lambda_{n,m} = \left(\tfrac{2n\pi}{b-a}\right)^2 + \left(\tfrac{2m\pi}{d-c}\right)^2 ; \\ & U_{n,m}(x,y) = \cos\left(\frac{2n\pi}{b-a}x\right)\cos\left(\frac{2m\pi}{d-c}y\right), \\ & \cos\left(\tfrac{2n\pi}{b-a}x\right)\sin\left(\tfrac{2m\pi}{d-c}y\right), \\ & \sin\left(\tfrac{2n\pi}{b-a}x\right)\cos\left(\tfrac{2m\pi}{d-c}y\right), \\ & \sin\left(\tfrac{2n\pi}{b-a}x\right)\sin\left(\tfrac{2m\pi}{d-c}y\right); n, m = 0, 1, \ldots \Big\}, \end{aligned}$$

dove le autofunzioni sono mutuamente ortogonali nel rettangolo $[a, b] \times [c, d]$.

Suggerimenti. Per ulteriori approfondimenti e applicazioni si vedano i libri [4, 19] oppure [27].

Capitolo 4

RISOLUZIONE ANALITICA DELLE PDEs

I metodi più utilizzati nella risoluzione analitica delle PDEs sono due: il metodo degli integrali generali e il metodo della separazione delle variabili.

4.1 Metodo degli integrali generali

Esso rappresenta la naturale estensione alle PDEs del metodo di D'Alembert, già utilizzato nella risoluzione delle ODEs. Tenuto conto delle finalità del libro, procediamo alla sua illustrazione mediante la sua applicazione a vari tipi di PDEs.

Iniziamo con il seguente problema di Cauchy per una PDE del primo ordine

$$\begin{cases} \dfrac{\partial u}{\partial x} + 3\,\dfrac{\partial u}{\partial y} = 0, \\ u(0,y) = 4\sin y. \end{cases}$$

Per la costruzione dell'integrale generale, osserviamo preliminarmente che una funzione del tipo

$$u(x,y) = e^{ax+by}$$

è soluzione dell'equazione differenziale soltanto se $a = -3b$. Di conseguenza l'integrale generale è della forma

$$u(x,y) = F(y-3x), \quad \text{con } F \text{ arbitraria e differenziabile rispetto a } x \text{ e } y.$$

La condizione al contorno richiede che

$$u(0,y) = F(y) = 4\sin y$$

il che implica che la soluzione del problema differenziale è

$$u(x,y) = F(y-3x) = 4\sin(y-3x).$$

Determiniamo ora la soluzione generale dell'equazione differenziale

$$3\,u_{tt} = 10\,u_{xx}. \tag{4.1}$$

Estendendo il ragionamento adottato per le ODEs cerchiamo una soluzione della forma $u(x,t) = e^{ax+bt}$ che, sostituita nella equazione differenziale, fornisce la seguente relazione tra le costanti reali a e b:

$$3b^2 = 10a^2 \Rightarrow b = \pm a\sqrt{\tfrac{10}{3}}.$$

Questo implica che nell'equazione differenziale data (equazione delle onde) la variabile spaziale e temporale sono correlate, nel senso che compaiono univocamente nella forma $a\left(x \mp t\sqrt{\frac{10}{3}}\right)$. Per questo motivo, ipotizziamo ora che la soluzione generale della (4.1) sia

$$u(x,t) = F\left(x - t\sqrt{\tfrac{10}{3}}\right) + G\left(x + t\sqrt{\tfrac{10}{3}}\right), \tag{4.2}$$

con F e G di classe C^2, peraltro arbitrarie. Per verificare se la (4.2) soddisfa l'Eq. (4.1) osserviamo che

$$\begin{cases} u_t = \sqrt{\frac{10}{3}}\left[-F'\left(x - t\sqrt{\frac{10}{3}}\right) + G'\left(x + t\sqrt{\frac{10}{3}}\right)\right], \\ u_x = F'\left(x - t\sqrt{\frac{10}{3}}\right) + G'\left(x + t\sqrt{\frac{10}{3}}\right), \end{cases}$$

e che

$$\begin{cases} u_{tt} = \frac{10}{3}\left[F''\left(x - t\sqrt{\frac{10}{3}}\right) + G''\left(x + t\sqrt{\frac{10}{3}}\right)\right], \\ u_{xx} = F''\left(x - t\sqrt{\frac{10}{3}}\right) + G''\left(x + t\sqrt{\frac{10}{3}}\right). \end{cases}$$

Da cui segue immediatamente che

$$3\,u_{tt} = 10\,u_{xx},$$

ossia che la (4.2) è la soluzione generale della (4.1).

Consideriamo quindi la PDE iperbolica

$$\begin{cases} \dfrac{\partial^2 u}{\partial x \partial y} = x^2 y, \\ u(x,0) = x^2, \\ u(1,y) = \cos y. \end{cases}$$

Integrando ordinatamente rispetto alla x e alla y si ottiene l'integrale generale

$$u(x,y) = \tfrac{1}{6}x^3y^2 + F(y) + G(x).$$

Per le condizioni al contorno deve essere

$$u(x,0) = F(0) + G(x) = x^2 \quad \Longrightarrow \quad G(x) = x^2 - F(0)$$

e

$$u(1,y) = \tfrac{1}{6}y^2 + F(y) + G(1) = \cos y \quad \Longrightarrow \quad F(y) = \cos y - \tfrac{1}{6}y^2 - G(1),$$

dalle quali si ottiene $u(x,y) = \frac{1}{6}x^3y^2 + \cos y - \frac{1}{6}y^2 - G(1) + x^2 - F(0)$, con $G(1) + F(0) = 1$, essendo $u(x,0) = x^2$.

La soluzione del problema differenziale è dunque

$$u(x,y) = \tfrac{1}{6}x^3y^2 + \cos y - \tfrac{1}{6}y^2 + x^2 - 1.$$

4.1 a Problemi di Cauchy per PDEs iperboliche

1. Equazione delle onde con velocità iniziale nulla

$$\begin{cases} 4\dfrac{\partial^2 u}{\partial t^2} = 25\dfrac{\partial^2 u}{\partial x^2}, \quad 0 \le x \le \pi,\ t \ge 0, \\ u(x,0) = \sin 2x, \\ u_t(x,0) = 0. \end{cases}$$

Ogni funzione

$$u(x,y) = e^{ax+bt}$$

è soluzione dell'equazione data se $b = \pm\frac{5}{2}a$.

L'integrale generale è dunque della forma

$$u(x,t) = F(2x+5t) + G(2x-5t), \qquad F \text{ e } G \text{ differenziabili rispetto a } x \text{ e } t.$$

Dalle condizioni al contorno

$$\begin{cases} u(x,0) = F(2x) + G(2x) = \sin 2x, \\ u_t(x,0) = 5[F'(2x) - G'(2x)] = 0, \end{cases}$$

segue che, derivando rispetto ad x la prima equazione e sommando alla seconda,

$$F'(2x) = \tfrac{1}{2}\cos 2x,$$

ossia

$$F(2x) = \tfrac{1}{2}\sin 2x + c,$$
$$G(2x) = \tfrac{1}{2}\sin 2x - c,$$

dove c è una costante arbitraria.

Infine, si ottiene la soluzione

$$u(x,t) = \tfrac{1}{2}\sin(2x+5t) + \tfrac{1}{2}\sin(2x-5t).$$

2. Equazione delle onde con posizione e velocità iniziali non nulle

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} = c^2 \dfrac{\partial^2 u}{\partial x^2}, \quad \text{con } 0 \le x \le \pi,\ t \ge 0, \\ u(x,0) = \sin(x), \\ u_t(x,0) = \cos(x), \end{cases}$$

dove $c > 0$.

Una funzione del tipo

$$u(x,t) = e^{ax+bt}$$

soddisfa l'equazione se $b = \pm ca$. Per cui $u(x,t) = e^{a(x\pm ct)}$ fornisce due soluzioni particolari, qualunque sia il valore di a. L'integrale generale sarà quindi della forma

$$u(x,t) = F(x+ct) + G(x-ct)$$

con F e G derivabili almeno due volte rispetto a x e a t.

Poiché $u_t(x,t) = c\,F'(x+ct) - c\,G'(x-ct)$, dalle condizioni iniziali ricaviamo il sistema

$$\begin{cases} u(x,0) = F(x) + G(x) = \sin x \\ u_t(x,0) = c\,[F'(x) - G'(x)] = \cos x. \end{cases}$$

Derivando la prima equazione si ha $F'(x) + G'(x) = \cos x$ per cui, tenuto conto della seconda equazione, è possibile ricavare $F'(x)$ e $G'(x)$. Più precisamente si trova $F'(x) = \frac{1}{2}\left(1 + \frac{1}{c}\right)\cos x$ e $G'(x) = \frac{1}{2}\left(1 - \frac{1}{c}\right)\cos x$. Integrando si ha che

$$F(x) = \tfrac{1}{2}\left(1 + \frac{1}{c}\right)\sin x + c_1,$$
$$G(x) = \tfrac{1}{2}\left(1 - \frac{1}{c}\right)\sin x + c_2,$$

dove c_1 e c_2 sono costanti arbitrarie che, per la condizione $F(x)+G(x) = \sin x$, devono soddisfare il vincolo $c_1 + c_2 = 0$. Possiamo quindi scrivere

$$F(x+ct) = \tfrac{1}{2}\left(1+\frac{1}{c}\right)\sin(x+ct) + c_1,$$
$$G(x-ct) = \tfrac{1}{2}\left(1-\frac{1}{c}\right)\sin(x-ct) - c_1,$$

perciò l'integrale generale diventa

$$u(x,t) = \tfrac{1}{2}\left(1+\frac{1}{c}\right)\sin(x+ct) + \tfrac{1}{2}\left(1-\frac{1}{c}\right)\sin(x-ct).$$

È inoltre immediato verificare che:

$$\begin{aligned} u(0,t)+u(\pi,t) = \frac{1}{2}\Bigg\{&\left(1+\frac{1}{c}\right)\sin(ct) - \left(1-\frac{1}{c}\right)\sin(ct) \\ &+\left(1+\frac{1}{c}\right)\sin(\pi+ct) + \left(1-\frac{1}{c}\right)\sin(\pi-ct)\Bigg\} = 0; \\ u_x(0,t)+u_x(\pi,t) = \frac{1}{2}\Bigg\{&\left(1+\frac{1}{c}\right)\cos(ct) + \left(1-\frac{1}{c}\right)\cos(ct) \\ &+\left(1+\frac{1}{c}\right)\cos(\pi+ct) + \left(1-\frac{1}{c}\right)\cos(\pi-ct)\Bigg\} = 0. \end{aligned}$$

Rimane da verificare che la soluzione trovata soddisfi il modello iniziale. Essendo

$$u_{tt}(x,t) = -\frac{c^2}{2}\left[\left(1+\frac{1}{c}\right)\sin(x+ct) + \left(1-\frac{1}{c}\right)\sin(x-ct)\right],$$
$$u_{xx}(x,t) = -\frac{1}{2}\left[\left(1+\frac{1}{c}\right)\sin(x+ct) + \left(1-\frac{1}{c}\right)\sin(x-ct)\right].$$

l'equazione $\frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial^2 u}{\partial x^2}$ è soddisfatta. La verifica delle due condizioni $u(x,0) = \sin(x)$ e $u_t(x,0) = \cos(x)$ è infine immediata.

3. Problema di Cauchy per una PDE iperbolica

$$\begin{cases} u_{xx} + 2u_{xy} - 3u_{yy} = 0, \\ u(x,0) = \sin x, \\ u_y(x,0) = \frac{4}{3}\cos x. \end{cases}$$

Si cercano soluzioni del tipo $u(x,y) \simeq e^{ax+by}$ la quale, per sostituzione, conduce all'equazione $a^2 + 2ab - 3b^2 = 0$. Dividendo tale equazione per $b \neq 0$

(se fosse $b = 0$ allora sarebbe anche $a = 0$ e tale caso è ovviamente da scartare) si ottiene

$$\left(\frac{a}{b}\right)^2 + 2\left(\frac{a}{b}\right) - 3 = 0.$$

Tale equazione ammette come soluzioni $\frac{a}{b} = -3$ e $\frac{a}{b} = 1$, per cui le soluzioni cercate sono del tipo $u(x, y) \simeq e^{-3x+y}$ e $u(x, y) \simeq e^{x+y}$. La soluzione generale è pertanto

$$u(x, y) = F(x + y) + G(-3x + y)$$

con le funzioni F e G derivabili almeno due volte rispetto a x e y. Dal teorema di derivazione delle funzioni composte si ha $u_y(x, y) = F'(x+y)+G'(-3x+y)$, per cui, imponendo le condizioni per $y = 0$, si trova il sistema

$$\begin{cases} F(x) + G(-3x) = \sin x \\ F'(x) + G'(-3x) = \frac{4}{3}\cos x. \end{cases}$$

Derivando la prima equazione si ottiene

$$\begin{cases} F'(x) - 3G'(-3x) = \cos x \\ F'(x) + G'(-3x) = \frac{4}{3}\cos x, \end{cases}$$

da cui, sottraendo dalla prima equazione la seconda, si trova $G'(-3x) = \frac{1}{12}\cos x$, da cui $-\frac{1}{3}G(-3x) = \frac{1}{12}\sin x + c$, ossia $G(t) = -\frac{1}{4}\sin\left(-\frac{t}{3}\right) + c$. Di conseguenza $F(x) = \sin x + \frac{1}{4}\sin x - c = \frac{5}{4}\sin x - c$. La soluzione generale del problema è dunque

$$u(x, y) = \tfrac{5}{4}\sin(x + y) - \tfrac{1}{4}\sin\left(-\frac{y - 3x}{3}\right) = \tfrac{5}{4}\sin(x + y) - \tfrac{1}{4}\sin\left(\frac{3x - y}{3}\right).$$

Tale funzione è la soluzione del problema differenziale proposto, come è facile verificare per sostituzione nell'equazione differenziale e nelle condizioni iniziali.

4.2 Metodo di separazione delle variabili

Il metodo consiste nel ricondurre un assegnato problema alle derivate parziali, *dotato di un'unica soluzione*, alla risoluzione di più problemi alle derivate ordinarie. Naturalmente esso non è sempre applicabile. Il metodo è infatti fondamentalmente utilizzato per risolvere problemi a derivate parziali con coefficienti costanti e con domini regolari. Quando è applicabile è in grado di fornire la soluzione esatta del problema, a differenza di quanto avviene con i metodi numerici. Nei casi più complicati il ricorso ai metodi numerici è inevitabile. Tuttavia, anche in questi casi, il ricorso ai metodi analitici è importante per verificare l'effettività dei codici di calcolo che si intende utilizzare. A tale scopo tipicamente si procede nel modo seguente:

(a) si risolvono analiticamente uno o più problemi, il più possibile vicini a quello da risolvere numericamente;

(b) si risolvono numericamente gli stessi problemi e si confrontano i risultati ottenuti con le soluzioni analitiche;

(c) si procede alla risoluzione numerica del problema di effettivo interesse con il codice disponibile, dopo aver verificato che l'errore commesso nei problemi semplificati corrisponda alle aspettative.

Iniziamo con il seguente problema di Cauchy per una PDE del primo ordine:

$$\begin{cases} \dfrac{\partial u}{\partial x} = 4\,\dfrac{\partial u}{\partial y}, \\ u(0,y) = 8\,e^{-3y}, \quad (x,y) \in \mathbb{R}^2. \end{cases}$$

Dalla teoria delle equazioni a derivate parziali è noto che il problema ammette una e una sola soluzione. La tecnica consiste nel cercare preliminarmente un integrale dell'equazione a variabili separate:

$$u(x,y) = X(x)Y(y), \qquad X \text{ e } Y \text{ differenziabili.}$$

Sostituendo, riordinando e dividendolo per XY si ha

$$\begin{aligned} X'Y &= 4XY', \\ \frac{X'}{4X} &= \frac{Y'}{Y}. \end{aligned} \tag{4.3}$$

La divisibilità per XY viene verificata a posteriori. Il primo membro della (4.3) dipende esclusivamente dalla variabile x, mentre il secondo membro dalla variabile y. Se due funzioni dipendenti rispettivamente da due variabili indipendenti hanno sempre lo stesso valore, sono inevitabilmente uguali ad una costante. Indicando con $\lambda \in \mathbb{R}$ tale costante, dalla (4.3) si ottiene il sistema

$$\begin{cases} X' - 4\lambda X = 0, \\ Y' - \lambda Y = 0, \end{cases} \qquad \text{da cui} \qquad \begin{cases} X = e^{4\lambda x}, \\ Y = e^{\lambda y}, \end{cases}$$

e pertanto

$$u(x,y) = c\,e^{\lambda(4x+y)}.$$

Imponendo la condizione $u(0,y) = 8\,e^{-3y}$ si ha $c = 8$ e $\lambda = -3$, da cui

$$u(x,y) = 8\,e^{-3(4x+y)}.$$

4.2 a Equazioni ellittiche

In questa sezione discutiamo alcuni problemi ellittici con diverse condizioni al bordo.

1. Consideriamo il problema di Dirichlet per l'equazione di Laplace

$$\begin{cases} \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = 0, & \\ u(x,0) = 4\,e^{-3x}. \qquad & u(x,\tfrac{\pi}{2}) = 0, \\ u(0,y) = 4\cos 3y, & u(\tfrac{\pi}{2},y) = 4\,e^{-\frac{3}{2}\pi}\cos 3y, \end{cases} \tag{4.4}$$

dove $(x,y) \in \left[0,\frac{\pi}{2}\right] \times \left[0,\frac{\pi}{2}\right]$.

Si tratta di un problema ellittico su un quadrato con assegnati i valori della soluzione lungo la frontiera del dominio. Il problema ha una sola soluzione che possiamo ottenere col metodo della separazione delle variabili, ponendo

$$u(x,y) = X(x)Y(y), \qquad X \text{ e } Y \text{ differenziabili.}$$

Sostituendo e riordinando si ha

$$\begin{cases} \dfrac{X''}{X} = -\dfrac{Y''}{Y} = \lambda, & \\ X(x)Y(0) = 4\,e^{-3x}. \qquad & X(x)Y(\tfrac{\pi}{2}) = 0, \\ X(0)Y(y) = 4\cos 3y, & X(\tfrac{\pi}{2})Y(y) = 4\,e^{-\frac{3}{2}\pi}\cos 3y. \end{cases} \tag{4.5}$$

Si hanno quindi tre casi a seconda del segno del parametro λ (autovalore).

- $\lambda > 0$.

  $$\begin{cases} X''(x) - \lambda X(x) = 0, \\ Y''(y) + \lambda Y(y) = 0, \end{cases} \implies \begin{cases} X(x) = c_1 e^{x\sqrt{\lambda}} + c_2 e^{-x\sqrt{\lambda}}, \\ Y(y) = d_1 \cos(y\sqrt{\lambda}) + d_2 \sin(y\sqrt{\lambda}). \end{cases}$$

  Le condizione $Y(0)X(x) = d_1 X(x) = 4e^{-3x}$ implica che

  $$\begin{cases} c_1 = 0, \\ d_1 c_2 = 4, \\ \sqrt{\lambda} = 3. \end{cases}$$

  Dalla condizione

  $$X(0)Y(y) = c_2[d_1 \cos(y\sqrt{\lambda}) + d_2 \sin(y\sqrt{\lambda})] = 4\cos(3y),$$

  ricordando che $d_1 c_2 = 4$ e $\sqrt{\lambda} = 3$, segue che $d_2 = 0$ e la soluzione del problema risulta essere

  $$u(x,y) = 4\,e^{-3x}\cos 3y,$$

  dato che essa soddisfa tutte le condizioni (4.4).

Poiché la soluzione è unica risulta del tutto inutile analizzare i casi $\lambda = 0$ e $\lambda < 0$.

2. Consideriamo il problema misto per l'equazione di Laplace

$$\begin{cases} \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = 0, \quad 0 \le x \le 5, \quad 0 \le y \le 3, \\ u_y(x,0) = 0, \quad u(x,3) = 0, \\ u_x(0,y) = 0, \quad u(5,y) = f(y). \end{cases}$$

Procedendo, come nel caso precedente, poniamo $u(x,y) = X(x)Y(y)$. Si ottiene così la condizione

$$\frac{X''}{X} = -\frac{Y''}{Y} = -\lambda, \quad \lambda \quad \text{indipendente da } x \text{ e da y.}$$

Da cui, tenendo conto delle condizioni al bordo,

$$\begin{cases} X'' + \lambda X = 0, \\ X'(0) = 0, \end{cases} \qquad \begin{cases} Y'' - \lambda Y = 0, \\ Y'(0) = 0, \quad Y(3) = 0. \end{cases}$$

Il problema in Y è un classico problema di Sturm-Liouville, il cui spettro è definito da un'infinità numerabile di autovalori e autofunzioni. Tali autovalori sono, ovviamente, i valori da attribuire a λ per il problema in Y. Posto $Y(y) = e^{\alpha y}$, l'equazione caratteristica, associata al problema spettrale in Y, è

$$\alpha^2 - \lambda = 0.$$

Si considerano tre casi, a seconda che λ sia positivo, nullo o negativo.

- $\lambda = \beta^2$, $\beta > 0$. In questo caso $Y(y) = c_1 e^{\beta y} + c_2 e^{-\beta y}$, essendo c_1 e c_2 soluzioni del sistema

$$\begin{cases} Y'(0) = 0, \\ Y(3) = 0, \end{cases} \implies \begin{cases} c_1 - c_2 = 0, \\ e^{3\beta} c_1 + e^{-3\beta} c_2 = 0, \end{cases}$$

  il cui determinante $\Delta = e^{-3\beta} + e^{3\beta} > 0$, per ogni $\beta > 0$. Di conseguenza $c_1 = c_2$ e la soluzione è inaccettabile.

- $\lambda = 0$. La soluzione è $Y(y) = a + by$, anch'essa inaccettabile in quanto comporta $a = b = 0$.

- $\lambda = -\beta^2$, $\beta > 0$. In quest'ultimo caso $Y(y) = a\,\cos\beta y + b\,\sin\beta y$, essendo a e b soluzione del sistema

$$\begin{cases} b = 0, \\ \cos 3\beta = 0, \end{cases}$$

da cui $\beta_k = (2k+1)\frac{\pi}{6}$, con $k = 0, 1, \ldots$.

Gli autovalori sono pertanto i $\lambda_k = -\frac{\pi^2}{36}(2k+1)^2$, $k = 0, 1, 2, \ldots$. Lo spettro relativo è

$$\left\{\lambda_k = -\left[(2k+1)\tfrac{\pi}{6}\right]^2,\ Y_k(y) = \cos(2k+1)\tfrac{\pi}{6}y;\ k = 0, 1, \ldots\right\}.$$

La k-esima soluzione dell'associato problema

$$\begin{cases} X_k'' - \left[(2k+1)\frac{\pi}{6}\right]^2 X_k = 0, \\ X_k'(0) = 0, \end{cases}$$

è $X_k(x) = a_k\, e^{(2k+1)\frac{\pi}{6}x} + b_k\, e^{-(2k+1)\frac{\pi}{6}x}$, con $a_k - b_k = 0$, ossia, a meno di un fattore moltiplicativo irrilevante,

$$X_k(x) = e^{(2k+1)\frac{\pi}{6}x} + e^{-(2k+1)\frac{\pi}{6}x} = 2\cosh(2k+1)\tfrac{\pi}{6}x.$$

Ciascuna delle funzioni $u_k(x, y) = \cosh\left[(2k+1)\frac{\pi}{6}x\right]\cos\left[(2k+1)\frac{\pi}{6}y\right]$, $k = 0, 1, \ldots$, soddisfa l'equazione di Laplace e tutte le condizioni assegnate, ad eccezione della $u(5, y) = f(y)$, con $f(y)$ funzione assegnata in $[0, 3]$. Si assume pertanto come soluzione la serie

$$u(x, y) = \sum_{k=0}^{\infty} c_k \cosh\left[(2k+1)\tfrac{\pi}{6}x\right]\cos\left[(2k+1)\tfrac{\pi}{6}y\right]$$

con i coefficenti c_k determinati in modo che risulti soddisfatta la condizione

$$\sum_{k=0}^{\infty} \hat{c}_k \cos(2k+1)\tfrac{\pi}{6}y = f(y), \quad \hat{c}_k = c_k \cosh(2k+1)\tfrac{5\pi}{6}.$$

Trattandosi di una serie di Fourier con $0 \le y \le 3$,

$$\hat{c}_k = \frac{\int_0^3 f(y)\cos[(2k+1)\frac{\pi y}{6}]\,dy}{\int_0^3 \cos^2[(2k+1)\frac{\pi y}{6}]\,dy} = 2\int_0^3 f(y)\cos[(2k+1)\tfrac{\pi y}{6}]\,dy,$$

$$c_k = \frac{\hat{c}_k}{\cosh[(2k+1)\frac{5}{6}\pi]}.$$

dove $k = 0, 1, \ldots$.

3. Consideriamo il problema di Dirichlet

$$\begin{cases} \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = 0, \\ u(0,y) = 0, \quad u(L,y) = f(y), \\ u(x,0) = u(0,M) = 0, \end{cases} \tag{4.6}$$

per $(x,y) \in [0,L] \times [0,M]$.

Applicando il metodo di separazione delle variabili, poniamo

$$u(x,y) = X(x)Y(y) \Longrightarrow \underbrace{\frac{X''(x)}{X(x)}}_{=\lambda} + \underbrace{\frac{Y''(y)}{Y(y)}}_{=-\lambda} = 0 \Longrightarrow \begin{cases} X''(x) - \lambda X(x) = 0, \\ Y''(y) + \lambda Y(y) = 0. \end{cases}$$

Imponendo le condizioni sulla frontiera è facile vedere che per la $Y(y)$ si ottiene un problema di Sturm-Liouville, essendo:

$$\begin{cases} X'' - \lambda X = 0, \\ X(0) = 0, \end{cases} \qquad \begin{cases} Y'' + \lambda Y = 0, \\ Y(0) = Y(M) = 0. \end{cases}$$

Per trovare gli autovalori e le autofunzioni risolvendo il problema relativo alla Y, otteniamo

$$\begin{cases} \lambda_k = \left(\frac{k\pi}{M}\right)^2, \\ Y_k(y) = \sin\left(\dfrac{k\pi y}{M}\right), \end{cases} \tag{4.7}$$

dove $k \in \mathbb{N}$. Il problema differenziale per la $X_k(x)$ diventa

$$\begin{cases} X_k''(x) - \lambda_k X(x) = 0, \\ X_k(0) = 0, \end{cases} \tag{4.8}$$

che, essendo $\lambda_k > 0$, essa ammette la seguente soluzione

$$X_k(x) = a\, e^{x\sqrt{\lambda_k}} + b\, e^{-x\sqrt{\lambda_k}}, \quad \text{da cui} \quad X_k(0) = 0 \Longrightarrow a + b = 0.$$

Pertanto la soluzione del (4.8), a meno di una costante moltiplicativa, è

$$X_k(x) = e^{\sqrt{\lambda_k}x} - e^{-\sqrt{\lambda_k}x} = 2\sinh(\sqrt{\lambda_k}x). \tag{4.9}$$

Facendo uso delle (4.8) e (4.9) la soluzione del problema è la serie trigonometrica:

$$u(x,y) = \sum_{k=1}^{\infty} a_k \sinh\left(\frac{k\pi x}{M}\right) \sin\left(\frac{k\pi y}{M}\right). \tag{4.10}$$

Dalla condizione $u(L,y) = f(y)$ si ottiene infine

$$\sum_{n=1}^{\infty} a_n \sinh \frac{n\pi L}{M} \sin \frac{n\pi}{M} y = f(y),$$

da cui segue che i coefficienti di Fourier a_n sono dati da

$$a_n = \frac{\frac{2}{M}\int_0^M f(x) \sin \frac{n\pi y}{M}\, dy}{\sinh \frac{\pi L}{M}}.$$

4. Equazione di Laplace sul cerchio unitario (con formula risolutiva di Poisson).

$$\begin{cases} u_{xx} + u_{yy} = 0, & x^2 + y^2 < 1, \\ u(x,y) = f(x,y), & x^2 + y^2 = 1. \end{cases} \tag{4.11}$$

Il problema (4.11) non è risolubile mediante il metodo di separazione delle variabili in quanto il dominio non è rettangolare. Pertanto si procede ad un cambiamento di coordinate al fine di avere (nel nuovo sistema di riferimento) un dominio che ci permetta di separare le variabili nella equazione differenziale trasformata. Vista la simmetria polare del dominio, trasformiamo il problema da coordinate cartesiane a coordinate polari, ponendo

$$\begin{cases} x = \rho\cos\theta, \\ y = \rho\sin\theta, \end{cases} \implies \begin{cases} \rho = \sqrt{x^2+y^2}, \\ \theta = \operatorname{arctg}\dfrac{y}{x}. \end{cases}$$

Applicando le regole di derivazione delle funzioni composte,

$$u_x = \frac{\partial u}{\partial \rho}\frac{\partial \rho}{\partial x} + \frac{\partial u}{\partial \theta}\frac{\partial \theta}{\partial x} = \frac{x}{\rho}\frac{\partial u}{\partial \rho} - \frac{y}{x^2+y^2}\frac{\partial u}{\partial \theta} = \cos\theta\frac{\partial u}{\partial \rho} - \frac{\sin\theta}{\rho}\frac{\partial u}{\partial \theta},$$
$$u_y = \frac{\partial u}{\partial \rho}\frac{\partial \rho}{\partial y} + \frac{\partial u}{\partial \theta}\frac{\partial \theta}{\partial y} = \frac{y}{\rho}\frac{\partial u}{\partial \rho} + \frac{x}{x^2+y^2}\frac{\partial u}{\partial \theta} = \sin\theta\frac{\partial u}{\partial \rho} + \frac{\cos\theta}{\rho}\frac{\partial u}{\partial \theta},$$

$$\begin{aligned} u_{xx} &= \left[\frac{\partial}{\partial \rho}\left(\cos\theta\frac{\partial u}{\partial \rho} - \frac{\sin\theta}{\rho}\frac{\partial u}{\partial \theta}\right)\right]\frac{\partial \rho}{\partial x} + \left[\frac{\partial}{\partial \theta}\left(\cos\theta\frac{\partial u}{\partial \rho} - \frac{\sin\theta}{\rho}\frac{\partial u}{\partial \theta}\right)\right]\frac{\partial \theta}{\partial x} \\ &= \left(\cos\theta\frac{\partial^2 u}{\partial \rho^2} + \frac{\sin\theta}{\rho^2}\frac{\partial u}{\partial \theta} - \frac{\sin\theta}{\rho}\frac{\partial^2 u}{\partial \rho \partial \theta}\right)\cos\theta \\ &+ \left(-\sin\theta\frac{\partial u}{\partial \rho} + \cos\theta\frac{\partial^2 u}{\partial \rho \partial \theta} - \frac{\cos\theta}{\rho}\frac{\partial u}{\partial \theta} - \frac{\sin\theta}{\rho}\frac{\partial^2 u}{\partial \theta^2}\right)\frac{-\sin\theta}{\rho} \\ &= \cos^2\theta\frac{\partial^2 u}{\partial \rho^2} + 2\frac{\cos\theta\sin\theta}{\rho^2}\frac{\partial u}{\partial \theta} - 2\frac{\cos\theta\sin\theta}{\rho}\frac{\partial^2 u}{\partial \rho \partial \theta} + \frac{\sin^2\theta}{\rho}\frac{\partial u}{\partial \rho} + \frac{\sin^2\theta}{\rho^2}\frac{\partial^2 u}{\partial \theta^2}, \end{aligned}$$

$$
\begin{aligned}
u_{yy} &= \left[\frac{\partial}{\partial\rho}\left(\sin\theta\frac{\partial u}{\partial\rho}+\frac{\cos\theta}{\rho}\frac{\partial u}{\partial\theta}\right)\right]\frac{\partial\rho}{\partial y}+\left[\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial u}{\partial\rho}+\frac{\cos\theta}{\rho}\frac{\partial u}{\partial\theta}\right)\right]\frac{\partial\theta}{\partial y}\\
&= \left(\sin\theta\frac{\partial^2 u}{\partial\rho^2}-\frac{\cos\theta}{\rho^2}\frac{\partial u}{\partial\theta}+\frac{\cos\theta}{\rho}\frac{\partial^2 u}{\partial\rho\partial\theta}\right)\sin\theta\\
&+\left(\cos\theta\frac{\partial u}{\partial\rho}+\sin\theta\frac{\partial^2 u}{\partial\rho\partial\theta}-\frac{\sin\theta}{\rho}\frac{\partial u}{\partial\theta}+\frac{\cos\theta}{\rho}\frac{\partial^2 u}{\partial\theta^2}\right)\frac{\cos\theta}{\rho}\\
&= \sin^2\theta\frac{\partial^2 u}{\partial\rho^2}-2\frac{\cos\theta\sin\theta}{\rho^2}\frac{\partial u}{\partial\theta}+2\frac{\cos\theta\sin\theta}{\rho}\frac{\partial^2 u}{\partial\rho\partial\theta}+\frac{\cos^2\theta}{\rho}\frac{\partial u}{\partial\rho}+\frac{\cos^2\theta}{\rho^2}\frac{\partial^2 u}{\partial\theta^2},
\end{aligned}
$$

da cui

$$
u_{xx}+u_{yy}=\frac{\partial^2 u}{\partial\rho^2}+\frac{1}{\rho}\frac{\partial u}{\partial\rho}+\frac{1}{\rho^2}\frac{\partial^2 u}{\partial\theta^2}=\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial u}{\partial\rho}\right)+\frac{1}{\rho^2}\frac{\partial^2 u}{\partial\theta^2}=0.
$$

Pertanto, tenendo conto delle condizioni di periodicità della u in $(1,0)$ e di limitatezza della soluzione nel cerchio, l'equazione di Laplace è ora esprimibile nella forma seguente:

$$
\begin{cases}
\dfrac{1}{\rho}\dfrac{\partial}{\partial\rho}\left(\rho\dfrac{\partial u}{\partial\rho}\right)+\dfrac{1}{\rho^2}\dfrac{\partial^2 u}{\partial\theta^2}=0,\\
u(1,2\pi)=u(1,0),\\
u_\theta(1,2\pi)=u_\theta(1,0),\qquad\text{(condizioni di periodicità)},\\
u(1,\theta)=f(\theta),\qquad\text{(condizione al contorno)},\\
\lim\limits_{\rho\to0^+}\ u(\rho,\theta)\quad\text{limitato}.
\end{cases}
$$

Essendo $\rho\in[0,1]$ e $\theta\in[0,2\pi]$, nelle nuove variabili il dominio è rettangolare, per cui possiamo porre:

$$
U(\rho,\theta)=U(\rho)V(\theta),
$$

da cui, sostituendo nella PDE,

$$
U''V+\frac{1}{\rho}U'V+\frac{1}{\rho^2}UV''=0\iff\rho^2U''V+\rho U'V+UV''=0,
$$

$$
\frac{\rho^2U''+\rho U'}{U}=-\frac{V''}{V}=\lambda.
$$

Da essa, tenuto conto delle condizioni di periodicità, seguono: il problema spettrale

$$
\begin{cases}
V''+\lambda V=0,\\
V(0)=V(2\pi),\\
V'(0)=V'(2\pi),
\end{cases}
$$

e la ODE

$$\begin{cases} \rho^2 U'' + \rho U' - \lambda U = 0, \\ \lim\limits_{\rho \to 0^+} U(\rho) \text{ limitato.} \end{cases}$$

Procedendo, in analogia con quanto già visto, si ottiene lo spettro

$$\left\{ \lambda_n = n^2;\ V_n(\theta) = a_n \cos n\theta + b_n \sin n\theta \right\}_{n=0}^{\infty}.$$

Per ogni $n \geq 0$, si deve ora risolvere l'equazione

$$\rho^2 U_n'' + \rho U_n' - \lambda U_n = 0, \tag{4.12}$$

che costituisce una classica equazione di Eulero-Cauchy ($x^2 y'' + xy' + by = 0$). Per la sua soluzione, posto $U_n = \rho^m$, dobbiamo risolvere l'equazione

$$\rho^m [m(m-1) + m - n^2] = 0,$$

da cui $m = \pm n$. Di conseguenza

$$U_n(\rho) = c_n \rho^n + d_n \rho^{-n}.$$

L'imposizione che il $\lim\limits_{\rho \to 0^+} U(\rho)$ sia limitato, implica che $d_n = 0$ per ogni $n \geq 1$ e di conseguenza, tenendo conto dell'espressione della V_n,

$$u(\rho, \theta) = a_0 + \sum_{n=1}^{\infty} \rho^n (a_n \cos n\theta + b_n \sin n\theta).$$

Dalla condizione $u(1, \theta) = f(\theta)$ si ricavano i coefficienti di Fourier

$$\begin{aligned} a_0 &= \frac{1}{2\pi} \int_0^{2\pi} f(\tau)\, d\tau \\ a_n &= \frac{1}{\pi} \int_0^{2\pi} f(\tau) \cos n\tau\, d\tau \\ b_n &= \frac{1}{\pi} \int_0^{2\pi} f(\tau) \sin n\tau\, d\tau. \end{aligned}$$

da cui

$$\begin{aligned} u(\rho, \theta) &= \frac{1}{2\pi} \int_0^{2\pi} \left[1 + 2 \sum_{n=1}^{\infty} \rho^n \left(\cos n\tau \cos n\theta + \sin n\tau \sin n\theta \right) \right] f(\tau)\, d\tau \\ &= \frac{1}{2\pi} \int_0^{2\pi} \left[1 + 2 \sum_{n=1}^{\infty} \rho^n \cos n(\theta - \tau)] f(\tau)\, d\tau \right] f(\tau)\, d\tau. \end{aligned}$$

A questo punto si osserva che la serie entro parentesi può essere espressa in forma chiusa, in quanto

$$1 + 2\sum_{n=1}^{\infty} r^n \cos n\theta = \mathrm{Re}\,(1 + re^{i\theta})(1 + re^{i\theta} + r^2 e^{2i\theta} + \ldots)$$
$$= \mathrm{Re}\,\frac{1 + re^{i\theta}}{1 - re^{i\theta}} = \frac{1 - r^2}{1 + r^2 - 2r\cos\theta}.$$

Si ottiene così la formula risolvente di Poisson

$$u(\rho, \theta) = \frac{1}{2\pi}\int_0^{2\pi} \frac{1 - \rho^2}{1 + \rho^2 - 2\rho\cos(\theta - \tau)} f(\tau)\, d\tau,$$

che definisce la $u(\rho, \theta)$ mediante la u sulla circonferenza.

4.2 b Equazioni paraboliche

In questa sezione illustriamo la risoluzione di alcune equazioni paraboliche con condizioni al bordo di vario tipo.

5. Equazione parabolica con condizioni iniziali e al bordo

$$\begin{cases} \dfrac{\partial u}{\partial t} = 2\dfrac{\partial^2 u}{\partial x^2}, \qquad 0 \le x \le 3,\ t \ge 0, \\ u(0, t) = u(3, t) = 0, \\ u(x, 0) = 5\sin 4\pi x - 3\sin 8\pi x + 2\sin 10\pi x. \end{cases} \tag{4.13}$$

Si tratta dell'equazione del calore in una dimensione spaziale che ammette una e una sola soluzione, visto che sono dati i valori della soluzione agli estremi dell'intervallo e la distribuzione della temperatura all'istante iniziale lungo tutto l'intervallo $(0, 3)$. Cerchiamo una soluzione del tipo:

$$u(x, y) = X(x)T(t), \qquad X \text{ e } T \text{ differenziabili.}$$

Sostituendo e riordinando si ha

$$XT' = 2X''T \quad \text{ossia} \quad \frac{T'}{2T} = \frac{X''}{X}.$$

Imponendo che entrambi i rapporti siano uguali ad uno stesso parametro λ, si ottiene il sistema

$$\begin{cases} X''(x) - \lambda X(x) = 0, \\ T'(t) - 2\lambda T(t) = 0. \end{cases}$$

Utilizzando le condizioni al contorno date agli estremi, otteniamo

$$\begin{cases} u(0,t) = X(0)T(t) = 0, \\ u(3,t) = X(3)T(t) = 0, \end{cases}$$

la condizione $T(t) = 0$ per ogni t non è accettabile, in quanto essa implicherebbe $u(x,t) = X(x)T(t) = 0$ per ogni (x,t) nel dominio, con la conseguenza che non risulterebbe soddisfatta la condizione iniziale

$$u(x,0) = 5\sin 4\pi x - 3\sin 8\pi x + 2\sin 10\pi x.$$

Di conseguenza la $X(x)$ deve soddisfare il problema spettrale

$$\begin{cases} X''(x) - \lambda X(x) = 0, \quad 0 \le x \le 3, \\ X(0) = 0, \quad X(3) = 0, \end{cases} \tag{4.14}$$

che rappresenta un tipico problema di Sturm-Liouville (Cap. 3). Per la sua risoluzione occorre distinguere tra i seguenti casi:

- $\lambda > 0$. In questo caso l'equazione caratteristica ammette due radici reali e distinte, con una soluzione esponenziale del tipo

  $$X(x) = a\, e^{x\sqrt{\lambda}} + b\, e^{-x\sqrt{\lambda}}.$$

  Imponendo le condizioni al contorno si trova il seguente sistema lineare omogeneo

  $$\begin{pmatrix} 1 & 1 \\ e^{3\sqrt{\lambda}} & e^{-3\sqrt{\lambda}} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

  Essendo il determinante della matrice $e^{-3\sqrt{\lambda}} - e^{3\sqrt{\lambda}} = -2\sinh(3\sqrt{\lambda}) \neq 0$ per ogni $\lambda > 0$, risulta $a = b = 0$, soluzione da scartare perché non risulterebbe soddisfatta la condizione iniziale.

- $\lambda = 0$. L'equazione caratteristica ammette due soluzioni reali e coincidenti, per cui

  $$X(x) = (a + bx)e^{x\sqrt{\lambda}}.$$

  Applicando le condizioni al contorno, si trova subito che $a = b = 0$, soluzione non accettabile.

- $\lambda < 0$. In questo caso l'equazione caratteristica possiede due soluzioni complesse coniugate, cui corrisponde la soluzione generale

  $$X(x) = a\, \cos(x\sqrt{-\lambda}) + b\, \sin(x\sqrt{-\lambda}).$$

Imponendo che siano soddisfatte le condizioni agli estremi, si ottiene

$$X(0) = X(3) = 0 \Longrightarrow \begin{cases} a = 0, \\ b \sin(3\sqrt{-\lambda}) = 0. \end{cases}$$

Di conseguenza si ha una infinità numerabile di autovalori λ_k, cui corrisponde una infinità numerabile di autofunzioni $X_k(x)$ e precisamente:

$$\lambda_k = -\left(\frac{k\pi}{3}\right)^2, \quad X_k(x) = \sin\left(\frac{k\pi x}{3}\right), \quad k = 1, 2, 3, \ldots. \tag{4.15}$$

Avendo determinato $X_k(x)$, per ogni λ_k dobbiamo determinare la corrispondente soluzione $T_k(t)$ della ODE

$$T_k'(t) - 2\lambda_k T_k(t) = 0. \tag{4.16}$$

La soluzione della (4.16), a meno di una costante moltiplicativa, è $e^{2\lambda_k t}$ e, di conseguenza, per ogni $k \in \mathbb{N}$, abbiamo la soluzione particolare

$$u_k(x,t) = e^{2\lambda_k t} \sin\left(\frac{k\pi x}{3}\right), \quad k \in \mathbb{N}.$$

La soluzione generale è una combinazione lineare delle $u_k(x,t)$, ossia la serie

$$u(x,t) = \sum_{k=1}^{\infty} c_k u_k(x,t) = \sum_{k=1}^{\infty} c_k \, e^{-2\left(\frac{k\pi}{3}\right)^2 t} \sin\left(\frac{k\pi x}{3}\right), \tag{4.17}$$

con i coefficienti $\{c_k\}_{k=1}^{\infty}$ da determinare in modo che risulti soddisfatta la condizione iniziale, l'unica non soddisfatta dalle singole $u_k(x,t)$.

Imponendo quest'ultimo vincolo si ottiene la seguente condizione sui coefficienti c_k:

$$\sum_{k=1}^{\infty} c_k \sin\left(\frac{k\pi x}{3}\right) = 5 \sin 4\pi x - 3 \sin 8\pi x + 2 \sin 10\pi x,$$

e questo implica che soltanto i tre seguenti coefficienti della serie siano diversi da zero:

$$\begin{cases} c_{12} = 5, \\ c_{24} = -3, \\ c_{30} = 2. \end{cases}$$

Di conseguenza l'unica soluzione del problema (4.13) è

$$u(x,t) = 5\, e^{-2(4\pi)^2 t} \sin 4\pi x - 3\, e^{-2(8\pi)^2 t} \sin 8\pi x + 2\, e^{-2(10\pi)^2 t} \sin 10\pi x.$$

6. Consideriamo il problema parabolico con condizioni di Dirichlet

$$\begin{cases} \dfrac{\partial u}{\partial t} = 2\dfrac{\partial^2 u}{\partial x^2}, \quad 0 \le x \le 3,\ t \ge 0, \\ u(0,t) = 10, \quad u(3,t) = 40, \\ u(x,0) = 6\sin \pi x + 3(x^2+1). \end{cases} \tag{4.18}$$

Si tratta di un problema parabolico con condizioni agli estremi non omogenee. In questi casi è sempre possibile convertire il problema dato in uno avente condizioni agli estremi omogenee. È sufficiente effettuare la sostituzione

$$u(x,t) = v(x,t) + \varphi(x), \tag{4.19}$$

con $\varphi \in C^2([0,3])$. Sostituendo la (4.19) nella (4.18), si ha

$$\begin{cases} \dfrac{\partial v}{\partial t} = 2\dfrac{\partial^2 v}{\partial x^2} + 2\varphi''(x), \\ v(0,t) + \varphi(0) = 10, \quad v(3,t) + \varphi(3) = 40, \\ v(x,0) + \varphi(x) = 6\sin \pi x + 3(x^2+1). \end{cases}$$

Di conseguenza, per avere un problema differenziale con la stessa PDE ma con tre condizioni agli estremi omogenee, è sufficiente scegliere $\varphi(x)$ in modo che sia

$$\begin{cases} \varphi''(x) = 0, \\ \varphi(0) = 10, \quad \varphi(3) = 40, \end{cases} \tag{4.20}$$

la cui soluzione è $\varphi(x) = 10x + 10$. Il problema (4.18) diventa quindi

$$\begin{cases} \dfrac{\partial v}{\partial t} = 2\dfrac{\partial^2 v}{\partial x^2}, \\ v(0,t) = 0, \quad v(3,t) = 0, \\ v(x,0) = 6\ \sin \pi x + 3\,x^2 + 1) - 10(x+1) = 6\ \sin \pi x + 3\,x^2 - 10x - 7. \end{cases} \tag{4.21}$$

Procedendo per separazione delle variabili, come nell'esempio precendente, si trova che la $v(x,t)$ è rappresentata dalla serie

$$v(x,t) = \sum_{k=1}^{\infty} c_k v_k(x,t) = \sum_{k=1}^{\infty} c_k\, e^{-\lambda_k t} \sin\left(\frac{k\pi x}{3}\right), \quad \lambda_k = \left(\frac{k\pi}{3}\right)^2. \tag{4.22}$$

Imponendo infine la condizione iniziale $v(x,0)$, si ottiene l'equazione

$$\sum_{k=1}^{\infty} c_k \sin\left(\frac{k\pi x}{3}\right) = 6\sin \pi x + 3x^2 - 10x - 7,$$

dalla quale segue che $c_3 = 6$, mentre le altre costanti c_k con $k \in \mathbb{N} \setminus \{3\}$ devono essere calcolate in modo che risulti soddisfatta l'equazione

$$\sum_{\substack{k=1 \\ k\neq 3}}^{\infty} c_k \sin \frac{k\pi}{3} x = 3x^2 - 10x - 7. \tag{4.23}$$

La (4.23) è una serie di Fourier e per isolare i coefficienti c_k occorre moltiplicare primo e secondo membro per $\sin(m\pi x/3)$ ($m \in \mathbb{N} \setminus \{3\}$) e integrare nell'intervallo $[0, 3]$, in modo da sfruttare l'ortogonalità delle funzioni $\sin(m\pi x/3)$ nell'intervallo $(0, 3)$. Così operando si trova che

$$c_m \int_0^3 \left(\sin \frac{m\pi x}{3}\right)^2 dx = \int_0^3 \left(\sin \frac{m\pi x}{3}\right) (3x^2 - 10x - 7)\, dx,$$

da cui, ricordando che $\sin^2(m\pi/3) = \frac{1}{2}[1 - \cos(2m\pi/3)]$, segue che

$$\tfrac{3}{2} c_m = \int_0^3 \sin\left(\frac{m\pi x}{3}\right) (3x^2 - 10x - 7)\, dx. \tag{4.24}$$

Calcolando, mediante integrazione per parti, l'integrale a secondo membro si trova che

$$c_m = \tfrac{2}{3}\left[2\frac{3^3}{(m\pi)^2}[(-1)^m - 1] + \frac{192}{m\pi}(-1)^{m+1} + \frac{21}{m\pi}\right], \tag{4.25}$$

con $m \in \mathbb{N} \setminus \{3\}$.

Possiamo dunque affermare che la soluzione del problema (4.18) è data da

$$u(x,t) = \sum_{m=1}^{\infty} c_m\, e^{-\lambda_m t} \sin\left(\frac{m\pi x}{3}\right) + 10(x+1),$$

dove $c_3 = 6$ e i coeffcenti c_m con $m \neq 3$ sono dati dalla (4.25).

7. Equazione del calore con estremi isolati

$$\begin{cases} \dfrac{\partial u}{\partial t} = a^2 \dfrac{\partial^2 u}{\partial x^2}, \quad 0 \le x \le L, \quad t \ge 0, \\ u(0,t) = u(L,t) = 0, \\ u(x,0) = f(x), \end{cases} \tag{4.26}$$

dove a è una costante reale e positiva.

Il problema (4.26) rappresenta un modello parabolico descrivente un filamento mantenuto a zero gradi agli estremi e avente una distribuzione di

temperatura iniziale data da $f(x)$ che, per semplicità, supponiamo continua. Tale problema ammette una e una sola soluzione che cerchiamo col metodo di separazione della variabili, ponendo

$$u(x,t) = X(x)T(t) \quad \Longrightarrow \frac{T'(t)}{a^2T(t)} = \frac{X''(x)}{X(x)} = -\lambda.$$

Dalle condizioni agli estremi del filamento ($u(0,t) = u(L,t) = 0$) segue che

$$\begin{cases} X''(x) + \lambda X(x) = 0, \\ X(0) = 0, \quad X(L) = 0. \end{cases}$$

Questo problema spettrale ammette i seguenti autovalori e autofunzioni:

$$\begin{cases} \lambda_k = \left(\frac{k\pi}{L}\right)^2, \\ X_k(x) = \sin\left(\frac{k\pi x}{L}\right), \end{cases} \tag{4.27}$$

dove $k \in \mathbb{N}$. Dall'equazione $T_k' + \lambda_k T_k = 0$, segue che

$$T_k(t) = e^{-a^2\lambda_k t} = e^{-\frac{k^2\pi^2a^2t}{L^2}}.$$

La soluzione $u(x,t)$ è dunque

$$u(x,t) = \sum_{k=1}^{\infty} c_k\, e^{-\frac{k^2\pi^2a^2t}{L^2}}\, \sin\frac{k\pi x}{L}.$$

Dalla condizione $u(x,0) = f(x)$ si ricava infine che

$$c_k = \frac{2}{L}\int_0^L f(x) \sin\frac{k\pi x}{L}.$$

8. Equazione del calore con un estremo radiante

$$\begin{cases} \dfrac{\partial u}{\partial t} = a^2\dfrac{\partial^2 u}{\partial x^2}. \quad 0 \le x \le L,\ t \ge 0, \\ u(0,t) = 0, \\ u_x(L,t) = -Au(L,t), \\ u(x,0) = f(x), \end{cases} \tag{4.28}$$

con a e A costanti reali positive.

Il problema (4.28) modellizza un filamento con una distribuzione di temperatura iniziale data da $f(x) \in C([0,L])$, con un estremo mantenuto a zero gradi e con l'altro estremo che irradia calore.

Dalla teoria sappiamo che il problema ammette una sola soluzione, che cerchiamo nella forma.

$$u(x,t) = X(x)T(t) \Longrightarrow \frac{T'}{a^2 T} = \frac{X''}{X} = -\lambda.$$

Le condizioni agli estremi della sbarra danno luogo alle seguenti condizioni per la $X(x)$:

$$\begin{cases} u(0,t) = 0, \\ u_x(L,t) = -Au(L,t), \end{cases} \Longrightarrow \begin{cases} X(0)T(t) = 0, \\ X'(L)T(t) = -A\,X(L)T(t), \end{cases}$$
$$\Longrightarrow \begin{cases} X(0) = 0, \\ X'(L) = -AX(L), \end{cases}$$

che generano il seguente problema di Sturm-Liouville con condizioni miste (Appendice A):

$$\begin{cases} X'' + \lambda X = 0, \\ X(0) = 0, \\ X'(L) + AX(L) = 0. \end{cases} \tag{4.29}$$

Avendo condizioni al contorno miste, il problema (4.29) va discusso al variare del segno dell'autovalore λ.

- $\lambda < 0$.

  $$X(x) = a\,e^{x\sqrt{|\lambda|}} + b\,e^{-x\sqrt{|\lambda|}},$$
  $$X'(x) = a\,\sqrt{|\lambda|}\,e^{x\sqrt{|\lambda|}} - b\,\sqrt{|\lambda|}\,e^{-x\sqrt{|\lambda|}}.$$

  Le condizioni al contorno implicano

  $$\begin{cases} a + b = 0, \\ a\sqrt{|\lambda|}\,e^{L\sqrt{|\lambda|}} - b\sqrt{|\lambda|}\,e^{-L\sqrt{|\lambda|}} + A(a\,e^{L\sqrt{|\lambda|}} + b\,e^{-L\sqrt{|\lambda|}}) = 0. \end{cases}$$

  Occorre stabilire se il seguente sistema omogeneo ammette soluzioni non banali:

  $$\begin{pmatrix} 1 & 1 \\ (\sqrt{|\lambda|} + A)e^{\sqrt{|\lambda|}L} & (-\sqrt{|\lambda|} + A)e^{-\sqrt{|\lambda|}L} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{4.30}$$

  Si trova facilmente che il determinante della matrice dei coefficienti della (4.30) è sempre diverso da zero perché l'equazione trascendente

  $$(A + \sqrt{|\lambda|})\,e^{L\sqrt{|\lambda|}} = (A - \sqrt{|\lambda|})e^{-L\sqrt{|\lambda|}}$$

  è soddisfatta solo per $\lambda = 0$. Di conseguenza $a = b = 0$, caso da scartare.

- $\lambda = 0$.

$$X(x) = ax + b,$$
$$X'(x) = a.$$

Le condizioni al contorno implicano

$$\begin{cases} b = 0, \\ a + A(aL + b) = 0, \end{cases}$$

cioè $a = b = 0$ e quindi anche questo caso è da scartare.

- $\lambda > 0$.

$$X(x) = a\,\cos(x\sqrt{\lambda}) + b\,\sin(x\sqrt{\lambda}),$$
$$X'(x) = -a\sqrt{\lambda}\,\sin(x\sqrt{\lambda}) + b\sqrt{\lambda}\,\cos(x\sqrt{\lambda}).$$

Le condizioni al contorno implicano

$$\begin{cases} a = 0, \\ b\sqrt{\lambda}\,\cos(L\sqrt{\lambda}) + Ab\,\sin(L\sqrt{\lambda}) = 0, \end{cases}$$

dove necessariamente $b \neq 0$, diversamente sarebbe $X(x) \equiv 0$. Da essa segue che $\cos(L\sqrt{\lambda}) \neq 0$, diversamente anche $\sin(L\sqrt{\lambda}) = 0$ e questo è impossibile perché $\sin\alpha$ e $\cos\alpha$ non si annullano mai simultaneamente. Di conseguenza

$$\operatorname{tg} z = -\frac{z}{AL}, \quad \text{dove} \quad z = L\sqrt{\lambda}. \tag{4.31}$$

La (4.31) è un'equazione trascendente che possiede una infinità numerabile di soluzioni $\{z_n\}_{n=1}^{\infty}$.

$$(2n-1)\tfrac{\pi}{2} \leq z_n \leq n\pi. \qquad n = 1, 2, \ldots.$$

Pertanto il problema di Sturm-Liouville studiato ammette le seguenti autofunzioni e autovalori

$$\begin{cases} X_n(x) = \sin\left(\dfrac{z_n x}{L}\right), \\ (2n-1)\tfrac{\pi}{2} \leq z_n \leq n\pi, \quad n \in \mathbb{N} \setminus \{0\}, \end{cases} \tag{4.32}$$

con $z_n = L\sqrt{\lambda_n}$. Da notare che tali autofunzioni sono ortogonali in $[0, L]$, dato che la funzione peso dell'equazione $X'' + \lambda X = 0$ è $r(x) = 1$.

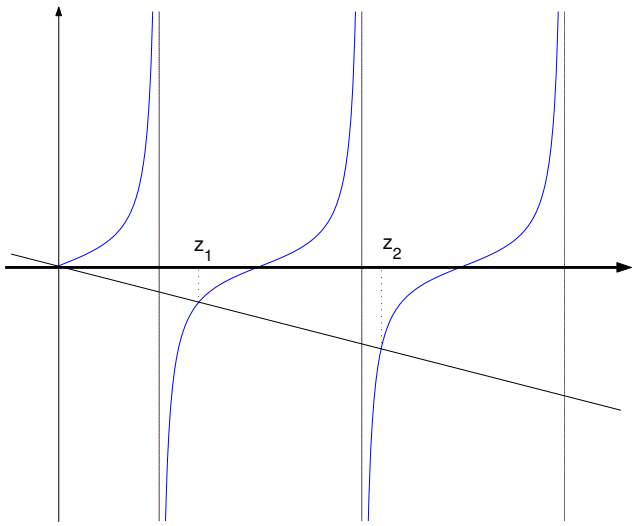


Figura 4.1: La figura mostra le intersezioni della curva $y = \operatorname{tg} z$ e della retta $y = -z/(AL)$.

Si trova facilmente che la soluzione per la $T_n(t)$ è (a meno di una costante moltiplicativa)

$$T_n(t) = e^{-\frac{z_n^2 a^2}{L^2} t},$$

e quindi

$$u(x,t) = \sum_{n=1}^{\infty} c_n \, e^{-\frac{z_n^2 a^2 t}{L^2}} \, \sin \frac{z_n x}{L}.$$

Dalla condizione $u(x,0) = f(x)$ si ottiene infine

$$\sum_{n=1}^{\infty} c_n \sin \frac{z_n x}{L} = f(x),$$

da cui, sfruttando l'ortogonalità delle autofunzioni $\sin \frac{z_n x}{L}$ in $[0, L]$, si ha che

$$c_n = \frac{\int_0^L f(x) \sin \frac{z_n x}{L} x \, dx}{\int_0^L \sin^2 \frac{z_n x}{L} \, dx}.$$

9. Equazione parabolica con condizioni di tipo misto

$$\begin{cases} u_t = 6u_{xx} + 3u_x + u, \quad \text{con } 0 \le x \le 2, \ t \ge 0, \\ u_x(0,t) + u(0,t) = 0, \\ u(2,t) = 0, \\ u(x,0) = x^2(2-x). \end{cases}$$

Sostituendo nell'equazione $u(x,t) = X(x)T(t)$ si trova

$$\frac{T'}{T} = 6\frac{X''}{X} + 3\frac{X'}{X} + 1 = -\lambda.$$

Si perviene così al seguente problema spettrale

$$\begin{cases} 6X'' + 3X' + (\lambda + 1)X = 0, \quad \text{con } 0 \leq x \leq 2, \\ X'(0) + X(0) = 0, \\ X(2) = 0. \end{cases}$$

Altre informazioni vengono poi ottenute dall'equazione

$$T' = -\lambda T.$$

Risolviamo il problema spettrale. Si cercano soluzioni del tipo $X(x) \simeq e^{\alpha x}$ che conducono all'equazione caratteristica

$$6\alpha^2 + 3\alpha + (1 + \lambda) = 0.$$

Questa equazione ha come soluzioni $\alpha_{1,2} = \frac{1}{12}[-3 \pm \sqrt{-(24\lambda + 15)}]$.

- $\lambda = -\frac{5}{8}$. In tal caso abbiamo due radici reali e coincidenti $\alpha_1 = \alpha_2 = -\frac{1}{4}$. La soluzione generale è $X(x) = e^{-\frac{1}{4}x}(c_1 + c_2 x)$. Essendo $X(0) = c_1$, $X'(0) = -\frac{1}{4}c_1 + c_2$ e $X(2) = e^{-\frac{1}{2}}(c_1 + 2c_2)$, le costanti c_1 e c_2 devono soddisfare il sistema
$$\begin{cases} \frac{3}{4}c_1 + c_2 = 0, \\ c_1 + 2c_2 = 0, \end{cases}$$
il quale ammette la sola soluzione banale.

- $\lambda < -\frac{5}{8}$. In tal caso abbiamo due radici reali e distinte $\alpha_{1,2} = \frac{1}{12}[-3 \pm \sqrt{-(24\lambda + 15)}]$. La soluzione generale è $X(x) = e^{-\frac{x}{4}}(c_1 e^{-\mu x} + c_2 e^{\mu x})$, dove si è posto $\mu = \frac{1}{12}\sqrt{-24\lambda - 15} > 0$. Le costanti c_1 e c_2 sono da determinare mediante il sistema di equazioni lineari e omogenee
$$\begin{cases} c_1 e^{-2\mu} + c_2 e^{2\mu} = 0, \\ (\frac{3}{4} - \mu)c_1 + (\frac{3}{4} + \mu)c_2 = 0. \end{cases}$$
Il determinante relativo è
$$(\frac{3}{4} + \mu)e^{-2\mu} - (\frac{3}{4} - \mu)e^{2\mu},$$

che si annulla quando

$$(\tfrac{3}{4}+\mu)e^{-2\mu}-(\tfrac{3}{4}-\mu)e^{2\mu}=0.$$

Tenuto conto che $\mu>0$, l'equazione può essere riscritta come

$$\left(\frac{3-4\mu}{3+4\mu}\right)e^{4\mu}=1.$$

Le eventuali soluzioni di questa equazione sono date dall'intersezione fra la retta di equazione $y=1$ e la funzione $y=\left(\frac{3-4\mu}{3+4\mu}\right)e^{4\mu}$ per $\mu>0$. Poiché $y'=\frac{4e^{4\mu}}{(3+4\mu)^2}(-16\mu^2+3)$ la funzione è crescente per $0<\mu<\frac{\sqrt{3}}{4}$, decrescente per $\mu>\frac{\sqrt{3}}{4}$ e raggiunge il massimo (assoluto) per $\mu=\frac{\sqrt{3}}{4}$. Inoltre poiché $\lim_{\mu\to 0^+} y=1$, $y(\frac{3}{4})=0$ e $\lim_{\mu\to+\infty} y=-\infty$ il grafico è quello riportato nella Fig. 4.2.

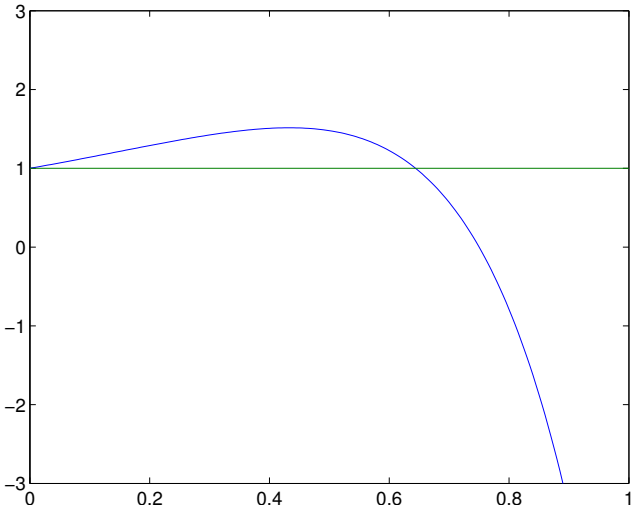


Figura 4.2: Grafico delle funzioni $y=\left(\frac{3-4x}{3+4x}\right)e^{4x}$ e $y=1$.

Poiché i due grafici si intersecano in un unico punto di ascissa $\tilde{\mu}$ con $\frac{\sqrt{3}}{4}<\tilde{\mu}<\frac{3}{4}$, essendo $\tilde{\mu}=\frac{1}{12}\sqrt{-24\lambda-15}$, il corrispondente autovalore $\tilde{\lambda}=-\frac{5}{8}-\frac{\tilde{\mu}^2}{6}$ e la corrispondente autofunzione è

$$\tilde{X}(x)=e^{-\frac{x}{4}}(e^{-\tilde{\mu}x}-e^{-4\tilde{\mu}}e^{\tilde{\mu}x}).$$

In corrispondenza all'autovalore $\tilde{\mu}$ abbiamo l'equazione

$$\tilde{T}'=-\tilde{\mu}\tilde{T},$$

da cui si ricava $\tilde{T} = e^{-\tilde{\mu}t}$. Otteniamo quindi l'autofunzione

$$\tilde{u}(x,t) = e^{-\left(\frac{1}{4}x - \tilde{\mu}t\right)}(e^{-\tilde{\mu}x} - e^{-4\tilde{\mu}}e^{\tilde{\mu}x}).$$

- $\lambda > -\frac{5}{8}$. In tal caso abbiamo due radici complesse coniugate $\alpha_{1,2} = \frac{1}{12}[-3 \pm i\sqrt{24\lambda + 15}]$. La soluzione generale è

$$X(x) = e^{-\frac{x}{4}}\left(c_1 \cos \mu x + c_2 \sin \mu x\right),$$

dove si è posto $\mu = \frac{\sqrt{24\lambda+15}}{12} > 0$. Poiché $X(0) = c_1$, $X'(0) = -\frac{1}{4}c_1 + c_2\mu$ e $X(2) = e^{-\frac{1}{2}}(c_1 \cos 2\mu + c_2 \sin 2\mu)$, le costanti c_1 e c_2 sono da determinare tramite il sistema

$$\begin{cases} \frac{3}{4}c_1 + c_2\mu = 0, \\ c_1 \cos 2\mu + c_2 \sin 2\mu = 0. \end{cases}$$

Dalla prima equazione del sistema si ricava $c_1 = -\frac{4}{3}c_2\mu$, che, sostituita nella seconda, fornisce l'equazione

$$c_2\left[-\frac{4}{3}\mu \cos 2\mu + \sin 2\mu\right] = 0.$$

Le soluzioni non banali si ottengono risolvendo l'equazione

$$-\frac{4}{3}\mu \cos 2\mu + \sin 2\mu = 0.$$

Potendo assumere $\cos 2\mu \neq 0$, ci si può ricondurre all'equazione

$$\operatorname{tg} 2\mu = \tfrac{2}{3}(2\mu) = \tfrac{4}{3}\mu, \quad \mu > 0.$$

Per risolvere questa equazione si procede per via grafica.

I grafici delle funzioni $y = \operatorname{tg} 2\mu$ e $y = \frac{2}{3}(2\mu)$ (con $\mu > 0$) si intersecano in un insieme numerabile di punti, in ciascuno dei quali l'equazione possiede una soluzione non banale. Dalla Fig. 4.3 si vede che le ascisse μ_k dei punti d'intersezione soddisfano la seguente condizione $\frac{\pi}{4} + k\frac{\pi}{2} < \mu_k < \frac{\pi}{4} + (k+1)\frac{\pi}{2}$, $k = 0, 1, 2, \ldots$. I corrispondenti autovalori sono $\lambda_k = -\frac{5}{8} + 6\mu_k^2$, cui corrispondono le autofunzioni

$$X_k(x) = e^{-\frac{x}{4}}\left(-\tfrac{4}{3}\mu_k \cos \mu_k x + \sin \mu_k x\right).$$

In corrispondenza abbiamo anche le equazioni $T_k' + \mu_k T = 0$, che conducono alle soluzioni

$$T_k(t) = e^{-\mu_k t}.$$

In definitiva troviamo le funzioni

$$u_k(x,t) = T_k(t)X_k(x) = e^{-\mu_k t}e^{-\frac{x}{4}}(-\tfrac{4}{3}\mu_k \cos \mu_k x + \sin \mu_k x).$$

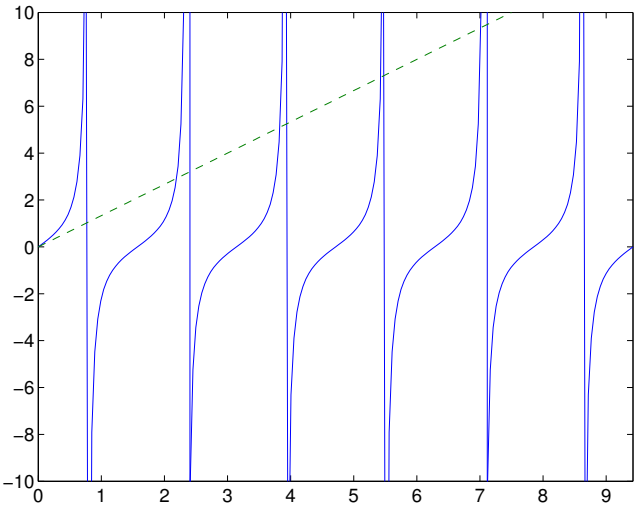


Figura 4.3: Grafico delle funzioni $y = \operatorname{tg} 2x$ e $y = \frac{4}{3}x$.

La soluzione del problema dunque risulta

$$u(x,t) = \tilde{a}\tilde{u}(x,t) + \sum_{k=1}^{\infty} c_k u_k(x,t) = \tilde{a}e^{-\tilde{\mu}t}e^{-\frac{x}{4}}(e^{-\tilde{\mu}x} - e^{-4\tilde{\mu}}e^{\tilde{\mu}x})$$
$$+ \sum_{k=1}^{\infty} c_k e^{-\mu_k t}e^{-\frac{x}{4}}(-\tfrac{4}{3}\mu_k \cos \mu_k x + \sin \mu_k x),$$

con $\tilde{a}$ e c_k costanti da determinare. A tal fine imponiamo la condizione $u(x,0) = x^2(2-x)$. Questa implica

$$\tilde{a}.^{-\frac{x}{4}}(e^{-\tilde{\mu}x} - e^{-4\tilde{\mu}}e^{\tilde{\mu}x}) + \sum_{k=1}^{\infty} c_k\, e^{-\frac{x}{4}}(-\tfrac{4}{3}\mu_k \cos \mu_k x + \sin \mu_k x) = x^2(2-x).$$

In questa condizione compaiono $\tilde{u}(x,0)$ e le $u_k(x,0)$ che sono soluzioni della equazione $X'' + \frac{1}{2}X' + \frac{1}{6}(\lambda+1)X = 0$, la quale può scriversi nella forma canonica

$$\tfrac{d}{dx}(e^{\frac{x}{2}}X') + e^{\frac{x}{2}}(\tfrac{1}{6} + \tfrac{1}{6}\lambda)X = 0.$$

La funzione $r(x) = \frac{1}{6}e^{\frac{x}{2}} > 0$ non cambia segno e questo garantisce che le funzioni $\tilde{u}(x,0)$ e $u_k(x,0)$ $(k = 1,2,3,\ldots)$ siano ortogonali in $[0,2]$ rispetto ad essa. Si ottengono così le seguenti espressioni dei coefficienti:

$$\tilde{a} = \frac{\int_0^2 e^{\frac{x}{2}}x^2(2-x)\tilde{u}(x,t)dx}{\int_0^2 e^{\frac{x}{2}}\tilde{u}(x,t)^2 dx},$$
$$c_k = \frac{\int_0^2 e^{\frac{x}{2}}x^2(2-x)u_k(x,t)dx}{\int_0^2 e^{\frac{x}{2}}u_k(x,t)^2 dx}, \quad k = 1,2,\ldots.$$

10. Equazione parabolica con condizioni miste

$$\begin{cases} u_t = u_{xx} + 2u_x + 3u, \\ u_x(0,t) + 2u(0,t) = 0, \\ u(\pi,t) = 3, \\ u(x,0) = \sin \pi x. \end{cases}$$

Si cercano soluzioni del tipo $u(x,t) = v(x,t) + \phi(x)$. Sostituendo si trova

$$\begin{cases} v_t = v_{xx} + \phi'' + 2v_x + 2\phi' + 3v + 3\phi, \\ v_x(0,t) + \phi'(0) + 2v(0,t) + 2\phi(0) = 0, \\ v(\pi,t) + \phi(\pi) = 3, \\ v(x,0) = \sin \pi x - \phi(x), \end{cases}$$

da cui seguono i due problemi

$$\begin{cases} v_t = v_{xx} + 2v_x + 3v, \\ v_x(0,t) + 2v(0,t) = 0, \\ v(\pi,t) = 0, \\ v(x,0) = \sin \pi x - \phi(x), \end{cases} \tag{4.33}$$

$$\begin{cases} \phi'' + 2\phi' + 3\phi = 0, \\ \phi'(0) + 2\phi(0) = 0, \\ \phi(\pi) = 3. \end{cases} \tag{4.34}$$

È immediato osservare che la soluzione cercata è del tipo

$$\phi(x) = e^{-x}\left(a\,\cos(x\sqrt{2}) + b\,\sin(x\sqrt{2})\right),$$

con a e b costanti da determinare tramite il sistema

$$\begin{cases} a + \sqrt{2}b = 0, \\ \left\{a\cos(\pi\sqrt{2}) + b\sin(\pi\sqrt{2})\right\} = 3e^{\pi}. \end{cases}$$

Indicata con $\left(\tilde{a}, \tilde{b}\right)$ la sua soluzione, la soluzione del problema (4.34) è

$$\phi(x) = e^{-x}\left(\tilde{a}\,\cos(x\sqrt{2}) + \tilde{b}\,\sin(x\sqrt{2})\right).$$

Possiamo adesso soffermarci sul problema (4.33). Cerchiamo soluzioni del tipo $v(x,t) = X(x)T(t)$. Sostituendo nell'equazione si trova

$$\frac{T'}{T} = \frac{X''}{X} + 2\frac{X'}{X} + 3 = -\lambda,$$

da cui segue il problema spettrale

$$\begin{cases} X'' + 2X' + (\lambda + 3)X = 0, \\ X'(0) + 2X(0) = 0, \\ X(\pi) = 0. \end{cases}$$

Altre informazioni vengono poi ottenute dall'equazione

$$T' = -\lambda T.$$

Per risolvere il problema spettrale, si cercano soluzioni del tipo $X(x) \simeq e^{\alpha x}$, le quali conducono all'equazione caratteristica

$$\alpha^2 + 2\alpha + (3 + \lambda) = 0.$$

Questa equazione ha come soluzioni $\alpha_{1,2} = -1 \pm \sqrt{-(\lambda + 2)}$. Occorre distinguere tre casi a seconda che sia $-2 - \lambda = 0$, $-2 - \lambda > 0$, oppure $-2 - \lambda < 0$.

- $\lambda = -2$. In questo caso la soluzione generale è $X(x) = e^{-x}(c_1 + c_2 x)$ con le costanti (c_1, c_2), soluzione del sistema

  $$\begin{cases} c_1 + c_2 = 0, \\ c_1 + c_2\pi = 0. \end{cases}$$

  Poiché l'unica soluzione è quella banale, questo caso non è da prendere in considerazione.

- Supponendo $\lambda < -2$, per semplicità, poniamo $\lambda + 2 = -\beta^2$, $\beta > 0$. In tal caso la soluzione generale è

  $$X(x) = e^{-x}(c_1 e^{\beta x} + c_2 e^{-\beta x}).$$

  Poiché $X(0) = c_1 + c_2$, $X'(0) = (\beta - 1)c_1 - (\beta + 1)c_2$ e $X(\pi) = e^{-\pi}(c_1 e^{\beta\pi} + c_2 e^{-\beta\pi})$, le costanti c_1 e c_2 sono da determinare attraverso il sistema di equazioni lineari (rispetto a c_1 e c_2) e omogenee

  $$\begin{cases} (\beta + 1)c_1 + (1 - \beta)c_2 = 0, \\ c_1 e^{\beta\pi} + c_2 e^{-\beta\pi} = 0. \end{cases}$$

Il determinante di questo sistema è

$$\beta + 1 - (1 - \beta)e^{2\beta\pi}.$$

A noi interessano i valori di β per cui

$$\beta + 1 - (1 - \beta)e^{2\beta\pi} = 0.$$

Tenuto conto che $\beta > 0$, si può riscrivere l'ultima equazione (dividendo ambo i membri per $1 + \beta > 0$) come

$$\left(\frac{1-\beta}{1+\beta}\right) e^{2\beta\pi} = 1.$$

Le soluzioni, se esistono, sono date dall'intersezione fra la funzione $y = 1$ e la funzione $y = \left(\frac{1-\beta}{1+\beta}\right) e^{2\beta\pi}$. Occorre quindi rappresentare graficamente tali funzioni (Fig. 4.4).

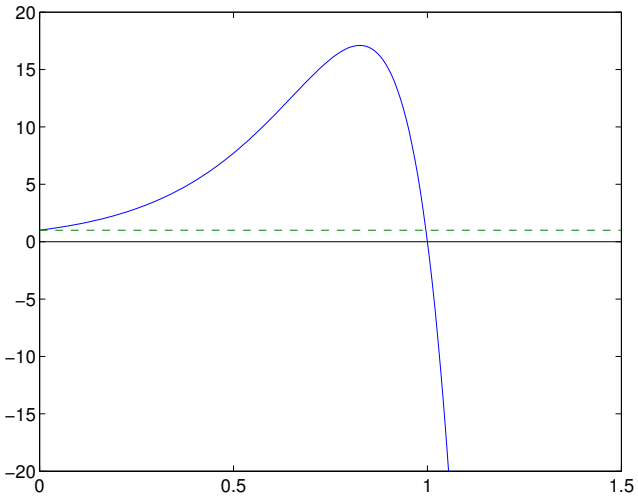


Figura 4.4: Grafici delle funzioni $y = \left(\frac{1-x}{1+x}\right) e^{2\pi x}$ e $y = 1$.

Poiché i due grafici si intersecano in un unico punto di ascissa $\tilde{\beta}$ con $0 < \tilde{\beta} < 1$, dall'equazione $\lambda + 2 = -\tilde{\beta}^2$ è possibile determinare il valore $\tilde{\lambda}$ di λ per cui il determinante si annulla. Quindi, in corrispondenza di $\tilde{\beta}$, si trova l'autofunzione

$$\tilde{X}(x) = e^{-x}(e^{\tilde{\beta}(2\pi - x)} + e^{\tilde{\beta}x}),$$

nella quale si è tenuto conto che $c_2 = -c_1 e^{2\tilde{\beta}\pi}$. In corrispondenza all'autovalore $\tilde{\beta}$ abbiamo inoltre l'equazione

$$\tilde{T}' = -\tilde{\beta}\tilde{T},$$

da cui si ricava $\tilde{T}(t) = e^{-\tilde{\beta}t}$. Otteniamo quindi la soluzione (del problema principale)

$$\tilde{v}(x,t) = e^{-\tilde{\beta}t-x}(e^{\tilde{\beta}(2\pi-x)} + e^{\tilde{\beta}x}).$$

- Per $\lambda > -2$, poniamo $\lambda + 2 = \beta^2$, $\beta > 0$. In tal caso abbiamo due radici complesse coniugate $\alpha_{1,2} = -1 \pm i\beta$ e la soluzione generale è

$$X(x) = e^{-x}\left(c_1 \cos \beta x + c_2 \sin \beta x)\right),$$

con le costanti c_1 e c_2 da determinare mediante il sistema

$$\begin{cases} c_1 + c_2\beta = 0, \\ c_1 \cos \beta\pi + c_2 \sin \beta\pi = 0. \end{cases}$$

Dalla prima equazione del sistema si ricava $c_1 = -c_2\beta$, che, sostituita nella seconda equazione, fornisce

$$c_2\left(-\beta \cos \beta\pi + \sin \beta\pi\right) = 0.$$

L'equazione è non banalmente soddisfatta quando

$$-\beta \cos \beta\pi + \sin \beta\pi = 0.$$

Osservato che $\cos \beta\pi \neq 0$, si perviene all'equazione

$$\operatorname{tg} \beta\pi = \beta,$$

che viene risolta per via grafica.

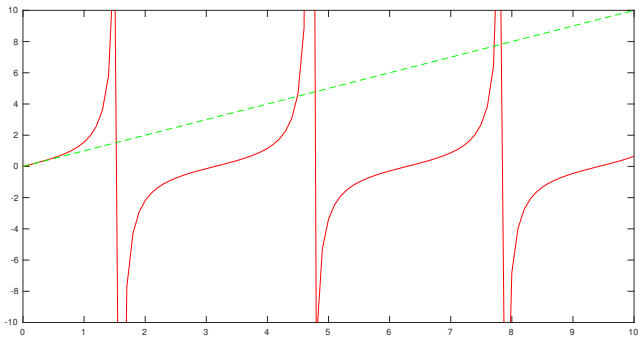


Figura 4.5: Grafici delle funzioni $y = \operatorname{tg} x$ e $y = (x/\pi)$.

I grafici delle funzioni $y = \operatorname{tg} \beta\pi$ e $y = \beta$ $(\beta > 0)$ si intersecano in un insieme numerabile di punti, in ciascuno dei quali il sistema di equazioni lineari possiede una soluzione non banale. Dalla Fig. 4.5 si vede che

le ascisse β_k dei punti d'intersezione soddisfano la seguente condizione $k < \beta_k < \frac{2k+1}{2}$, $k = 1, 2, \ldots$. Una volta noti questi valori di β_k è possibile determinare i corrispondenti valori di λ_k per cui il sistema possiede soluzioni non banali attraverso la relazione $\lambda_k = \beta_k^2 - 2$. Quindi, in corrispondenza ai valori $k < \beta_k < \frac{2k+1}{2}$, $k = 1, 2, \ldots$ del problema spettrale si trovano le seguenti autofunzioni

$$X_k(x) = e^{-x} \left(-\beta_k \cos \beta_k x + \sin \beta_k x\right).$$

Per ciascun valore di β_k si considera l'equazione $T_k' + \beta_k T = 0$, che ha la soluzione

$$T_k(t) = e^{-\beta_k t}.$$

In definitiva troviamo le autofunzioni

$$v_k(x,t) = T_k(t) X_k(x) = e^{-\beta_k t} e^{-x} \left(-\beta_k \cos \beta_k x + \sin \beta_k x\right).$$

Siamo quindi in grado di scrivere la soluzione del problema, che risulta

$$u(x,t) = \tilde{a}\, \tilde{v}(x,t) + \sum_{k=1}^{\infty} c_k v_k(x,t) = \tilde{a}\, e^{-\tilde{\beta} t - x} (e^{\tilde{\beta}(2\pi - x)} + e^{\tilde{\beta} x})$$
$$+ \sum_{k=1}^{\infty} c_k e^{-\beta_k t} e^{-x} \left(-\beta_k \cos \beta_k x + \sin \beta_k x\right),$$

con $\tilde{a}$ e c_k costanti da determinare. A tal fine imponiamo la condizione $u(x,0) = \sin \pi x - \phi(x)$. Questa implica

$$\tilde{a}\, e^{-x} (e^{\tilde{\beta}(2\pi - x)} + e^{\tilde{\beta} x}) + \sum_{k=1}^{\infty} c_k e^{-x} \left(-\beta_k \cos \beta_k x + \sin \beta_k x\right) = \sin \pi x - \phi(x).$$

In questa condizione compaiono $\tilde{u}(x,0)$ e le $u_k(x,0)$ che sono soluzioni della eequazione $X'' + 2X' + (\lambda + 3)X = 0$, la quale può scriversi nella forma canonica di Sturm-Liouville

$$\frac{d}{dx}(e^{2x} X') + e^{2x}(3 + \lambda) X = 0.$$

La funzione $r(x) = e^{2x} > 0$ non cambia segno e questo garantisce che le funzioni $\tilde{u}(x,0)$ e $u_k(x,0)$ $(k = 1, 2, 3 \ldots)$ siano ortogonali tra loro rispetto ad essa in $[0, \pi]$, e questo determina le seguenti espressioni dei coefficienti

$$\tilde{a} = \frac{\int_0^{\pi} e^{2x} (\sin \pi x - \phi(x)) \tilde{u}(x,t)\, dx}{\int_0^{\pi} e^{2x} \tilde{u}(x,t)^2\, dx},$$
$$c_k = \frac{\int_0^{\pi} e^{2x} (\sin \pi x - \phi(x)) u_k(x,t)\, dx}{\int_0^{\pi} e^{2x} u_k(x,t)^2\, dx}, \quad k = 1, 2, \ldots.$$

4.2 c Equazioni iperboliche

Illustriamo ora la risoluzione di alcune PDEs iperboliche con condizioni di vario tipo.

11. Equazione delle onde con spostamento iniziale nullo

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} = 4\dfrac{\partial^2 u}{\partial x^2}, \quad 0 \le x \le 5,\ t \ge 0, \\ u(0,t) = u(5,t) = 0, \\ u(x,0) = 0, \quad u_t(x,0) = 2\sin \pi x - \sin 4\pi x. \end{cases} \tag{4.35}$$

Si tratta di un problema iperbolico in una dimensione spaziale (equazione delle onde descrivente una corda vibrante) con condizioni iniziali e alla frontiera che garantiscono l'unicità della soluzione. Cerchiamo ora tale soluzione con il metodo di separazione delle variabili:

$$u(x,t) = X(x)T(t), \qquad X \text{ e } T \text{ differenziabili.}$$

Sostituendo e riordinando si ha:

$$XT'' = 4X''T \quad \Longrightarrow \frac{T''}{4T} = \frac{X''}{X} = -\lambda \quad \Longrightarrow \begin{cases} X''(x) + \lambda X(x) = 0, \\ T''(t) + 4\lambda T(t) = 0. \end{cases}$$

Applicando le condizioni agli estremi della corda $u(0,t) = u(5,t) = 0$, si trova il problema di Sturm-Liouville per la $X(x)$:

$$\begin{cases} X'' + \lambda X = 0, \\ X(0) = 0, \quad X(5) = 0, \end{cases} \tag{4.36}$$

con $x \in [0,5]$.

Tale problema ammette diversi casi a seconda del segno dell'autovalore λ ma, in analogia a ciò visto negli esempi precedenti, sappiamo che l'unico caso accettabile è quello con soluzioni oscillanti ($\lambda > 0$) del tipo

$$X(x) = a\,\cos(x\sqrt{\lambda}) + b\,\sin(x\sqrt{\lambda}), \ \begin{cases} X(0) = 0, \\ X(5) = 0, \end{cases} \Rightarrow \begin{cases} \lambda_k = \left(\dfrac{k\pi}{5}\right)^2, \\ X_k(x) = \sin\left(\dfrac{k\pi x}{5}\right), \end{cases}$$

con $k \in \mathbb{N}$.

Di conseguenza il problema differenziale per la $T_k(t)$ diventa:

$$\begin{cases} T_k''(t) + 4\lambda_k T_k(t) = 0, \\ T_k(0) = 0, \end{cases} \tag{4.37}$$

che ha come soluzione

$$T_k(t) = \sin\left(\frac{2k\pi t}{5}t\right), \quad k \in \mathbb{N} \setminus \{0\}.$$

Pertanto

$$u_k(x,t) = \sin\left(\frac{k\pi}{5}\right)\sin\left(\frac{2k\pi t}{5}\right),$$

e

$$u(x,t) = \sum_{k=1}^{\infty} a_k \sin\left(\frac{k\pi x}{5}\right)\sin\left(\frac{2k\pi t}{5}\right). \tag{4.38}$$

Possiamo ora applicare la condizione $u_t(x,0) = 2\sin\pi x - \sin 4\pi x$, sfruttando il fatto che la serie (4.38) è uniformemente convergente rispetto alla variabile t e che quindi è possibile permutare l'operatore di serie con quello di derivata parziale rispetto al tempo. Di conseguenza

$$u_t(x,t) = \sum_{k=1}^{\infty} a_k \frac{2k\pi}{5}\sin\left(\frac{k\pi x}{5}\right)\cos\left(\frac{2k\pi t}{5}\right),$$

da cui

$$\sum_{k=1}^{\infty} a_k \frac{2k\pi}{5}\sin\left(\frac{k\pi x}{5}\right) = 2\sin\pi x - \sin 4\pi x. \tag{4.39}$$

Dalla (4.39) si capisce chiaramente che sono diversi da zero soltanto i coefficienti a_5 e a_{20}. Da tale osservazione segue che

$$\begin{cases} a_5\frac{10\pi}{5} = 2, \\ a_{20}\frac{40\pi}{5} = -1, \end{cases} \Longrightarrow \begin{cases} a_5 = \frac{1}{\pi}, \\ a_{20} = -\frac{1}{8\pi}. \end{cases}$$

La soluzione del problema (4.35) risulta dunque

$$u(x,t) = \tfrac{1}{\pi}\sin(\pi x)\sin(2\pi t) - \tfrac{1}{8\pi}\sin(4\pi x)\sin(8\pi t). \tag{4.40}$$

Facendo uso delle formule di prostaferesi la (4.40) può essere scritta nel seguente modo

$$u(x,t) = \frac{\cos\pi(x-2t) - \cos\pi(x+2t)}{2\pi} - \frac{\cos 4\pi(x-2t) - \cos 4\pi(x+2t)}{16\pi},$$

nel quale viene evidenziato che nella soluzione della equazione delle onde, le variabili spazio e tempo compaiono sempre come $x \pm ct$ essendo c la velocità di propagazione dell'onda nel mezzo considerato.

12. Equazione delle onde con velocità iniziale nulla

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} = a^2 \dfrac{\partial^2 u}{\partial x^2}, \quad 0 \le x \le L, \quad t \ge 0, \\ u(0,t) = u(L,t) = 0, \\ u(x,0) = f(x), \\ u_t(x,0) = 0, \end{cases} \tag{4.41}$$

dove a è una costante reale positiva.

Il problema ammette una e una sola soluzione che possiamo ottenere con il metodo della separazione delle variabili.

La procedura è analoga a quella vista negli esempi precedenti, per cui, sorvolando sui dettagli, possiamo scrivere:

$$u(x,t) = X(x)T(t)$$

$$\begin{cases} X'' + \lambda X = 0, \\ X(0) = X(L) = 0, \end{cases} \implies \begin{cases} \lambda_n = \dfrac{n^2\pi^2}{L^2}, \qquad n = 1, 2, \ldots, \\ X_n = b_n \sin \dfrac{n\pi}{L}. \end{cases}$$

$$\begin{cases} T'' + \lambda a^2 T = 0, \\ T(0) = 0, \end{cases} \implies T_n = a_n \cos \frac{na\pi}{L}.$$

$$\begin{aligned} u(x,t) &= \sum_{n=1}^{\infty} c_n \sin \frac{n\pi x}{L} \cos \frac{na\pi t}{L} \\ &= \sum_{n=1}^{\infty} \frac{c_n}{2} \left[\sin \frac{n\pi}{L}(x - at) + \sin \frac{n\pi}{L}(x + at) \right]. \end{aligned}$$

Dalla condizione $u(x,0) = \sum_{n=1}^{\infty} c_n \sin \frac{n\pi x}{L} = f(x)$ si ricava infine

$$c_n = \frac{2}{L} \int_0^L f(x) \sin \frac{n\pi x}{L}\, dx.$$

13. Equazione iperbolica con condizioni non nulle agli estremi

$$\begin{cases} u_{tt} = 3\, u_{xx} + u_x + 2x, \quad \text{con } 0 \le x \le 1 \text{ e } t \ge 0, \\ u(0,t) = u(1,t) = 1, \\ u(x,0) = x(1-x), \quad u_t(x,0) = 0. \end{cases}$$

Poiché le condizioni per $x = 0$ e $x = 1$ non sono di tipo omogeneo, si cerca una soluzione del tipo $u(x,t) = v(x,t) + \phi(x)$. Sostituendo nell'equazione si trova

$$v_{tt} = 3\,(v_{xx} + \phi'') + v_x + \phi' + 2x.$$

Si impone quindi che siano soddisfatti i problemi differenziali

$$\begin{cases} v_{tt} = 3\,v_{xx} + v_x, \\ v(0,t) = v(1,t) = 0, \\ v(x,0) = x(1-x) - \phi(x), \quad v_t(x,0) = 0, \end{cases} \tag{4.42}$$

$$\begin{cases} 3\,\phi'' + \phi' + 2x = 0, \\ \phi(0) = \phi(1) = 1. \end{cases} \tag{4.43}$$

La soluzione generale dell'equazione (4.43) è

$$\phi(x) = c_1 + c_2 e^{-\frac{1}{3}x} + x(-x+6).$$

Imponendo le condizioni agli estremi si trova che c_1 e c_2 devono soddisfare il sistema

$$\begin{cases} c_1 + c_2 = 1, \\ c_1 + e^{-\frac{1}{3}} c_2 = -4. \end{cases}$$

Indicata con $(\tilde{c}_1, \tilde{c}_2)$ la soluzione di tale sistema, la soluzione del problema (4.43) è

$$\phi(x) = \tilde{c}_1 + \tilde{c}_2 e^{-\frac{1}{3}x} + x(-x+6). \tag{4.44}$$

Passiamo ora alla risoluzione del problema (4.42). Procedendo con il metodo di separazione delle variabili, Cerchiamo soluzioni del tipo $v(x,t) = X(x)T(t)$. Sostituendo si perviene ai seguenti problemi:

$$\begin{cases} 3X'' + X' + \lambda X = 0, \\ X(0) = X(1) = 0, \end{cases} \tag{4.45}$$

$$\begin{cases} T'' + \lambda T = 0, \\ T'(0) = 0. \end{cases} \tag{4.46}$$

Iniziamo a risolvere il problema di Sturm Liouville (4.45). Cerchiamo soluzioni del tipo $X(x) \simeq e^{\alpha x}$. Si ottiene così l'equazione caratteristica $3\alpha^2 + \alpha + \lambda = 0$ che ammette come soluzioni

$$\alpha_{1,2} = \frac{-1 \pm \sqrt{1 - 12\lambda}}{6}.$$

- $\lambda = \frac{1}{12}$. In tal caso l'equazione caratteristica possiede due radici reali coincidenti $\alpha_1 = \alpha_2 = -\frac{1}{6}$. La soluzione dell'equazione è $X(x) = e^{-\frac{1}{6}x}(a + bx)$, con a e b costanti da determinare. Essendo, $X(0) = a$ e $X(1) = e^{-\frac{1}{6}}(a + b)$, a e b debbono soddisfare il sistema di equazioni lineari e omogenee
$$\begin{cases} a = 0, \\ a + b = 0. \end{cases}$$
Quest'ultimo ammette come unica soluzione quella banale $a = b = 0$.

- $\lambda < \frac{1}{12}$. In tal caso l'equazione caratteristica possiede due radici reali e distinte $\alpha_{1,2} = \frac{1}{6}[-1 \pm \sqrt{1 - 12\lambda}\,]$. La soluzione dell'equazione è
$$X(x) = e^{-\frac{1}{6}x}\left(a\, e^{-\mu x} + b\, e^{\mu x}\right),$$
dove si è posto $\mu = \frac{1}{6}\sqrt{1 - 12\lambda} > 0$, con a e b da determinare mediante il sistema di equazioni lineari e omogenee
$$\begin{cases} a + b = 0, \\ e^{-\mu}\, a + e^{\mu}\, b = 0. \end{cases}$$
Il determinante di questo sistema di equazioni lineari è $\det = e^{-\mu} - e^{\mu}$ che si annula solo se $\mu = 0$. Poiché $\mu > 0$, il sistema ammette come unica soluzione quella banale $a = b = 0$.

- $\lambda > \frac{1}{12}$. In tal caso l'equazione caratteristica possiede due radici reali e distinte $\alpha_{1,2} = \frac{1}{6}[-1 \pm i\sqrt{-1 + 12\lambda}\,]$. La soluzione dell'equazione è
$$X(x) = e^{-\frac{1}{6}x}\left(a\, \cos \mu x + b\, \sin \mu x\right),$$
dove si è posto $\mu = \frac{1}{6}\sqrt{-1 + 12\lambda} > 0$, con a e b costanti da determinare dal sistema di equazioni lineari e omogenee
$$\begin{cases} a = 0, \\ a\, \cos \mu + b\, \sin \mu = 0. \end{cases}$$
Le soluzioni non banali di questo sistema si ricavano dall'equazione $\sin \mu = 0$. Si trova $\frac{1}{6}\sqrt{-1 + 12\lambda_k} = k\pi$, $k = 1, 2, \ldots$. In corrispondenza agli autovalori λ_k si hanno le autofunzioni $X_k(x) = e^{-\frac{1}{6}x}\, \sin k\pi x$, $k = 1, 2, \ldots$. In corrispondenza degli autovalori sopra trovati occorre risolvere
$$\begin{cases} T_k'' + \lambda_k T_k = 0, \\ T_k'(0) = 0. \end{cases}$$

Questa ammette come soluzione

$$T_k(t) = a_k \cos(t\sqrt{\lambda_k}) + b_k \sin(t\sqrt{\lambda_k}).$$

Poiché $T_k'(0) = b_k\sqrt{\lambda_k} = 0$, ossia $b_k = 0$, si ottiene $T_k(t) = a_k \cos(t\sqrt{\lambda_k})$. Quindi, relativamente agli autovalori $\frac{1}{6}\sqrt{-1+12\lambda_k} = k\pi,\ k = 1, 2, \ldots$, si trovano le seguenti funzioni $v_k(x,t) = e^{-\frac{1}{6}x} \sin(k\pi x) \cos(t\sqrt{\lambda_k})$.

Rimane da imporre la condizione $v(x,0) = x(1-x) - \phi(x)$ con $\phi(x)$ definita dalla (4.44). A tal fine osserviamo che

$$v(x,t) = \sum_{k=1}^{\infty} c_k e^{-\frac{1}{6}x} \sin(k\pi x) \cos(t\sqrt{\lambda_k}),$$

con $\sum_{k=1}^{\infty} c_k e^{-\frac{1}{6}x} \sin(k\pi x) = x(1-x) - \phi(x)$. Questa equazione può essere riscritta come

$$\sum_{k=1}^{\infty} c_k \sin k\pi x = e^{\frac{1}{6}x} \left\{x(1-x) - \phi(x)\right\},$$

dalla quale è possibile ricavare i coefficienti c_k attaverso l'analisi di Fourier (a primo membro compare infatti una serie di Fourier). Infatti, moltiplicando primo e secondo membro per $\sin(h\pi x)$ e integrando (tenendo conto dell'ortogonalità delle funzioni $\sin(k\pi x)\ \sin(h\pi x)$) si trova

$$c_k = 2\int_0^1 e^{\frac{1}{6}x}(x(1-x) - \phi(x)) \sin k\pi x, \quad k = 1, 2, \ldots. \tag{4.47}$$

Finalmente si può scrivere la soluzione del problema come

$$v(x,t) = \sum_{k=1}^{\infty} c_k e^{-\frac{1}{6}x} \sin(k\pi x) \cos(t\sqrt{\lambda_k}) + \tilde{c}_1 + \tilde{c}_2 e^{-\frac{1}{3}x} + x(-x+6),$$

con i coefficienti c_k dati dalle equazioni (4.47).

4.2 d PDEs in tre variabili

Vediamo ora come si estende al caso tridimensionale la tecnica di separazione delle variabili illustrata nel caso bidimensionale.

14. Equazione delle onde sul rettangolo

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} = a^2 \left(\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} \right), \quad 0 \le x \le L,\ 0 \le y \le M,\ t \ge 0, \\ u(x,0,t) = u(x,M,t) = 0, \\ u(0,y,t) = u(L,y,t) = 0, \\ u(x,y,0) = f(x,y), \\ u_t(x,y,0) = 0. \end{cases} \tag{4.48}$$

Il problema (4.48) rappresenta la propagazione delle onde su una membrana elastica di forma rettangolare (lunghezza dei lati M e L) con lati bloccati, posizione iniziale data dalla superficie regolare $f(x,y) \in C([0,L]\times[0,M])$ e con velocità iniziale nulla.

Vista la regolarità del dominio, è possibile applicare il metodo di separazione delle variabili

$$u(x,y,t) = X(x)Y(y)T(t).$$

Tenuto conto delle condizioni al contorno e, in particolare, della condizione $u(0,y,t) = u(L,y,t) = 0$, otteniamo $X(0) = X(L) = 0$. Iniziamo ora con il separare la variabile x e t, operando nel modo seguente:

$$\frac{T''}{a^2T} - \frac{Y''}{Y} = \frac{X''}{X} = -\lambda,$$

da cui

$$\frac{T''}{a^2T} - \frac{Y''}{Y} = -\lambda \quad \text{e} \quad X'' + \lambda X = 0.$$

La seconda equazione, tenendo conto della condizione $u(0,y,t) = u(L,y,t) = 0$, genera il problema spettrale classico

$$\begin{cases} X'' + \lambda X = 0, \\ X(0) = X(L) = 0. \end{cases}$$

Dalla prima equazione è ora possibile separare le variabili y e t, ponendo

$$\frac{T''}{a^2T} + \lambda = \frac{Y''}{Y} = -\mu,$$

dove μ è un secondo parametro spettrale. Da questa, ricordando che $u(x,0,t) = u(x,M,t) = 0$, otteniamo il problema spettrale

$$\begin{cases} Y'' + \mu Y = 0, \\ Y(0) = Y(M) = 0, \end{cases}$$

e l'equazione di raccordo

$$T'' + (\lambda + \mu)a^2T = 0.$$

Dal problema di Sturm-Liouville in X segue che

$$\lambda_n = \left(\frac{n\pi}{L}\right)^2 \quad \text{e} \quad X_n(x) = \sin\left(\frac{n\pi x}{L}\right), \quad n = 1,2,\ldots.$$

Analogamente, dal problema di Sturm-Liouville in y segue che

$$\mu_n = \frac{m^2\pi^2}{M^2} \quad \text{e} \quad Y_m(y) = \sin\frac{m\pi y}{M}, \quad m = 1,2,\ldots.$$

Di conseguenza resta da risolvere l'equazione

$$T''_{n,m} + \left(\frac{n^2}{L^2} + \frac{m^2}{M^2}\right) a^2 T_{n,m} = 0.$$

Da essa segue che

$$T_{n,m}(t) = a_{n,m} \cos\left(\pi a t \sqrt{\frac{n^2}{L^2} + \frac{m^2}{M^2}}\right) + b_{n,m} \sin\left(\pi a t \sqrt{\frac{n^2}{L^2} + \frac{m^2}{M^2}}\right),$$

da cui, ricordando che $u_t(x, y, 0) = 0$ segue che $b_{n,m} = 0$, $n, m = 1, 2, \ldots$.

Pertanto avendo ottenuto, a meno di fattori moltiplicativi arbitrari,

$$\begin{cases} X_n(x) = \sin \frac{n\pi x}{L}, \\ Y_m(y) = \sin \frac{m\pi y}{M}, \\ T_{n,m}(t) = \cos\left(\pi a t \sqrt{\frac{n^2}{L^2} + \frac{m^2}{M^2}}\right), \end{cases}$$

si ha che, a meno di un fattore moltiplicativo arbitrario,

$$u_{n,m}(x, y, t) = \sin \frac{n\pi x}{L} \sin \frac{m\pi y}{M} \cos\left(\pi a t \sqrt{\frac{n^2}{L^2} + \frac{m^2}{M^2}}\right),$$

e dunque

$$u(x, y, t) = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} c_{n,m} \sin \frac{n\pi x}{L} \sin \frac{m\pi y}{M} \cos\left(\pi a t \sqrt{\frac{n^2}{L^2} + \frac{m^2}{M^2}}\right),$$

da cui, per l'ortogonalità delle autofunzioni $\sin \frac{n\pi x}{L}$ in $[0, L]$ e delle autofunzioni $\sin \frac{m\pi y}{M}$ in $[0, M]$,

$$c_{n,m} = \frac{4}{LM} \int_0^M \int_0^L f(x, y) \sin \frac{n\pi x}{L} \sin \frac{m\pi y}{M} \, dxdy.$$

Come già osservato in casi simili, la tecnica è analoga se $u(x, y, 0) = 0$ e $u_t(x, y, 0) = g(x, y)$.

Inoltre, se $u(x, y, 0) = f(x, y)$ e $u_t(x, y, 0) = g(x, y)$, per la linearità del problema omogeneo a parità delle altre condizioni, si risolvono separatamente i due casi con $u(x, y, 0) = 0$, $u_t(x, y, 0) = g(x, y)$, $u(x, y, 0) = f(x, y)$ e $u_t(x, y, 0) = 0$, rispettivamente, e si assume come $u(x, y)$ la somma delle due soluzioni così ottenute.

15. Equazione iperbolica con posizione iniziale nulla

$$\begin{cases} u_{tt} = 3u_{xx} + 2u_{yy} + 4u_x + 3u_y, \quad 0 \le x, y \le \pi,\ t \ge 0, \\ u(0, y, t) = u(\pi, y, t) = 0, \\ u(x, 0, t) = u(x, \pi, t) = 0, \\ u(x, y, 0) = 0, \\ u_t(x, y, 0) = \sin(xy). \end{cases}$$

Si cercano soluzioni del tipo $u(x, y, t) = X(x)Y(y)T(t)$. Sostituendo nell'equazione, si trova $XYT'' = 3X''YT + 2XY''T + 4X'YT + 3XY'T$, da cui

$$\frac{T''}{T} = 3\frac{X''}{X} + 2\frac{Y''}{Y} + 4\frac{X'}{X} + 3\frac{Y'}{Y}.$$

Potendo scrivere l'ultima equazione come

$$3\frac{X''}{X} + 4\frac{X'}{X} = \frac{T''}{T} - 2\frac{Y''}{Y} - 3\frac{Y'}{Y} = -\lambda,$$

si perviene al primo problema spettrale

$$\begin{cases} 3X'' + 4X' + \lambda X = 0, \\ X(0) = 0, \\ X(\pi) = 0. \end{cases} \tag{4.49}$$

Inoltre, poiché il rapporto $\dfrac{T''}{T}$ non dipende da y, introducendo un secondo parametro spettrale, si può scrivere

$$2\frac{Y''}{Y} + 3\frac{Y'}{Y} = \frac{T''}{T} + \lambda = -\mu,$$

da cui si ottengono i problemi differenziali

$$\begin{cases} 2Y'' + 3Y' + \mu Y = 0, \\ Y(0) = 0, \\ Y(\pi) = 0, \end{cases} \tag{4.50}$$

$$\begin{cases} T'' + (\lambda + \mu)T = 0, \\ T(0) = 0. \end{cases} \tag{4.51}$$

Discutiamo brevemente la soluzione di ciascuno di questi tre problemi, iniziando con il problema (4.49). Si cercano soluzioni del tipo $X(x) \simeq e^{\alpha x}$ che conducono all'equazione caratteristica

$$3\alpha^2 + 4\alpha + \lambda = 0,$$

la quale ha come soluzioni $\alpha_{1,2} = \frac{-2 \pm \sqrt{4 - 3\lambda}}{3}$.

- $\lambda = \frac{4}{3}$. In questo caso abbiamo due radici reali e coincidenti $\alpha_1 = \alpha_2 = -\frac{2}{3}$. La soluzione generale è $X(x) = e^{-\frac{2}{3}x}(c_1 + c_2 x)$. Poiché $X(0) = c_1$ e $X(\pi) = e^{-\frac{2}{3}\pi}(c_1 + c_2\pi)$, le costanti c_1 e c_2 devono soddisfare il sistema

$$\begin{cases} c_1 = 0, \\ c_1 + c_2\pi = 0, \end{cases}$$

che ammette come unica soluzione quella banale.

- Nel caso $\lambda < \frac{4}{3}$ abbiamo due radici reali e distinte $\alpha_{1,2} = \frac{-2}{3} \pm \beta$, $\beta = \frac{\sqrt{4-3\lambda}}{3} > 0$, per cui la soluzione generale è

$$X(x) = e^{-\frac{2}{3}x}(c_1 e^{\beta x} + c_2 e^{-\beta x}).$$

Poiché $X(0) = c_1 + c_2$ e $X(\pi) = e^{-\frac{2}{3}\pi}(c_1 e^{\beta\pi} + c_2 e^{-\beta\pi})$, le costanti c_1 e c_2 sono da determinare mediante il sistema

$$\begin{cases} c_1 + c_2 = 0, \\ c_1 e^{\beta\pi} + c_2 e^{-\beta\pi} = 0, \end{cases}$$

che possiede, come unica soluzione, quella banale.

- Sia ora $\lambda > \frac{4}{3}$. In tal caso abbiamo la soluzione generale

$$X(x) = e^{-\frac{2}{3}x}\left(c_1 \cos\beta x + c_2 \sin\beta x)\right), \quad \beta = \frac{\sqrt{3\lambda - 4}}{3} > 0.$$

Poiché $X(0) = c_1$ e $X(\pi) = e^{-\frac{2}{3}\pi}(c_1 \cos\beta\pi + c_2 \sin\beta\pi)$, le costanti c_1 e c_2 sono da determinare mediante il sistema

$$\begin{cases} c_1 = 0, \\ c_1 \cos\beta\pi + c_2 \sin\beta\pi = 0. \end{cases}$$

L'equazione è non banalmente soddisfatta quando

$$\sin\beta\pi = 0.$$

Questa equazione ammette come soluzioni $\beta_k = k$, $k = 1, 2, \ldots$, cui corrisponde $\lambda_k = \frac{4}{3} + k^2$. Lo spettro per il problema (4.49) è pertanto

$$\left\{\lambda_k = \tfrac{4}{3} + k^2,\ X_k(x) = e^{-\frac{2}{3}x}\sin kx,\ k = 1, 2, \ldots\right\}.$$

Discutiamo ora il problema (4.50). Si cercano soluzioni del tipo $Y(y) \simeq e^{\alpha y}$ che conducono all'equazione caratteristica

$$2\alpha^2 + 3\alpha + \mu = 0.$$

Questa equazione ha come soluzioni $\alpha_{1,2} = \frac{1}{4}[-3 \pm \sqrt{9 - 8\mu}]$. Occorre distinguere tre casi a seconda che sia $9 - 8\mu = 0$, $9 - 8\mu > 0$, oppure $9 - 8\mu < 0$.

- $\mu = \frac{9}{8}$. In tal caso la soluzione generale è $Y(y) = e^{-\frac{3}{4}y}(c_1 + c_2 y)$. Poiché $Y(0) = c_1$ e $Y(\pi) = e^{-\frac{3}{4}\pi}(c_1 + c_2\pi)$, le costanti c_1 e c_2 devono soddisfare il sistema
$$\begin{cases} c_1 = 0, \\ c_1 + c_2\pi = 0, \end{cases}$$
che ammette come unica soluzione quella banale. In maniera analoga si trova che anche il caso $9 - 8\mu > 0$ è da scartare in quanto conduce alla soluzione $Y(y) \equiv 0$.

- Sia ora $9 - 8\mu < 0$. In tal caso abbiamo due radici complesse coniugate $\alpha_{1,2} = \frac{-3}{4} \pm i\beta$, $\beta = \frac{\sqrt{8\mu - 9}}{4}$, per cui la soluzione generale è
$$Y(y) = e^{-\frac{3}{4}y}\left(c_1 \cos\beta y + c_2 \sin\beta y\right).$$
Poiché $Y(0) = c_1$ e $Y(\pi) = e^{-\frac{3}{4}\pi}(c_1 \cos\beta\pi + c_2 \sin\beta\pi)$, le costanti c_1 e c_2 sono da determinare mediante il sistema
$$\begin{cases} c_1 = 0, \\ c_1 \cos\beta\pi + c_2 \sin\beta\pi = 0. \end{cases}$$
Il sistema è non banalmente soddisfatto soltanto quando
$$\sin\beta\pi = 0.$$
Questa equazione ammette come soluzioni $\beta_k = k, . = 1, 2, \ldots$ e quindi $\mu_n = \frac{9}{8} + 2n^2$. Lo spettro per il problema (4.49) è pertanto
$$\left\{\mu_n = \tfrac{9}{8} + \tfrac{1}{2}n^2,\ Y_n(y) = e^{-\frac{3}{4}y}\sin(ny),\ n = 1, 2, \ldots\right\}.$$

Resta da risolvere l'associato problema (4.51). Esso è del tipo

$$\begin{cases} T''_{nk} + (\lambda_k + \mu_n)T_{nk} = 0, \\ T_{nk}(0) = 0, \quad k, n = 1, 2, \ldots. \end{cases}$$

Poiché $\lambda_k + \mu_n > 0$, $T_{nk} = a_{nk}\cos(t\sqrt{\lambda_k+\mu_n}) + b_{nk}\sin(t\sqrt{\lambda_k+\mu_n})$ e dalla condizione $T_{nk}(0) = 0$ si ricava $a_{nk} = 0$. Lo spettro è quindi

$$\left\{\mu_n + \lambda_k,\ T_{nk}(t) = \sin(t\sqrt{\lambda_k+\mu_n}),\ n,k = 1,2,\ldots\right\}.$$

Pertanto $u_{kn}(x,y,t) = (e^{-\frac{2}{3}x}\sin kx)(e^{-\frac{3}{4}y}\sin ny)\ \sin(t\sqrt{\lambda_k+\mu_n})$ e la soluzione del problema iniziale è

$$u(x,y,t) = \sum_{k,n=1}^{\infty} c_{kn}\, e^{-(\frac{2}{3}x+\frac{3}{4}y)}\sin kx\ \sin ny\ \sin(t\sqrt{\lambda_k+\mu_n}).$$

Imponendo la condizione $u_t(x,y,0) = \sin xy$ si trova

$$\sum_{k,n=1}^{\infty} \sqrt{\lambda_k+\mu_n}\, c_{kn}\, e^{-(\frac{2}{3}x+\frac{3}{4}y)}\sin kx\ \sin ny = \sin xy.$$

Sfruttando l'ortogonalità delle $X_k(x)$ in $[0,\pi]$ rispetto al peso $e^{\frac{4}{3}x}$, si ottiene la relazione

$$\sum_{n=1}^{\infty} \sqrt{\lambda_k+\mu_n}\, c_{kn}\left(\int_0^\pi \sin^2 kx\,dx\right) e^{-\frac{3}{4}y}\sin ny = \int_0^\pi e^{\frac{2}{3}x}\sin kx\ \sin(xy)\,dx.$$

Infine imponendo l'ortogonalità delle $Y_n(y)$ rispetto al peso $e^{\frac{3}{2}y}$, ugualmente in $[0,\pi]$, si trova

$$\begin{aligned}\sqrt{\lambda_k+\mu_n}\, c_{kn}&\left(\int_0^\pi \sin^2 kx\,dx\right)\left(\int_0^\pi \sin^2 ny\,dy\right)\\ &= \int_0^\pi e^{\frac{3}{4}x}\sin ny\int_0^\pi e^{\frac{2}{3}x}\sin kx\ \sin(xy)\,dxdy.\end{aligned}$$

Dall'ultima equazione ricaviamo quindi la relazione

$$\frac{1}{4}\sqrt{\lambda_k+\mu_n}\, c_{kn} = \int_0^\pi e^{\frac{3}{4}x}\sin ny\int_0^\pi e^{\frac{2}{3}x}\sin kx\ \sin(xy)\,dxdy$$

per $k,n = 1,2,\ldots$, la quale identifica, per ogni coppia (k,n), il valore del coefficiente c_{kn}.

Osservazione. Come abbiamo visto negli esempi precedenti, l'applicazione del metodo di separazione delle variabili è notevolmente semplificata nel caso in cui in tutte le condizioni al bordo e iniziali, tranne una, siano funzioni identicamente nulle. Questa tuttavia, almeno per le equazioni differenziali omogenee, non rappresenta una oggettiva difficoltà in quanto il problema iniziale può sempre essere ricondotto alla risoluzione di uno o più problemi che presentano le caratteristiche desiderate. L'affermazione è una diretta conseguenza della seguente proprietà di linearità, immediatamente verificabile.

Proposizione 4.1 *Se $u_1(x,y)$ e $u_2(x,y)$ sono soluzioni di una PDE lineare, lo è anche una qualunque loro combinazione lineare.*

Per semplicità, limitiamoci a considerare l'equazione

$$u_{tt} = A(x,t)u_{xx} + B(x,t)u_x + C(x,t)u_t + D(x,t)u.$$

Dobbiamo verificare che se $u_1(x,t)$ e $u_2(x,t)$ sono soluzioni, lo è anche $u(x,t) = c_1u_1(x,t) + c_2u_2(x,t)$, qualunque siano le costanti c_1 e c_2.

Per la verifica, basta osservare che

$$\begin{aligned} c_1(u_1)_{tt} + c_2(u_2)_{tt} = A(x,t)\left[c_1(u_1)_{xx} + c_2(u_2)_{xx}\right] + B(x,t)\left[c_1(u_1)_x + c_2(u_2)_x\right] \\ + C(x,t)\left[c_1(u_1)_t + c_2(u_2)_t\right] + D(x,t)\left[c_1u_1 + c_2u_2\right] \end{aligned}$$

in quanto, per ipotesi,

$$\begin{aligned} c_1(u_1)_{tt} &= c_1\left[A(x,t)(u_1)_{xx} + B(x,t)(u_1)_x + C(x,t)(u_1)_t + D(x,t)u_1\right], \\ c_2(u_2)_{tt} &= c_2\left[A(x,t)(u_2)_{xx} + B(x,t)(u_2)_x + C(x,t)(u_2)_t + D(x,t)u_2\right]. \end{aligned}$$

Supponiamo ora di dover risolvere una PDE a coefficienti costanti del tipo:

$$\begin{cases} u_{tt} = Au_{xx} + Bu_x + Cu_t + Du, & \text{con } a \le x \le b \text{ e } t \ge 0, \\ u(a,t) = f_1(t), \quad u(b,t) = f_2(t), \\ u(x,0) = g_1(x), \quad u_t(x,0) = g_2(x). \end{cases}$$

Supponiamo inoltre di saper risolvere la stessa equazione differenziale in ciascuna delle seguenti situazioni

(1) $f_1(t) \not\equiv 0$, $f_2(t) \equiv 0$, $g_1(x) \equiv 0$, $g_2(x) \equiv 0$;

(2) $f_1(t) \equiv 0$, $f_2(t) \not\equiv 0$, $g_1(x) \equiv 0$, $g_2(x) \equiv 0$;

(3) $f_1(t) \equiv 0$, $f_2(t) \equiv 0$, $g_1(x) \not\equiv 0$, $g_2(x) \equiv 0$;

(4) $f_1(t) \equiv 0$, $f_2(t) \equiv 0$, $g_1(x) \equiv 0$, $g_2(x) \not\equiv 0$.

Indichiamo con $u_i(x,t)$, $i = 1,2,3,4$ rispettivamente le soluzioni dell'equazione differenziale con assegnate le condizioni (1), (2), (3) e (4). È allora immediato verificare che

$$u(x,t) = u_1(x,t) + u_2(x,t) + u_3(x,t) + u_4(x,t).$$

A tale scopo basta osservare che:

(a) in conseguenza della linearità dell'equazione differenziale, $u(x,t)$ è una soluzione;

(b) $u(x,t)$ soddisfa sia le condizioni al bordo sia quelle iniziali, dato che su ciascuna di esse tre soluzioni sono nulle e la quarta assume i valori desiderati.

4.3 Esercizi proposti

(a) Risolvere con il metodo degli integrali generali i seguenti problemi differenziali:

$$\begin{cases} u_{tt} = c^2 u_{xx}, \quad \text{con } 0 \le x \le \pi \text{ e } t \ge 0, \\ u(0,t) + u(\pi,t) = 0, \\ u(x,0) = \sin x, \quad u_t(x,0) = \cos x. \end{cases}$$

$$\begin{cases} u_{xy} = x + y\cos x, \\ u(x,t) = 3x + 2, \\ u(0,y) = 2. \end{cases}$$

$$\begin{cases} u_{tt} = 3u_{xx} + 3u_{xt}, \\ u(x,0) = x, \\ u_t(x,0) = 1. \end{cases}$$

(b) Risolvere mediante separazione delle variabili i seguenti problemi differenziali:

$$\begin{cases} u_{tt} = 2u_{xx} + 3u_x + 4u_t + 5u, \qquad \text{con } 0 \le x \le 2 \text{ e } t \ge 0, \\ u(x,0) = 1 + \frac{x}{2}, \quad u_t(x,0) = 0, \\ u(0,t) = 1, \quad u(2,t) = 2. \end{cases}$$

$$\begin{cases} u_t = 4u_{xx} + 6u_x - 2u, \\ u_x(0,t) + u(0,t) = 0, \\ u(1,t) = 0, \\ u(x,0) = \sin x. \end{cases}$$

$$\begin{cases} u_{xx} + 2u_{yy} + u_x + 2u_y = 0, \\ u(x,0) = u(x,1) = 0, \\ u(0,y) = y, \quad u(1,y) = \sin \pi y. \end{cases}$$

$$\begin{cases} u_{tt} = 3u_{xx} + 2u, \quad \text{con } 0 \le x \le 1 \text{ e } t \ge 0, \\ u(0,t) = 2, \qquad u(1,t) = 3, \\ u(x,0) = x\sin x, \quad u_t(x,0) = 0. \end{cases}$$

$$\begin{cases} u_t = 3u_{xx} - 2u_x + 4u, \quad \text{con } 0 \le x \le \pi \text{ e } t \ge 0, \\ u_x(0,t) - 2u(0,t) = 0, \quad u(\pi,t) = 0, \\ u(x,0) = x(\pi - x)^2. \end{cases}$$

$$\begin{cases} u_{tt} = 3u_{xx} + 5u_x - 4u_t + \sin x, \quad \text{con } 0 \le x \le 5 \text{ e } t \ge 0, \\ u(0,t) = 1, \quad u(5,t) = 3, \\ u(x,0) = \cos \frac{2}{5}x, \qquad u_t(x,0) = x^2 + 1. \end{cases}$$

$$\begin{cases} \dfrac{\partial u}{\partial t} = \dfrac{\partial^2 u}{\partial x^2} - 2u, \qquad 0 \le x \le 3,\ t \ge 0, \\ u(0,t) = u(3,t) = 0, \\ u(x,0) = 2\sin \pi x - \sin 4\pi x. \end{cases}$$

$$\begin{cases} u_t = c^2 u_{xx} + 2u_x, \quad \text{con } 0 \le x \le 1,\ t \ge 0, \\ u(1,t) = 0, \\ u(0,t) + 4u_x(0,t) = 0, \\ u(x,t) = f(x), \end{cases}$$

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} = a^2 \dfrac{\partial^2 u}{\partial x^2}, \qquad 0 \le x \le L,\ t \ge 0, \quad a > 0, \\ u(0,t) = u(L,t) = 0, \\ u(x,0) = 0, \\ u_t(x,0) = g(x). \end{cases}$$

Suggerimenti per ulteriori approfondimenti: Gli argomenti trattati in questo capitolo si ritrovano nella generalità dei libri (Advanced Engineering Mathematics) utilizzati nelle università statunitensi. Si veda in particolare l'ottimo libro di O'Neil [19], oppure, in italiano, il libro di Spiegel [27].

Capitolo 5

ALGEBRA LINEARE ESSENZIALE

L'algebra lineare e la teoria delle matrici forniscono attualmente le metodologie e gli strumenti di calcolo più importanti in molti settori della Matematica Applicata. L'affermazione è vera, oltre che per problemi tipici dell'algebra lineare quali: la risoluzione dei sistemi lineari e il calcolo di autovalori e autovettori nella risoluzione delle equazioni differenziali, per la teoria dell'approssimazione, per il controllo ottimale e per la risoluzione dei sistemi nonlineari [29, 11, 31, 24, 17].

5.1 Proprietà di base

Iniziamo questa sezione con qualche richiamo sulle notazioni.

Con $\mathbb{R}^n$ indichiamo lo spazio reale n-dimensionale dei vettori colonna $\boldsymbol{x}$ con componenti $x_1, x_2, \ldots, x_n$ e con $\mathbb{C}^n$ il corrispondente spazio complesso.

Assegnato $\boldsymbol{x} \in \mathbb{R}^n$, con $\boldsymbol{x}^T$ intendiamo il trasposto di $\boldsymbol{x}$, ossia il vettore riga $(x_1, x_2, \ldots, x_n)$, mentre, per $\boldsymbol{x} \in \mathbb{C}^n$, con $\boldsymbol{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$ intendiamo il suo aggiunto, ossia il suo trasposto coniugato.

Ricordiamo che n vettori $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ di $\mathbb{R}^n$ sono *linearmente indipendenti* se la relazione $\alpha_1\boldsymbol{x}_1 + \alpha_2\boldsymbol{x}_2 + \ldots + \alpha_n\boldsymbol{x}_n = \boldsymbol{0}$ vale unicamente per $\alpha_1 = \alpha_2 = \ldots = \alpha_n = 0$. Ogni n-pla di vettori linearmente indipendenti di $\mathbb{R}^n$ rappresenta una base di $\mathbb{R}^n$. Di conseguenza affermare che gli n vettori $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ di $\mathbb{R}^n$ sono linearmente indipendenti equivale ad affermare che ogni vettore $\boldsymbol{b} \in \mathbb{R}^n$ si può esprimere univocamente come loro combinazione lineare.

L'insieme delle matrici reali A da $\mathbb{R}^m$ a $\mathbb{R}^n$ ($A : \mathbb{R}^m \to \mathbb{R}^n$), per evidenziare che rappresentano operatori lineari da $\mathbb{R}^m$ a $\mathbb{R}^n$, viene indicato con $L(\mathbb{R}^m, \mathbb{R}^n)$; più semplicemente $L(\mathbb{R}^n)$ nel caso $m = n$. Analogamente con $L(\mathbb{C}^m, \mathbb{C}^n)$, viene indicato lo spazio delle matrici complesse da $\mathbb{C}^m$ a $\mathbb{C}^n$; semplicemente $L(\mathbb{C}^n)$

nel caso $m = n$. Gli elementi di $A \in L(\mathbb{R}^m, \mathbb{R}^n)$, come anche di $A \in L(\mathbb{C}^m, \mathbb{C}^n)$, vengono indicati con a_{ij}, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$.

Nel caso reale, con A^T indichiamo la trasposta di A e, nel caso complesso, con $A^* = \overline{A}^T = \overline{(A^T)}$ la trasposta coniugata.

Per ogni matrice quadrata, $\det A$ indica il determinante di A e per ogni matrice A non singolare A^{-1} indica la sua inversa. Per *rango* di una matrice $\mathrm{Rang}(A)$ indichiamo il numero massimale delle sue colonne (o righe) linearmente indipendenti.

5.1. Relativamente all'invertibilità di una matrice è fondamentale aver presente l'equivalenza delle seguenti affermazioni:

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

(b) A è non singolare;

(c) il sistema lineare omogeneo $A\boldsymbol{x} = \mathbf{0}$ possiede l'unica soluzione $\boldsymbol{x} = \mathbf{0}$;

(d) qualunque sia il vettore $\boldsymbol{b}$, il sistema lineare $A\boldsymbol{x} = \boldsymbol{b}$ possiede una sola soluzione;

(e) le colonne (e le righe) di A sono linearmente indipendenti;

(f) il rango della matrice A è n.

Se $A \in L(\mathbb{C}^n)$, un numero (reale o complesso) λ e un vettore $\boldsymbol{x} \neq \mathbf{0}$ vengono definiti autovalore e autovettore di A se

$$A\boldsymbol{x} = \lambda\boldsymbol{x}.$$

Per le equivalenze richiamate al punto 5.1, indicata (come usuale) con I la matrice identità, questo avviene se e solo se

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

ossia se e solo se λ soddisfa l'equazione caratteristica di A. $P_A(\lambda)$ rappresenta il "polinomio caratteristico" di A. Essendo $P_A(\lambda)$ un polinomio di grado n, A ha esattamente n autovalori non necessariamente distinti che, tenuto conto delle loro molteplicità, sono esattamente gli zeri (con relative molteplicità) di $P_A(\lambda)$. Di conseguenza, essendo $P_A(\lambda)$ monico,

$$P_A(\lambda) = \prod_{i=1}^{n} (\lambda - \lambda_i) \quad \text{con} \quad \det A = (-1)^n P_A(0) = (-1)^n \prod_{i=1}^{n} \lambda_i.$$

I vettori relativi agli autovalori sono gli autovettori e la coppia $(\lambda, \boldsymbol{x})$ indica una coppia autovalore-autovettore.

Dalla definizione di polinomio caratteristico, essendo $\det(A) = \det(A^T)$, segue immediatamente che gli autovalori di A e A^T sono uguali. Per convincersene basta osservare che

$$P_{A^T}(\lambda) = \det(\lambda I - A^T) = \det[(\lambda I - A)^T] = \det(\lambda I - A) = P_A(\lambda),$$

da cui discende che i due polinomi hanno gli stessi zeri con le stesse molteplicità.

Per *spettro* di A si intende l'insieme degli autovalori $\lambda_1, \lambda_2, \ldots, \lambda_n$ di A, mentre con

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

intendiamo il raggio spettrale di A, ossia il raggio del minimo cerchio (con centro nell'origine) contenente lo spettro di A.

Calcolare lo spettro di una matrice è, in generale, difficile. Non lo è tuttavia in casi particolari. Se, per esempio, A è *triangolare* (superiore o inferiore),

$$A = \begin{pmatrix} a_{11} & 0 & \ldots & \ldots & 0 \\ a_{21} & a_{22} & 0 & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & 0 \\ a_{n1} & a_{n2} & \ldots & \ldots & a_{nn} \end{pmatrix}, \quad A = \begin{pmatrix} a_{11} & a_{12} & \ldots & \ldots & a_{1n} \\ 0 & a_{22} & \ldots & \ldots & a_{2n} \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \ldots & \ldots & 0 & a_{nn} \end{pmatrix},$$

non è necessario nessun calcolo, in quanto gli autovalori coincidono con gli elementi diagonali di A. Ovviamente questo vale anche se A è una matrice diagonale ($A = \mathrm{diag}(\alpha_1, \alpha_2, \ldots, \alpha_n)$).

Essendo gli autovalori di una matrice gli zeri dell'associato polinomio caratteristico, la loro distribuzione nel piano complesso è del tutto arbitraria. Esistono tuttavia delle importanti situazioni nelle quali gli autovalori sono distribuiti in zone meglio precisate del piano. Valgono infatti le seguenti particolarità:

(1) Se A è reale e simmetrica ($A = A^T$), gli autovalori sono reali;

(2) Se A è complessa e Hermitiana ($A^* = A$), gli autovalori sono reali;

(3) Se A è una matrice reale semi-definita positiva ($\boldsymbol{x}^T A \boldsymbol{x} \geq 0$, per ogni $\boldsymbol{x} \in \mathbb{R}^n$), gli autovalori sono reali non negativi;

(4) Se A è reale e definita positiva, ($\boldsymbol{x}^T A \boldsymbol{x} > 0$, per ogni $\mathbf{0} \neq \boldsymbol{x} \in \mathbb{R}^n$), gli autovalori sono tutti positivi.

Due matrici A e B di $L(\mathbb{R}^n)$ (o di $L(\mathbb{C}^n)$) sono *simili* se esiste una matrice non singolare $P \in L(\mathbb{R}^n)$ (o in $L(\mathbb{C}^n)$) tale che

$$B = P^{-1}AP.$$

Questa definizione appare del tutto naturale se si osserva che le due matrici condividono la stessa trasformazione lineare in due diversi sistemi di riferimento. Supponiamo infatti che in un dato sistema di riferimento

$$y = Ax.$$

Un cambiamento di variabili comporta trasformazioni del tipo

$$y = P\hat{y} \quad \text{e} \quad y = P\hat{x}, \quad \text{con } P \text{ non singolare.}$$

Di conseguenza,

$$P\hat{y} = AP\hat{x},$$

dalla quale (essendo P non singolare)

$$\hat{y} = P^{-1}AP\hat{x} = B\hat{x}, \qquad B = P^{-1}AP,$$

ossia la trasformazione è la stessa, anche se riferita ad assi diversi.

5.2. Due matrici simili hanno gli stessi autovalori.
Per la dimostrazione basta osservare

$$\begin{aligned}\det(\lambda I - P^{-1}AP) &= \det\left[P^{-1}(\lambda P - AP)\right] \\ &= \det\left[P^{-1}(\lambda I - A)P\right] = \det(P^{-1})\det(\lambda I - A)\det(P).\end{aligned}$$

5.3. In molte applicazioni non è necessario calcolare gli autovalori; è sufficiente darne una buona localizzazione. A questo serve, in particolare, il seguente

Teorema 5.1 (*di Gershgorin*) *Gli autovalori di una matrice n-dimensionale A (reale o complesso) sono tutti contenuti nell'unione dei cerchi*

$$|a_{ii} - \lambda| \le \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|, \qquad i = 1, 2, \ldots, n. \tag{5.1}$$

Dimostrazione. Sia λ un autovalore e $\boldsymbol{x} \neq \mathbf{0}$ il corrispondente autovettore. Indicata con x_i la componente di massimo modulo di $\boldsymbol{x}$ ($|x_i| = \max_{j=1,\ldots,n} |x_j|$), è evidente che $x_i \neq 0$, dato che $\boldsymbol{x} \neq \mathbf{0}$. Di conseguenza, considerando la i-esima

riga di $A\boldsymbol{x} = \lambda\boldsymbol{x}$ e ricordando che $|x_j| \leq |x_i|$, $j = 1, \ldots, n$, possiamo affermare che

$$|a_{ii} - \lambda|\,|x_i| = |a_{ii}x_i - x_i| = \left|\sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}x_j\right| \leq \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|\,|x_j|$$

$$\leq \left(\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|\right)|x_i|$$

dalla quale segue la (5.1). □

Poiché, come osservato precedentemente, A e A^T hanno gli stessi autovalori, il teorema di localizzazione di Gershgorin può riferirsi alle colonne di A, oltre che alle righe. In altre parole il teorema è ugualmente valido nei cerchi

$$|a_{ii} - \lambda| \leq \sum_{\substack{i=1\\ i\neq j}}^{n} |a_{ij}|, \qquad j = 1, 2, \ldots, n.$$

La prima e praticamente più importante conseguenza del teorema riguarda le matrici diagonalmente dominanti.

Una matrice A è definita *diagonalmente dominante in senso stretto* se, per ogni $i = 1, 2, \ldots, n$, vale la disuguaglianza

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

Dal Teorema 5.1 segue immediatamente che ogni matrice diagonalmente dominante in senso stretto è non singolare. Per rendersene conto basta osservare che nessuno dei cerchi (5.1) contiene il valore $\lambda = 0$ e dunque $\lambda = 0$ non è un autovalore, il che equivale ad affermare che A è non singolare. Se infatti A fosse singolare,indicato con $r < n$ il rango della matrice, l'equazione

$$A\boldsymbol{x} = \lambda\boldsymbol{x}, \qquad \lambda = 0,$$

avrebbe ∞^{n-r} soluzioni non nulle.

Una matrice A si dice diagonalmente dominante in senso debole se

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

con il maggiore stretto valido almeno una volta. La matrice

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

ad esempio, è diagonalmente dominante in senso debole, ma non in senso stretto.

5.4. Matrici ortogonali. Una matrice $A \in L(\mathbb{R}^n)$ è ortogonale se

$$A^T A = AA^T = I \iff A^T = A^{-1}.$$

Esempio 5.2 La matrice $A = \frac{1}{\sqrt{6}} \begin{pmatrix} \sqrt{2} & \sqrt{3} & 1 \\ \sqrt{2} & 0 & -2 \\ \sqrt{2} & -\sqrt{3} & 1 \end{pmatrix}$ è ortogonale, in quanto $A^T A = AA^T = I$. Di conseguenza:

(a) $A^{-1} = A^T = \frac{1}{\sqrt{6}} \begin{pmatrix} \sqrt{2} & \sqrt{2} & \sqrt{2} \\ \sqrt{3} & 0 & -\sqrt{3} \\ 1 & -2 & 1 \end{pmatrix}$;

(b) Se A è ortogonale, la soluzione del sistema $A\boldsymbol{x} = \boldsymbol{b}$ è $\boldsymbol{x} = A^T\boldsymbol{b}$. Se, ad es., A è la matrice precedente e $\boldsymbol{b}^T = (0, 1, 1)$ e $\boldsymbol{x}^T = \frac{1}{\sqrt{6}}(2\sqrt{2}, -\sqrt{3}, -1)$.

In $L(\mathbb{C}^n)$ proprietà analoghe valgono per le matrici unitarie, ossia per le matrici per le quali

$$A^* A = AA^* = I, \qquad A^* = \overline{(A^T)} = (\overline{A})^T,$$

nel qual caso

$$A^{-1} = A^* \text{ e } A\boldsymbol{x} = \boldsymbol{b} \Longrightarrow \boldsymbol{x} = A^*\boldsymbol{b}.$$

5.5. Autovalori di una matrice ortogonale. Tutti gli autovalori di una matrice Hermitiana $[A^* = A^{-1}]$ hanno modulo 1, conseguentemente sono tutti localizzati nella circonferenza con centro l'origine e raggio 1.

Dimostrazione. $A\boldsymbol{x} = \lambda\boldsymbol{x}$ e $\boldsymbol{x} \neq \boldsymbol{0} \Longrightarrow \boldsymbol{x}^* A\boldsymbol{x} = \lambda \boldsymbol{x}^*\boldsymbol{x} \Longrightarrow \boldsymbol{x}^*\boldsymbol{x} = \lambda^{-1}\boldsymbol{x}^* A\boldsymbol{x}$. D'altronde, $\boldsymbol{x}^* A^* = \overline{\lambda}\boldsymbol{x}^* \Longrightarrow \boldsymbol{x}^* A^*\boldsymbol{x} = \overline{\lambda}\boldsymbol{x}^*\boldsymbol{x} = \overline{\lambda}\lambda^{-1}\boldsymbol{x}^* A\boldsymbol{x} \Longrightarrow |\lambda| = 1$. Questo implica che anche gli autovalori di una matrice *ortogonale* $(A^T = A^{-1})$ sono $\lambda = \pm 1$.

Altre utili proprietà:

(a) gli autovalori di una matrice simmetrica $A^T = A$ sono reali;

(b) gli autovalori di una matrice antisimmetrica $(A^T = -A)$ sono immaginari puri;

(c) in una matrice simmetrica gli autovettori corrispondenti ad autovalori distinti sono ortogonali;

(d) se $A \in L(\mathbb{R}^n)$ è ortogonale, le sue colonne rappresentano una base ortonormale di $\mathbb{R}^n$.

In tal caso è immediato calcolare le proiezioni sui vettori di base. Infatti, indicato con $\boldsymbol{x}$ un generico vettore di $\mathbb{R}^n$ e con $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n$ le colonne di una matrice ortogonale A, dalla relazione

$$\boldsymbol{x} = \alpha_1 \boldsymbol{a}_1 + \alpha_2 \boldsymbol{a}_2 + \ldots + \alpha_n \boldsymbol{a}_n$$

segue, come immediata conseguenza della ortogonalità dei vettori $\{\boldsymbol{a}_i\}_{i=1}^n$, che

$$\boldsymbol{x}^T \boldsymbol{a}_i = (\boldsymbol{x}, \boldsymbol{a}_i) = \alpha_i, \qquad i = 1, 2, \ldots, n.$$

Questo significa che l'i-esimo coefficiente α_i di $\boldsymbol{x}$ è la proiezione ortogonale di $\boldsymbol{x}$ su $\boldsymbol{a}_i$.

5.6. Ortonormalizzazione di una base. Anche se teoricamente le basi di $\mathbb{R}^n$ sono equivalenti, dal punto di vista computazionale è di gran lungo preferibile disporre di una base ortonormale. Tale obiettivo può essere raggiunto ortonormalizzando la base iniziale con il metodo di Gram-Schmidt. Assegnata una base $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n\}$ di $\mathbb{R}^n$ il metodo di ortonormalizzazione di Gram-Schmidt genera una base ortonormale $\{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n\}$, procedendo nel modo seguente:

$$\begin{aligned} &\boldsymbol{q}_1 = \frac{\boldsymbol{a}_1}{\|\boldsymbol{a}_1\|}, \qquad (5.2)\\ &\hat{\boldsymbol{q}}_2 = \boldsymbol{a}_2 - r_{12}\boldsymbol{q}_1, \quad r_{12} = \boldsymbol{q}_1^T \boldsymbol{a}_2 \Longrightarrow \boldsymbol{q}_1^T \boldsymbol{q}_2 = 0,\\ &\boldsymbol{q}_2 = \frac{\hat{\boldsymbol{q}}_2}{\|\hat{\boldsymbol{q}}_2\|} \qquad \text{e, in generale,} \end{aligned}$$

$$\begin{cases} \hat{\boldsymbol{q}}_j = \boldsymbol{a}_j - \displaystyle\sum_{i=1}^{j-1} r_{ij}\boldsymbol{q}_i, \quad r_{ij} = \boldsymbol{q}_i^T \boldsymbol{q}_j \\ \boldsymbol{q}_j = \frac{\hat{\boldsymbol{q}}_j}{\|\hat{\boldsymbol{q}}_j\|}, \quad j = 2, 3, \ldots, n, \end{cases} \qquad (5.3)$$

dove il simbolo $\|\cdot\|$ indica la norma Euclidea.

Questo algoritmo può essere considerato il prototipo dei metodi QR, ossia dei metodi di fattorizzazione di una matrice A nel prodotto di una ortogonale Q per una triangolare superiore R.

Dimostrazione. Posto $A = \begin{pmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \dots & \boldsymbol{a}_n \end{pmatrix}$, vogliamo realizzare la fattorizzazione

$$A = \begin{pmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \dots & \boldsymbol{a}_n \end{pmatrix} = \begin{pmatrix} \boldsymbol{q}_1 & \boldsymbol{q}_2 & \dots & \boldsymbol{q}_n \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} & \dots & \dots & r_{1n} \\ & r_{22} & \dots & \dots & r_{2n} \\ & & \ddots & & \vdots \\ & & & \ddots & \\ & & & & r_{nn} \end{pmatrix} = QR, \tag{5.4}$$

dove $Q^T Q = I$. Dalla (5.3), procedendo per colonne, otteniamo

$$\boldsymbol{a}_1 = r_{11}\boldsymbol{q}_1, \text{ dalla quale, posto } r_{11} = \|\boldsymbol{a}_1\|, \text{ segue che } \boldsymbol{q}_1 = \frac{\boldsymbol{a}_1}{r_{11}},$$

$$\begin{cases} \boldsymbol{a}_2 = r_{12}\boldsymbol{q}_1 + r_{22}\boldsymbol{q}_2, \text{ dalla quale segue che } r_{12} = \boldsymbol{q}_1^T \boldsymbol{a}_2; \\ \text{osservato quindi che } r_{22}\boldsymbol{q}_2 = \boldsymbol{a}_2 - r_{12}\boldsymbol{q}_1, \ r_{22} = \|\boldsymbol{a}_2 - r_{12}\boldsymbol{q}_1\| \\ \text{e conseguentemente } \boldsymbol{q}_2 = \dfrac{\boldsymbol{a}_2 - r_{12}\boldsymbol{q}_1}{r_{22}}. \end{cases}$$

In generale, per $j = 2, 3, \dots, n$

$$\boldsymbol{a}_j = \sum_{i=1}^{j} r_{ij}\boldsymbol{q}_i,$$

dove $\boldsymbol{q}_1, \boldsymbol{q}_2, \dots, \boldsymbol{q}_{j-1}$ sono vettori noti, normalizzati e tra loro ortogonali. Da esse segue che

$$r_{ij} = \boldsymbol{q}_i^T \boldsymbol{a}_j, \quad i = 1, \dots, j-1, \text{ con } r_{jj}\boldsymbol{q}_j = \boldsymbol{a}_j - \sum_{i=1}^{j-1} r_{ij}\boldsymbol{q}_j.$$

Da quest'ultima, la normalizzazione di $\boldsymbol{q}_j$ richiede che $r_{jj} = \|\boldsymbol{a}_j - \sum_{i=1}^{j-1} r_{ij}\boldsymbol{q}_j\|$ e conseguentemente

$$\boldsymbol{q}_j = \frac{\boldsymbol{a}_j - \sum_{i=1}^{j-1} r_{ij}\boldsymbol{q}_j}{r_{jj}}.$$

Di questo metodo esistono diverse varianti che hanno come obiettivo la sua stabilità numerica [28, 31, 11].

Le matrici ortogonali svolgono un ruolo molto importante nel calcolo dello spettro di una matrice, come nella risoluzione dei sistemi lineari, come risulta dalle seguenti proprietà:

Se $A \in L(\mathbb{C}^n)$ è simmetrica, essa è diagonalizzabile con una matrice ortogonale. Esiste cioè una matrice ortogonale P tale che

$$A = P^{-1}DP = P^T DP, \text{ essendo } D = \operatorname{diag}(\alpha_1, \dots, \alpha_n).$$

Da essa segue che, (essendo $P^{-1} = P^T$) lo spettro di A è rappresentato dagli elementi diagonali di D ($\lambda_i = \alpha_i$, $i = 1, \ldots, n$).

5.7. Teorema di Schur [20]. Se $A \in L(\mathbb{C}^n)$, esiste una matrice ortogonale P in grado di triangolarizzare A, tale cioè da ottenersi

$$A = P^T T P, \qquad T \text{ triangolare.}$$

Essendo A simile a T, gli autovalori di A sono gli elementi diagonali di T.

È importante anche notare che per calcolare gli autovalori di un polinomio matriciale

$$P_m(A) = a_0 A^m + a_1 A^{m-1} + \ldots + a_{m-1} A + a_m I,$$

è sufficiente conoscere gli autovalori di A.

Allo scopo basta osservare che

$$A\boldsymbol{x} = \lambda \boldsymbol{x},\ \boldsymbol{x} \neq \boldsymbol{0} \Rightarrow A^2\boldsymbol{x} = A(A\boldsymbol{x}) = A(\lambda \boldsymbol{x}) = \lambda A\boldsymbol{x} = \lambda^2 \boldsymbol{x},\ \boldsymbol{x} \neq \boldsymbol{0},$$

$$A\boldsymbol{x} = \lambda \boldsymbol{x},\ \boldsymbol{x} \neq \boldsymbol{0} \Rightarrow \begin{cases} A^k\boldsymbol{x} = A^{k-1}(A\boldsymbol{x}) = A^{k-1}(\lambda \boldsymbol{x}) = \lambda A^{k-1}\boldsymbol{x} = \ldots = \lambda^k \boldsymbol{x}, \\ \boldsymbol{x} \neq \boldsymbol{0}, \end{cases}$$

da cui segue che

$$P_m(A)\boldsymbol{x} = (a_0\lambda^m + a_1\lambda^{m-1} + \ldots + a_{m-1}\lambda + a_m)\boldsymbol{x}, \quad \boldsymbol{x} \neq \boldsymbol{0}.$$

Di conseguenza, lo spettro di $P_m(A)$ è

$$\{a_0\lambda_i^m + a_1\lambda_i^{m-1} + \ldots + a_{m-1}\lambda_i + a_m,\ i = 1, \ldots, n\},$$

dove $\lambda_1, \ldots, \lambda_n$ sono gli autovalori di A.

5.8. Se $A \in L(\mathbb{R}^n)$ è una matrice simmetrica, gli autovettori relativi a due autovalori diversi sono ortogonali.

Dimostrazione. Siano $A\boldsymbol{x} = \lambda_1 \boldsymbol{x}$, $\boldsymbol{x} \neq \boldsymbol{0}$ e $A\boldsymbol{y} = \lambda_2 \boldsymbol{y}$, $\boldsymbol{y} \neq \boldsymbol{0}$ con $\lambda_1 \neq \lambda_2$. Come conseguenza,

$$A\boldsymbol{x} = \lambda_1 \boldsymbol{x} \Longrightarrow \boldsymbol{y}^T A\boldsymbol{x} = \lambda_1 \boldsymbol{y}^T \boldsymbol{x} \text{ e } \boldsymbol{x}^T A\boldsymbol{y} = \lambda_2 \boldsymbol{x}^T \boldsymbol{y},$$

da cui, essendo $\boldsymbol{y}^T A\boldsymbol{x} = \boldsymbol{x}^T A\boldsymbol{y}$ e $\boldsymbol{y}^T\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{y}$, segue immediatamente che

$$(\lambda_1 - \lambda_2)\boldsymbol{x}^T\boldsymbol{y} = 0 \Longrightarrow \boldsymbol{x}^T\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{x} = 0, \text{ dato che } \lambda_1 \neq \lambda_2.$$

Da questa proprietà segue che, se gli autovalori di A sono distinti, gli autovettori formano una base ortogonale di $\mathbb{R}^n$.

5.2 Norme vettoriali e matriciali

In questa sezione introduciamo le definizioni e proprietà delle norme vettoriali e matriciali, basilari nella risoluzione numerica dei sistemi algebrici.

5.9. Norme vettoriali. Per norma vettoriale in $\mathbb{R}^n$ (o in $\mathbb{C}^n$) si intende la funzione $\|\cdot\|$ soddisfacente le seguenti proppprietà:

(a) $\|\boldsymbol{x}\| \geq 0$, qualunque sia $\boldsymbol{x}$ con $\|\boldsymbol{x}\| = 0$ se e solo se $\boldsymbol{x} = \mathbf{0}$;

(b) $\|\alpha\boldsymbol{x}\| = |\alpha|\,\|\boldsymbol{x}\|$, per ogni vettore $\boldsymbol{x}$ e ogni scalare α;

(c) $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$, qualunque siano $\boldsymbol{x}$ e $\boldsymbol{y}$.

Le tre norme più frequentemente utilizzate nelle applicazioni sono le seguenti:

$$\begin{aligned}
\|\boldsymbol{x}\|_2 &= \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}, && \text{norma Euclidea (norma 2);}\\
\|\boldsymbol{x}\|_1 &= \sum_{i=1}^{n} |x_i|, && \text{norma in } \ell_1 \text{ (norma 1);}\\
\|\boldsymbol{x}\|_\infty &= \max_{i=1,\ldots,n} |x_i|, && \text{norma in } \ell_\infty \text{ (norma infinito).}
\end{aligned}$$

Ciascuna di esse appartiene alla famiglia delle norme p essendo, per $1 \leq p < \infty$,

$$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}.$$

La norma ∞ è così denominata, in quanto limite delle $\|\cdot\|_p$ per $p \to \infty$.

Per cerchio unitario in $\|\cdot\|_p$ si intende l'insieme dei vettori $\boldsymbol{x} \in \mathbb{R}^n$ con $\|\boldsymbol{x}\|_p \leq 1$. È facile dimostrare che, relativamente alla norma 2, 1 e ∞ abbiamo le rappresentazioni geometriche riportate nella Fig. 5.1: Osserviamo che il terzo cerchio contiene il primo, che a sua volta contiene il secondo. È facile altresì dimostrare che $\|\cdot\|_p$ è una funzione decrescente di p, da cui segue, in particolare, che

$$\|\boldsymbol{x}\|_\infty \leq \|\boldsymbol{x}\|_2 \leq \|\boldsymbol{x}\|_1, \qquad \text{per ogni } \boldsymbol{x} \in \mathbb{R}^n.$$

Tutte le precedenti considerazioni sono immediatamente estendibili al campo complesso $\mathbb{C}^n$.

5.10. Prodotto interno. Per prodotto interno in $\mathbb{R}^n$ si intende un'applicazione da $\mathbb{R}^n$ a $\mathbb{R}^n$ che soddisfa le seguenti proprietà:

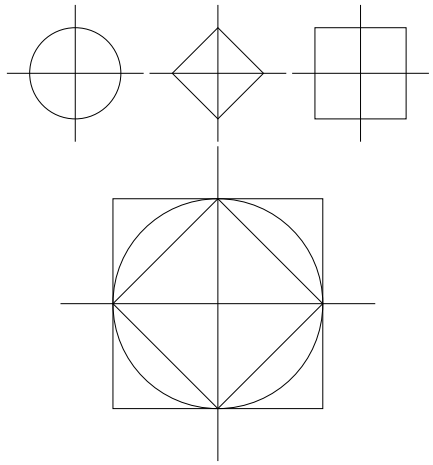


Figura 5.1: Nella prima riga: I cerchi unitari tipici delle norme 2, 1 e ∞, rispettativamente. Nella seconda riga: i tre cerchi unitari in una singola figura.

(a) $(\boldsymbol{x}, \boldsymbol{x}) \geq 0$, per ogni $\boldsymbol{x}$, con $(\boldsymbol{x}, \boldsymbol{x}) = 0$ se e solo se $\boldsymbol{x} = \boldsymbol{0}$;

(b) $(\alpha\boldsymbol{x}, \boldsymbol{y}) = \alpha(\boldsymbol{x}, \boldsymbol{y})$ per ogni coppia di vettori in $\mathbb{R}^n$ e ogni numero reale α;

(c) $(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{y}, \boldsymbol{x})$, qualunque siano $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$;

(d) $(\boldsymbol{x}, \boldsymbol{y} + \boldsymbol{z}) = (\boldsymbol{x}, \boldsymbol{y}) + (\boldsymbol{x}, \boldsymbol{z})$, qualunque siano i vettori $\boldsymbol{x}$, $\boldsymbol{y}$ e $\boldsymbol{z}$.

Le stesse proprietà valgono per vettori in $\mathbb{C}^n$ a condizione di sostituire la (c) con la seguente:

(c') $(\boldsymbol{x}, \boldsymbol{y}) = \overline{(\boldsymbol{y}, \boldsymbol{x})}$, qualunque siano $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^n$.

Ad ogni prodotto interno si può associare una norma così definita

$$\|\boldsymbol{x}\| = \sqrt{(\boldsymbol{x}, \boldsymbol{x})}\,. \tag{5.5}$$

La verifica delle proprietà (a) e (b) è immediata. La verifica della (c) [o della (c')] richiede il ricorso alla seguente

Disuguaglianza di Cauchy-Schwartz:

$$|(\boldsymbol{x}, \boldsymbol{y})| \leq \|\boldsymbol{x}\|\,\|\boldsymbol{y}\|, \qquad \text{qualunque siano } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

Per la sua dimostrazione basta osservare che, per le proprietà (a) e (b) del prodotto interno, il polinomio

$$P(\alpha) = (\alpha\boldsymbol{x} + \boldsymbol{y}, \alpha\boldsymbol{x} + \boldsymbol{y}) = \alpha^2(\boldsymbol{x}, \boldsymbol{x}) + 2\alpha(\boldsymbol{x}, \boldsymbol{y}) + (\boldsymbol{y}, \boldsymbol{y})$$

è non negativo per ogni valore $\alpha \in \mathbb{R}$. Di conseguenza, il suo discriminante

$$\Delta(P) = 4\left\{|(\boldsymbol{x},\boldsymbol{y})|^2 - (\boldsymbol{x},\boldsymbol{x})(\boldsymbol{y},\boldsymbol{y})\right\} \leq 0$$

è non positivo e questo equivale a dire che la disuguaglianza è valida.

Per dimostrare la disuguaglianza triangolare basta infine osservare che (per ogni coppia $\boldsymbol{x}$, $\boldsymbol{y}$)

$$\begin{aligned}\|\boldsymbol{x}+\boldsymbol{y}\|^2 &= (\boldsymbol{x}+\boldsymbol{y},\boldsymbol{x}+\boldsymbol{y}) = (\boldsymbol{x},\boldsymbol{x}) + 2(\boldsymbol{x},\boldsymbol{y}) + (\boldsymbol{y},\boldsymbol{y}) \\ &\leq (\boldsymbol{x},\boldsymbol{x}) + 2\|\boldsymbol{x}\|\,\|\boldsymbol{y}\| + (\boldsymbol{y},\boldsymbol{y}) = (\|\boldsymbol{x}\| + \|\boldsymbol{y}\|)^2 \\ &\Longrightarrow \|\boldsymbol{x}+\boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|\end{aligned}$$

e dunque la (5.5) definisce una norma.

Implicazione della disuguaglianza triangolare. Nella disuguaglianza triangolare sulla norma segue immediatamente che, qualunque siano $\boldsymbol{x}$, $\boldsymbol{y}$,

$$|\|\boldsymbol{x}\| - \|\boldsymbol{y}\|| \leq \|\boldsymbol{x}-\boldsymbol{y}\|. \tag{5.6}$$

Per verificare ls (5.6), basta osservare che (per la disuguaglianza triangolare)

$$\begin{aligned}\|\boldsymbol{x}\| = \|(\boldsymbol{x}+\boldsymbol{y})-\boldsymbol{y}\| \leq \|\boldsymbol{x}+\boldsymbol{y}\| + \|\boldsymbol{y}\| &\Longrightarrow \|\boldsymbol{x}\| - \|\boldsymbol{y}\| \leq \|\boldsymbol{x}+\boldsymbol{y}\|. \\ \|\boldsymbol{y}\| = \|(\boldsymbol{y}+\boldsymbol{x})-\boldsymbol{x}\| \leq \|\boldsymbol{x}+\boldsymbol{y}\| + \|\boldsymbol{x}\| &\Longrightarrow \|\boldsymbol{y}\| - \|\boldsymbol{x}\| \leq \|\boldsymbol{x}+\boldsymbol{y}\|,\end{aligned}$$

dalle quali segue immediatamente la (5.6).

Matrice di Gram. Il seguente risultato è molto utile nello stabilire la non singolarità delle matrici provenienti dalla risoluzione numerica delle equazioni differenziali e di altri problemi applicativi.

Teorema 5.3 *Gli n vettori $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n$ in $\mathbb{R}^n$ sono linearmente indipendenti se e solo se la corrispondente matrice di Gram*

$$G = \{g_{ij}\}_{i,j=1}^n, \qquad g_{ij} = (\boldsymbol{a}_i, \boldsymbol{a}_j),$$

è non singolare.

Dimostrazione. È facile verificare che

$$\|c_1\boldsymbol{a}_1 + \ldots + c_n\boldsymbol{a}_n\|^2 = \sum_{i,j=1}^n c_i c_j (\boldsymbol{a}_i, \boldsymbol{a}_j) = \begin{pmatrix} c_1 & \ldots & c_n \end{pmatrix} G \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix},$$

qualunque siano le costanti reali $c_1, \ldots, c_n$. Di conseguenza, esistono tali costanti c_i non tutte nulle tali che $c_1\boldsymbol{a}_1 + \ldots + c_n\boldsymbol{a}_n = \boldsymbol{0}$ se e solo se il vettore colonna $\begin{pmatrix} c_1 & \ldots & c_n \end{pmatrix}^T$ è autovettore di G corrispondente all'autovalore G, cioè gli n vettori $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n$ sono linearmente dipendenti se e solo se la matrice di Gram G è singolare. □

Un risultato simile vale per n vettori in $\mathbb{C}^n$.

5.11. Norme matriciali (indotte, naturali). Ad ogni norma vettoriale $\|\cdot\|$ in $\mathbb{R}^n$ si può associare una norma matriciale, per tale motivo, definita indotta o anche naturale.

Definizione. Per ogni $A \in L(\mathbb{R}^n)$ si ha:

$$\|A\| = \sup_{\boldsymbol{x}\neq\mathbf{0}} \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|}. \tag{5.7a}$$

Ponendo $\boldsymbol{y} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$, la (5.7a), osservato che $\|A\boldsymbol{x}\| = \|A(\|\boldsymbol{x}\|)\boldsymbol{y}\| = \|\boldsymbol{x}\|\,\|A\boldsymbol{y}\|$ e che $\|\boldsymbol{y}\| = 1$, diventa

$$\|A\| = \max_{\|y\|=1} \|A\boldsymbol{y}\|. \tag{5.7b}$$

La sostituzione del sup con il max discende dal fatto che la $\|\cdot\|$ è una funzione continua e l'insieme dei vettori $\boldsymbol{y} \in \mathbb{R}^n$ con $\|\boldsymbol{y}\| = 1$ è chiuso.

Le norme matriciali sono caratterizzate dalla validità delle seguenti proprietà:

(a) $\|A\| \geq 0$ con $\|A\| = 0$ se e solo se $A = O$;

(b) $\|\alpha A\| = |\alpha|\,\|A\|$, qualunque sia il numero $\alpha \in \mathbb{R}$ ($\alpha \in \mathbb{C}$);

(c) $\|A + B\| \leq \|A\| + \|B\|$, qualunque siano le matrici A e B;

(d) $\|AB\| \leq \|A\|\,\|B\|$, qualunque siano le matrici A e B.

Le prime tre seguono immediatamente dalle (5.7). Per la (d) si deve prima osservare che per la (5.7a)

$$\frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \|A\|, \qquad \boldsymbol{x} \neq \mathbf{0},$$

dato che $\|A\|$ (per definizione) è il più piccolo maggiorante di $\frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|}$, $\|\boldsymbol{x}\| \neq 0$, da cui

$$\|A\boldsymbol{x}\| \leq \|A\|\,\|\boldsymbol{x}\|.$$

Conseguentemente, per ogni $\boldsymbol{x} \neq \mathbf{0}$,

$$\|A(B\boldsymbol{x})\| \leq \|A\|\,\|B\boldsymbol{x}\| \leq \|A\|\,\|B\|\,\|\boldsymbol{x}\| \Longrightarrow$$

$$\Longrightarrow \frac{\|(AB)\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \|A\|\,\|B\| \text{ per ogni } \boldsymbol{x} \neq \mathbf{0} \text{ e dunque vale la (c), dato che}$$

$$\|AB\| = \sup_{\boldsymbol{x}\neq\mathbf{0}} \frac{\|(AB)\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \|A\|\,\|B\|.$$

Matrice identità. È evidente che, indicata con I la matrice identità in $L(\mathbb{C}^n)$, qualunque sia la norma matriciale indotta, $\|I\| = 1$.

5.12. Relazione tra norma matriciale e raggio spettrale. Qualunque sia la $\|\cdot\|$ matriciale indotta e qualunque sia la matrice $A \in L(\mathbb{C}^n)$,

$$\rho(A) \leq \|A\|. \tag{5.8}$$

Verifica. Dalla definizione

$$A\boldsymbol{x} = \lambda\boldsymbol{x}, \qquad \boldsymbol{x} \neq \mathbf{0}$$

deriva infatti che, qualunque sia l'autovalore λ e il corrispondente autovettore $\boldsymbol{x}$,

$$\|A\boldsymbol{x}\| = |\lambda|\,\|\boldsymbol{x}\| \Longrightarrow |\lambda| = \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \|A\| = \sup_{\boldsymbol{x}\neq\mathbf{0}} \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|}.$$

Anche se, in generale vale il $<$, in casi particolari può valore l'uguale. Per $A = I$, vale l'uguale dato che, per ogni norma, $\|A\| = 1$ e $\lambda_1 = \lambda_2 = \ldots = \lambda_n = 1 \Longrightarrow \rho(I) = 1$.

Calcolo delle $\|A\|_2$, $\|A\|_\infty$ e $\|A\|_1$.

(a) Se $A \in L(\mathbb{R}^n)$, $\|A\|_2 = \sqrt{\rho(A^TA)}$.

Per la sua dimostrazione, indicati con $(\lambda, \boldsymbol{u})$ una coppia autovalore-autovettore di A^TA, possiamo affermare che

$$A^TA\boldsymbol{u} = \lambda\boldsymbol{u}, \qquad \boldsymbol{u} \neq \mathbf{0}.$$

Da essa segue che

$$\boldsymbol{u}^TA^TA\boldsymbol{u} = \|A\boldsymbol{u}\|_2^2 = \lambda\|\boldsymbol{u}\|_2^2 \Longrightarrow \lambda \geq 0$$

e conseguentemente

$$\rho(A^TA) = \frac{\|A\boldsymbol{u}\|_2^2}{\|\boldsymbol{u}\|_2^2} \leq \sup_{\boldsymbol{x}\neq\mathbf{0}} \frac{\|A\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2} = \|A\|_2^2,$$

ossia $\rho(A^TA) \leq \|A\|_2^2$.

Per dimostrare che vale l'uguale, posto $\mu = \rho(A^TA)$ e indicato con $\boldsymbol{v}$ l'autovettore corrispondente, osserviamo che

$$\frac{\|A\boldsymbol{v}\|_2^2}{\|\boldsymbol{v}\|_2^2} = \frac{\boldsymbol{v}^TA^TA\boldsymbol{v}}{\boldsymbol{v}^T\boldsymbol{v}} = \frac{\boldsymbol{v}^T(\mu\boldsymbol{v})}{\boldsymbol{v}^T\boldsymbol{v}} = \mu = \rho(A^TA).$$

Resta così dimostrato che $\|A\|_2 = \sqrt{\rho(A^TA)}$.

Se la matrice A è *simmetrica*, essendo $\rho(A^TA) = \rho(A^2) = \rho(A)^2$,

$$\|A\|_2 = \rho(A).$$

Se A è *ortogonale*, $\|A\|_2 = 1$, dato che $A^TA = I$ e $\rho(I) = 1$.

Anche se esistono vari metodi per il calcolo del raggio spettrale di una matrice, il calcolo della $\|A\|_2$ è, in generale, piuttosto complicato. Per questo motivo nei metodi iterativi, in particolare, si utilizzano più frequentemente la $\|\cdot\|_1$ e la $\|\cdot\|_\infty$, molto più semplici da calcolare.

(b) Se $A \in L(\mathbb{R}^n)$, $\|A\|_\infty = \max\limits_{i=1,\dots,n} \sum\limits_{j=1}^{n} |a_{ij}|$.

5.13. Iniziamo con l'osservazione che, per ogni $\boldsymbol{x} \in \mathbb{R}^n$ con $\|\boldsymbol{x}\|_\infty = \max\limits_{i=1,\dots,n} |x_i| = 1$, si ha:

$$\begin{aligned}\|A\boldsymbol{x}\|_\infty &= \max_{i=1,\dots,n} \left| \sum_{j=1}^{n} a_{ij}x_j \right| \le \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}|\,|x_j| \\ &\le \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}|.\end{aligned}$$

Per dimostrare che vale l'uguale, basta trovare un vettore $\boldsymbol{x}$ con $\|\boldsymbol{x}\|_\infty = 1$ per cui vale l'uguale. A tale scopo, indicata con r la riga della matrice con

$$\sum_{j=1}^{n} |a_{rj}| = \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}|,$$

definiamo $\boldsymbol{x}$ ponendo per $j = 1, 2, \dots, n$

$$x_j = \begin{cases} \frac{a_{rj}}{|a_{rj}|}, & \text{se } a_{rj} \neq 0 \\ 0, & \text{se } a_{rj} = 0. \end{cases}$$

La dimostrazione risulta pertanto completa, dato che

$$\|A\boldsymbol{x}\|_\infty = \left| \sum_{j=1}^{n} a_{rj}x_j \right| = \sum_{j=1}^{n} |a_{rj}|.$$

Calcolare la $\|\cdot\|_\infty$ di una matrice è molto semplice, a differenza di quanto avviene per il calcolo della $\|\cdot\|_2$. Come risulta dalla proprietà seguente, il costo computazionale della $\|A\|_\infty$ e $\|A\|_1$ è lo stesso, per ogni $A \in L(\mathbb{C}^n)$.

5.14. Per ogni $A \in L(\mathbb{R}^n)$, $\|A\|_1 = \max\limits_{j=1,\dots,n} \sum\limits_{i=1}^{n} |a_{ij}|$.

Dimostrazione.

$$\|A\boldsymbol{x}\|_1 = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \leq \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|\,|x_j|$$
$$= \sum_{j=1}^{n} |x_j| \sum_{i=1}^{n} |a_{ij}| \leq \left(\max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}| \right) \|\boldsymbol{x}\|_1.$$

Si tratta ora di dimostrare che esiste un vettore $\boldsymbol{x} \neq \mathbf{0}$ per cui vale l'uguale. A tale scopo, supponendo che il massimo della sommatoria valga per $j = r$, basta prendere $\boldsymbol{x} = \boldsymbol{e}_k$ (versore con 1 nella k-esima posizione e zero altrove). Con tale scelta risulta infatti $A\boldsymbol{e}_k = \boldsymbol{a}_k$, colonna k-esima di A, e la dimostrazione è completa (essendo $\|\boldsymbol{e}_k\| = 1$).

La (5.8) evidenzia che $\rho(A)$ è minore di una qualsiasi norma indotta. In realtà si può affermare che $\rho(A)$ è l'estremo inferiore delle norme indotte in quanto, alla (5.8) può essere aggiunta la seguente proprietà (di cui si tralascia la dimostrazione): prefissato ad arbitrario $\varepsilon > 0$, esiste una norma $\|A\|$ per la quale

$$\rho(A) \leq \|A\| < \rho(A) + \varepsilon. \tag{5.9}$$

In una matrice simmetrica $\rho(A) = \|A\|_2$, dato che $\|A\|_2 = \sqrt{\rho(A^T A)} = \sqrt{\rho(A)^2} = \rho(A)$, dato che $A^T = A$ e $\rho(A^2) = \rho(A)^2$.

Nella seguente matrice simmetrica $A = \left(\begin{smallmatrix} 1 & 1 & 0 \\ 1 & -1 & 1 \\ 0 & 1 & 1 \end{smallmatrix}\right)$, $\rho(A) = \sqrt{3}$ e, conseguentemente $\|A\|_2 = \rho(A)$, mentre:

$$\|A\|_\infty = \max_{i=1,2,3} \sum_{j=1}^{3} |a_{ij}| = \sum_{j=1}^{3} |a_{2j}| = 3,$$
$$\|A\|_1 = \max_{j=1,2,3} \sum_{i=1}^{3} |a_{ij}| = \sum_{i=1}^{3} |a_{i2}| = 3.$$

È importante osservare, per la rilevanza del risultato nella risoluzione dei sistemi lineari, che $\rho(A)$ può essere < 1, anche se $\|A\|_1 > 1$ e $\|A\|_\infty > 1$. Per evidenziarlo consideriamo il seguente esempio riportato in[20, pag. 55]

$$A = \begin{pmatrix} 0.1 & 0 & 1.0 & 0.2 \\ 0.2 & 0.3 & 0.1 & 0.3 \\ 0.1 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.2 & 0.2 & 0.1 \end{pmatrix},$$

per la quale $\|A\|_\infty = 1.3$ e $\|A\|_1 = 1.4$. Il suo raggio spettrale $\rho(A) < 1$. Per verificarlo, indicata con D la matrice diagonale $D = \mathrm{diag}(6, 6, 10, 6)$,

consideriamo la matrice simile alla A

$$DAD^{-1} = \begin{pmatrix} 0.1 & 0 & 0.6 & 0.2 \\ 0.2 & 0.3 & 0.06 & 0.3 \\ \frac{1}{6} & \frac{1}{6} & 0.1 & \frac{1}{3} \\ 0.1 & 0.2 & 0.12 & 0.1 \end{pmatrix}.$$

Applicando ad essa il teorema di localizzazione di Gershgorin si ha:

$$\begin{gathered} |\lambda - 0.1| \leq 0.8 \Longrightarrow -0.7 \leq \lambda \leq 0.9 \\ |\lambda - 0.3| \leq 0.56 \Longrightarrow -0.26 \leq \lambda \leq 0.86 \\ |\lambda - 0.1| \leq \tfrac{2}{3} \Longrightarrow -\tfrac{17}{30} \leq \lambda \leq \tfrac{23}{30} \\ |\lambda - 0.1| \leq 0.42 \Longrightarrow -0.32 \leq \lambda \leq 0.52. \end{gathered}$$

Di conseguenza, $\rho(A) < 1$, dato che $|\lambda| < 1$ in ciascuno dei cerchi di Gershgorin. Alla stessa conclusione si può arrivare applicando il teorema di localizzazione di Gershgorin alle colonne di A.

Convergenza di una successione. Una successione $\{\boldsymbol{x}_k\}_{k=1}^{\infty}$ in $\mathbb{R}^n$ è convergente a $\boldsymbol{x} \in \mathbb{R}^n$, secondo una prefissata $\|\cdot\|$, se $\lim_{k\to\infty} \|\boldsymbol{x}_k - \boldsymbol{x}\| = 0$. Analogamente, la successione $\{\boldsymbol{x}_k\}_{k=1}^{\infty}$ è di Cauchy (per la norma in considerazione) se, prefissato $\varepsilon > 0$, esiste un $N(\varepsilon)$ tale che, per ogni coppia $k, l > N(\varepsilon)$,

$$\|\boldsymbol{x}_k - \boldsymbol{x}_l\| < \varepsilon.$$

Per la completezza dello spazio $\mathbb{R}^n$, rispetto a una sua qualunque norma, ogni successione di Cauchy è anche convergente. Pertanto, essendo l'inverso ovvio, le due definizioni sono equivalenti.

Nei processi iterativi studiare la convergenza rispetto a una norma può essere molto più complicato che rispetto a un'altra norma. Per fortuna, come evidenziato nel seguente teorema, la convergenza secondo una norma (la più semplice da utilizzare) implica la convergenza con un'altra norma (quella di maggior interesse).

Teorema 5.4 (*Equivalenza delle norme*) *Prefissata due qualsiasi norme* $\|\cdot\|'$ *e* $\|\cdot\|''$*, esistono due costanti* $c_2 \geq c_1 > 0$*, tali che*

$$c_1\|\boldsymbol{x}\|' \leq \|\boldsymbol{x}\|'' \leq c_2\|\boldsymbol{x}\|', \quad \textit{qualunque sia il vettore } \boldsymbol{x}$$

Dimostrazione. Per rendersi conto dell'equivalenza basta osservare che se una successione $\{\boldsymbol{x}_k\}_{k=1}^{\infty}$ converge a $\boldsymbol{x}$ secondo la $\|\cdot\|'$, essendo c_2 indipendente dai vettori in considerazione,

$$\|\boldsymbol{x}_k - \boldsymbol{x}\|'' \leq c_2\|\boldsymbol{x}_k - \boldsymbol{x}\|' \Longrightarrow \|\boldsymbol{x}_k - \boldsymbol{x}\|'' \to 0.$$

Se, viceversa $\{\boldsymbol{x}^k\}_{k=1}^{\infty}$ converge e se secondo la $\|\cdot\|''$, per l'indipendenza di $\boldsymbol{e}_1$ dei vettori in considerazione,

$$c_1\|\boldsymbol{x}_k - \boldsymbol{x}\|' \leq \|\boldsymbol{x}_k - \boldsymbol{x}\|'' \Longrightarrow \|\boldsymbol{x}_k - \boldsymbol{x}\|' \to 0.$$

È questo il motivo per il quale, nei metodi iterativi, in virtù della facilità del loro calcolo, le norme più usate sono $\|\cdot\|_\infty$ e la $\|\cdot\|_1$. La convergenza secondo una delle due implica, in particolare, la convergenza secondo la $\|\cdot\|_2$. □

Prima di discutere della risoluzione numerica dei sistemi lineari, è bene soffermarsi sulla *rappresentazione spettrale* della soluzione di un sistema lineare. La sua rilevanza trascende il settore specifico, in quanto una analoga rappresentazione viene utilizzata nella risoluzione analitica dei sistemi di ODEs lineari. Iniziamo con il supporre che la matrice $A \in L(\mathbb{R}^n)$ del sistema $A\boldsymbol{x} = \boldsymbol{b}$ sia non singolare e simmetrica. In tal caso gli autovalori $\lambda_1, \ldots, \lambda_n$ sono reali e gli autovettori $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ possono essere scelti mutuamente ortonormali. Di conseguenza, si può scrivere

$$\begin{cases} A\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i, \ i = 1, \ldots, n \text{ con } |\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n| > 0 \\ \boldsymbol{u}_i^T \boldsymbol{u}_j = \delta_{ij}, \ i, j = 1, 2, \ldots, n. \end{cases} \tag{5.10}$$

Tale proprietà, in forma matriciale, può essere così rappresentata

$$AU = UD, \quad U = \begin{pmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \ldots & \boldsymbol{u}_n \end{pmatrix}, \quad D = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n).$$

Da quest'ultima, tenendo conto della ortonormalità delle colonne di U, segue che

$$A = UDU^T \Longrightarrow UDU^T\boldsymbol{x} = \boldsymbol{b}, \tag{5.11a}$$

e conseguentemente, essendo D non singolare

$$\boldsymbol{x} = UD^{-1}U^T\boldsymbol{b} = \begin{pmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \ldots & \boldsymbol{u}_n \end{pmatrix} D^{-1} \begin{pmatrix} (\boldsymbol{u}_1, \boldsymbol{b}) \\ (\boldsymbol{u}_2, \boldsymbol{b}) \\ \vdots \\ (\boldsymbol{u}_n, \boldsymbol{b}) \end{pmatrix},$$

che, in forma scalare diventa

$$\boldsymbol{x} = \sum_{j=1}^{n} \frac{(\boldsymbol{u}_j, \boldsymbol{b})}{\lambda_j} \boldsymbol{u}_j. \tag{5.11b}$$

La (5.11) fornisce la rappresentazione spettrale della soluzione. Il risultato, più che numericamente, è importante teoricamente per la proprietà da essa desumibile e per le analogie che ispira in altri settori della matematica.

Decomposizione a valori singolari (Singular Value Decomposition, SVD). Se A è non singolare, le matrici A^TA e AA^T sono simmetriche e definite positive. Di conseguenza, indicati con $\{\lambda_i = \sigma_i^2,\ i = 1, \ldots, n\}$ gli autovalori di A^TA e con $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ i relativi autovettori ortonormali, possiamo scrivere

$$A^TA\boldsymbol{u}_i = \sigma_i^2\boldsymbol{u}_i, \qquad i = 1, 2, \ldots, n. \tag{5.12}$$

Da essa, ponendo (per definizione)

$$A\boldsymbol{u}_i = \sigma_i\boldsymbol{v}_i, \qquad i = 1, 2, \ldots, n,$$

otteniamo i due sistemi

$$\left\{\begin{matrix} A^T\boldsymbol{v}_i = \sigma_i\boldsymbol{u}_i \\ A\boldsymbol{u}_i = \sigma_i\boldsymbol{v}_i \end{matrix}\right\} \qquad i = 1, 2, \ldots, n. \tag{5.13}$$

Dalla (5.12) segue immediatamente che i $\sigma_1^2, \ldots, \sigma_n^2$ sono gli autovalori di AA^T (oltre che di A^TA), i cui autovettori relativi sono i $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$. Per verificarlo basta osservare che, molteplicando la prima delle (5.13) per A

$$AA^T\boldsymbol{v}_i = \sigma_iA\boldsymbol{u}_i, \qquad i = 1, \ldots, n, \tag{5.14}$$

dalla quale, ricordando la seconda, segue che

$$A^TA\boldsymbol{v}_i = \sigma_i^2\boldsymbol{v}_i, \qquad i = 1, \ldots, n.$$

Di conseguenza, A^TA e AA^T hanno gli stessi autovalori, mentre gli autovettori sono diversi. Anche gli autovettori $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ sono ortonormali, dato che (indicato con $\delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases}$ il simbolo di Kronecker) per la seconda delle (5.13)

$$\boldsymbol{v}_j^T\boldsymbol{v}_i = \left(\frac{1}{\sigma_j}\boldsymbol{u}_j^TA^T\right)\left(\frac{1}{\sigma_i}A\boldsymbol{u}_i\right) = \frac{\sigma_i}{\sigma_j}\boldsymbol{u}_j^T\boldsymbol{u}_i = \frac{\sigma_i}{\sigma_j}\delta_{ij}, \qquad i, j = 1, \ldots, n,$$

e dunque anche i $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ sono ortonormali.

Ponendo

$$V = \begin{pmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \ldots & \boldsymbol{v}_n \end{pmatrix}, \quad U = \begin{pmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \ldots & \boldsymbol{u}_n \end{pmatrix}, \quad \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n),$$

dalla prima delle (5.13) segue che

$$AU = V\Sigma$$

e da questa, ricordando che $U^{-1} = U^T$ e che $V^{-1} = V^T$, segue la SVD di A

$$A = V\Sigma U^T \Longrightarrow A^{-1} = U\Sigma^{-1}V^T.$$

I $\sigma_1, \ldots, \sigma_n$ sono definiti i *valori singolari* di A e i vettori $\{\boldsymbol{u}_i, \boldsymbol{v}_i\}_{i=1}^n$ i suoi vettori singolari. Di conseguenza, la soluzione $\boldsymbol{x} = A^{-1}\boldsymbol{b}$ può essere rappresentata nella forma vettoriale

$$\boldsymbol{x} = U\Sigma^{-1}V^T\boldsymbol{b} = \begin{pmatrix}\boldsymbol{u}_1 & \boldsymbol{u}_2 & \ldots & \boldsymbol{u}_n\end{pmatrix} \Sigma^{-1} \begin{pmatrix}(\boldsymbol{v}_1, \boldsymbol{b}) \\ (\boldsymbol{v}_2, \boldsymbol{b}) \\ \vdots \\ (\boldsymbol{v}_n, \boldsymbol{b})\end{pmatrix} \tag{5.15a}$$

o, nella forma scalare

$$\boldsymbol{x} = \sum_{j=1}^{n} \frac{(\boldsymbol{v}_j, \boldsymbol{b})}{\sigma_j}\boldsymbol{u}_j. \tag{5.15b}$$

La (5.15) fornisce la rappresentazione singolare della soluzione di un sistema con matrice non singolare. La (5.12) implica che, ordinati i valori singolari per valori decrescenti ($\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$),

$$\|A\|_2 = \sqrt{\rho(A^TA)} = \sigma_1, \qquad \|A^{-1}\|_2 = \sqrt{\rho((A^TA)^{-1})} = \frac{1}{\sigma_n},$$

da cui segue immediatamente che (relativamente alla stessa norma)

$$\text{cond}(A) \equiv \|A\|_2\|A^{-1}\|_2 = \frac{\sigma_1}{\sigma_n}.$$

Esistono nella letteratura molti metodi per il calcolo della SVD di A e del cond(A). Algoritmi molto efficienti per il calcolo dei valori singolari e dei vettori singolari e conseguentemente anche del cond(A), rispetto alla $\|\cdot\|_2$, si trovano in MATLAB.

5.3 Sistemi lineari

Per la sua particolare rilevanza, tutti i libri di Analisi Numerica trattano l'argomento con la illustrazione di numerosi metodi diretti e iterativi [28]. Esistono altresì più riviste specializzate dedicate alla risoluzione numerica dei sistemi lineari e all'analisi delle matrici.

Si definiscono *diretti* i metodi che, con un numero finito di trasformazioni (anche se elevato) e in assenza di errori di arrotondamento, consentono di calcolare la soluzione esatta del sistema. Tra le varie tipologie esistenti, la più

classica è quella definita di tipo LU. Si tratta di metodi che riconducono la risoluzione di un sistema non singolare

$$A\boldsymbol{x} = \boldsymbol{b}$$

a quello di un sistema equivalente del tipo

$$LU\boldsymbol{x} = \boldsymbol{c}, \tag{5.16}$$

dove L è una matrice triangolare inferiore e U è triangolare superiore. Esse sono dunque del tipo

$$L = \begin{pmatrix} \ell_{11} & 0 & \dots & \dots & 0 \\ \ell_{21} & \ell_{22} & 0 & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & 0 \\ \ell_{n1} & \ell_{n2} & \dots & \dots & \ell_{nn} \end{pmatrix}, \quad U = \begin{pmatrix} u_{11} & u_{12} & \dots & \dots & u_{1n} \\ 0 & u_{22} & \dots & \dots & u_{2n} \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & u_{nn} \end{pmatrix},$$

con $\ell_{ii} \neq 0$, $i = 1, \dots, n$ e $u_{ii} \neq 0$, $i = 1, \dots, n$.

La risoluzione del sistema viene quindi ottenuta risolvendo, in seguenza, i due sistemi

$$L\boldsymbol{y} = \boldsymbol{c} \quad \text{e} \quad U\boldsymbol{x} = \boldsymbol{y}.$$

La motivazione (dovuta a Gauss) è che risolvere un sistema triangolare inferiore/superiore è piuttosto facile (per discesa nel primo caso e per risali nel secondo caso). Lo schema di risoluzione di $L\boldsymbol{x} = \boldsymbol{c}$ è infatti il seguente:

$$y_1 = \frac{c_1}{\ell_{11}}, \qquad y_i = \frac{1}{\ell_{ii}} \left(c_i - \sum_{j=1}^{i-1} \ell_{ij} y_j \right), \qquad i = 2, 3, \dots, n,$$

mentre per il secondo vale lo schema

$$x_n = \frac{y_n}{u_{nn}}, \qquad x_i = \frac{1}{u_{ii}} \left(y_i - \sum_{j=i+1}^{n} u_{ij} x_j \right), \qquad i = 1, 2, \dots, n-1.$$

Il prototipo di tale famiglia di metodi è il "metodo di eliminazione di Gauss", così definito in quanto basato su $n-1$ trasformazioni sequenziali nella prima delle quali si elimina la x_1 della seconda, terza,..., n-esima equazione, nella seconda si elimina la x_2 dalla terza, quarta, ..., n-esima e nella $(n-1)$-esima la x_{n-1}. La matrice così ottenuta è triangolare superiore e la soluzione viene quindi ottenuta con un procedimento di risalita. Naturalmente la tecnica di calcolo effettivo (come riportata in MATLAB) contiene molti accorgimenti

computazionali che hanno come obiettivo primario il contenimento della propagazione degli errori di arrotondamento. Nel metodo di eliminazione di Gauss il costo computazionale, ossia il numero delle operazioni aritmetiche necessarie per ottenere la soluzione, è dell'ordine di $n^3/3$.

Altra importante famiglia di metodi diretti è rappresentata dei "metodi QR". Si tratta di metodi nei quali la matrice A del sistema viene fattorizzata nella forma

$$A = QR, \text{ dove } Q \text{ è una matrice ortogonale } Q^TQ = QQ^T = I$$

e R è una matrice triangolare superiore con gli elementi diagonali $r_{ii} \neq 0$, $i = 1, \ldots, n$, nell'ipotesi che A sia non singolare. In questo modo la risoluzione del sistema $A\boldsymbol{x} = \boldsymbol{b}$ viene ricondotta a quella del sistema triangolare $R\boldsymbol{x} = Q^T\boldsymbol{b}$. Considerazioni del tutto analoghe valgono nel caso di un sistema a termini complessi. La fattorizzazione QR di A è ugualmente possibile con la precisione che Q è unitaria ($Q^*Q = QQ^* = I$).

Numero di condizione di una matrice. L'aritmetica con cui i computers eseguono i calcoli è finita, in quanto ogni numero reale viene sempre rappresentato con un numero finito di cifre decimali anche se, fortunatamente, elevato. Lo status attuale è di 16 cifre decimali. Numero elevato ma non infinito, come richiesto per la rappresentazione esatta di un numero irrazionale. Il procedimento di calcolo risulta, conseguentemente, del tipo: arrotondamento dei numeri, esecuzione delle equazioni aritmetiche, nuovo arrotondamento ecc.. La sequenza arrotondamento-operazione-arrotondamento comporta una più o meno marcata propagazione degli errori di arrotondamento, fortemente dipendente dalle caratteristiche della matrice. La stima dell'attitudine di una matrice A viene "misurata" del suo numero di condizione $\text{cond}(A)$, numero dipendente della norma matriciale adottata.

5.15. Definizione. Sia A una matrice non singolare in $L(\mathbb{C}^n)$ e $\|\cdot\|$ il simbolo di una norma indotta. Per *numero di condizione* di A, rispetto alla norma $\|\cdot\|$, si intende il numero

$$\text{cond}(A) = \|A\| \, \|A^{-1}\|. \tag{5.17}$$

Dalla definizione segue immediatamente che $\text{cond}(A) \geq 1$. Per rendersene conto basta osservare che, essendo $I = AA^{-1}$, e $\text{cond}(I) = 1$, qualunque sia la norma in considerazione

$$1 \leq \|A\| \, \|A^{-1}\| = \text{cond}(A).$$

Per capire come la perturbazione dei dati influenzi la soluzione del sistema, in dipendenza del suo condizionamento, consideriamo il sistema perturbato

$$(A + \delta A)(\boldsymbol{x} + \delta\boldsymbol{x}) = \boldsymbol{b} + \delta\boldsymbol{b}. \tag{5.18}$$

Una stima della variazione relativa delle soluzione può essere ottenuta, in funzione delle variazioni relative sui dati mediante il seguente

Teorema 5.5 *Se A è non singolare e la perturbazione δA non è eccessiva, nel senso che*

$$\|\delta A\| \, \|A^{-1}\| < 1, \tag{5.19}$$

supponendo che $\boldsymbol{x}$ e $\delta\boldsymbol{x}$ soddisfino il sistema iniziale e quello perturbato, si ha

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \frac{\mu}{1 - \mu\frac{\|\delta A\|}{\|A\|}} \left\{ \frac{\|\delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|} + \frac{\|\delta A\|}{\|A\|} \right\}, \text{ dove } \mu = cond(A). \tag{5.20}$$

Dimostrazione. Essendo $\|A^{-1}\delta A\| \leq \|A^{-1}\| \, \|\delta A\| < 1$, come conseguenza del teorema di Gershgorin sulla localizzazione degli autovalori, $I + A^{-1}\delta A$ è non singolare e inoltre

$$\|(I + A^{-1}\delta A)^{-1}\| \leq \frac{1}{1 - \|A^{-1}\delta A\|} \leq \frac{1}{1 - \|A^{-1}\| \, \|\delta A\|}. \tag{5.21}$$

Molteplicando la (5.18) per A^{-1} e ricordando che $A\boldsymbol{x} = \boldsymbol{b}$, si ottiene che

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

Dividendo per $\|\boldsymbol{x}\|$ e utilizzando la disuguaglianza (5.21) si trova che

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \frac{\|A\| \, \|A^{-1}\|}{1 - \|\delta A\| \, \|A^{-1}\|} \left\{ \frac{\|\delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|} + \frac{\|\delta A\|}{\|A\|} \right\}, \tag{5.22}$$

dalla quale, posto $\mu = \|A\| \, \|A^{-1}\|$, segue la (5.19). Da notare che, per l'ipotesi (5.17), aumentando il cond(A) l'errore relativo $\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|}$ cresce, sia perché viene incrementato il numeratore, sia perché decresce il denominatore. Quando la perturbazione δA è trascurabile rispetto a $\delta\boldsymbol{b}$ (come spesso avviene nelle applicazioni) la (5.20), più semplicemente, diventa

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \text{cond}(A)\frac{\|\delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}. \tag{5.23}$$

□

La (5.23) evidenzia che, sotto tale ipotesi, il cond(A) indica il massimo possibile fattore di amplificazione della perturbazione relativa dei dati sull'errore relativo della soluzione. Dall'osservazione cond$(A) \geq 1$, deriva che il condizionamento ottimale di una matrice è cond$(A) = 1$. È importante osservare che esiste una importantissima classe di matrici che godono di tale proprietà.

Matrici ortogonali. Tutte le matrici unitarie, rispetto alla norma Euclidea (norma 2) hanno numero di condizione uguale a 1.

Dimostrazione. Se U è una matrice unitaria di $L(\mathbb{C}^n)$, per definizione

$$U^*U = UU^* = I.$$

Conseguentemente, per la definizione di norma Euclidea,

$$\|U\|_2^2 = \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|U\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2} = \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\boldsymbol{x}^*U^*U\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}} = 1.$$
$$\|U^{-1}\|_2^2 = \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|U^*\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2} = \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\boldsymbol{x}^*UU^*\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}} = 1,$$

e dunque $\mathrm{cond}(U) = 1$.

Lo stesso risultato vale per le matrici ortogonali, ossia per le matrici $U \in L(\mathbb{R}^n)$ con $U^TU = UU^T = I$, come è immediato osservare adattando la dimostrazione precedente.

Qualche osservazione sul condizionamento di una matrice.

(1) Per le matrici $A, B \in L(\mathbb{C}^n)$ si ha:

$$\mathrm{cond}(AB) \leq \mathrm{cond}(A)\,\mathrm{cond}(B).$$

La disuguaglianza segue immediatamente dalle proprietà delle norme indotte, in quanto

$$\begin{aligned}\mathrm{cond}(AB) &= \|AB\|\,\|(AB)^{-1}\| = \|AB\|\,\|B^{-1}A^{-1}\| \\ &\leq \|A\|\,\|B\|\,\|B^{-1}\|\,\|A^{-1}\| = \mathrm{cond}(A)\,\mathrm{cond}(B).\end{aligned}$$

(2) Nelle fattorizzazioni QR di una matrice (Q unitraria, R triangolare superiore)

$$\mathrm{cond}(A) = \mathrm{cond}(R).$$

Basta osservare che $\mathrm{cond}(A) \leq \mathrm{cond}(Q)\,\mathrm{cond}(R) = \mathrm{cond}(R)$, come anche $\mathrm{cond}(R) = \mathrm{cond}(Q^*QR)$, dato che $Q^*Q = I$, per cui

$$\mathrm{cond}(R) \leq \mathrm{cond}(Q^*)\,\mathrm{cond}(QR) = \mathrm{cond}(A),$$

dato che $\mathrm{cond}(Q^*) = 1$. È questo il motivo per cui, nella risoluzione dei sistemi lineari, vengono molto utilizzati i metodi QR. In tal caso,

$$A\boldsymbol{x} = \boldsymbol{b} \Longleftrightarrow QR\boldsymbol{x} = \boldsymbol{b} \Longleftrightarrow R\boldsymbol{x} = Q^*\boldsymbol{b},$$

sistema facile da risolvere in quanto R è triangolare superiore. Da notare che $\text{cond}(R) = \text{cond}(A)$, proprietà importante non valida nei metodi LU, il prototipo del quale è il metodo di eliminazione di Gauss. In questo caso il sistema iniziale $A^{(1)}\boldsymbol{x} = \boldsymbol{b}^{(1)}$ viene trasformato in uno del tipo

$$A^{(n)}\boldsymbol{x} = \boldsymbol{b}^{(n)}, \text{ dove } A^{(n)} \text{ è triangolare superiore,}$$

con $\text{cond}(A^{(n)})$ talvolta molto superiore a quello di $\text{cond}(A^{(1)})$.

Osservazione. Stimare il condizionamento di una matrice, sulla base della definizione, è generalmente impossibile, dato che non si conosce la sua inversa. Ci sono comunque classi di matrici importanti per le quali non è difficile. Se, ad esempio, la matrice A è simmetrica e la norma indotta è quella Euclidea,

$$\text{cond}(A) = \frac{\max\{|\lambda_i|,\ i = 1, 2, \ldots, n\}}{\min\{|\lambda_i|,\ i = 1, 2, \ldots, n\}} = \frac{|\lambda_1|}{|\lambda_n|}$$

nell'ipotesi che gli autovalori di A siano ordinati per autovalori in modulo decrescente ($|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n| > 0$). Per la dimostrazione basta osservare che

$$\begin{aligned} \|A\|_2 &= \rho(A) = |\lambda_1|, \\ \|A^{-1}\| &= \rho(A^{-1}) = \max_{i=1,\ldots,n} \frac{1}{|\lambda_i|} = \frac{1}{|\lambda_n|}, \\ &\text{dato che lo spettro di } A^{-1} = \{\tfrac{1}{\lambda_i}, i = 1, \ldots, n\}. \end{aligned}$$

In molte situazioni applicative, l'analisi tecnica del problema fisico iniziale, o del modello matematico che lo rappresenta, fornisce una importante indicazione sul condizionamento del problema. Questo è, tipicamente, dipendente dalla stabilità della soluzione rispetto ai dati del problema (es. continuità della soluzione dai dati iniziali).

Spesso, empiricamente, si ritiene di poter evitare la stima del condizionamento della matrice, nella risoluzione del sistema $A\boldsymbol{x} = \boldsymbol{b}$, ricorrendo ad una valutazione a posteriori della approssimazione ottenuta. Più precisamente, indicata con $\tilde{\boldsymbol{x}}$ la "soluzione numerica" fornita del metodo, si ritiene $\tilde{\boldsymbol{x}}$ accettabile se la norma ∞ del residuo

$$r(\tilde{\boldsymbol{x}}) = \boldsymbol{b} - A\tilde{\boldsymbol{x}}$$

è sufficientemente piccola.

Il seguente esempio evidenzia che, quando il $\text{cond}(A)$ è elevato, tale criterio non ha alcun fondamento.

Esempio 5.6 Consideriamo il sistema lineare

$$\begin{cases} x + y = 2 \\ x + (1+\alpha)y = 2, \end{cases}$$

dove $0 < \alpha < 2$, con soluzione esatta $\boldsymbol{x} = \left(\begin{smallmatrix} 2 \\ 0 \end{smallmatrix}\right)$. Sulla base del suddetto criterio empirico, i due vettori

$$\boldsymbol{x}' = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \boldsymbol{x}'' = \begin{pmatrix} 3 \\ -1 \end{pmatrix}$$

(anche se molto diversi) hanno la stessa attendibilità, dato che

$$r(\boldsymbol{x}') = r(\boldsymbol{x}'') = \begin{pmatrix} 0 \\ -\alpha \end{pmatrix}.$$

Nel caso α sia sufficientemente piccolo ambedue i vettori dovrebbero essere considerati accettabili, nonostante siano ambedue molto distanti da $\boldsymbol{x}$. Questo dipende dal fatto che

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1+\alpha \end{pmatrix} \quad \text{e} \quad A^{-1} = \frac{1}{\alpha} \begin{pmatrix} 1+\alpha & -1 \\ -1 & 1 \end{pmatrix}.$$

Per cui essendo, relativamente alla norma $\|\cdot\|_\infty$,

$$\mathrm{cond}(A) = \frac{(2+\alpha)^2}{\alpha},$$

il condizionamento di A cresce tra 8 e ∞ al decrescere di α tra 2 e 0. Questo tipo di errore di valutazione non si presenta se α è "abbastanza" grande. Infatti se α non è piccolo, lo stesso criterio porta ad escludere sia $\boldsymbol{x}'$ sia $\boldsymbol{x}''$, dato che $\|r(\boldsymbol{x}')\|_\infty = \|r(\boldsymbol{x}'')\|_\infty = \alpha$.

Altro errore frequente consiste nel considerare "sperimentalmente provato" che il $\det(A)$ è un buon stimatore del $\mathrm{cond}(A)$. Questa ipotesi, pur avendo una certa attendibilità sperimentale, non è sempre valida, come dimostra il seguente semplicissimo esempio:

$$A = \begin{pmatrix} 10^{-6} & 0 \\ 0 & 10^{-6} \end{pmatrix} \Longrightarrow \det(A) = 10^{-12} \quad \text{e} \quad \mathrm{cond}(A) = 1.$$

Come l'esempio precedente dimostra (per α molto piccolo) una matrice può avere determinante piccolo ed essere mal condizionata. L'esempio seguente [11,

pag. 82] dimostra che il determinante può essere molto piccolo e la matrice ben condizionata.

$$A = \begin{pmatrix} 1 & -1 & \dots & \dots & \dots & -1 \\ 0 & 1 & -1 & & & -1 \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & & & \ddots & \vdots \\ \vdots & & & & \ddots & -1 \\ 0 & \dots & \dots & \dots & 0 & 1 \end{pmatrix}$$

In questo caso $\det(A) = 1$ e $\text{cond}(A) = n\,2^{n-1}$ relativamente alla $\|\cdot\|_\infty$.

Come risulterà chiaro nella risoluzione dei sistemi lineari con metodi iterativi, altra definizione fondamentale sulle matrici è quella di matrice convergente.

Definizione. Una matrice A è definita convergente se, indicata con O la matrice nulla (matrice con elementi tutti nulli),

$$\lim_{n\to\infty} A^n = O.$$

Il seguente teorema dimostra che essa può essere espressa anche in altre forme, spesso più utili.

Teorema 5.7 *Le seguenti affermazioni sono infatti equivalenti:*

(a) A *è convergente;*

(b) $\lim_{n\to\infty} \|A^n\| = 0$*, per una qualche norma matriciale;*

(c) $\rho(A) < 1$.

Dimostrazione. $(a) \Longrightarrow (b)$ dato che, per la continuità delle norme, $A^n \to O \Longrightarrow \|A^n\| \to 0$.

$(b) \Longrightarrow (a)$. Infatti se esiste una norma $\|\cdot\|$ per la quale $\|A^n\| \to 0$, per l'equivalenza delle norme, esiste una costante positiva c con $\|A^n\|_\infty \le c\|A^n\|$ la quale implica che $\|A^n\|_\infty \to 0$ e dunque $A^n \to 0$.

$(b) \Longrightarrow (c)$. Per l'equivalenza delle norme, basta limitarsi a considerare le norme indotte. Ricordando allora che se λ è un autovalore di A, λ^n lo è di A^n e che il raggio spettrale di una matrice è $\le$ di una sua qualsiasi norma,

$$\|A^n\| \ge \rho(A^n) = \rho(A)^n \quad \text{e pertanto } \rho(A)^n \to 0, \text{ ossia } \rho(A) < 1.$$

$(c) \Longrightarrow (b)$. nfatti, per la (5.8), esiste una norma indotta per la quale (essendo $\rho(A) < 1$),

$$\|A\| \le \rho(A) + \varepsilon < 1$$

e questo implica che $\|A^n\| \to 0$, dato che

$$\|A^n\| \leq \|A\|^n \quad \text{e} \quad \|A\| < 1.$$

L'equivalenza $(a) \implies (c)$ è automaticamente verificata, essendo già stato dimostrato che $(a) \implies (b) \implies (c)$. □

Strettamente connessa con la definizione di matrice convergente è la convergenza delle serie geometriche che intervengono frequentemente nei metodi iterativi.

Teorema 5.8 *La serie geometrica*

$$I + A + A^2 + \ldots + A^n + \ldots$$

è convergente se e solo se A è convergente. In tal caso la matrice $I - A$ è non singolare e

$$(I - A)^{-1} = I + A + A^2 + \ldots + A^n + \ldots. \tag{5.24}$$

Dimostrazione. La necessità della convergenza della matrice per la convergenza della serie è immediata. La sufficienza deriva dal fatto che, se A è convergente, esiste una norma per la quale $\|A\| < 1$ e conseguentemente (qualunque sia n)

$$\begin{aligned} \|S_n\| &= \|I + A + A^2 + \ldots + A^n\| \leq 1 + \|A\| + \ldots + \|A\|^n \\ &= \frac{1 - \|A\|^{n+1}}{1 - \|A\|} \overset{n\to\infty}{\longrightarrow} \frac{1}{1 - \|A\|}. \end{aligned}$$

Essendo A convergente per ipotesi, $\rho(A) < 1$ e quindi $I - A$ è non singolare, dato che ad ogni autovalore λ di A corrisponde l'autovalore $1 - \lambda$ di $I - A$ e, ovviamente,

$$|1 - \lambda| > |1 - |\lambda|| > 0.$$

Inoltre, osservato che (qualunque sia n)

$$(I - A)(I + A + \ldots + A^n) = I - A^{n+1},$$

la convergenza della A implica che

$$\lim_{n\to\infty} (I - A)(I + A + \ldots + A^n) = I$$

e questo completa la dimostrazione. □

Corollario 5.9 *Se esiste una norma indotta per la quale* $\|A\| < 1$, $I - A$ *è non singolare e inoltre*

$$\frac{1}{1+\|A\|} \leq \|(I-A)^{-1}\| \leq \frac{1}{1-\|A\|}. \tag{5.25}$$

Dimostrazione. $\|A\| < 1 \Longrightarrow \rho(A) < 1$, per cui $I - A$ è non singolare. Poiché per ogni norma indotta $\|I\| = 1$, dalla identità

$$I = (I-A)(I-A)^{-1}$$

segue che

$$1 \leq \|I-A\|\,\|(I-A)^{-1}\| \leq (1+\|A\|)\|(I-A)^{-1}\|$$

e questo dimostra la prima disuguaglianza. Inoltre, dalla identità

$$(I-A)^{-1} = I + A(I-A)^{-1}$$

segue che

$$\|(I-A)^{-1}\| \leq 1 + \|A\|\,\|(I-A)^{-1}\|$$

dalla quale, essendo $\|A\| < 1$,

$$\|(I-A)^{-1}\| \leq \frac{1}{1-\|A\|}$$

la cui validità completa la dimostrazione. □

Naturalmente, essendo $\|-A\| = \|A\|$, per ogni matrice A con $\|A\| < 1$, risulta anche

$$\frac{1}{1+\|A\|} \leq \|(I+A)^{-1}\| \leq \frac{1}{1-\|A\|}.$$

Osservazione. La (5.25) viene spesso utilizzata per stimare la norma ∞ dell'inversa di matrici tridiagonali che intervengono nella risoluzione numerica dei problemi agli estremi per ODEs lineari del secondo ordine. La sua valutazione permette quindi la stima, rispetto alla norma ∞, del suo numero di condizione. Sia, ad esempio,

$$A = \begin{pmatrix} 9 & -3 & 0 & \dots & \dots & \dots & 0 \\ 1 & -6 & 1 & 0 & & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & & \ddots & 1 & 6 & 1 \\ 0 & \dots & \dots & \dots & 0 & -3 & 9 \end{pmatrix}.$$

Indicata con D' la matrice diagonale con $(D')_{ii} = d'_{ii} = \frac{1}{a_{ii}}$, si considera la decomposizione $D'A = I + A'$ e si osserva che $\|A'\|_\infty = \frac{1}{3}$, dato che

$$A' = \begin{pmatrix} 0 & \frac{1}{3} & 0 & \dots & \dots & \dots & 0 \\ -\frac{1}{6} & 0 & -\frac{1}{6} & 0 & & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & & \ddots & -\frac{1}{6} & 0 & -\frac{1}{6} \\ 0 & \dots & \dots & \dots & 0 & -\frac{1}{3} & 0 \end{pmatrix}.$$

Di conseguenza per la (5.25)

$$\|(D'A)^{-1}\|_\infty = \|(I + A')^{-1}\|_\infty \leq \frac{1}{1 - \|A'\|_\infty} = \frac{3}{2}$$

ed infine, dato che $A^{-1} = (I + A')^{-1}D'$,

$$\|A^{-1}\|_\infty \leq \|(I + A')^{-1}\|_\infty \|D'\|_\infty \leq \frac{3}{2} \cdot \frac{1}{6} = \frac{1}{4}.$$

La matrice è pertanto ben condizionata in quanto, utilizzando la norma ∞,

$$\text{cond}(A) \leq \frac{1}{4} \cdot 12 = 3.$$

Tale procedimento è utilizzabile per calcolare, con semplici adattamenti formali, la norma dell'inversa di una qualsiasi matrice diagonalmente dominante, ossia di una matrice A per la quale

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

Procedimento che, ovviamente, consenta anche il calcolo del cond(A) relativamente alla norma ∞.

5.4 Metodi iterativi

L'idea alla base dei metodi iterativi è la seguente: indicato con

$$A\boldsymbol{x} = \boldsymbol{b} \tag{5.26}$$

il sistema non singolare da risolvere, si decompone la matrice A nella forma

$$A = N - P,$$

con l'unica limitazione che N sia non singolare. Il sistema (5.26) diventa pertanto

$$\boldsymbol{x} = N^{-1}P\boldsymbol{x} + N^{-1}\boldsymbol{b}. \tag{5.27}$$

Indicato allora con $\boldsymbol{x}^{(0)}$ un vettore iniziale, si costruisce la successione $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ generata dalla relazione di ricorrenza

$$\boldsymbol{x}^{(k)} = N^{-1}P\boldsymbol{x}^{(k-1)} + N^{-1}\boldsymbol{b}. \tag{5.28}$$

Sostituendo $P = N - A$ nella (5.28), si ottiene la relazione di ricorrenza, spesso più rapidamente convergente,

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} - N^{-1}A\boldsymbol{x}^{(k-1)} + N^{-1}\boldsymbol{b} = (I - N^{-1}A)\boldsymbol{x}^{(k-1)} + N^{-1}\boldsymbol{b}. \tag{5.29}$$

Teorema 5.10 *Posto $M = N^{-1}P$, la successione $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ converge alla soluzione esatta, secondo una qualsiasi norma, se e solo se la matrice M ha raggio spettrale inferiore ad uno.*

Dimostrazione. Sia

$$\boldsymbol{e}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}, \qquad k = 0, 1, 2, \ldots,$$

la successione dei vettori errore generata dal metodo iterativo. Indicato con $\boldsymbol{e}^{(0)}$ il vettore errore iniziale, dalle relazioni (5.27)-(5.28) deriva che

$$\boldsymbol{e}^{(k)} = M\boldsymbol{e}^{(k-1)} = M^2\boldsymbol{e}^{(k-2)} = \ldots = M^k\boldsymbol{e}^{(0)}, \qquad k = 1, 2, \ldots. \tag{5.30}$$

Dalla (5.30) segue facilmente che $\|\boldsymbol{e}^{(k)}\| \to 0$, qualunque sia $\boldsymbol{x}^{(0)}$ se e solo se $\rho(M) < 1$, essendo $\rho(M)$ il raggio spettrale di M.

La dimostrazione è conseguenza delle seguenti tre proprietà basilari:

(1) In $\mathbb{C}^n$ (oppure $\mathbb{R}^n$) tutte le norme sono equivalenti, nel senso che se una successione $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ converge a zero secondo una norma, converge a zero seconda una qualsiasi altra norma. In altre parole, qualunque siano le norme $\|\cdot\|'$ e $\|\cdot\|''$, $\|\boldsymbol{x}^{(k)}\|' \to 0 \Longrightarrow \|\boldsymbol{x}^{(k)}\|'' \to 0$ e viceversa $\|\boldsymbol{x}^{(k)}\|'' \to 0 \Longrightarrow \|\boldsymbol{x}^{(k)}\|' \to 0$.

(2) Il raggio spettrale $\rho(A)$ di una qualsiasi matrice A è l'estremo inferiore delle norme indotte, ossia, qualunque sia la norma indotta $\|\cdot\|$,

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

e inoltre, qualunque sia $\varepsilon > 0$, esiste una norma con $\|A\| < \rho(A) + \varepsilon$.

(3) Per ogni norma matriciale indotta e qualunque siano $\boldsymbol{x} \in \mathbb{C}^n$ ($\mathbb{R}^n$) e $A \in L(\mathbb{C}^n)$ ($L(\mathbb{R}^n)$)

$$\|A\boldsymbol{x}\| \leq \|A\| \, \|\boldsymbol{x}\|.$$

Per la conclusione osserviamo che, se $\rho(M) < 1$, per la (2) esiste una norma indotta $\|\cdot\|$ con $\|M\| < 1$ e pertanto

$$\|\boldsymbol{e}^{(k)}\| \leq \|M\| \, \|\boldsymbol{e}^{(k-1)}\| \leq \ldots \leq \|M\|^k \, \|\boldsymbol{e}^{(0)}\|,$$

da cui $\|\boldsymbol{e}^{(k)}\| \to 0$, qualunque sia $\|\boldsymbol{e}^{(0)}\|$, dato che $\|M\|^k \to 0$ per $k \to \infty$. Se, viceversa, $\|\boldsymbol{e}^{(k)}\| \to 0$, qualunque sia $\|\boldsymbol{e}^{(0)}\|$, necessariamente $\|M^k\| \to 0$ (cioè, M è convergente) e conseguentemente $\rho(M) < 1$. Da cui deriva che, essendo $\rho(M)$ l'estremo inferiore delle norme indotte di M, esiste una norma con $\|M\| < 1$. □

Quando un metodo iterativo è convergente, come indicatore della velocità di convergenza si considera il numero

$$R = -\log_{10} \rho(M)$$

definito *indice di convergenza*. La conoscenza di R permette di stabilire il numero minimo di iterazioni necessarie per ridurre la norma di un vettore errore di un prefissato fattore. Infatti, in base alla (5.30), qualunque norma naturale si consideri, per ogni k

$$\|\boldsymbol{e}^{(k)}\| \leq \|M\|^k \|\boldsymbol{e}^{(0)}\|.$$

Pertanto, poiché $\rho(M)$ è l'estremo inferiore delle norme, il numero minimo k di iterazioni richieste per ridurre la norma dell'errore iniziale di un fattore 10^{-m} è così definito:

$$|\rho(M)|^k \leq 10^{-m},$$

ossia

$$k \geq \frac{m}{R},$$

con $R = -\log_{10} \rho(M)$.

Metodo di Jacobi e delle iterazioni simultanee. Il metodo di Jacobi è caratterizzato dalla scelta

$$N_{ij} = a_{ij}\delta_{ij}, \qquad i, j = 1, \ldots, n,$$

dove δ_{ij} è il simbolo di Kronecker. Poiché N deve essere non singolare, il metodo è applicabile solo se tutti gli elementi diagonali di A sono non nulli. In tale ipotesi la matrice di iterazione M è definita dal seguente schema:

$$M = N^{-1}P = I - N^{-1}A,$$

da cui

$$M_{ij} = \begin{cases} 0, & \text{per } i = j, \\ -\dfrac{a_{ij}}{a_{ii}}, & \text{per } i \neq j. \end{cases}$$

Pertanto, in base alla (5.27), il relativo schema iterativo è definitivo dalla seguente relazione di ricorrenza:

$$x_i^{(k)} = \frac{1}{a_{ii}} \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(k-1)} \right), \qquad i = 1, \ldots, n, \ k = 1, 2, \ldots. \tag{5.31}$$

Teorema 5.11 *Condizione sufficiente perché il metodo di Jacobi sia convergente è che la matrice A sia diagonalmente dominante in senso stretto.*

Dimostrazione. La suddetta ipotesi esprime il fatto che

$$\|M\|_\infty = \max_{i=1,\ldots,n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| < 1.$$

Poiché $\rho(M) \leq \|M\|_\infty$, anche $\rho(M) < 1$, e di conseguenza il metodo è convergente, qualunque sia il vettore iniziale $\boldsymbol{x}^{(0)}$.

Dal momento che $\rho(M) \leq \|M\|_1$, la stessa conclusione vale se $\|M\|_1 < 1$, cioè se la matrice A è strettamente diagonalmente dominante per colonne. □

Per la semplicità con cui $\|M\|_\infty$ e $\|M\|_1$ possono essere calcolate, oltre che per dimostrare la convergenza del metodo, $\|M\|_\infty$ e $\|M\|_1$ vengono spesso usate per stimare l'indice di convergenza del metodo. A tale scopo ci si serve della seguente disuguaglianza:

$$R = \log_{10} \frac{1}{\rho(M)} \geq \min \left\{ \log_{10} \frac{1}{\|M\|_\infty}, \frac{1}{\|M\|_1} \right\}.$$

Nel metodo di Jacobi il numero di molteplicazioni/divisioni per iterazione è dell'ordine di n^2 e pertanto, se il metodo converge, per ridurre di un fattore di 10^{-m} la norma del vettore errore iniziale, occorrono circa $n^2 \frac{m}{R}$ molteplicazioni/divisioni. Per questo motivo si considera il metodo di Jacobi computazionalmente competitivo, rispetto ai metodi diretti, soltanto se $\frac{m}{R} \leq \frac{n}{3}$.

Esercizio 5.12 Risolvere con il metodo di Jacobi il seguente sistema:

$$\begin{cases} 10x_1 + x_2 + x_3 = 12 \\ 2x_1 + 10x_2 + x_3 = 13 \\ 2x_1 + 2x_2 + 10x_3 = 14. \end{cases}$$

Il metodo è convergente perché la matrice dei coefficienti

$$A = \begin{pmatrix} 10 & 1 & 1 \\ 2 & 10 & 1 \\ 2 & 2 & 10 \end{pmatrix}$$

è diagonalmente dominante in senso stretto. Infatti gli elementi diagonali di A sono, in valore assoluta, maggiori della somma dei valori assoluti degli elementi non diagonali lungo le tre righe.

La soluzione esatta del sistema è $x_1 = x_2 = x_3 = 1$. Sia $\boldsymbol{x}^{(0)} = (1.2, 0, 0)^T$ il vettore iniziale. Iterando si ottiene:

$$\begin{cases} x_1^{(1)} = \frac{1}{10}(12 - 0 - 0) = 1.2 \\ x_2^{(1)} = \frac{1}{10}(13 - 2.4 - 0) = 1.06 \\ x_3^{(1)} = \frac{1}{10}(14 - 2.4 - 0) = 1,16, \end{cases} \qquad \begin{cases} x_1^{(2)} = \frac{1}{10}(12 - 1.06 - 1.16) = 0.978 \\ x_2^{(2)} = \frac{1}{10}(13 - 2.4 - 1.16) = 0.972 \\ x_3^{(2)} = \frac{1}{10}(14 - 2.4 - 2.12) = 0.948, \end{cases}$$

$$\begin{cases} x_1^{(3)} = \frac{1}{10}(12 - 0.972 - 0.948) = 1.008 \\ x_2^{(3)} = \frac{1}{10}(13 - 1.956 - 0.948) = 1.0096 \\ x_3^{(3)} = \frac{1}{10}(14 - 1.956 - 1.944) = 1.01. \end{cases}$$

Iterando ancora ed arrontondando alla quarta cifra si ottiene:

$$\begin{aligned} &x_1^{(4)} = 0.9980, \quad x_2^{(4)} = 0.9974, \quad x_3^{(4)} = 0.9965, \\ &x_1^{(5)} = 1.0006, \quad x_2^{(5)} = 1.0008, \quad x_3^{(5)} = 1.0009, \\ &x_1^{(6)} = 0.9998, \quad x_2^{(6)} = 0.9998, \quad x_3^{(6)} = 0.9997, \\ &x_1^{(7)} = 1.0000, \quad x_2^{(7)} = 1.0000, \quad x_3^{(7)} = 1.0000. \end{aligned}$$

Da cui, osservato che la successione esatta del sistema è $x_1 = x_2 = x_3 = 1$, segue che in *norma infinito*, gli errori relativi ai suddetti vettori di iterazione sono:

$$\begin{aligned} &\|\boldsymbol{e}^{(0)}\| = 1, \quad \|\boldsymbol{e}^{(1)}\| = 0.2, \quad \|\boldsymbol{e}^{(2)}\| = 0.052, \quad \|\boldsymbol{e}^{(3)}\| = 0.01, \\ &\|\boldsymbol{e}^{(4)}\| = 0.0035, \quad \|\boldsymbol{e}^{(5)}\| = 0.0009, \quad \|\boldsymbol{e}^{(6)}\| = 0.0003, \quad \|\boldsymbol{e}^{(7)}\| = 0. \end{aligned}$$

Metodo di Gauss-Seidel e delle iterazioni successive. Il metodo di Gauss-Seidel può essere interpretato come una modifica di quello di Jacobi. Infatti è da esso ottenibile con la sola variante che ogni componente di $\boldsymbol{x}^{(k)}$ già calcolata, viene utilizzata nella stessa iterazione per calcolare le componenti successive. Invece nel metodo di Jacobi le componenti di $\boldsymbol{x}^{(k)}$ calcolate al passo k vengono utilizzate soltanto nella $(k+1)$-esima iterazione. Pertanto lo schema iterativo di Gauss-Seidel è così definito:

$$x_i^{(k)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} \right), \tag{5.32}$$

dove, per ogni $k = 1, 2, \ldots$, $i = 1, 2, \ldots, n$. Si può dimostrare che la decomposizione di A che genera lo schema (5.32) è la seguente:

$$N = \begin{pmatrix} a_{11} & 0 & \ldots & 0 \\ a_{21} & a_{22} & & 0 \\ \vdots & & \ddots & \vdots \\ a_{n1} & \ldots & \ldots & a_{nn} \end{pmatrix}, \qquad P = N - A.$$

Poiché $\det(N) = \prod_{i=1}^{n} a_{ii}$, tale decomposizione è utilizzabile soltanto se tutti gli elementi diagonali di A sono non nulli, come avviene nel metodo di Jacobi.

Dalla (5.32) si ricava immediatamente la seguente rappresentazione del k-esimo vettore errore

$$e_i^{(k)} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} e_j^{(k)} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} e_j^{(k-1)}, \tag{5.33}$$

con $i = 1, 2, \ldots, n$.

Le (5.32) e (5.33) per $i = 1$ e $i = n$ devono essere così interpretate:

$$\begin{cases} x_1^{(k)} = \dfrac{1}{a_{11}} \left(b_1 - \displaystyle\sum_{j=2}^{n} a_{ij} x_j^{(k-1)} \right) \\ e_1^{(k)} = -\displaystyle\sum_{j=2}^{n} \dfrac{a_{ij}}{a_{11}} e_j^{(k-1)}, \end{cases} \qquad \begin{cases} x_n^{(k)} = \dfrac{1}{a_{nn}} \left(b_n - \displaystyle\sum_{j=1}^{n-1} a_{nj} x_j^{(k)} \right) \\ e_n^{(k)} = -\displaystyle\sum_{j=1}^{n-1} \dfrac{a_{nj}}{a_{nn}} e_j^{(k)}. \end{cases}$$

Teorema 5.13 *Qualunque sia il vettore iniziale $\boldsymbol{x}^{(0)}$, se la matrice A è strettamente diagonalmente dominante, cioè se*

$$r = \max_{1 \le i \le n} r_i < 1, \qquad r_i = \sum_{\substack{j=1 \\ j \ne i}}^{n} \left| \frac{a_{ij}}{a_{ii}} \right|$$

risulta

$$\|\boldsymbol{e}^{(k)}\|_\infty \le r^k \|\boldsymbol{e}^{(0)}\|_\infty, \qquad k = 1, 2, \ldots, n, \tag{5.34}$$

e pertanto il metodo è convergente.

Dimostrazione. La validità della (5.34) può essere accertata per induzione relativamente alle componenti di $\boldsymbol{e}^{(k)}$. Nella suddetta ipotesi, per la prima componente di $\boldsymbol{e}^{(k)}$ si ha che

$$\begin{aligned} |e_1^{(k)}| &\le \sum_{j=2}^{n} \left| \frac{a_{1j}}{a_{11}} \right| |e_j^{(k-1)}| \le \|\boldsymbol{e}^{(k-1)}\|_\infty \sum_{j=2}^{n} \left| \frac{a_{1j}}{a_{11}} \right| \\ &= \|\boldsymbol{e}^{(k-1)}\|_\infty r_1 \le r \|\boldsymbol{e}^{(k-1)}\|_\infty. \end{aligned}$$

Supponendo ora che $|e_h^{(k)}| \leq r\|\boldsymbol{e}^{(k-1)}\|_\infty$, $h = 1, \ldots, i-1$, si deve dimostrare che $|e_i^{(k)}| \leq r\|\boldsymbol{e}^{(k-1)}\|_\infty$.

Dalla (5.33), in conseguenza della suddetta ipotesi, si ha che

$$\begin{aligned}
|e_i^{(k)}| &\leq \sum_{j=1}^{i-i} \left|\frac{a_{ij}}{a_{ii}}\right| |e_j^{(k)}| + \sum_{j=i+1}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| |e_j^{(k-1)}| \\
&\leq \|\boldsymbol{e}^{(k-1)}\|_\infty \left\{ \sum_{j=1}^{i-i} r \left|\frac{a_{ij}}{a_{ii}}\right| + \sum_{j=i+1}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| \right\} \\
&\leq \|\boldsymbol{e}^{(k-1)}\|_\infty \sum_{\substack{j=1 \\ j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| = r_i \|\boldsymbol{e}^{(k-1)}\|_\infty < r\|\boldsymbol{e}^{(k-1)}\|_\infty.
\end{aligned}$$

La (5.34) rimane dunque dimostrata, essendo

$$\|\boldsymbol{e}^{(k)}\|_\infty < r\|\boldsymbol{e}^{(k-1)}\|_\infty < \ldots < r^k\|\boldsymbol{e}^{(0)}\|_\infty.$$

□

Pertanto, quando la matrice del sistema è diagonalmente dominante in senso stretto, risultanto convergenti sia il metodo di Jacobi che quello di Gauss-Seidel (si veda Es. (5.12)). Esistono tuttavia degli esempi, con matrici non diagonalmenti dominanti, per i quali si ha convergenza con il metodo di Jacobi e non quello di Gauss-Seidel e viceversa. Quando entrambi i metodi convergono, pur non essendovi alcuna dimostrazione in questo senso, si ritiene generalmente più velocemente convergente il metodo di Gauss-Seidel.

L'ipotesi della dominanza diagonale in senso stretto può essere indebolita a condizione che la matrice sia irriducibile. Per introdurre tale definizione permettiamo le definizioni di matrice di permutazione e di matrice riducibile.

Una matrice P è detta di *permutazione* se le sue righe o le sue colonne sono semplice permutazioni della matrice identità.

Esempio 5.14

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad P = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

Nel primo caso P è ottenuta scambiando la seconda riga della matrice identità con la prima e nel secondo caso scambiando la prima con la terza. È evidente che P non è singolare e che $P^2 = I$ e $P^{-1} = P$.

Una matrice A è detta *riducibile* se esiste una matrice di permutazione P tale che PAP^{-1} risulti triangolare superiore o triangolare inferiore a blocchi.

Questo significa che permutando opportunamente le righe e le colonne di A si può pervenire ad una delle due seguenti forme:

$$PAP^{-1} = \begin{pmatrix} \hat{A}_{11} & \hat{A}_{12} \\ O & \hat{A}_{22} \end{pmatrix}, \text{ oppure } PAP^{-1} = \begin{pmatrix} \hat{A}_{11} & O \\ \hat{A}_{21} & \hat{A}_{22} \end{pmatrix},$$

dove le matrici diagonali sono quadrate e O indica una matrice (non necessariamente quadrata) di elementi tutti nulli.

Una matrice non riducibile, ossia per la quale non esiste una matrice di permutazione che operando su righe e colonne possa ridurla ad una forma triangolare a blocchi, è detta irriducibile. La matrice

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

è riducibile, perché la matrice di permutazione che scambia la seconda con la terza riga la riduce alla seguente

$$PAP^{-1} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \text{ essendo } P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Una classe importante di matrici irriducibili è rappresentata dalle matrici tridiagonali del tipo

$$A = \begin{pmatrix} a_1 & b_1 & & & & & & \\ c_2 & a_2 & b_2 & & & & & \\ & c_3 & a_3 & b_3 & & & & \\ & & c_4 & a_4 & b_4 & & & \\ & & & c_5 & a_5 & b_5 & & \\ & & & & c_6 & a_6 & b_6 & \\ & & & & & c_7 & a_7 & b_7 \\ & & & & & & c_8 & a_8 \end{pmatrix}$$

con $b_i \neq 0$, $i = 1, \ldots, n-1$ e $c_j \neq 0$, $j = 2, \ldots, n$.

Questo ovviamente implica la irriducibilità di ogni matrice nella quale siano nulli tutti gli elementi della prima sopra e sottodiagonale. Esistono tecniche standard per verificare in modo automatico la eventuale riducibilità di una matrice, ormai presenti nel software algebrico di qualità per la risoluzione dei sistemi lineari. Accertare questo è importante in quanto la riducibilità di una matrice permette di ricondurre la risoluzione di un sistema lineare a

quella di due sistemi lineari, ciascuno di dimensione ridotta. Per evidenziarlo, supponiamo che la matrice A del sistema

$$A\boldsymbol{x} = \boldsymbol{b}, \qquad A \in \mathbb{R}^{n\times m},$$

sia riducibile alla forma triangolare superiore mediante una matrice di permultazione P. Utilizzando P il sistema può essere trasformato nella forma equivalente

$$\hat{A}\boldsymbol{y} = \boldsymbol{c}, \quad \hat{A} = PAP^{-1}, \quad \boldsymbol{y} = P\boldsymbol{x}, \quad \boldsymbol{c} = P\boldsymbol{b},$$

con $\hat{A}_{11}$ e $\hat{A}_{22}$ non singolari.

Ponendo $\boldsymbol{c} = \begin{pmatrix} \boldsymbol{c}_1 \\ \boldsymbol{c}_2 \end{pmatrix}$, $\boldsymbol{c}_1 \in \mathbb{R}^p$, $\boldsymbol{c}_2 \in \mathbb{R}^{n-p}$ e $\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix}$, $\boldsymbol{y}_1 \in \mathbb{R}^p$, $\boldsymbol{y}_2 \in \mathbb{R}^{n-p}$, il sistema $\hat{A}\boldsymbol{c} = \boldsymbol{c}$ può essere espresso nella forma

$$\begin{cases} \hat{A}_{11}\boldsymbol{y}_1 + \hat{A}_{12}\boldsymbol{y}_2 = \boldsymbol{c}_1 \\ \hat{A}_{22}\boldsymbol{y}_2 = \boldsymbol{c}_2, \end{cases}$$

da cui, ottenuto $\boldsymbol{y}_2 = \hat{A}_{22}^{-1}\boldsymbol{c}_2$, per risalita si calcola $\boldsymbol{y}_1$ dal primo, ottenendo $\boldsymbol{y}_1 = \hat{A}_{11}^{-1}(\boldsymbol{c}_1 - \hat{A}_{12}\boldsymbol{y}_2)$. Il procedimento è del tutto analogo nel caso la matrice sia riducibile ad una forma triangolare inferiore. La sola differenza è che si procede per discesa, ossia prima si calcola $\boldsymbol{y}_1$ e successivamente $\boldsymbol{y}_2$. Le considerazioni precedenti sulla riducibilità di una matrice implicano l'inutilità di dare condizioni sufficienti per la convergenza dei metodi iterativi per matrici riducibili. Per tale motivo esse vengono date unicamente per matrici irriducibili.

Teorema 5.15 *Condizione sufficiente perché il metodo di Jacobi, come anche quello di Gauss-Seidel, siano convergenti qualunque sia il vettore iniziale, è che la matrice A sia diagonalmente dominante in senso stretto oppure sia irriducibile e diagonalmente dominante in senso debole.*

Metodo del gradiente coniugato. Esso rappresenta una pietra miliare del calcolo scientifico. Introdotto nel 1952 da Hestenes e Stiefel [13] viene utilizzato per risolvere sistemi lineari con matrici definite positive. Il metodo risulta tanto più efficiente quanto più gli autovalori sono ben distribuiti. Se la matrice del sistema non singolare $A\boldsymbol{x} = \boldsymbol{b}$ non è simmetrica e definita positiva, esso viene applicato al sistema equivalente

$$A^TA\boldsymbol{x} = \boldsymbol{c}, \qquad \boldsymbol{c} = A^T\boldsymbol{b},$$

nel quale A^TA è simmetrica ($(A^TA)^T = A^TA$) e definita positiva in quanto $\boldsymbol{x}^TA^TA\boldsymbol{x} = \|A\boldsymbol{x}\|_2^2 > 0$ qualunque sia $\boldsymbol{x}$. Sistemi lineari del tipo

$$A^TA\boldsymbol{x} = A^T\boldsymbol{b} \tag{5.35}$$

con $A \in L(\mathbb{R}^m, \mathbb{R}^n)$ e $m > n$, si presentano nella risoluzione del classico problema ai minimi quadrati

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|A\boldsymbol{x} - \boldsymbol{b}\|_2, \tag{5.36}$$

dove $A \in L(\mathbb{R}^m, \mathbb{R}^n)$, $\boldsymbol{x} \in \mathbb{R}^n$ e $\boldsymbol{b} \in \mathbb{R}^m$, essendo $m > n$.

Si può infatti dimostrare il seguente teorema.

Teorema 5.16 *Il vettore $\boldsymbol{x} \in \mathbb{R}^n$ è la sola soluzione del problema* (5.36) *se e solo se esso è soluzione del sistema* (5.35)*, ossia se e solo se $\boldsymbol{x}$ soddisfa la seguente condizione di ortogonalità*

$$A^T(\boldsymbol{b} - A\boldsymbol{x}) = \boldsymbol{0}. \tag{5.37}$$

Dimostrazione. Supponiamo che $\boldsymbol{x}^*$ soddisfi la condizione di ortogonalità

$$A^T\boldsymbol{r}^* = \boldsymbol{0}, \text{ essendo } \boldsymbol{r}^* \text{ il residuo } \boldsymbol{r}^* = \boldsymbol{b} - A\boldsymbol{x}^*.$$

Allora, per ogni $\boldsymbol{x} \in \mathbb{R}^n$, posto $\boldsymbol{\ell} = \boldsymbol{x}^* - \boldsymbol{x}$,

$$\boldsymbol{r} = \boldsymbol{b} - A\boldsymbol{x} = \boldsymbol{r}^* + A(\boldsymbol{x}^* - \boldsymbol{x}) = \boldsymbol{r}^* + A\boldsymbol{\ell}.$$

Di conseguenza, essendo $A^T\boldsymbol{r}^* = \boldsymbol{0}$,

$$\boldsymbol{r}^T\boldsymbol{r} = (\boldsymbol{r}^* + A\boldsymbol{\ell})^T(\boldsymbol{r}^* + A\boldsymbol{\ell}) = (\boldsymbol{r}^*)^T\boldsymbol{r}^* + \|A\boldsymbol{\ell}\|_2^2$$

che è minima se $\boldsymbol{x} = \boldsymbol{x}^*$.

Supponendo ora che $A^T\boldsymbol{r}^* = \boldsymbol{y} \neq \boldsymbol{0}$, prendiamo $\boldsymbol{x} = \boldsymbol{x}^* + \alpha\boldsymbol{y}$, con α parametro reale, cui corrisponde il residuo

$$\boldsymbol{r} = \boldsymbol{r}^* - \alpha A\boldsymbol{y}, \text{ con } \boldsymbol{r}^T\boldsymbol{r} = (\boldsymbol{r}^*)^T\boldsymbol{r}^* - 2\alpha\boldsymbol{y}^T\boldsymbol{y} + \alpha^2(A\boldsymbol{y})^TA\boldsymbol{y} < (\boldsymbol{r}^*)^T\boldsymbol{r}^*$$

per α sufficientemente piccolo. Di conseguenza, $\boldsymbol{x}^*$ è soluzione del problema ai minimi quadrati se e solo se $A^T\boldsymbol{r}^* = \boldsymbol{0}$. □

Esso è il più importante tra i metodi i cui iterati forniscono approssimanti della soluzione in sottospazi di Krylov di dimensione crescente. Famiglia di metodi noti come metodi di proiezione nei sottospazi di Krylov.

Sia $A\boldsymbol{x} = \boldsymbol{b}$, con $A \in L(\mathbb{R}^m)$, il sistema non singolare da risolvere la cui soluzione esatta è $\boldsymbol{x}^* = A^{-1}\boldsymbol{b}$. Per sottospazio di Krylov generato da $\boldsymbol{b}$ e da A, di dimensione $n \leq m$, si intende il sottospazio di $\mathbb{R}^m$

$$K_n = \langle \boldsymbol{b}, A\boldsymbol{b}, \ldots, A^{n-1}\boldsymbol{b} \rangle, \tag{5.38}$$

ossia il sottospazio n-dimensionale di $\mathbb{R}^m$ avente come base $\boldsymbol{b}, A\boldsymbol{b}, \ldots, A^{n-1}\boldsymbol{b}$.

Poiché $K_1 \subset K_2 \subset \ldots \subset K_m = \mathbb{R}^m$, l'idea base dei metodi di proiezione nei sottospazi di Krylov è approssimare ricorsivamente la soluzione in sottospazi di dimensione crescente. È evidente che, per la proprietà di inclusione dei sottospazi, la distanza in norma delle approssimanti da $\boldsymbol{x}^*$ decresce all'aumentare di n e risulterà "tecnicamente" esatta in K_m.

Essendo studiati per risolvere sistemi di grandi dimensioni, la risoluzione in K_m è praticamente esclusa, ci si limita ad approssimare la soluzione in sottospazi di dimensione nettamente inferiore ad m. Gli algoritmi che si basano su questa logica si caratterizzano, in particolare, per la tecnica di costruzione di una base ortogonale per K_n, in quanto quella indicata nella (5.38) è poco utile. L'ipotesi A definita positiva equivale a supporre che tutti gli autovalori di A sono positivi o che, equivalentemente, qualunque sia il vettore (non nullo) $\boldsymbol{x}$ risulti $\boldsymbol{x}^T A\boldsymbol{x} > 0$. Sotto questa ipotesi, la funzione $\|\cdot\|_A$, così definita

$$\|\boldsymbol{x}\|_A = \sqrt{\boldsymbol{x}^T A\boldsymbol{x}}$$

è una norma, nota come A-norma. Per convincersene basta osservare che soddisfa le proprietà caratterizzanti le norme vettoriali. Questo permette di stimare l'errore di approssimazione di $\boldsymbol{x}^*$ in K_n, al crescere di n, mediante la

$$\|\boldsymbol{e}_n\|_A = \sqrt{\boldsymbol{e}_n^T A\boldsymbol{e}_n}, \qquad \boldsymbol{e}_n = \boldsymbol{x}^* - \boldsymbol{x}_n,$$

essendo $\boldsymbol{x}_n$ l'approssimante di $\boldsymbol{x}^*$ in K_n.

In sintesi, il metodo del gradiente coniugato (CG method) genera l'unica sequenza di iterati $\boldsymbol{x}_n \in K_n$ che gode delle seguenti proprietà: ad ogni passo n, $\|e_n\|_A$ è minima. Indicato con $r(\hat{\boldsymbol{x}})$ il residuo relativo ad $\hat{\boldsymbol{x}}$ ($r(\hat{\boldsymbol{x}}) = \boldsymbol{b} - A\hat{\boldsymbol{x}}$), il residuo relativo all'approssimante $\boldsymbol{x}_n$ calcolato nell'n-esimo iterato viene indicato con $\boldsymbol{r}_n = \boldsymbol{b} - A\boldsymbol{x}_n$, $n = 0, 1, \ldots$, essendo $\boldsymbol{r}_0$ il residuo corrispondente al vettore iniziale (vettore d'innesco) $\boldsymbol{x}_0$. Scelto un vettore di innesco e calcolato il residuo relativo, il metodo richiede: la scelta di una direzione di discesa (rispetto alla $\|\cdot\|_A$), la ottimizzazione del passo lungo tale direzione, il calcolo della nuova approssimazione alla soluzione e del relativo residuo e, infine, la determinazione della nuova direzione di discesa e la ottimizzazione del suo passo.

Algoritmo. Presentiamo ora i diversi passi dell'algoritmo:

(1) vettore d'innesco ($\boldsymbol{x}_0$) e della direzione di discesa relativa ($\boldsymbol{p}_0$)

$$\boldsymbol{x}_0 = \boldsymbol{0} \Longrightarrow \boldsymbol{r}_0 = \boldsymbol{b}, \quad \boldsymbol{p}_0 = \boldsymbol{r}_0;$$

(2) per $n = 1, 2, \ldots$, n-esima approssimante della soluzione

$$\boldsymbol{x}_n = \boldsymbol{x}_{n-1} + \alpha_n \boldsymbol{p}_{n-1},$$

con $\alpha_n = (\boldsymbol{r}_{n-1}^T \boldsymbol{r}_{n-1})/(\boldsymbol{p}_{n-1}^T A \boldsymbol{p}_{n-1})$, ottimizzazione del passo nella direzione di discesa $\boldsymbol{p}_{n-1}$;

(3) nuovo reiduo

$$\boldsymbol{r}_n = \boldsymbol{r}_{n-1} - \alpha_n \boldsymbol{p}_{n-1};$$

(4) nuova direzione di discesa

$$\boldsymbol{p}_n = \boldsymbol{r}_n + \beta_n \boldsymbol{p}_{n-1};$$

(5) ottimizzazione del passo nella nuova direzione

$$\beta_n = \frac{\boldsymbol{r}_n^T \boldsymbol{r}_n}{\boldsymbol{r}_{n-1}^T \boldsymbol{r}_{n-1}}.$$

La prima osservazione è la semplicità della programmazione che, ad ogni passo, richiede varie operazioni sui vettori e il calcolo di $A\boldsymbol{p}_{n-1}$. Di conseguenza, se la matrice è densa e non strutturata, la complessità di calcolo è $\approx 2m^2$ per ogni passo. Se invece è sparsa o dotata di strutture (utilizzabili nel calcolo) la sua complessità è un $O(m)$. Questo si verifica, in particolare, se la matrice è a banda (tri-, penta- o eptadiagonale).

Le principali proprietà dell'algoritmo sono riassunte nel seguente Teorema, per le cui dimostrazione si rinvia alla referenza [31].

Teorema 5.17 *Se la matrice A del sistema $A\boldsymbol{x} = \boldsymbol{b}$ è simmetrica e definita positiva, fintanto che il metodo non ha generato la soluzione del sistema ($\boldsymbol{r}_{n-1} \neq \boldsymbol{0}$), l'algoritmo procede senza divisioni per zero e i vettori posizione costruiti $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, come i residui $\boldsymbol{r}_0, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_{n-1}$ e le direzioni di discesa $\boldsymbol{p}_0, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_{n-1}$ sono basi del sottospazio di Krylov K_n.*

In simboli

$$\begin{aligned} K_n &= \langle \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \rangle = \langle \boldsymbol{p}_0, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_{n-1} \rangle \\ &= \langle \boldsymbol{r}_0, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_{n-1} \rangle = \langle \boldsymbol{b}, A\boldsymbol{b}, \ldots, A^{n-1}\boldsymbol{b} \rangle. \end{aligned} \tag{5.39}$$

Inoltre, i residui sono ortogonali,

$$\boldsymbol{r}_n^T \boldsymbol{r}_j, \qquad j = 0, 1, \ldots, n-1,$$

mentre le direzioni di ricerca sono "A-coniugate",

$$\boldsymbol{p}_n^T A \boldsymbol{p}_j = 0, \qquad j = 0, 1, \ldots, n-1.$$

È proprio quest'ultima proprietà ad aver indotto Hestenes e Stiefel a definire il metodo "CG-method".

Altra proprietà importante, che dimostra che (in assenza di errori di arrotondamento) gli iterati convergono alla soluzione esatta in non più di m passi è il seguente:

Teorema 5.18 *Se la matrice del sistema $A\boldsymbol{x} = \boldsymbol{b}$ è simmetrica e definita positiva e l'iterazione non ha già determinato la soluzione esatta ($\boldsymbol{r}_{n-1} \neq \boldsymbol{0}$), il vettore $\boldsymbol{x}_n$ è l'unico vettore di K_n che minimizza la $\|\boldsymbol{e}_n\|_A$. Inoltre, la successione è monotona decrescente, nel senso che*

$$\|\boldsymbol{e}_n\|_A \leq \|\boldsymbol{e}_{n-1}\|_A,$$

con $\boldsymbol{e}_n = \boldsymbol{0}$ per un indice di iterazione $n \leq m$.

Errori di arrotondamento. Gli importanti risultati dei teoremi (5.17) e (5.18) sono esatti solo in aritmetica infinita, ossia in assenza di errori di arrotondamento. In presenza di errori di arrotondamento, le proprietà di convergenza dipendono molto dal condizionamento della matrice. Anche senza entrare nel merito dei molti risultati esistenti sull'argomento, vogliamo evidenziare il seguente risultato.

Indicato con μ il numero di condizione della matrice A simmetrica e definita positiva,

$$\frac{\|\boldsymbol{e}_n\|_A}{\|\boldsymbol{e}_0\|_A} \leq 2\left(\frac{\sqrt{\mu}-1}{\sqrt{\mu}+1}\right)^n = 2\left(1 - \frac{2}{\sqrt{\mu}+1}\right)^n, \tag{5.40}$$

dove, con riferimento alla norma Euclidea,

$$\mu = \mathrm{cond}(A) = \|A\|_2 = \frac{\lambda_{\max}}{\lambda_{\min}},$$

essendo $\lambda_{\max}$ e $\lambda_{\min}$ il massimo e il minimo autovalore di A.

La (5.40) evidenzia che il numero degli iterati da calcolare per avere un risultato accettabile è una funzione crescente con μ. Per μ moderatamente elevato, il numero degli iterati previsto per avere un buon risultato è un $O(\sqrt{\mu})$. Per approfondimenti sul CG method, come per la illustrazione dei vari metodi di proiezione di tipo Krylov, si rinvia ai libri sull'algebra lineare numerica [11, 10].

Precondizionamento. La convergenza effettiva dei metodi iterativi (convergenza in presenza degli errori di arrotondamento) dipende dalle caratteristiche della matrice del sistema: numero di condizione, decadimento dei valori singolari e altre proprietà. Fortunatamente, negli ultimi 40 anni sono stati introdotti vari metodi in grado di migliorare significativamente le caratteristiche delle matrici, ai fini della convergenza dei metodi iterativi. Per tale motivo, vengono definiti "metodi di precondizionamento".

L'idea di base "teoricamente semplice" è la seguente: Qualunque sia la matrice non singolare M, il sistema non singolare $A\boldsymbol{x} = \boldsymbol{b}$, $A \in L(\mathbb{R}^n)$, è equivalente al sistema

$$M^{-1}A\boldsymbol{x} = \boldsymbol{c}, \qquad \boldsymbol{c} = M^{-1}\boldsymbol{b}. \tag{5.41}$$

Nonostante ciò, un metodo iterativo può essere convergente se applicato al sistema (5.41), anche se non lo è se applicato al sistema iniziale. Il motivo è che nel primo sistema la convergenza dipende dalle caratteristiche di A e nel secondo da quelle di $M^{-1}A$. Di conseguenza, se M è "scelta bene", un metodo iterativo può essere rapidamente convergente se applicato al sistema $M^{-1}A\boldsymbol{x} = M^{-1}\boldsymbol{b}$ e non convergente o convergente molto lentamente se applicato al sistema iniziale. Se M non è distante da A, $M^{-1}A$ non lo è da I e il sistema (5.41) è ben condizionato. Poiché invertire una matrice è molto più oneroso che risolvere un sistema, una tale operazione non è numericamente fattibile. Nel caso tutti gli elementi diagonali di A siano non nulli, un possibile condizionamento è dato da $M = \mathrm{diag}(a_{11}, a_{22}, \ldots, a_{nn})$.

Per evidenziare questa proprietà, in [31, pag. 325] è riportato il seguente **esempio**: A è una matrice pentadiagonale simmetrica di ordine 1000 con

$a_{ii} = 0.5 + \sqrt{i}$, sulla diagonale principale;
$a_{ij} = 1$ nella prima subdiagonale e nella prima sopradiagonale;
$a_{ij} = 1$ nella 100-esima subdiagonale e nella 100-esima sopradiagonale.

In questo caso il CG method converge lentamente, anche se la matrice è fortemente sparsa. Se invece si precondiziona ponendo $M = \mathrm{diag}(a_{11}, a_{22}, \ldots, a_{nn})$, il metodo converge molto più rapidamente.

Per gli approfondamenti si rinvia alle referenze indicate per il CG method [11, 31].

Capitolo 6

METODI ALLE DIFFERENZE FINITE

Questo capitolo è dedicato ai metodi alle differenze finite, una famiglia di tecniche numeriche per la soluzione di equazioni differenziali alle derivate ordinarie e parziali molto nota e consolidata [12, 18].

Principali vantaggi:

- facilità di implementazione;
- buona efficienza computazionale.

Principale difficoltà:

- scarsa capacità di risoluzione in domini con geometria irregolare.

L'ultimo punto è dovuto all'utilizzo di griglie di calcolo *strutturate* nei metodi alle differenze finite. Queste ultime sono griglie i cui nodi possono essere messi in corrispondenza biunivoca con una matrice. In altre parole, si tratta di griglie i cui nodi sono identificati in modo univoco da una coppia (terna) di indici locali, e quindi possono al più ottenersi da griglie regolari per deformazione continua.

I metodi ad elementi finiti hanno invece la capacità di utilizzare griglie *non strutturate*, in cui i nodi sono liberi di disporsi senza nessun vincolo topologico: caratteristica che conferisce una flessibilità d'uso notevolmente superiore rispetto alle differenze finite. Un esempio di griglia strutturata e uno di griglia non strutturata per una geometria semplice (un quadrato) sono mostrate nella figura 6.1.

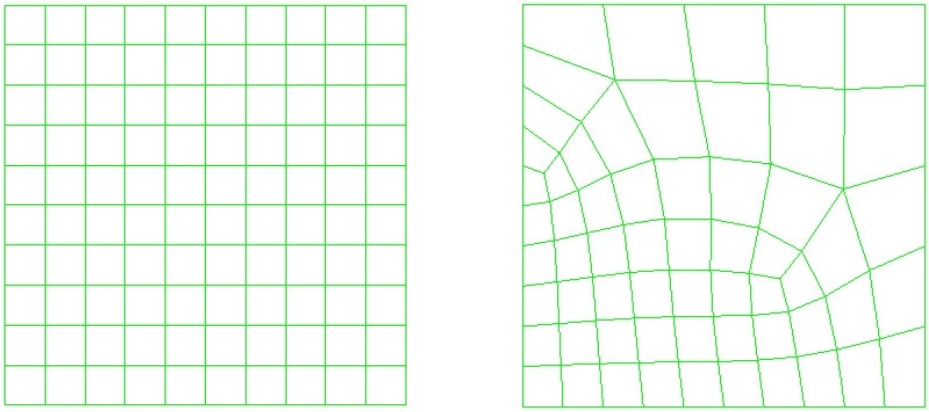


Figura 6.1: Un esempio di griglia strutturata e uno di griglia non strutturata per una semplice geometria di riferimento.

Problema differenziale con valori agli estremi

Risolviamo preliminarmente il problema differenziale

$$\begin{cases} y''(x) = p(x)\,y'(x) + q(x)\,y + r(x), \\ p, q \ \text{ e } \ r \text{ funzioni continue in } [a,b], \\ y(a) = \alpha, \quad y(b) = \beta. \end{cases}$$

Supponiamo, per semplicità, di discretizzare l'intervallo $[a,b]$ con punti equidistanti, ponendo pertanto $x_i = a + ih$, dove $i = 0, 1, \ldots, n+1$ e $h = \frac{b-a}{n+1}$. Indichiamo inoltre con y_i il valore in x_i, $i = 0, 1, \ldots, n+1$, della soluzione del modello discretizzato. Lo schema di discretizzazione è valido se

$$\max_{i=1,\ldots,n} |y_i - y(x_i)|$$

è "sufficientemente piccolo", essendo $y(x_i)$ il valore esatto della soluzione in x_i.

Schema di discretizzazione alle differenze centrali del secondo ordine:

$$y''(x_i) \approx \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2}, \qquad y''(x_i) - \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} = O(h^2),$$
$$y'(x_i) \approx \frac{y_{i+1} - y_{i-1}}{2h}, \qquad y'(x_i) - \frac{y_{i+1} - y_{i-1}}{2h} = O(h^2),$$

dove, come al solito, il simbolo $O(h^p)$, $p > 0$, indica un infinitesimo di ordine p rispetto ad h, per $h \to 0$. La collocazione in x_i , $i = 1, \ldots, n$, dell'equazione

differenziale genera il sistema lineare:

$$\frac{y_{i-1}-2y_i+y_{i+1}}{h^2}=p_i\frac{y_{i+1}-y_{i-1}}{2h}+q_iy_i+r_i \qquad \left|\begin{array}{l} p_i=p(x_i)\\ q_i=q(x_i)\\ r_i=r(x_i)\end{array}\right.$$

dove $i=0,1,\ldots,n+1$, essendo $y_0=\alpha$ e $y_{n+1}=\beta$.

In questo caso, nell'ipotesi che la soluzione sia sufficientemente regolare, è immediato valutare il residuo differenziale, ossia il residuo che si ottiene nel modello discretizzato sostituendo gli $\{y_i\}$ con gli $\{y(x_i)\}$. La formula di Taylor di ordine 4 permette infatti di dimostrare che

$$\begin{aligned}\tau(x,h)=&\frac{y(x-h)-2y(x)+y(x+h)}{h^2}-p(x)\frac{y(x+h)-y(x-h)}{2h}\\ &-q(x)y(x)-r(x)=O(h^2)\,.\end{aligned}$$

Lo schema è detto del secondo ordine in quanto $\tau(h)=\max\limits_{a\le x\le b}|\tau(x,h)|=O(h^2)$. La risoluzione del modello discretizzato richiede dunque la risoluzione del seguente sistema lineare di n equazioni in n incognite:

$$\begin{cases}\left(1+\dfrac{p_i}{2}h\right)y_{i-1}-\left(2+h^2q_i\right)y_i+\left(1-\dfrac{p_i}{2}h\right)y_{i+1}=h^2r_i\\ y_0=\alpha;\quad y_{n+1}=\beta.\end{cases} \tag{6.1}$$

Posto

$$a_i=-\left(2+h^2q_i\right),\qquad b_i=\left(1+\frac{p_i}{2}h\right),\qquad c_i=\left(1-\frac{p_i}{2}h\right)$$

per $i=1,2,\ldots,n$, si deve dunque risolvere il sistema lineare tridiagonale

$$\begin{pmatrix} a_1 & c_1 & & & & \\ b_2 & a_2 & c_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & & & \ddots & c_{n-1}\\ & & & & b_n & a_n\end{pmatrix}\begin{pmatrix} y_1\\ y_2\\ \vdots\\ \\ \vdots\\ y_n\end{pmatrix}=\begin{pmatrix} h^2r_1-b_1\alpha\\ h^2r_2\\ \vdots\\ \vdots\\ h^2r_{n-1}\\ h^2r_n-c_n\beta\end{pmatrix}.$$

Il caso più semplice si ha quando $q(x)\ge 0$ per $x\in(a,b)$, in quanto la matrice del sistema risulta allora diagonalmente dominante e irriducibile se $ph\le 2$, essendo $p=\max\limits_{a\le x\le b}|p(x)|$, ossia pur di prendere h sufficientemente piccolo. In tal caso il sistema possiede una e una sola soluzione, qualunque siano le condizioni iniziali. Per dimostrare la irriducibilità del sistema basta osservare che, sotto tali ipotesi tutti gli elementi della sopra e sotto-diagonale sono non nulli. Per la dominanza diagonale in senso debole, osserviamo preliminarmente che $ph<2$ implica la positività sia dei c_i sia dei b_i. Di conseguenza si ha la dominanza diagonale, in quanto:

(1) nella seconda, terza e $(n-1)$-esima riga, il valore assoluto dell'elemento diagonale risulta maggiore o almeno uguale alla somma dei valori assoluti dei due fuori diagonale ($|a_i| \geq |b_i| + |c_i|$, $i = 2, \ldots, n-1$);

(2) per $i = 1$, $|a_i| = 2 + h^2 q_i > 1 - \frac{p}{2}h$, dato che $1 + \frac{p}{2}h > 0$ e $q_i \geq 0$ e, per $i = n$, $|a_n| > |b_n|$.

Per quanto concerne la precisione dei risultati, vale il seguente

Teorema 6.1 (Gerschgorin [20]) *Nelle ipotesi precedenti, l'errore sulla soluzione possiede una maggiorazione del tipo:*

$$|y(x_i) - y_i| \leq c\,\tau(h), \qquad i = 1, 2, \ldots, n,$$

essendo c una costante positiva indipendente da h e $\tau(h) = \max_{a \leq x \leq b} |\tau(x,h)| = O(h^2)$.

Questo significa che nelle suddette ipotesi di regolarità, l'ordine dell'errore di approssimazione della soluzione coincide con quello del residuo differenziale.

6.1 Equazioni ellittiche

Consideriamo ora il problema differenziale[1]

$$\begin{cases} u_{xx} + u_{yy} + p(x,y)u_x + q(x,y)u_y + r(x,y)u + s(x,y) = 0, & (x,y) \in \Omega, \\ u(x,y) = f(x,y), & (x,y) \in \partial\Omega, \end{cases}$$

dove $\Omega = [a,b] \times [c,d]$ e $\partial\Omega$ indica il contorno di Ω, $f(a,y) = f_0(y)$, $f(b,y) = f_1(y)$; $f(x,c) = g_0(x)$ e $f(x,d) = g_1(x)$. Anche se la trattazione può essere adattata a situazioni più generali, per semplicità, supponiamo che le funzioni p, q, r e s siano continue in Ω e che la f lo sia su $\partial\Omega$. *L'assenza del termine* u_{xy}, in un problema ellittico, non è una limitazione effettiva in quanto è noto che [12, 18] con una rotazione degli assi, esso può essere azzerato.

Considerando, per semplicità, una reticolazione con nodi equidistanti in ciascuno dei due intervalli $[a,b]$ e $[c,d]$, si ottengono i nodi $\{(x_i, y_i)\}$, essendo

$$x_i = a + ih, \qquad i = 0, 1, \ldots, n+1, \quad \text{con } h = \frac{b-a}{n+1},$$
$$y_j = c + jk, \qquad j = 0, 1, \ldots, m+1, \quad \text{con } k = \frac{d-c}{m+1}.$$

[1] Nel caso di geometria più complessa, è talvolta possibile adottare una efficace strategia di adattamento delle condizioni al contorno nei nodi "prossimi" alla frontiera.

La discretizzazione con il metodo alle differenze centrali (schema a 5 punti) e la collocazione dell'equazione differenziale nei punti interni generano il sistema:

$$\frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h^2} + \frac{u_{i,j-1} - 2u_{i,j} + u_{i,j+1}}{k^2} + p_{i,j}\frac{u_{i+1,j} - u_{i-1,j}}{2h} + q_{i,j}\frac{u_{i,j+1} - u_{i,j-1}}{2k} + r_{i,j}u_{i,j} + s_{i,j} = 0$$

per $i = 1, \ldots, n$ e $j = 1, \ldots, m$.

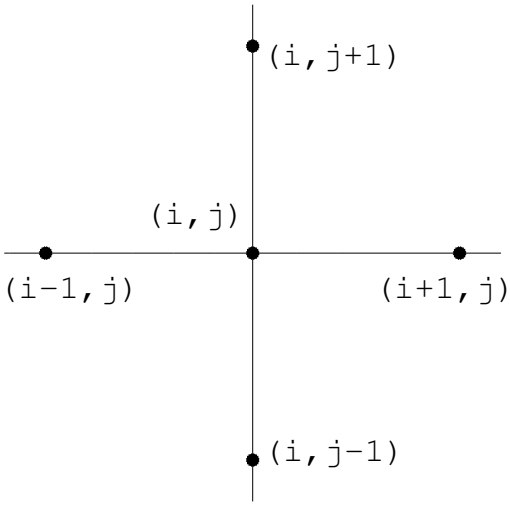


Figura 6.2: Schema di discretizzazione a 5 punti.

Ordinando gli $\{u_{i,j}\}_{\substack{i=1,\ldots,n \\ j=1,\ldots,m}}$ per linee (per j crescente e, a parità di j, per i crescente), la discretizzazione precedente genera il seguente sistema lineare

$$h^2(2 - k\,q_{i,j})u_{i,j-1} + k^2(2 - hp_{i,j})u_{i-1,j} - 2\left[2(h^2 + k^2) - h^2k^2r_{i,j}\right]u_{i,j} + k^2(2 + hp_{i,j})u_{i+1,j} + h^2(2 + kq_{i,j})u_{i,j+1} = -2h^2k^2s_{ij} \qquad (6.2)$$

per $i = 1, \ldots, n$ e $j = 1, \ldots, m$. Il sistema di ordine nm così ottenuto è chiaramente pentadiagonale, avente come elementi diagonali della matrice i coefficienti di u_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, m$. Fuori diagonale, per ogni riga abbiamo due elementi sottodiagonali, formate dai coefficienti di $u_{i-1,j}$ e $u_{i,j-1}$, e due sopradiagonali, formate dai coefficienti di $u_{i+1,j}$ e $u_{i,j+1}$.

Da notare che i termini contenenti u_{i0}, $u_{i,m+1}$, u_{0j} e $u_{n+1,j}$ sono noti, essendo $u_{i0} = g_0(x_i)$, $u_{i,m+1} = g_1(x_i)$ per $i = 0, 1, \ldots, n+1$ e $u_{0,j} = f_0(y_j)$, $u_{n+1,j} = f_1(y_j)$ per $j = 0, 1, \ldots, m+1$.

A titolo esemplificativo la (6.3) riporta schematicamente la matrice dei coefficiente del sistema lineare (6.2), nell'ipotesi che $n = 4$ e $m = 3$, nel quale

abbiamo indicato con a_{ij}, b_{ij}, $\hat{b}_{ij}$, c_{ij} e $\hat{c}_{ij}$ i coefficienti di u_{ij}, $u_{i-1,j}$, $u_{i+1,j}$, $u_{i,j-1}$ e $u_{i,j+1}$, rispettivamente, per $i = 1, 2, 3, 4$ e $l = 1, 2, 3$.

$$\left(\begin{array}{cccccccccccc}
a_{11} & \hat{b}_{11} & 0 & 0 & \hat{c}_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
b_{21} & a_{21} & \hat{b}_{21} & 0 & 0 & \hat{c}_{21} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & b_{31} & a_{31} & \hat{b}_{31} & 0 & 0 & \hat{c}_{31} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & b_{41} & a_{41} & 0 & 0 & 0 & \hat{c}_{41} & 0 & 0 & 0 & 0 \\
c_{12} & 0 & 0 & 0 & a_{12} & \hat{b}_{12} & 0 & 0 & \hat{c}_{12} & 0 & 0 & 0 \\
0 & c_{22} & 0 & 0 & b_{22} & a_{22} & \hat{b}_{22} & 0 & 0 & \hat{c}_{22} & 0 & 0 \\
0 & 0 & c_{32} & 0 & 0 & b_{32} & a_{32} & \hat{b}_{32} & 0 & 0 & \hat{c}_{32} & 0 \\
0 & 0 & 0 & c_{42} & 0 & 0 & b_{42} & a_{42} & 0 & 0 & 0 & \hat{c}_{42} \\
0 & 0 & 0 & 0 & c_{13} & 0 & 0 & 0 & a_{13} & \hat{b}_{13} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & c_{23} & 0 & 0 & b_{23} & a_{23} & \hat{b}_{23} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & c_{33} & 0 & 0 & b_{33} & a_{33} & \hat{b}_{33} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & c_{43} & 0 & 0 & b_{43} & a_{43}
\end{array}\right) \tag{6.3}$$

Nel caso $r_{ij} \leq 0$, $i = 1, \ldots, n$ e $j = 1, \ldots, m$, la matrice del sistema è irriducibile e diagonalmente dominante in senso debole a condizione che i valori di h e k siano sufficientemente piccoli. Per la irriducibilità è sufficiente l'esistenza di due diagonali (le due formate dagli elementi b_{ij} e $\hat{b}_{ij}$, rispettivamente) con elementi tutti non nulli. Nel caso $r_{ij} < 0$, la dominanza in senso stretto è assicurata dall'ipotesi che tutti gli elementi fuori diagonali siano non negativi, in quanto, in tal caso, la somma dei loro valori assoluti sarebbe inferiore al valore assoluto dell'elemento diagonale, essendo

$$4(h^2 + k^2) < 4(h^2 + k^2) - h^2 r_{ij}, \quad i = 1, \ldots, n \text{ e } j = 1, \ldots, m.$$

Nel caso $r_{ij} \leq 0$ (come anche $r_{ij} = 0$) sotto le stesse ipotesi di non negatività degli elementi fuori diagonale, si avrebbe la dominanza diagonale in senso debole per le seguenti ragioni:

(1) in tutte le righe intermedie (quelle aventi come termini noti i soli f_{ij}), il valore assoluto del termine diagonale risulta uguale alla somma dei corrispondenti fuori diagonale;

(2) in tutte le altre risulta non inferiore, con il maggiore in senso stretto nelle righe estremi (per es. per $i = j = 1$). Tali condizioni sono chiaramente soddisfatte se h e k soddisfano i seguenti vincoli:

$$\begin{aligned}
-2 \leq hp_{ij} \leq 2 \quad &\Longleftrightarrow \quad h|p_{ij}| \leq 2, \\
-2 \leq kq_{ij} \leq 2 \quad &\Longleftrightarrow \quad h|q_{ij}| \leq 2,
\end{aligned}$$

certamente soddisfatti per $hp \leq 2$ e $kq \leq 2$, essendo

$$p = \max_{(x,y)\in\Omega} |p(x,y)| \quad \text{e} \quad q = \max_{(x,y)\in\Omega} |q(x,y)|.$$

Da tali considerazioni segue immediatamente che nelle menzionate ipotesi di regolarità dei coefficienti in Ω e della soluzione, $r(x,y) \leq 0$ in Ω, se h e k soddisfano le condizioni $ph \leq 2$ e $qk \leq 2$, essendo $|p(x,y)| \leq p$ e $|q(x,y)| \leq q$ per $(x,y) \in \Omega$, il sistema ottenuto con il metodo alle differenze centrali possiede una e una sola soluzione.

In questo caso, se p, q, r, e s sono funzioni continue in Ω e la f lo è in $\partial\Omega$, il residuo differenziale

$$\tau(x,y;h,k) = O(h^2 + k^2)$$

e inoltre l'errore sulla soluzione, in ogni punto nodale, soddisfa la condizione [12]

$$\max_{i,j} |u(x_i,y_j) - u_{i,j}| \leq C(h^2 + k^2),$$

con C non dipendente da h e k.

Metodo upwind. Qualora le condizioni $ph \leq 2$ e $qk \leq 2$ risultino molto restrittive, nel senso che h e/o k siano molto piccoli e dunque n e/o m molto grandi, onde evitare la risoluzione di un sistema di dimensioni eccessivamente grandi spesso si ricorre al cosiddetto metodo upwind [12].

In tal caso, mentre u_{xx} e u_{yy} vengono discretizzate con il precedente schema alle differenze centrali, u_x e u_y vengono discretizzate nel modo seguente:

$$\begin{aligned}
u_x(x_i,y_j) &\approx \frac{u_{i+1,j} - u_{i,j}}{h}, \qquad \text{se } p_{ij} \geq 0 \\
u_x(x_i,y_j) &\approx \frac{u_{i,j} - u_{i-1,j}}{h}, \qquad \text{se } p_{ij} \leq 0 \\
u_y(x_i,y_j) &\approx \frac{u_{i,j+1} - u_{i,j}}{k}, \qquad \text{se } q_{ij} \geq 0 \\
u_y(x_i,y_j) &\approx \frac{u_{i,j} - u_{i,j-1}}{k}, \qquad \text{se } q_{ij} \leq 0.
\end{aligned}$$

Questo equivale a porre

$$\begin{aligned}
p(x_i,y_j)u_x &\approx \frac{(|p_{ij}| + p_{ij})\, u_{i+1,j} - 2|p_{ij}|u_{ij} + (|p_{ij}| - p_{ij})\, u_{i-1,j}}{2h} \\
q(x_i,y_j)u_y &\approx \frac{(|q_{ij}| + q_{ij})\, u_{i,j+1} - 2|q_{ij}|u_{ij} + (|q_{ij}| - q_{ij})\, u_{i,j-1}}{2k}.
\end{aligned}$$

Tale schema genera il sistema

$$\left[\frac{1}{k^2}+\frac{1}{2k}\left(|q_{ij}|-q_{ij}\right)\right]u_{i,j-1}+\left[\frac{1}{h^2}+\frac{1}{2h}\left(|p_{ij}|-p_{ij}\right)\right]u_{i-1,j}$$
$$-\left[\frac{2}{h^2}+\frac{2}{k^2}+\frac{|p_{ij}|}{h}+\frac{|q_{ij}|}{k}-r_{ij}\right]u_{ij}+\left[\frac{1}{h^2}+\frac{1}{2h}\left(|p_{ij}|+p_{ij}\right)\right]u_{i+1,j}$$
$$+\left[\frac{1}{k^2}+\frac{1}{2k}\left(|q_{ij}|+q_{ij}\right)\right]u_{i,j+1}=-s_{ij}$$

per $i=1,\ldots,n$ e $j=1,\ldots,m$. Questo sistema, se $r(x,y)\leq 0$, è diagonalmente dominante e dunque univocamente risolubile per ogni coppia (h,k).

Questo fatto non implica tuttavia che h e k possano essere presi non piccoli, ossia che si debba risolvere un sistema piccolo. Infatti si deve tenere conto dell'errore di discretizzazione che, mentre per le differenze centrali è un $O(h^2+k^2)$, nel caso upwind è un $O(h+k)$. Per questo motivo, non necessariamente, il metodo upwind deve riguardare ambedue i termini $p(x,y)u_x$ e $q(x,y)u_y$, in quanto non è detto che ambedue i vincoli su h e k siano troppo restrittivi nel caso delle differenze centrali. Nel caso lo sia soltanto su h, si applica al termine $p(x,y)u_x$ e non al termine $q(x,y)u_y$ e viceversa, nel caso il vincolo sia troppo restrittivo per k, esso si applica al termine $q(x,y)u_y$ e non a $p(x,y)u_x$. Nel primo caso l'errore di discretizzazione $\tau(h,k)=O(h+k^2)$ e nel secondo $\tau(h,k)=O(h^2+k)$. Nelle ipotesi precedenti, posto ad esempio $a=b=0$ e $b=d=10$, la scelta di $h=\frac{1}{100}$ e $k=\frac{1}{50}$, comporta la risoluzione di un sistema pentadiagonale (diagonalmente dominante) di ordine $n\cdot m=999\cdot 499\simeq 5\cdot 10^5$, comunque risolubile con i metodi iterativi di Jacobi e Gauss-Seidel.

Osservazione 6.2 *Il metodo precedentemente descritto si estende in modo immediato, nell'ipotesi che i coefficienti di u_{xx} e u_{yy} non siano costanti, naturalmente nell'ipotesi che siano ambedue positivi o negativi, onde evitare che esistano punti del dominio in cui l'equazione non sia ellittica. Questo implica che, come risulta nell'esempio 6.3, le condizioni $ph\leq 2$ e $qk\leq 2$ non sono sufficienti ad assicurare la dominanza diagonale della matrice del sistema. In questo caso conviene discretizzare e verificare sull'esempio le condizioni per la dominanza diagonale.*

Esempio 6.3 Discutere la risoluzione numerica, mediante il metodo alle differenze finite, del seguente problema differenziale:

$$\begin{cases}(x+y)^2u_{xx}+5u_{yy}+(\sin\sqrt{x^2+y^2})\,u_x+\sqrt{x^2+y^2}\,u_y=x^2+y^2,\\ 0\leq x,\,y\leq 5\\ u(0,\,y)=f_0(y),\quad u(5,\,y)=f_1(y)\\ u(x,\,0)=g_0(x),\quad u(x,\,5)=g_1(x).\end{cases}$$

Nell'ipotesi che i dati al contorno si raccordino, ossia che risulti $f_0(0) = g_0(0)$, $g_0(5) = f_1(0)$, $f_1(5) = g_1(5)$ e $f_0(5) = g_1(5)$.

Considerando una reticolazione con nodi equidistanti in ciascuno dei due intervalli $[0, 5]$, si ottengono i nodi $\{(x_i, y_j)\}$, essendo $x_i = ih$, $i = 0, 1, 2, \ldots, n+1$ e $y_j = jk$, $j = 0, 1, 2, \ldots, m+1$ con $h = \frac{5}{n+1}$ e $k = \frac{5}{m+1}$. Consideriamo inoltre le seguenti approssimazioni

$$u_{xx}(x_i,\, y_j) \simeq \frac{u_{i+1,\,j} - 2u_{i,\,j} + u_{i-1,\,j}}{h^2},\ u_x(x_i,\, y_j) \simeq \frac{u_{i+1,\,j} - u_{i-1,\,j}}{2h}$$
$$u_{yy}(x_i,\, y_j) \simeq \frac{u_{i,\,j+1} - 2u_{i,\,j} + u_{i,\,j-1}}{k^2}, u_y(x_i,\, y_j) \simeq \frac{u_{i,\,j+1} - u_{i,\,j-1}}{2k}.$$

La discretizzazione con il metodo alle differenze centrali (*schema a* 5 *punti*) e la collocazione dell'equazione differenziale nei punti interni generano il sistema:

$$(x_i + y_j)^2\, \frac{u_{i+1,\,j} - 2u_{i,\,j} + u_{i-1,\,j}}{h^2} + 5\frac{u_{i,\,j+1} - 2u_{i,\,j} + u_{i,\,j-1}}{k^2}$$
$$+ (\sin\sqrt{x_i^2 + y_j^2})\frac{u_{i+1,\,j} - u_{i-1,\,j}}{2h} + \sqrt{x_i^2 + y_j^2}\frac{u_{i,\,j+1} - u_{i,\,j-1}}{2k} = x_i^2 + y_j^2,$$

per $i = 1, \ldots, n$ e $j = 1, \ldots, m$. Da notare che tale procedimento comporta un errore di discretizzazione $\tau = O(h^2 + k^2)$. Ordinando gli $\{u_{i,j}\}_{\substack{i=1,\ldots,n\\ j=1,\ldots,m}}$ per linee (per j crescente e, a parità di j, per i crescente), la discretizzazione precedente genera il seguente sistema lineare:

$$\left(\frac{5}{k^2} - \frac{\sqrt{x_i^2 + y_j^2}}{2k}\right) u_{i,\,j-1} + \left(\frac{(x_i + y_j)^2}{h^2} - \frac{\sin\sqrt{x_i^2 + y_j^2}}{2h}\right) u_{i-1,\,j}$$
$$- \left(\frac{2(x_i + y_j)^2}{h^2} + \frac{10}{k^2}\right) u_{i,\,j} + \left(\frac{(x_i + y_j)^2}{h^2} + \frac{\sin\sqrt{x_i^2 + y_j^2}}{2h}\right) u_{i+1,\,j}$$
$$+ \left(\frac{5}{k^2} + \frac{\sqrt{x_i^2 + y_j^2}}{2k}\right) u_{i,\,j+1} = x_i^2 + y_j^2,$$

nel quale, come è immediato verificare, i coefficienti del termine u_{ij} per $i = 1, \ldots, n$ e $j = 1, \ldots, m$, forniscono i valori della diagonale della matrice del sistema. La matrice di ordine nm è chiaramente pentadiagonale. È anche immediato osservare che se tutti i coefficienti fuori diagonale fossero non negativi, il valore assoluto dell'elemento diagonale sarebbe uguale alla somma dei valori assoluti di quelli fuori diagonale. Per la precisione questo avverrebbe per tutte le equazioni *interne*, ossia nelle quali i termini noti sono dati unicamente da

$f_{ij} = x_i^2 + y_j^2$. Per quelle *non interne* il valore assoluto dell'elemento diagonale è superiore, dato che almeno un termine, essendo noto, viene spostato alla destra dell'uguale. Questo avviene in particolare per $i = j = 1$.

Sotto tali condizioni la matrice risulta debolmente diagonalmente dominante e pertanto, se irriducibile, si ha l'unicità della soluzione ottenibile con metodi iterativi standard come i metodi di Jacobi e Gauss-Seidel. Per la irriducibilità basta osservare che la diagonale principale ha tutti gli elementi non nulli, come lo sono le due diagonali, formate dai coefficienti $u_{i,j-1}$ e $u_{i,j+1}$ per k sufficientemente piccolo, ossia per k soddisfacente il vincolo

$$\frac{\sqrt{x_i^2 + y_j^2}}{2k} < \frac{5}{k^2},$$

essendo sempre positivo il coefficiente di $u_{i,j+1}$. Essendo $x_i^2 + y_j^2 \leq 50$, tale vincolo è certamente soddisfatto, qualunque siano i, j, per $k < \frac{1}{5}$. I coefficienti di $u_{i-1,j}$ e $u_{i+1,j}$ sono non negativi se

$$-\frac{(x_i + y_j)^2}{h^2} \leq \frac{\sin\sqrt{x_i^2 + y_j^2}}{2h} \leq \frac{(x_i + y_j)^2}{h^2},$$

ossia per

$$\frac{\left|\sin\sqrt{x_i^2 + y_j^2}\right|}{2h} \leq \frac{(x_i + y_j)^2}{h^2},$$

essendo $\sin\sqrt{x_i^2 + y_j^2} < \sqrt{x_i^2 + y_j^2}$. Da quest'ultima, osservato che, essendo x_i e y_j positivi, $\sqrt{x_i^2 + y_j^2} < x_i + y_j$, segue la disuguaglianza

$$\frac{1}{2} \leq \frac{x_i + y_j}{h},$$

chiaramente soddisfatta per ogni coppia di valori positivi $x_i = ih$ e $y_j = jk$. In definitiva per tutte le equazioni *interne* si ha la dominanza diagonale a condizione che $k < \frac{1}{5}$. A maggior ragione questa vale quando il termine noto passa a destra dell'uguale. Per $i = j = 1$, in particolare, si ha la dominanza diagonale qualora

$$\frac{2(x_1 + y_1)^2}{h^2} + \frac{10}{k^2} > \frac{(x_1 + y_1)^2}{h^2} + \frac{\sin(\sqrt{x_1^2 + y_1^2})}{2h} + \frac{5}{k^2} + \frac{\sqrt{x_1^2 + y_1^2}}{2k},$$

condizione ovviamente soddisfatta in quanto, per $k < \frac{1}{5}$ e qualunque sia h,

$$\frac{5}{k^2} > \frac{\sqrt{x_1^2 + y_1^2}}{2k} \quad \text{e} \quad \frac{(x_1 + y_1)^2}{k^2} > \frac{\sin(\sqrt{x_1^2 + y_1^2})}{2h}.$$

Tenuto conto dell'errore di discretizzazione e del fatto che la matrice è pentadiagonale, una buona scelta per h e k potrebbe essere la seguente:

$$h = k = \frac{1}{50}.$$

In tal caso $n+1 = m+1 = 250$, scelta che comporta la risoluzione, con il metodo di Jacobi o di Gauss-Seidel, di un sistema pentadiagonale, irriducibile e diagonalmente dominante dell'ordine di 250×250 equazioni in altre incognite.

Per quando concerne la scelta (nelle iterazioni) del vettore iniziale, un'ottima scelta è rappresentata dai valori di interpolazione $u(x_i, y_j) = \Phi(x_i, y_j)$, $i = 0, 1, \ldots, n+1$, $j = 0, 1 \ldots, m+1$, dove

$$\Phi(x,y) = \frac{x^2(x-5)^2 G(x,y) + y^2(y-5)^2 F(x,y)}{x^2(x-5)^2 + y^2(y-5)^2},$$

essendo

$$F(x,y) = \frac{x^2 f_1(y) + (x-5)^2 f_0(y)}{x^2 + (x-5)^2}$$

la funzione che interpola $f_0(y)$ e $f_1(y)$ per ogni prefissato y, e

$$G(x,y) = \frac{y^2 g_1(x) + (y-5)^2 g_0(x)}{y^2 + (y-5)^2}$$

l'interpolante di $g_1(x)$ e $g_0(x)$, per ogni prefissato x.

6.2 Equazioni paraboliche

Consideriamo ora il problema differenziale

$$\begin{cases} u_t = c^2\, u_{xx} + p(x,t)u_x + q(x,t)u + r(x,t), & a \le x \le b,\ 0 \le t \le T, \\ u(a,t) = f_1(t), \quad u(b,t) = f_2(t), & 0 \le t \le T, \\ u(x,0) = \phi(x), & a \le x \le b. \end{cases}$$

In questo caso, si può sfruttare la conoscenza della condizione iniziale per ricondurre la soluzione del problema bidimensionale alla risoluzione di una sequenza di problemi *monodimensionali*, con conseguente miglioramento della precisione dei risultati e riduzione della complessità di calcolo. A tale scopo si adatta lo schema a 4 punti riportato nella Fig. 6.3. Il punto di partenza è la costruzione della mesh:

$$\begin{aligned} x_i &= a + ih, \quad i = 0, 1, \ldots, n+1, \quad h = \frac{b-a}{n+1} \\ t_j &= jk, \qquad j = 0, 1, \ldots, m+1, \quad k = \frac{T}{m+1}, \end{aligned}$$

e successiva collocazione dell'equazione nel punto nodale (x_i, t_j), la quale, per ogni $j = 1, \ldots, m$ genera il sistema n-dimensionale

$$\frac{u_{i,j} - u_{i,j-1}}{k} = c^2 \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2} + p_{i,j} \frac{u_{i+1,j} - u_{i-1,j}}{2h} + q_{i,j} u_{i,j} + r_{i,j},$$

per $i = 1, \ldots, n$.

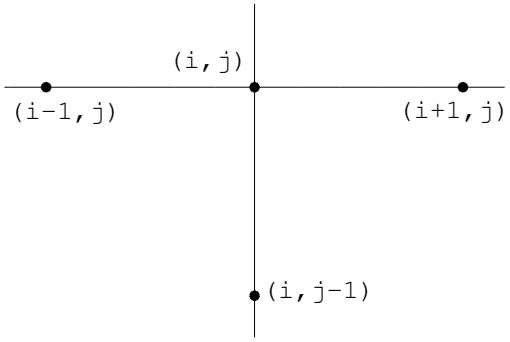


Figura 6.3: Schema di discretizzazione a 4 punti.

Da notare che per $j = 1$, le incognite sono $u_{1,1}, \ldots, u_{n,1}$, le quali possono essere determinate risolvendo il sistema di ordine n ottenuto per $i = 1, \ldots, n$, in quanto sono noti sia i valori iniziali $\{u_{i0},\ i = 0, 1, \ldots, n+1\}$ sia i valori al bordo u_{01} e $u_{n+1,1}$. Più in generale, supponendo noti i valori $u_{i,j-1}$ al livello $j-1$ come pure i due valori al bordo coinvolti, i valori incogniti $u_{i,j}$, $i = 1, \ldots, n$, (al livello j) possono essere calcolati risolvendo il sistema di ordine n ottenuto ponendo $i = 1, \ldots, n$.

Di conseguenza, lo schema precedente implica che l'avanzamento temporale richiede la risoluzione del sistema:

$$\left(\frac{c^2}{h^2} - \frac{p_{i,j}}{2h}\right) u_{i-1,j} - \left(\frac{1}{k} + 2\frac{c^2}{h^2} - q_{i,j}\right) u_{i,j} + \left(\frac{c^2}{h^2} + \frac{p_{i,j}}{2h}\right) u_{i+1,j} = -r_{i,j} - \frac{u_{i,j-1}}{k}$$

per $i = 1, \ldots, n$, prefissato $j = 1, \ldots, m$.

Se $q(x,t) < 0$ per $a \le x \le b$ e $0 \le t \le T$, tale sistema è diagonalmente dominante in senso stretto per $hp \le 2c^2$, essendo $p = \max_{\substack{a \le x \le b \\ 0 \le t \le T}} |p(x,t)|$. Se $q(x,t) \le 0$, la matrice è irriducibile e diagonalmente dominante per $hp \le 2c^2$. In ambedue i casi, per ogni prefissato j, il sistema possiede una e una sola soluzione, ottenibile con i metodi di Jacobi o di Gauss-Seidel.

Se esistono valori $q_{ij} > 0$, tutte le precedenti considerazioni possono essere ripetute a condizione che si assuma k con $\frac{1}{k} \ge q_{ij}$, $i = 1, \ldots, n$, $j =$

$1, \ldots, m$. Condizione che risulta certamente soddisfatta se $k \leq \frac{1}{q}$, essendo $q = \max_{(x,t)\in\Omega} q(x,t)$.

Osservazione. Poiché $u_t(x_i, t_j)$ è stata approssimata con uno schema alle differenze del primo ordine in t e uno del secondo in x, l'errore di discretizzazione è un $O(h^2 + k)$.

Nel caso $a = 0$, $b = 20$ e $T = 10$, con la scelta di $h = \frac{1}{20}$, $k = \frac{1}{100}$, si debbono risolvere $m + 1 = 1000$ sistemi tridiagonali (diagonalmente dominanti), ciascuno di ordine $n = 399$. Essi possono essere risolti con buona precisione, utilizzando i metodi di Jacobi e di Gauss-Seidel. Per tali metodi iterativi un'ottima scelta dei valori iniziali è rappresentata dai seguenti valori di interpolazione

$$u_{ij}^{(0)} = f_1(t_j) + \frac{x_i - a}{b - a}\,[f_2(t_j) - f_1(t_j)]$$

per ogni $j = 0, 1, \ldots, m + 1$ e $i = 0, 1, 2 \ldots, n + 1$.

Esempio 6.4 Discutere la risoluzione numerica, mediante il metodo alle differenze finite, del seguente problema differenziale di tipo parabolico:

$$\begin{cases} u_t = x^2t^2u_{xx} + 10^{2x+t}(\cos xt)u_x + (xt - 1)u, \quad 0 \leq x \leq 10,\ 0 \leq t \leq 5, \\ u(0,\, t) = f_0(t), \quad u(10,\, t) = f_1(t), \\ u(x,\, 0) = g(x). \end{cases}$$

Per la risoluzione si inizia con la costruzione della mesh

$$x_i = ih, \quad i = 0, 1, \ldots, n + 1, \quad h = \frac{b - a}{n + 1},$$
$$t_j = jk, \quad j = 0, 1, \ldots, m + 1, \quad k = \frac{T}{m + 1}.$$

Poiché la funzione 10^{2x+t} assume valori particolarmente elevati in prossimità dei valori $x = 10$ e $t = 5$, conviene ricorrere, nella discretizzazione della funzione $10^{2x+t}(\cos xt)\, u_x$, al metodo up-wind. Questo significa che nella discretizzazione dei termini differenziali si dovranno usare approssimazioni del tipo

$$u_{xx}(x_i,\, t_j) \simeq \frac{u_{i+1,\,j} - 2u_{i,\,j} + u_{i-1,\,j}}{h^2},$$
$$u_x(x_i, t_j) \simeq \frac{u_{i+1,j} - u_{i,j}}{h} \quad \text{se } \cos x_i t_j \geq 0,$$
$$u_x(x_i, t_j) \simeq \frac{u_{i,j} - u_{i-1,j}}{h} \quad \text{se } \cos x_i t_j \leq 0,$$
$$u_t(x_i, t_j) \simeq \frac{u_{i,j} - u_{i,j-1}}{k}.$$

La discretizzazione con lo schema a 4 punti e la collocazione dell'equazione differenziale nei punti interni generano il sistema:

$$\frac{u_{i,j}-u_{i,j-1}}{k} = (x_i t_j)^2 \frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{h^2}$$
$$+ 10^{2x_i+t_j} \frac{(|\cos x_i t_j|+\cos x_i t_j)u_{i+1,j} - 2|\cos x_i t_j|u_{i,j} + (|\cos x_i t_j|-\cos x_i t_j)u_{i-1,j}}{2h}$$
$$+ (x_i t_j - 1)u_{i,j}$$

In questo caso l'errore di discretizzazione è $\tau = O(h+k)$. Per $j = 1$, le incognite sono $u_{1,1}, \ldots, u_{n,1}$, le quali possono essere determinate risolvendo il sistema di ordine n ottenuto per $i = 1, \ldots, n$, in quanto le $\{u_{i0}\}$ sono note, come lo sono i valori agli estremi u_{01} e $u_{n+1,1}$. Più in generale, supponendo noti i valori $u_{i,j-1}$ al livello $j-1$, i valori $u_{i,j}$, $i = 1, \ldots, n$, al livello j, possono essere calcolati risolvendo il sistema di ordine n ottenuto per $i = 1, \ldots, n$. Quindi assegnato j, il vettore $\{u_{i,j-1}\}$ è noto, e lo schema precedente implica che l'avanzamento temporale richiede la risoluzione del sistema:

$$\left[\frac{(x_i t_j)^2}{h^2} + 10^{2x_i+t_j}\frac{|\cos x_i t_j| - \cos x_i t_j}{2h}\right] u_{i-1,j}$$
$$- \left[\frac{1}{k} + \frac{2(x_i t_j)^2}{h^2} + 10^{2x_i+t_j}\frac{|\cos x_i t_j|}{h} + (1 - x_i t_j)\right] u_{i,j}$$
$$+ \left[\frac{(x_i t_j)^2}{h^2} + 10^{2x_i+t_j}\frac{|\cos x_i t_j| + \cos x_i t_j}{2h}\right] u_{i+1,j} = \frac{u_{i,j-1}}{k},$$

per $i = 1, \ldots, n$, prefissato $j = 1, \ldots, m$. Poiché

1. Per $\frac{1}{k} \geq x_i t_j - 1$, che implica $k \leq \frac{1}{49}$, il coefficiente del termine $u_{i,j}$ è positivo,

2. Sotto tale condizione il sistema è debolmente diagonalmente dominante in quanto

   (a) per $i = 2, \ldots, n-1$

   $$\frac{1}{k} + \frac{2(x_i t_j)^2}{h^2} + 10^{2x_i+t_j}\frac{|\cos x_i t_j|}{h} + (1 - x_i t_j)$$
   $$\geq \frac{2(x_i t_j)^2}{h^2} + 10^{2x_i+t_j}\frac{|\cos x_i t_j|}{h};$$

   (b) per $k \leq \frac{1}{49}$

   $$\frac{1}{k} + \frac{2(x_1 + t_1)^2}{h^2} + 10^{2x_1+t_1}\frac{|\cos x_1 t_1|}{h} + (1 - x_1 t_1)$$
   $$> \frac{(x_1 t_1)^2}{h^2} + 10^{2x_1 t_1}\frac{\cos x_1 t_1}{h};$$

(c) condizione di dominanza diagonale, analoga alle (b), vale anche per $i = n$.

Poiché la matrice è anche irriducibile, dato che tutti i termini delle tre diagonali sono diversi da zero, il sistema possiede un'unica soluzione. Essa è univocamente risolubile con il metodo di Jacobi o di Gauss-Seidel. Nel caso $h = k = \frac{1}{100}$, la soluzione del problema richiede la risoluzione di $m + 1 = 500$ sistemi tridiagonali, diagonalmente dominanti, ciascuno di ordine $n = 999$. Per la scelta del vettore iniziale vale quanto detto precedentemente nel caso parabolico.

6.3 Equazioni iperboliche

Consideriamo ora il problema differenziale

$$\begin{cases} u_{tt} = c^2\, u_{xx} + p(x,t)u_x + q(x,t)u_t + r(x,t)u + s(x,t), & a \le x \le b,\ t \ge 0, \\ u(a,t) = f_1(t), \quad u(b,t) = f_2(t), & 0 \le t \le T, \\ u(x,0) = \phi_1(x), \quad u_t(x,0) = \phi_2(x), & a \le x \le b. \end{cases}$$

In questo problema, oltre alla soluzione in $x = a$ e $x = b$, per ogni $0 \le t \le T$, sono note la u e la u_t all'istante iniziale in ogni punto dell'intervallo $[a, b]$. Questa informazione, utilizzando opportunamente la formula di Taylor, permette di calcolare la soluzione in un passo temporale successivo. La conoscenza della u in due livelli iniziali consente di calcolare ricorsivamente la soluzione mediante l'utilizzo dello schema a 7 punti riportato della Fig. 6.4.

Come al solito, il primo passo consiste nella generazione della mesh

$$\begin{aligned} x_i &= a + ih, \qquad i = 0, 1, \ldots, n+1, \qquad h = \frac{b-a}{n+1} \\ t_j &= jk, \qquad\quad j = 0, 1, \ldots, m+1, \qquad k = \frac{T}{m+1}, \end{aligned}$$

e nella successiva collocazione in (x_i, t_j), prefissato $j = 1, \ldots, m$ per $i = 1, \ldots, n$. Si perviene così al sistema

$$\begin{aligned} \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{k^2} = c^2 \frac{1}{2} \Bigg[& \frac{u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1}}{h^2} \\ & + \frac{u_{i+1,j-1} - 2u_{i,j-1} + u_{i-1,j-1}}{h^2} \Bigg] \\ & + \frac{p_{i,j}}{2} \left[\frac{u_{i+1,j+1} - u_{i-1,j+1}}{2h} + \frac{u_{i+1,j-1} - u_{i-1,j-1}}{2h} \right] \\ & + q_{i,j} \frac{u_{i,j+1} - u_{i,j-1}}{2k} + r_{i,j} u_{i,j} + s_{i,j} \end{aligned}$$

per $i = 1, \ldots, n$ e $j = 1, \ldots, m$. In esso le derivate u_x e u_{xx} sono approssimate mediante la medie delle loro approssimazioni centrali nei livelli $j+1$ e $j-1$ per ogni $j = 1, \ldots, m$. Poiché la discretizzazione è stata effettuata con le differenze centrali, sia in x che in t, l'errore di discretizzazione $\tau(h, k) = O(h^2 + k^2)$.

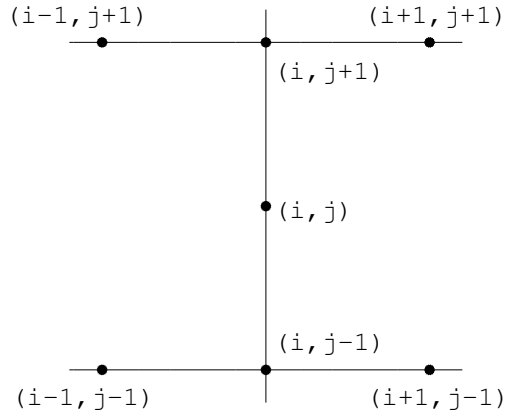


Figura 6.4: Schema di discretizzazione a 7 punti.

Se, in aggiunta agli $u_{i,0}$, fossero noti in valori $u_{i,1}$, lo schema consentirebbe di calcolare gli $u_{i,2}$, $i = 1, \ldots, n$, risolvendo un sistema lineare di ordine n. Di conseguenza, iterando il processo, noti i valori in 2 livelli di j, gli n valori del livello superiore verrebbero calcolati risolvendo un semplice sistema lineare di ordine n.

Tale calcolo può essere effettuato utilizzando la formula di Taylor, troncata al termine di 2° ordine. Da notare che così procedendo si ottiene, per $i = 1, \ldots, n$,

$$\begin{aligned} u_{i,1} = u(x_i, t_1) &\approx u_{i,0} + k u_t(x_i, 0) + \frac{k^2}{2} u_{tt}(x_i, 0) \\ &= \phi_1(x_i) + k\phi_2(x_i) + \frac{k^2}{2} u_{tt}(x_i, 0), \end{aligned}$$

da cui, utilizzando l'equazione stessa,

$$\begin{aligned} u_{i1} &\approx \phi_1(x_i) + k\phi_2(x_i) \\ &\quad + \tfrac{k^2}{2} \left[c^2 u_{xx}(x_i, 0) + p_{i0} u_x(x_i, 0) + q_{i0}\phi_2(x_i) + r_{i0} u_{i0} + s_{i0} \right]. \end{aligned}$$

I valori incogniti di quest'ultima formula, ossia $u_{xx}(x_i, 0)$ e $u_x(x_i, 0)$, possono essere infine valutati mediante lo schema alle differenze centrali

$$\begin{aligned} u_{xx}(x_i, 0) &\approx \frac{u_{i+1,0} - 2u_{i,0} + u_{i-1,0}}{h^2} = \frac{\phi_1(x_{i+1}) - 2\phi_1(x_i) + \phi_1(x_{i-1})}{h^2} \\ u_x(x_i, 0) &\approx \frac{u_{i+1,0} - u_{i-1,0}}{2h} = \frac{\phi_1(x_{i+1}) - \phi_1(x_{i-1})}{2h}. \end{aligned}$$

Così procedendo, se le funzioni assegnate sono tutte continue, l'errore di discretizzazione è complessivamente un $O(h^2 + k^2)$.

Riordinando di conseguenza, ossia tenendo conto del fatto che gli unici valori incogniti sono $u_{i-1,j+1}$, $u_{i,j+1}$ e $u_{i,j+1}$, si ottiene il sistema

$$\begin{aligned}
&\frac{1}{2h}\left(\frac{c^2}{h} - \frac{p_{ij}}{2}\right) u_{i-1,j+1} - \left(\frac{1}{k^2} + \frac{c^2}{h^2} - \frac{q_{ij}}{2k}\right) u_{i,j+1} + \frac{1}{2h}\left(\frac{c^2}{h} + \frac{p_{ij}}{2}\right) u_{i+1,j+1} \\
&= \frac{1}{2h}\left(\frac{p_{ij}}{2} - \frac{c^2}{h}\right) u_{i-1,j-1} + \left(\frac{1}{k^2} + \frac{c^2}{h^2} + \frac{q_{ij}}{2k}\right) u_{i,j-1} \\
&- \frac{1}{2h}\left(\frac{p_{ij}}{2} + \frac{c^2}{h}\right) u_{i+1,j-1} - \left(r_{ij} + \frac{2}{k^2}\right) u_{i,j} - s_{ij},
\end{aligned}$$

dove $i = 1, \ldots, n$ per ogni prefissato $j = 1, \ldots, m$.

Per calcolare $u_{i,j}$ in tutti i punti nodali si debbono dunque risolvere m sistemi lineari tridiagonali, ciascuno di ordine n. Tali sistemi, se $q(x,t) \leq 0$, qualunque sia il valore del peso k, sono tutti diagonalmente dominanti in senso stretto purché risulti $hp \leq 2c^2$, essendo $p = \max\limits_{\substack{a\leq x\leq b \\ 0\leq t\leq T}} |p(x,t)|$.

Nel caso $q(x,t) \geq 0$, valgono le stesse considerazioni, purchè $\frac{1}{k^2} \geq \frac{q_{ij}}{2k}$, per $i = 1, \ldots, n$ e $j = 1, \ldots, m$. Tale condizione è certamente soddisfatta nell'ipotesi che si consideri $k \leq \frac{1}{q}$, essendo $q = \max\{q(x,t),\ a \leq x \leq b,\ 0 \leq t \leq T\}$.

Sotto tali ipotesi per la risoluzione numerica degli m sistemi lineari, si può procedere come nella risoluzione di quelli derivanti dai problemi di tipo parabolico. Si possono cioè utilizzare i metodi di Jacobi e Gauss-Seidel, utilizzando, per ogni livello j, l'interpolante lineare tra $f_1(t_j)$ e $f_2(t_j)$, ossia ponendo

$$u_{ij}^{(0)} = f_1(t_j) + \frac{x_i - a}{b - a}\left[f_2(t_j) - f_1(t_j)\right],$$

dove $j = 0, 1, \ldots, m+1$ e $i = 0, 1, 2 \ldots, n+1$.

Esempio 6.5 Discutere la risoluzione numerica del seguente problema differenziale di tipo iperbolico:

$$\begin{cases}
(x^2+1)u_{tt} = (t^2+2)u_{xx} + 3^{x^2+t^2}u_x + (\cos t)u_t + (x+t)^2 u, \\
0 < x < 10,\ 0 < t < 5, \\
u(0,\, t) = f_1(t), \quad u(10,\, t) = f_2(t), \\
u(x,\, 0) = g_1(x), \quad u_t(x,\, 0) = g_2(x).
\end{cases}$$

Seguendo il procedimento illustrato, si genera la mesh

$$\begin{aligned}
x_i &= ih, \quad i = 0, 1, \ldots, n+1, \quad h = \tfrac{10}{n+1}, \\
t_j &= jk, \quad j = 0, 1, \ldots, m+1, \quad k = \tfrac{5}{m+1},
\end{aligned}$$

e poi si colloca in (x_i, t_j) utilizzando lo schema a 7 punti. Usando il *metodo upwind* per il termine contenente u_x (dato che $3^{x^2+t^2}$ assume nel dominio valori molto elevati) verranno utilizzate le seguenti approssimazioni

$$
\begin{aligned}
&u_{xx}(x_i,\, t_j) \simeq \frac{u_{i+1,\,j+1} - 2u_{i,\,j+1} + u_{i-1,\,j+1}}{2h^2} + \frac{u_{i+1,\,j-1} - 2u_{i,\,j-1} + u_{i-1,\,j-1}}{2h^2},\\
&u_x(x_i, t_j) \simeq \frac{u_{i+1,j+1} - u_{i,j+1}}{2h} + \frac{u_{i+1,j-1} - u_{i,j-1}}{2h},\\
&u_t(x_i, t_j) \simeq \frac{u_{i,j+1} - u_{i,j-1}}{2k},\\
&u_{tt}(x_i,\, t_j) \simeq \frac{u_{i,\,j+1} - 2u_{i,\,j} + u_{i,\,j-1}}{k^2}.
\end{aligned}
$$

La discretizzazione con lo schema a 7 punti e la collocazione dell'equazione differenziale nei punti interni generano il sistema:

$$
\begin{aligned}
&(x_i^2 + 1)\frac{u_{i,\,j+1} - 2u_{i,\,j} + u_{i,\,j-1}}{k^2}\\
&= (t_j^2 + 2)\,\frac{1}{2}\left[\frac{u_{i+1,\,j+1} - 2u_{i,\,j+1} + u_{i-1,\,j+1}}{h^2} + \frac{u_{i+1,\,j-1} - 2u_{i,\,j-1} + u_{i-1,\,j-1}}{h^2}\right]\\
&+ 3^{x_i^2+t_j^2}\,\frac{1}{2}\left[\frac{u_{i+1,j+1} - u_{i,j+1}}{h} + \frac{u_{i+1,j-1} - u_{i,j-1}}{h}\right] + (\cos t_j)\frac{u_{i,j+1} - u_{i,j-1}}{2k}\\
&+ (x_i + t_j)^2\, u_{ij}
\end{aligned}
$$

In questo caso l'errore di discretizzazione è $\tau = O(h + k^2)$. Riordinando l'equazione precedente e tenendo conto che le incognite sono soltanto al livello $j + 1$, si ottiene

$$
\begin{aligned}
&\left(\frac{t_j^2 + 2}{2h^2}\right) u_{i-1,j+1} - \left(\frac{x_i^2 + 1}{k^2} + \frac{t_j^2 + 2}{h^2} + \frac{3^{x_i^2+t_j^2}}{2h} - \frac{\cos t_j}{2k}\right) u_{i,j+1}\\
&+ \left(\frac{t_j^2 + 2}{2h^2} + \frac{3^{x_i^2+t_j^2}}{2h}\right) u_{i+1,j+1} = \phi(u_{i-1,j-1},\, u_{i,j-1},\, u_{i+1,j-1},\, u_{i,j}),
\end{aligned}
$$

dove ϕ indica una funzione nota, dipendente dai soli argomenti esplicitamente indicati.

Tale sistema risulta essere diagonalmente dominante in senso stretto purché $\frac{x_i^2+1}{k^2} > \frac{|\cos t_j|}{2k}$, condizione soddisfatta se $k < 1$. Prendendo ad esempio $k = \frac{1}{50}$ e $h = \frac{1}{100}$, si debbono risolvere 250 sistemi tridiagonali e diagonalmente dominanti ciascuno dell'ordine di 1000. Per la determinazione della soluzione al livello 1, ossia per $t = k$ e per ogni x_i, $i = 1, \ldots, n$, si procede nel modo

seguente:

$$u(x_i,\, k) = u_{i1} = u(x_i, 0) + k u_t(x_i, 0) + \frac{k^2}{2} u_{tt}(x_i, 0) = g_1(x_i) + k g_2(x_i)$$
$$+ \frac{k^2}{2} \left[2 \frac{u_{i+1,0} - 2u_{i,0} + u_{i-1,0}}{h^2} + 3^{x_i^2} \frac{u_{i+1,0} - u_{i-1,0}}{2h} + g_1(x_i) + x_i^2 g_2(x_i) \right]$$
$$= g_1(x_i) + k g_2(x_i) + \frac{k^2}{2} \left[2 \frac{g_1(x_{i+1}) - 2g_1(x_i) + g_1(x_{i-1})}{h^2} \right.$$
$$\left. + 3^{x_i^2} \frac{g_1(x_{i+1}) - g_1(x_{i-1})}{2h} + g_1(x_i) + x_i^2 g_2(x_i) \right],\ i = 1,\, 2,\, \ldots, n.$$

In questo modo si calcola la $u(x_i,\, k)$ con un errore di discretizzazione dell'ordine di k^2, così che l'errore complessivo di discretizzazione risulta ancora un $O(h + k^2)$. Per la risoluzione degli m sistemi lineari si procede come precedentemente specificato.

6.4 Modelli debolmente non lineari

Consideriamo inizialmente il problema differenziale

$$\begin{cases} y''(x) + p(x)\, y'(x) + q(x, y) + r(x) = 0, & a \le x \le b, \\ y(a) = \alpha, \quad y(b) = \beta, \end{cases}$$

con $q(x, y)$ non lineare in y e soddisfacente la condizione $\frac{\partial q}{\partial y} \le 0$, per $a \le x \le b$, qualunque sia y. Tale condizione [12] garantisce l'esistenza e l'unicità della soluzione del problema differenziale. Posto $x_i = a + ih$ e $y_i \approx y(x_i)$, per $i = 0, 1, \ldots, n+1$ e $h = \frac{b-a}{n+1}$, la collocazione in x_i dell'equazione differenziale con lo schema alle differenze centrali trasforma il problema iniziale nel sistema non lineare

$$\begin{cases} \dfrac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + p_i \dfrac{y_{i+1} - y_{i-1}}{2h} + q(x_i, y_i) + r_i = 0, & i = 1, \ldots, n \\ y_0 = \alpha, \quad y_{n+1} = \beta. \end{cases}$$

Questo sistema è dunque esprimibile nella forma

$$f_i(y_{i-1}, y_i, y_{i+1}) = 0, \quad i = 1, 2, \ldots, n, \qquad y_0 = \alpha, \quad y_{n+1} = \beta,$$

essendo

$$f_i(y_{i-1}, y_i, y_{i+1}) = \left(1 - \frac{h}{2} p_i\right) y_{i-1} - 2y_i + \left(1 + \frac{h}{2} p_i\right) y_{i+1} + q(x_i, y_i) h^2 + h^2 r_i.$$

Condizione sufficiente [12, 20] affinché tale sistema sia univocamente risolubile è che h sia tale da soddisfare la condizione

$$ph \leq 2,$$

essendo $p = \max_{a \leq x \leq b} |p(x)|$.

Da notare che se il sistema fosse lineare, ossia $q(x, y) = q(x)y$, questa condizione implicherebbe la dominanza diagonale del sistema, nell'ipotesi che sia $q(x) \leq 0$ per $a \leq x \leq b$ con $ph \leq 2$.

Nelle suddette ipotesi la soluzione può essere ottenuta mediante il metodo iterativo di Newton-Jacobi consistente nel calcolare, per ogni $k = 0, 1, \ldots$, gli iterati

$$y_i^{(k+1)} = y_i^{(k)} - \frac{f_i(y_{i-1}^{(k)}, y_i^{(k)}, y_{i+1}^{(k)})}{\frac{\partial f_i}{\partial y_i}(y_{i-1}^{(k)}, y_i^{(k)}, y_{i+1}^{(k)})}, \qquad i = 1, \ldots, n,$$

dove

$$f_i(y_{i-1}^{(k)}, y_i^{(k)}, y_{i+1}^{(k)}) = \left[1 - \frac{h}{2}p_i\right] y_{i-1}^{k} - 2y_i^{(k)} + \left[1 + \frac{h}{2}p_i\right] y_{i+1}^{(k)} + q(x_i, y_i^{(k)})h^2 + h^2 r_i,$$

e

$$\frac{\partial f_i}{\partial y_i}(y_{i-1}^{(k)}, y_i^{(k)}, y_{i+1}^{(k)}) = -2 + \frac{\partial q}{\partial y_i}(x_i, y_i^{(k)}),$$

essendo $y_0^{(k)} = \alpha$ e $y_{n+1}^{(k)} = \beta$, $k = 0, 1, \ldots$. Nelle citate ipotesi il metodo è **globalmente convergente**, dunque teoricamente indipendente dei valori $y_i^{(0)}$, $i = 2, \ldots, n$. Poiché $y_0^{(0)} = \alpha$ e $y_{n+1}^{(0)} = \beta$, la scelta più frequente del vettore iniziale consiste nell'assumere

$$y_i^{(0)} = \alpha + \frac{i}{n+1}(\beta - \alpha), \qquad i = 0, 1, \ldots, n+1,$$

ossia nell'interpolare linearmente i valori α e β. Questo allo scopo di evitare che un elevato numero di iterati possa pregiudicare i risultati, in conseguenza della propagazione degli errori di arrotondamento.

Consideriamo ora il seguente problema debolmente nonlineare di tipo ellittico

$$\begin{cases} u_{xx} + u_{yy} + p(x, y)u_x + q(x, y)u_y + r(x, y, u) + s(x, y) = 0, & (x, y) \in \Omega, \\ u(x, y) = f(x, y), & (x, y) \in \partial\Omega, \end{cases}$$

dove, per semplicità $\Omega = [a, b] \times [c, d]$ e la $r(x, y, u)$, non dipendente linearmente da u, soddisfa la condizione $\frac{\partial r}{\partial u} \leq 0$, per $(x, y) \in \Omega$, qualunque sia u. Ipotesi che garantisce [12] l'esistenza e l'unicità del problema differenziale.

Posto

$$x_i = a + ih, \quad i = 0, 1, \ldots, n+1, \quad h = \frac{b-a}{n+1},$$
$$y_j = c + jk, \quad j = 0, 1, \ldots, m+1, \quad k = \frac{d-c}{m+1},$$

la discretizzazione dell'equazione differenziale con il metodo alle differenze centrali genera il sistema debolmente non lineare

$$\frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2} + \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{k^2} + p_{i,j}\frac{u_{i+1,j} - u_{i-1,j}}{2h} + q_{i,j}\frac{u_{i,j+1} - u_{i,j-1}}{2k} + r(x_i, y_j, u_{i,j}) + s(x_i, y_j) = 0$$

per $i = 1, \ldots, n$ e $j = 1, \ldots, m$. Tale sistema può essere espresso nella forma più compatta

$$\phi_{i,j}(u_{i,j-1}, u_{i-1,j}, u_{i,j}, u_{i+1,j}, u_{i,j+1}) = 0, \qquad i = 1, \ldots, n, \quad j = 1, \ldots, m,$$

essendo

$$\phi_{i,j}(u_{i,j-1}, u_{i-1,j}, u_{i,j}, u_{i+1,j}, u_{i,j+1}) = h^2(2 - kq_{ij})u_{i,j-1} + k^2(2 - hp_{ij})u_{i-1,j} - 4(h^2 + k^2)u_{i,j} + k^2(2 + hp_{ij})u_{i+1,j} + h^2(2 + kq_{ij})u_{i,j+1} + 2h^2k^2[r(x_i, y_j, u_{i,j}) + s(x_i, y_j)],$$

con $u_{i,0} = f(x_i, c)$ e $u_{i,m+1} = f(x_i, d)$, $i = 0, 1, \ldots, n+1$; $u_{0,j} = f(a, y_j)$ e $u_{n+1,j} = f(b, y_j)$ per $j = 0, 1, \ldots, m+1$.

Condizione sufficiente affinché questo sistema sia univocamente risolubile, è che h e k siano tali da soddisfare la seguente condizione:

$$Ph \leq 2 \quad \text{e} \quad Qk \leq 2,$$

essendo $P = \max |p(x, y)|$ e $Q = \max |q(x, y)|$ per $(x, y) \in \Omega$.

In caso di linearità di s rispetto ad u, le condizioni precedenti assicurano che la matrice del sistema, relativo alla parte lineare, sia diagonalmente dominante, come già rilevato in precedenza.

Nelle suddette ipotesi il metodo di Newton/Jacobi applicato al sistema debolmente lineare ottenuto per discretizzazione come sopra specificato, risulta **globalmente convergente**. Di conseguenza, prefissato $l = 0, 1, \ldots$, la soluzione può essere ottenuta mediante il processo iterativo

$$u_{i,j}^{(l+1)} = u_{i,j}^{(l)} - \frac{\phi_{i,j}(u_{i,j-1}^{(l)}, u_{i-1,j}^{(l)}, u_{i,j}^{(l)}, u_{i+1,j}^{(l)}, u_{i,j+1}^{(l)})}{\frac{\partial \phi_{i,j}}{\partial u_{i,j}}(u_{i,j-1}^{(l)}, u_{i-1,j}^{(l)}, u_{i,j}^{(l)}, u_{i+1,j}^{(l)}, u_{i,j+1}^{(l)})},$$

per $i = 1, \ldots, n$ e $j = 1, \ldots, m$, essendo

$$\frac{\partial \phi_{i,j}}{\partial u_{i,j}} = -4(h^2 + k^2) + 2h^2k^2 \frac{\partial r}{\partial u_{i,j}},$$

con $u^{(l)}_{i,0} = f(x_i, c)$ e $u^{(l)}_{i,m+1} = f(x_i, d)$, $i = 0, 1, \ldots, n+1$; $u^{(l)}_{0,j} = f(a, y_j)$ e $u^{(l)}_{n+1,j} = f(b, y_j)$, $j = 0, 1, \ldots, m+1$.

Poiché la u è nota al contorno (vettore inziale), i valori $u^{(0)}_{ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, m$, possono essere efficacemente ottenuti per interpolazione rispetto ai valori assunti sul contorno. Un punto importante del procedimento riguarda infatti la scelta del vettore d'innesco $u^{(0)}_{ij}$. Per spiegare la tecnica, iniziamo con l'indicare le condizioni al contorno in questo modo:

$$\begin{cases} u(a,\, y) = f_1(y), \quad u(b,\, y) = f_2(y), \\ u(x,\, c) = \varphi_1(x), \quad u(x,\, d) = \varphi_2(x). \end{cases}$$

A questo punto interpoliamo $f_1(y)$, $f_2(y)$ e $\varphi_1(x)$ e $\varphi_2(x)$ ponendo

$$f(x, y) = \frac{(x-a)^2 f_2(y) + (x-b)^2 f_1(y)}{(x-a)^2 + (x-b)^2},$$
$$\varphi(x, y) = \frac{(y-c)^2 \varphi_2(x) + (y-d)^2 \varphi_1(x)}{(y-c)^2 + (y-d)^2}.$$

Ottenuti $f(x, y)$ e $\varphi(x, y)$, la scelta del vettore d'innesco può essere fatta ricorrendo alla funzione di interpolazione *bidimensinale*

$$F(x,\, y) = \frac{(x-a)^2(x-b)^2 \varphi(x,\, y) + (c-y)^2(d-y)^2 f(x,\, y)}{(x-a)^2(x-b)^2 + (c-y)^2(d-y)^2}.$$

Tale funzione è effettivamente interpolante, dato che

$$\begin{aligned} F(a, y) &= f(a, y) = f_1(y), \\ F(b, y) &= f(b, y) = f_2(y), \\ F(x, c) &= \varphi(x, c) = \varphi_1(x), \\ F(x, d) &= \varphi(x, d) = \varphi_2(x). \end{aligned}$$

Il vettore d'innesco $u^{(0)}_{i,j}$, $i = 1, \ldots, n$ e $j = 1, \ldots, m$ è dato da $u^{(0)}_{i,j} = F(x_i, y_j)$ in ciascuno dei punti nodali prefissati.

6.5 Problema spettrale di Helmholtz

Allo scopo di mostrare come si possono utilizzare gli schemi alle differenze finite nella trattazione delle PDEs con coefficienti discontinui, consideriamo l'equazione 2D (bidimensionale) di Helmholtz

$$-\nabla \cdot \left(\frac{1}{\varepsilon}\nabla\psi\right) = \eta\psi \tag{6.4}$$

su un rettangolo $a \leq x \leq b$, $c \leq y \leq d$ [6]. Tenuto conto del significato dei simboli di gradiente di una funzione e di divergenza di un campo vettoriale, la (6.4), in forma più esplicita, può essere espressa nel modo seguente:

$$-\frac{\partial}{\partial x}\left(\frac{1}{\varepsilon}\frac{\partial\psi}{\partial x}\right) - \frac{\partial}{\partial y}\left(\frac{1}{\varepsilon}\frac{\partial\psi}{\partial y}\right) = \eta\psi, \tag{6.5}$$

dove $\psi = \psi(x, y)$, $\varepsilon = \varepsilon(x, y)$ e $\eta = \eta(x, y)$ con $(x, y) \in \Omega = [a, b] \times [c, d]$. La (6.5) rappresenta il problema spettrale associato all'operatore 2D di Helmholtz nel dominio rettangolare Ω.

Supponiamo ora che la funzione nota ε soddisfi le seguenti condizioni di periodicità:

$$\begin{aligned} &\varepsilon(a, y) = \varepsilon(b, y), \quad \varepsilon_x(a, y) = \varepsilon_x(b, y), \quad c \leq y \leq d; \\ &\varepsilon(x, c) = \varepsilon(x, d), \quad \varepsilon_y(x, c) = \varepsilon_y(x, d), \quad a \leq x \leq b. \end{aligned}$$

Sotto tali ipotesi possiamo richiedere che la ψ sia periodica, ossia che risulti:

$$\begin{aligned} &\psi(a, y) = \psi(b, y), \quad \psi_x(a, y) = \psi_x(b, y), \quad c \leq y \leq d; \\ &\psi(x, c) = \psi(x, d), \quad \psi_y(x, c) = \psi_y(x, d), \quad a \leq x \leq b. \end{aligned}$$

La funzione $\eta(x, y)$ rappresenta l'autovalore dell'operatore di Helmholtz corrispondente all'autofunzione $\psi(x, y)$. Dal punto di vista fisico, il modello governa la propagazione della componente elettrica del campo elettromagnetico creato dalla luce che attraversa un mezzo periodico. Questo problema è diventato molto attuale nello studio delle nanostrutture (strutture dell'ordine di 10^{-6} mm) di tipo periodico, come lo sono i cristalli fotonici. In tale ambito la funzione $\varepsilon(x, y)$ indica la costante dielettrica del mezzo in (x, y). La teoria stabilisce che gli autovalori, definiti modi nell'elettromagnetismo, costituiscono una infinità numerabile η_{ij}, $i, j = 1, 2, \ldots$.

Supponiamo ora che, come avviene in varie applicazioni ingegneristiche, il dominio Ω sia formato dall'unione di un numero finito di domini omogenei, in ciascuno dei quali la $\varepsilon(x, y)$ sia continua. Più precisamente, come indicato nella

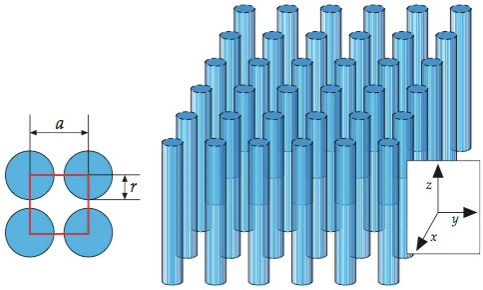


Figura 6.5: La figura mostra lo schema di un cristallo fotonico bidimensionale costituito da una distribuzione cilindrica della costante dielettrica ε [15].

Fig. 6.5 supponiamo che all'interno di Ω siano presenti dei cerchi, ciascuno di raggio r, e che in esso

$$\varepsilon(x,y) = \begin{cases} c_1, & \text{in ciascuno dei cerchi;} \\ c_2, & \text{nella parte restante del dominio,} \end{cases}$$

essendo $c_1 > c_2 \geq 1$. Procedendo come al solito, reticoliamo il dominio introducendo un insieme di punti nodali

$$x_i = a + ih_x, \quad i = 0, 1, \ldots, n+1, \quad h_x = \frac{b-a}{n+1};$$
$$y_j = b + jh_y, \quad j = 0, 1, \ldots, m+1, \quad h_y = \frac{d-c}{m+1}.$$

Al fine di pervenire ad un sistema lineare con il massimo delle simmetrie possibili, collochiamo l'equazione differenziale in (x_i, y_j) e discretizziamo con le differenze centrali con passi $h_x/2$ e $h_y/2$ rispettivamente. Questo equivale a considerare in (x_i, y_j) l'equazione

$$-\frac{\left(\frac{1}{\varepsilon}\frac{\partial\psi}{\partial x}\right)_{i+1/2,j} - \left(\frac{1}{\varepsilon}\frac{\partial\psi}{\partial x}\right)_{i-1/2,j}}{h_x} - \frac{\left(\frac{1}{\varepsilon}\frac{\partial\psi}{\partial y}\right)_{i+1/2,j} - \left(\frac{1}{\varepsilon}\frac{\partial\psi}{\partial y}\right)_{i-1/2,j}}{h_y} = \eta^{(i,j)}\psi_{i,j},$$

dove $i = 1, \ldots, n$ e $j = 1, \ldots, m$. Per tenere conto delle discontinuità della $\varepsilon(x,y)$, approssimiamo il valore della ε in un punto con la media dei due punti contigui. Discretizzando ancora con le differenze centrali e tenendo conto di tale osservazione, otteniamo la seguente (i,j)-ma equazione:

$$-\frac{\frac{1}{2}\left(\frac{1}{\varepsilon_{i+1,j}} + \frac{1}{\varepsilon_{i,j}}\right)\frac{\psi_{i+1,j}-\psi_{i,j}}{h_x} - \frac{1}{2}\left(\frac{1}{\varepsilon_{i,j}} + \frac{1}{\varepsilon_{i-1,j}}\right)\frac{\psi_{i,j}-\psi_{i-1,j}}{h_x}}{h_x}$$
$$-\frac{\frac{1}{2}\left(\frac{1}{\varepsilon_{i,j+1}} + \frac{1}{\varepsilon_{i,j}}\right)\frac{\psi_{i,j+1}-\psi_{i,j}}{h_y} - \frac{1}{2}\left(\frac{1}{\varepsilon_{i,j}} + \frac{1}{\varepsilon_{i,j-1}}\right)\frac{\psi_{i,j}-\psi_{i,j-1}}{h_y}}{h_y} = \eta^{(i,j)}\psi_{i,j}.$$

Riordinando opportunamente essa diventa

$$
\begin{aligned}
&-\frac{1}{2}\left(\frac{1}{\varepsilon_{i,j}}+\frac{1}{\varepsilon_{i,j-1}}\right)\frac{1}{h_y^2}\psi_{i,j-1}-\frac{1}{2}\left(\frac{1}{\varepsilon_{i,j}}+\frac{1}{\varepsilon_{i-1,j}}\right)\frac{1}{h_x^2}\psi_{i-1,j}\\
&\left[\frac{1}{2}\left(\frac{1}{\varepsilon_{i+1,j}}+\frac{2}{\varepsilon_{i,j}}+\frac{1}{\varepsilon_{i-1,j}}\right)\frac{1}{h_x^2}+\frac{1}{2}\left(\frac{1}{\varepsilon_{i,j+1}}+\frac{2}{\varepsilon_{i,j}}+\frac{1}{\varepsilon_{i,j-1}}\right)\frac{1}{h_y^2}\right]\psi_{i,j}\\
&-\frac{1}{2}\left(\frac{1}{\varepsilon_{i+1,j}}+\frac{1}{\varepsilon_{i,j}}\right)\frac{1}{h_x^2}\psi_{i+1,j}-\frac{1}{2}\left(\frac{1}{\varepsilon_{i,j+1}}+\frac{1}{\varepsilon_{i,j}}\right)\frac{1}{h_y^2}\psi_{i,j+1}=\eta^{(i,j)}\psi_{i,j},
\end{aligned}
$$

dove $i=1,\ldots,n$ e $j=1,\ldots,m$.

Le condizioni di periodicità delle funzioni ε e ψ, in termini delle differenze finite, implicano che

$$
\begin{cases}
\varepsilon_{0,j}=\varepsilon_{n,j} \quad \text{e} \quad \varepsilon_{1,j}=\varepsilon_{n+1,j},\\
\psi_{0,j}=\psi_{n,j} \quad \text{e} \quad \psi_{1j}=\psi_{n+1,j}, \quad \text{per} \quad j=0,1,\ldots,m+1;
\end{cases}
$$

$$
\begin{cases}
\varepsilon_{i,0}=\varepsilon_{i,m} \quad \text{e} \quad \varepsilon_{i,1}=\varepsilon_{i,m+1},\\
\psi_{i,0}=\psi_{i,m} \quad \text{e} \quad \psi_{i,1}=\psi_{i,m+1}, \quad \text{per} \quad i=0,1,\ldots,n+1;
\end{cases}
$$

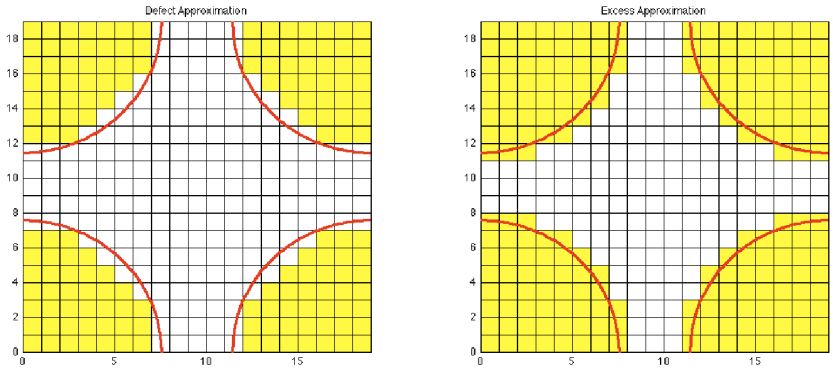


Figura 6.6: La figura mostra una discretizzazione regolare del dominio Ω dove la costante dielettrica ε viene campionata per difetto (immagine a sinistra) e per eccesso (immagine a destra).

Queste considerazioni sulla periodicità consentono di ridurre il numero delle incognite del sistema con le sostituzioni $\psi_{n+1,j}=\psi_{1,j}$, $\psi_{n,j}=\psi_{0,j}$, $\psi_{i,m+1}=\psi_{i,1}$ e $\psi_{i,m}=\psi_{i,0}$. In tal modo il sistema risulta di ordine $(n-1)(m-1)$. Naturalmente ci si deve ricordare di effettuare le analoghe sostituzioni sulla ε, ponendo $\varepsilon_{i,m+1}=\varepsilon_{i,1}$, $\varepsilon_{i,m}=\varepsilon_{i,0}$ e $\varepsilon_{n+1,j}=\varepsilon_{1,j}$ e $\varepsilon_{n,j}=\varepsilon_{0,j}$. Poiché la matrice è simmetrica e debolmente diagonale, i suoi autovalori sono tutti non negativi. Più precisamente essa possiede un autovalore semplice uguale a zero, mentre gli altri sono strettamente positivi. È altresì possibile osservare che gli

autovalori sono debolmente decrescenti al crescere della costante dielettrica. Per questo motivo si considerano due mesh differenti, una contenente valori di ε approssimati per difetto e una contenente una valutazione per eccesso, come evidenziato nella Fig. 6.6. Come valori effettivi si considerano le loro medie aritmetiche. Tenuto conto delle caratteristiche della matrice, gli autovalori si possono calcolare, con buona precisione, utilizzando le opportune routines del MatLab.

A titolo esemplificativo, posto $c_1 = 1$ il valore della costante dielettrica al di fuori dei cerchi, sono state effettuate le seguenti scelte di c_2:

$$c_2 = 3, 5, 7, 9,$$

avendo posto $r = 0.2\,L$, come raggio dei cerchi riportati nel dominio, rappresentato da un quadrato di lato L. I valori dei primi 10 autovalori (ordinati per valori crescenti) sono stati riportati nella Fig. 6.7.

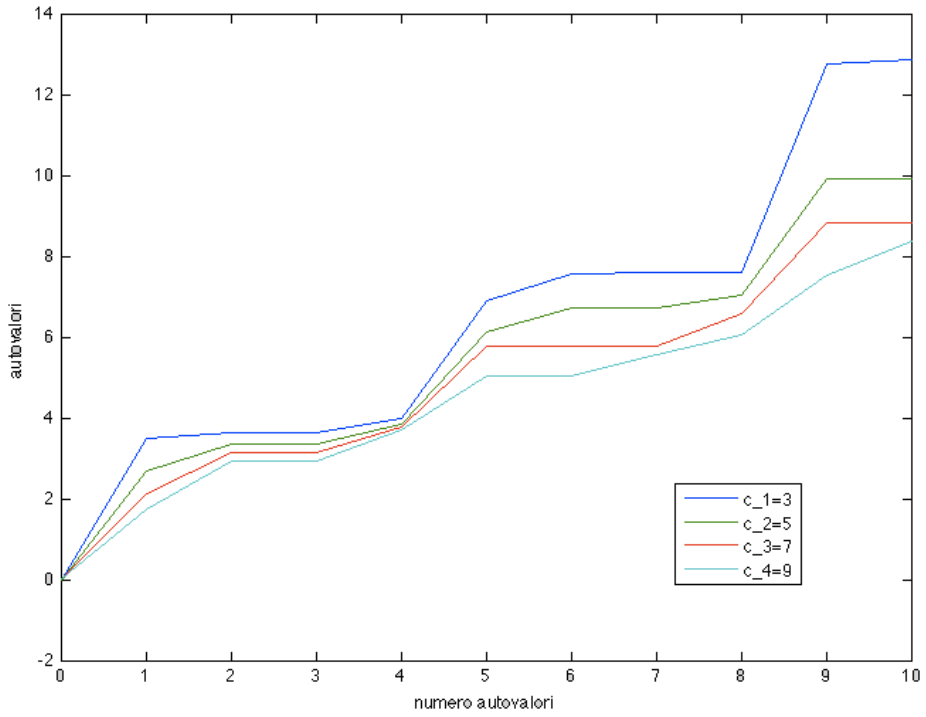


Figura 6.7: La figura mostra i primi dieci autovalori del problema spettrale (6.4) con condizioni al contorno periodiche, nel caso in cui il valore della costante dielettrica al di fuori dei cerchi sia uno e i valori all'interno dei cerchi siano $c_2 = 3, 5, 7, 9$, rispettivamente.

6.6 Una tipica applicazione industriale

Una delle più note applicazioni delle differenze finite, nella geofisica applicata, è la *seismic migration*, una metodologia per la ricostruzione della struttura geologica del sottosuolo, orientata alla localizzazione di giacimenti di acqua o idrocarburi. L'industria petrolifera ha sviluppato questa tecnologia a partire dagli anni '70. Camion speciali, dal peso di 40 tonnellate e dotati di piattaforme vibranti generano onde acustiche ed elastiche che propagandosi nel sottosuolo, vengono riflesse da eventuali strutture geologiche e registrate alla superficie mediante appositi ricevitori. I risultati ottenuti vengono opportunamente trattati e utilizzati dagli analisti come strumento di supporto alle decisioni per l'identificazione di eventuali giacimenti. A seconda della profondità e del tipo di applicazione, possono essere usate anche altre sorgenti d'onda, come gli *air-gun* (cannoni ad aria compressa) o cariche esplosive.

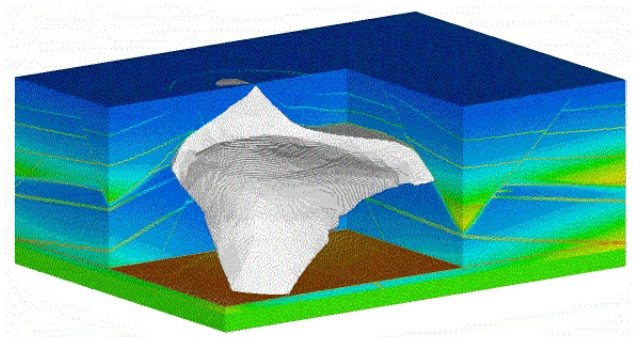


Figura 6.8: Un tipico modello 3D a differenze finite utilizzato per la seismic migration. Immagine del Center for Economic Migration and Tomography, Reno, Nevada.

La simulazione della propagazione delle onde acustiche/elastiche è particolarmente efficace con metodi alle differenze finite, perché il dominio computazionale è estremamente regolare: una porzione di suolo a forma di parallelepipedo, in cui in generale anche la faccia superiore corrispondente alla superficie libera è piana, perché, tenuto conto dell'estensione del dominio sotto esame, si considera non rilevante la topologia di superficie. Nella Fig. 6.8 è mostrato un tipico modello 3D a differenze finite nel quale i confini tra regioni con diverse velocità di propagazione delle onde, se obliqui, vengono discretizzati tramite una struttura "a scaletta", con un rapido alternarsi di segmenti orizzontali e verticali. Al fine di contenere al massimo la conseguente perdita di accuratezza si infittisce la mesh di riferimento, naturalmente aumentando la dimensione dei problemi algebrici da risolvere.

La Figura 6.9 mostra il risultato di una seismic migration relativa ad un sito nel Nevada (USA), utilizzato per la produzione di energia geotermica, nel quale è evidenziato un possibile giacimento di acqua calda intrappolato all'interno di una serie di faglie, ad una profondità di circa 4 km.

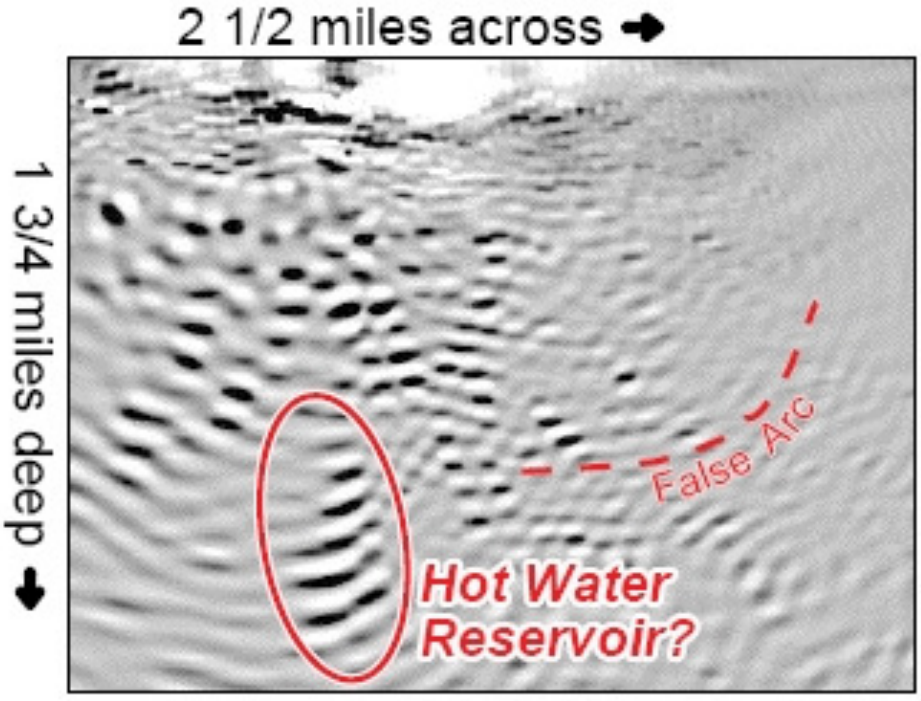


Figura 6.9: Seismic migration relativa ad un sito nel Nevada (USA) utilizzato per la produzione di energia geotermica nel quale è evidenziato un possibile giacimento di acqua calda ad una profondità di circa 4 km. Immagine del Center for Economic Migration and Tomography, Reno, Nevada.

6.7 Esercizi proposti

(a) Discutere la risoluzione numerica dei seguenti problemi differenziali lineari:

$$\begin{cases} (2+\sin x)u_{xx} + (y^2+1)u_{yy} + (xy)u_x + (x\cos y)u_y, \\ -(3-\sin xy)u = 0, \\ -\pi \le x \le \pi, \quad 0 \le y \le 3\pi, \\ u(-\pi, y) = f_1(y), \quad u(\pi, y) = f_2(y), \\ u(x, 0) = g_1(x), \quad u(x, 3\pi) = g_2(x). \end{cases}$$

$$\begin{cases} u_t = (xt)u_{xx} + (\sin xt)u_x + (t\cos x)u + x\sin xt, \\ -3 \le x \le 4, \quad 0 \le t \le 6, \\ u(-3,t) = f_1(t), \quad u(4,t) = f_2(t), \\ u(x,0) = g(x). \end{cases}$$

$$\begin{cases} u_{tt} = (1 + x^2 t)u_{xx} + (xt)u_x + 2^{xt}u + \cos xt, \\ -2 \le x \le 3, \quad 0 \le t \le 5, \\ u(-2,t) = f_1(t), \quad u(3,t) = f_2(t), \\ u(x,0) = g_1(x), \quad u_t(x,0) = g_2(x). \end{cases}$$

$$\begin{cases} u_{tt} = (2 + x^2 t^2)u_{xx} + (x+t)u_t - (x^2 + t)u_x + \cos(xt), \\ -3 \le x \le 4, \quad 0 \le t \le 5, \\ u(-3,t) = f_1(t), \quad u(4,t) = f_2(t), \\ u(x,0) = g_1(x), \quad u_t(x,0) = g_2(x). \end{cases}$$

(b) Discutere la risoluzione numerica dei seguenti problemi differenziali, debolmente nonlineari

$$\begin{cases} 2y'' + (x\sin x)y' - (2 - \cos x)y^7 = 0, \\ -3 \le x \le 3, \\ y(-3) = 1, \quad y(3) = 5. \end{cases}$$

$$\begin{cases} (2 + \sin x)u_{xx} + (y^2 + 1)u_{yy} + (xy)u_x + (x\cos y)u_y, \\ -(3 - \sin xy)u^7 = 0, \\ -\pi \le x \le \pi, \quad 0 \le y \le 3\pi, \\ u(-\pi, y) = f_1(y), \quad u(\pi, y) = f_2(y), \\ u(x,0) = g_1(x), \quad u(x, 3\pi) = g_2(x). \end{cases}$$

$$\begin{cases} u_{xx} + 2u_{yy} + (xy)^2 u_x + (x+y)u_y + (\sin u - 2u)^3 = x^2 y, \\ -4 \le x \le 5, \quad 3 \le y \le 6, \\ u(-4,y) = f_1(y), \quad u(5,y) = f_2(y), \\ u(x,3) = g_1(x), \quad u(x,6) = g_2(x). \end{cases}$$

Suggerimenti: Per una trattazione più generale e per ulteriori approfondimenti si rinvia ai libri [12] e [18] della bibliografia.

Capitolo 7

SISTEMI NON LINEARI

In questo capitolo verranno illustrati alcuni risultati base sui sistemi non lineari evidenziando, altresì, le principali differenze tra i metodi di risoluzione dei sistemi lineari e di quelli non lineari.

7.1 Definizioni e risultati basilari

Per sistema non lineare, intendiamo un sistema di equazioni non lineari

$$f_i(x_1, x_2, \ldots, x_n) = 0, \qquad i = 1, 2, \ldots, n, \tag{7.1a}$$

che, con notazione vettoriale, indichiamo nella forma

$$\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{0}, \tag{7.1b}$$

dove $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^n$, la cui i-esima componente è $f_i(x_1, x_2, \ldots, x_n)$ e $\boldsymbol{x}$ è il vettore $(x_1, x_2, \ldots, x_n)^T$. Per $n = 1$, il sistema si riduce ad una equazione non lineare

$$f(x) = 0.$$

In questo ambito, l'analogo della $f'(x)$ è rappresentato dalla *matrice Jacobiana* della $\boldsymbol{F}(\boldsymbol{x})$, così definita

$$\boldsymbol{F}'(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix}, \tag{7.2}$$

dove

$$\frac{\partial f_i}{\partial x_j} = \frac{\partial}{\partial x_j} f_i(x_1, x_2, \ldots, x_n), \qquad i, j = 1, 2, \ldots, n.$$

Nel caso vettoriale l'analogo della derivabilità di una $f : \mathbb{R} \to \mathbb{R}$ è stabilito dalla seguente

Definizione. Una funzione vettoriale $F : \mathbb{R}^n \to \mathbb{R}^n$ è *totalmente differenziabile* (differenziabile nel senso di Fréchet) in $\boldsymbol{x}$ se la matrice Jacobiana (7.2) esiste in $\boldsymbol{x}$ e

$$\lim_{\boldsymbol{h}\to\boldsymbol{0}} \frac{\|\boldsymbol{F}(\boldsymbol{x}+\boldsymbol{h}) - \boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}'(\boldsymbol{x})\boldsymbol{h}\|}{\|\boldsymbol{h}\|} = 0. \tag{7.3}$$

Come è immediato osservare la (7.3) si riduce alla differenziabilità ordinaria nel caso $n = 1$. Analogamente a quanto avviene per la differenziabilità (nel caso $n = 1$), la totale differenziabilità implica la continuità. Basta infatti osservare che, per la disuguaglianza triangolare della norma,

$$\begin{aligned}\|\boldsymbol{F}(\boldsymbol{x}+\boldsymbol{h}) - \boldsymbol{F}(\boldsymbol{x})\| &= \|[\boldsymbol{F}(\boldsymbol{x}+\boldsymbol{h}) - \boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}'(\boldsymbol{x})\boldsymbol{h}] + \boldsymbol{F}'(\boldsymbol{x})\boldsymbol{h}\| \\ &\le \|\boldsymbol{F}(\boldsymbol{x}+\boldsymbol{h}) - \boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}'(\boldsymbol{x})\boldsymbol{h}\| + \|\boldsymbol{F}'(\boldsymbol{x})\boldsymbol{h}\|\end{aligned}$$

che tende a zero (per la (7.3)) per $\|\boldsymbol{h}\| \to 0$.

Non sempre l'estensione dei risultati dal caso $n = 1$ al caso generale è semplice, come bene evidenziato in [20, pp. 141-142]. L'importante teorema del valor medio per una funzione unidimensionale f, ad esempio, non è immediatamente estendibile al caso n-dimensionale.

Nel caso unidimensionale esso stabilisce che se f è differenziabile in $[a, b]$, per ogni coppia di valori $x, y \in [a, b]$, esiste un punto z tra x e y con

$$f(y) - f(x) = f'(z)(y - x).$$

La sua estensione per funzioni $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^n$, differenziabili nel senso di Fréchet, è la seguente:

$$\boldsymbol{F}(\boldsymbol{y}) - \boldsymbol{F}(\boldsymbol{x}) = \int_0^1 \boldsymbol{F}'[\boldsymbol{x} + \theta(\boldsymbol{y} - \boldsymbol{x})](\boldsymbol{y} - \boldsymbol{x})\, d\theta. \tag{7.4}$$

Per la dimostrazione basta osservare che, posto

$$g_i(\theta) = f_i[\boldsymbol{x} + \theta(\boldsymbol{y} - \boldsymbol{x})], \qquad 0 < \theta < 1, \quad i = 1, 2, \ldots, n,$$

$$f_i(\boldsymbol{y}) - f_i(\boldsymbol{x}) = g_i(1) - g_i(0) = \int_0^1 g_i'(\theta)\, d\theta = \int_0^1 \sum_{j=1}^{n} \partial_j f_i[\boldsymbol{x} + \theta(\boldsymbol{y} - \boldsymbol{x})](y_j - y_i)\, d\theta,$$

che rappresenta la i-esima componente della relazione vettoriale (7.4).

Dalla (7.4) deriva immediatamente che

$$\boldsymbol{F}(\boldsymbol{y}) - \boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}'(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) = \int_0^1 [\boldsymbol{F}'[\boldsymbol{x} + \theta(\boldsymbol{y} - \boldsymbol{x}) - \boldsymbol{F}'(\boldsymbol{x})](\boldsymbol{y} - \boldsymbol{x})\, d\theta$$

da cui, se la $\boldsymbol{F}'$ è continua in un insieme Ω e $\boldsymbol{x}, \boldsymbol{y} \in \Omega$ sono abbastanza vicini,

$$\boldsymbol{F}(\boldsymbol{y}) - \boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}'(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) \simeq \boldsymbol{0}. \tag{7.5}$$

Poiché solo in casi molto particolari si riesce a risolvere tali problemi in modo esplicito, la generalità dei metodi numerici è basata su procedimenti iterativi. I metodi iterativi si compongono delle due fasi seguenti:

a) ricerca di una funzione di iterazione Φ con la proprietà che ogni soluzione ξ del sistema $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{0}$ è un punto fisso per la Φ, tale cioè da aversi

$$\Phi(\xi) = \xi$$

e viceversa, ogni punto fisso ξ della funzione di iterazione Φ è una soluzione del sistema $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{0}$;

b) generazione di una successione di iterati $\boldsymbol{x}^{(i)}$ a partire da un punto iniziale $\boldsymbol{x}^{(0)}$.

In questo contesto si pongono i seguenti problemi:

1) scelta di una funzione di iterazione computazionalmente valida;

2) scelta di un punto iniziale $\boldsymbol{x}^{(0)}$ e verifica della convergenza della successione $\{\boldsymbol{x}^{(i)}\}_{i=0}^{\infty}$;

3) valutazione della velocità di convergenza della successione $\{\boldsymbol{x}^{(i)}\}_{i=0}^{\infty}$.

Nel seguito, sia nel caso scalare sia in quello vettoriale, faremo sempre riferimento a metodi iterativi del tipo

$$\boldsymbol{x}^{(i+1)} = \Phi(\boldsymbol{x}^{(i)}), \qquad i = 0, 1, \ldots, \tag{7.6}$$

dove $\boldsymbol{x}^{(0)}$ è il punto iniziale e Φ la funzione di iterazione definita in un dominio chiuso $\overline{\Omega}$ in $\mathbb{R}^n$. Supporremo inoltre che la Φ possieda almeno un punto fisso ξ ed indicheremo con I_ξ un suo intorno. Come abbiamo già osservato, quest'ultima ipotesi equivale a supporre che ξ sia una soluzione del problema iniziale.

Supponiamo ora che ξ sia una soluzione del sistema (7.1b) e I_ξ un intorno sufficientemente piccolo di ξ. Indicato con $\boldsymbol{x}^{(0)}$ un punto di tale intorno, la (7.5) indica che

$$\boldsymbol{F}(\xi) = \boldsymbol{0} \simeq \boldsymbol{F}(\boldsymbol{x}^{(0)}) + \boldsymbol{F}'(\boldsymbol{x}^{(0)})(\xi - \boldsymbol{x}^{(0)}).$$

Pertanto, se la matrice Jacobiana $\boldsymbol{F}'(\boldsymbol{x})$ è non singolare in I_ξ, possiamo definire $\boldsymbol{x}^{(1)}$ come soluzione del sistema

$$\boldsymbol{F}'(\boldsymbol{x}^{(0)})(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}) = -\boldsymbol{F}(\boldsymbol{x}^{(0)}),$$

ponendo

$$\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} - (\boldsymbol{F}'(\boldsymbol{x}^{(0)}))^{-1}\boldsymbol{F}(\boldsymbol{x}^{(0)}).$$

Questa osservazione suggerisce di introdurre la funzione di iterazione

$$\Phi(\boldsymbol{x}) = \boldsymbol{x} - (\boldsymbol{F}'(\boldsymbol{x}))^{-1}\boldsymbol{F}(\boldsymbol{x}), \tag{7.7}$$

dalla quale, iterando, deriva che

$$\boldsymbol{x}^{(i+1)} = \boldsymbol{x}^{(i)} - \boldsymbol{F}'(\boldsymbol{x}^{(i)})^{-1}\boldsymbol{F}(x^{(i)}), \qquad i = 0, 1, \ldots.$$

La $\Phi(\boldsymbol{x})$ così definita è la funzione di iterazione del metodo di Newton, che è il più famoso dei metodi iterativi.

Definizioni. (a) Indicati con ξ un punto fisso di Φ, con I_ξ un suo intorno e con $\boldsymbol{x}^{(0)}$ un punto di I_ξ se, per la successione (7.6), vale la disuguaglianza

$$\|\boldsymbol{x}^{(i+1)} - \xi\| \leq c\|\boldsymbol{x}^{(i)} - \xi\|^p, \qquad i = 0, 1, \ldots, \tag{7.8}$$

il metodo iterativo è *almeno di ordine* $p > 1$. Se la (7.8) è valida per $p = 1$, con $0 < c < 1$, il metodo è *almeno del primo ordine*. Il metodo è esattamente di ordine p se non esiste un valore più elevato di p per cui la (7.8) sia valida; (b) Se esiste un numero $0 < c < 1$ tale che

$$\|\Phi(\boldsymbol{x}') - \Phi(\boldsymbol{x}'')\| \leq c\|\boldsymbol{x}' - \boldsymbol{x}''\|, \qquad \boldsymbol{x}', \boldsymbol{x}'' \in \overline{\Omega}, \tag{7.9}$$

la funzione Φ è definita *contrattiva*.

Teorema 7.1 (*delle contrazioni*) *Supponiamo che* $\Phi : \overline{\Omega} \to \overline{\Omega}$ *sia una funzione contrattiva soddisfacente la* (7.9). *Allora la successione* $\{\boldsymbol{x}^{(i)}\}_{i=0}^{\infty}$ *converge ad un solo punto fisso* $\xi \in \overline{\Omega}$, *qualunque sia il punto d'innesco* $\boldsymbol{x}^{(0)}$, *nell'ipotesi che anche* $\boldsymbol{x}^{(1)} \in \overline{\Omega}$.

Dimostrazione. Basta osservare che, utilizzando la (7.6) e la (7.9)

$$\begin{aligned}\|\boldsymbol{x}^{(i+p)} - \boldsymbol{x}^{(i)}\| &\leq \|\boldsymbol{x}^{(i+p)} - \boldsymbol{x}^{(i+p-1)}\| + \ldots + \|\boldsymbol{x}^{(i+1)} - \boldsymbol{x}^{(i)}\| \\ &\leq (c^{i+p-1} + c^{i+p-2} + \ldots + c^i)\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\| \\ &\leq \frac{c^i}{1-c}\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\|.\end{aligned}$$

Di conseguenza, $\{\boldsymbol{x}^{(i)}\}_{i=0}^{\infty}$ è una successione di Cauchy in $\overline{\Omega}$ e quindi ha limite $\xi \in \overline{\Omega}$. Facendo tendere p all'infinito, risulta

$$\|\boldsymbol{x}^{(i)} - \xi\| \leq \frac{c^i}{1-c}\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\|.$$

Inoltre, facendo tendere i all'infinito nella (7.9) risulta $\xi = \Phi(\xi)$, cioè ξ è un punto fisso della Φ.

L'unicità del punto fisso si dimostra facilmente. Infatti, se ne esistessero due, ξ e η, allora la stima $\|\xi - \eta\| = \|\Phi(\xi) - \Phi(\eta)\| \leq c\|\xi - \eta\|$ per $c \in (0,1)$ implicherebbe $\xi = \eta$. □

Corollario 7.2 *Supponiamo che esistano una costante $c \in (0,1)$ e un intero $r \in \mathbb{N}$ tali che*

$$\|\Phi^r(\boldsymbol{x}') - \Phi^r(\boldsymbol{x}'')\| \leq c\|\boldsymbol{x}' - \boldsymbol{x}''\|.$$

Allora la successione $\{\boldsymbol{x}^{(i)}\}_{i=0}^{\infty}$ converge ad un solo punto fisso $\xi \in \overline{\Omega}$, qualunque sia il punto d'innesco $\boldsymbol{x}^{(0)}$.

Dimostrazione. Secondo il Teorema delle contrazione esiste un unico punto fisso ξ della mappa Φ^r. Definiamo, per $j = 0, 1, \ldots, r-1$, $\xi_j = \Phi^j(\xi)$. Allora

$$\Phi^r(\xi_j) = \Phi^{r+j}(\xi) = \Phi^j(\Phi^r(\xi)) = \Phi^j(\xi) = \xi_j, \qquad j = 0, 1, \ldots, r-1.$$

Poichè ciascun ξ_j è punto fisso della Φ^r e tale mappa ha un solo punto fisso, risulta $\xi_0 = \xi_1 = \ldots = \xi_{r-1} = \xi$. Di conseguenza, $\Phi(\xi) = \xi_1 = \xi_0 = \xi$. Siccome, per $j = 0, 1, \ldots, r-1$ e dato il punto d'innesco $\boldsymbol{x}^{(0)}$, le r successioni $\{\boldsymbol{x}^{(j+ir)}\}_{i=0}^{\infty}$ tendono tutte a ξ, anche la successione $\{\boldsymbol{x}^{(i)}\}_{i=0}^{\infty}$ tende a ξ. □

Applicando la (7.8) r volte, si ottiene la stima

$$\|\boldsymbol{x}^{(i+r)} - \xi\| \leq c^{p^{r-1}+p^{r-2}+\ldots+p+1}\|\boldsymbol{x}^{(i)} - \xi\|^{p^r}. \tag{7.10}$$

Infatti, essendo valida la (7.10) per $r = 1$ e supponendo la sua validità per un certo r, si ha che

$$\begin{aligned}\|\Phi^r(\boldsymbol{x}^{(i)}) - \Phi^r(\xi)\| = \|\boldsymbol{x}^{(i+r+1)} - \xi\| &\leq c\|\boldsymbol{x}^{(i+r)} - \xi\|^p \\ &\leq c\left(c^{p^{r-1}+p^{r-2}+\ldots+p+1}\|\boldsymbol{x}^{(i)} - \xi\|^{p^r}\right)^p \\ &\leq c^{p^r+p^{r-1}+\ldots+p+1}\|\boldsymbol{x}^{(i)} - \xi\|^{p^{r+1}}.\end{aligned}$$

Quindi, per induzione, la stima (7.10) è valida per $r = 1, 2, \ldots$. Ovviamente, $\boldsymbol{x}^{(i)} \to \xi$ se $0 < c < 1$, qualunque sia il punto d'innesco $\boldsymbol{x}^{(0)}$. D'altra parte,

scegliendo, per $c \geq 1$, il punto d'innesco $\boldsymbol{x}^{(0)}$ tale che $\|\boldsymbol{x}^{(0)} - \xi\| < c^{-1/(p-1)}$, la stima

$$\|\boldsymbol{x}^{(r)} - \xi\| \leq C_r \|\boldsymbol{x}^{(0)} - \xi\|^{p^r}, \qquad C_r = c^{\frac{p^r-1}{p-1}},$$

implica che $\boldsymbol{x}^{(r)} \to \xi$. Di conseguenza, nell'ipotesi (7.8) esiste un intorno di ξ tale che, qualunque sia il punto d'innesco $\boldsymbol{x}^{(0)}$ appartenente all'intorno, gli iterati $\boldsymbol{x}^{(i)}$ tendono a ξ.

Nei metodi iterativi per i sistemi non lineari, nei quali la convergenza è tipicamente locale, è di grande importanza la seguente definizione di *punto di attrazione*.

Definizione. Un punto $\boldsymbol{x}^* \in \mathbb{R}^n$ è di attrazione, per la successione degli iterati $\boldsymbol{x}^{(k+1)} = \Phi(\boldsymbol{x}^{(k)})$, se esiste un intorno $I_{\boldsymbol{x}^*}$ di $\boldsymbol{x}^*$ tale che, qualunque sia il punto iniziale $\boldsymbol{x}^{(0)} \in I_{\boldsymbol{x}^*}$, la successione degli iterati $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^* = \Phi(\boldsymbol{x}^*)$, ossia alla soluzione del sistema $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{0}$, dato che

$$\boldsymbol{x}^* = \Phi(\boldsymbol{x}^*) \Longleftrightarrow \boldsymbol{F}(\boldsymbol{x}^*) = \boldsymbol{0}.$$

L'intorno $I_{\boldsymbol{x}^*}$ è detto dominio di attrazione. Nel caso particolare in cui $I_{\boldsymbol{x}^*} = \mathbb{R}^n$, $\boldsymbol{x}^*$ è definito punto di attrazione globale.

Relativamente al punto di attrazione, il risultato più importante è stabilito dal **Teorema di Ostrowski** [20].

Teorema 7.3 *Se la funzione di iterazione Φ è differenziabile in $\boldsymbol{x}^* = \Phi(\boldsymbol{x}^*)$ e inoltre $\rho(\Phi'(\boldsymbol{x}^*)) < 1$, $\boldsymbol{x}^*$ è un punto di attrazione per gli iterati $\boldsymbol{x}^{(k)}$.*

Dimostrazione. Essendo la Φ differenziabile in $\boldsymbol{x}^*$, per ogni $\varepsilon > 0$, esiste un intorno $I_{\boldsymbol{x}^*}$ di $\boldsymbol{x}^*$ per ogni punto $\boldsymbol{x}$ del quale

$$\|\Phi(\boldsymbol{x}) - \Phi(\boldsymbol{x}^*) - \Phi'(\boldsymbol{x}^*)(\boldsymbol{x} - \boldsymbol{x}^*)\| \leq \varepsilon \|\boldsymbol{x} - \boldsymbol{x}^*\|.$$

Inoltre, essendo $\rho(\Phi'(\boldsymbol{x}^*)) < 1$, qualunque sia $\varepsilon > 0$, esiste una norma tale che

$$\|\Phi'(\boldsymbol{x}^*)\| \leq \rho(\Phi'(\boldsymbol{x}^*)) + \varepsilon.$$

Di conseguenza, per ogni $x \in I_{x^*}$, risulta

$$\begin{aligned} \|\Phi(\boldsymbol{x}) - \Phi(\boldsymbol{x}^*)\| &\leq \|\Phi(\boldsymbol{x}) - \Phi(\boldsymbol{x}^*) - \Phi'(\boldsymbol{x}^*)(\boldsymbol{x} - \boldsymbol{x}^*)\| + \|\Phi'(\boldsymbol{x}^*)(\boldsymbol{x} - \boldsymbol{x}^*)\| \\ &\leq (\rho(\boldsymbol{x}^*) + 2\varepsilon)\|\boldsymbol{x} - \boldsymbol{x}^*\|. \end{aligned}$$

Da tale disuguaglianza, ricordando che $\rho(\Phi'(\boldsymbol{x}^*)) < 1$ e che $\varepsilon > 0$ è arbitrario, deriva che ε può essere scelto in modo da avere $\alpha = \rho(\Phi'(\boldsymbol{x}^*)) + 2\varepsilon < 1$, ossia che

$$\|\Phi(\boldsymbol{x}) - \Phi(\boldsymbol{x}^*)\| \leq \alpha \|\boldsymbol{x} - \boldsymbol{x}^*\|, \quad \alpha < 1.$$

La convergenza con tale norma segue immediatamente in quanto, essendo

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\| = \|\Phi(\boldsymbol{x}^{(k-1)}) - \Phi(\boldsymbol{x}^*)\| \leq \alpha\|\boldsymbol{x}^{(k-1)} - \boldsymbol{x}^*\| \leq \ldots \leq \alpha^k\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|,$$

con $0 < \alpha < 1$, è evidente che $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^*$, per $k \to \infty$.

L'equivalenza tra le norme garantisce che la convergenza vale con qualsiasi norma indotta. □

Metodo di Newton. Il Teorema di Ostrowski consente di dimostrare che ogni soluzione ξ del sistema $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{0}$ è un punto di attrazione per gli iterati del metodo di Newton, a condizione che $\boldsymbol{F}'(\xi)$ sia non singolare. Basta infatti osservare che essendo $\Phi(\boldsymbol{x}) = \boldsymbol{x} - (\boldsymbol{F}'(\boldsymbol{x}))^{-1}\boldsymbol{F}(\boldsymbol{x})$ e $\boldsymbol{F}(\xi) = \boldsymbol{0}$,

$$\Phi'(\xi) = I - (\boldsymbol{F}'(\xi))^{-1}\boldsymbol{F}'(\xi) = O,$$

da cui segue che $\rho(\Phi'(\xi)) = 0$.

Esempio 7.4

$$\begin{cases} x_1^2 + x_2^2 - 1 = 0, \\ 2x_1 + x_2 - 1 = 0, \end{cases} \tag{7.11}$$

di cui $\boldsymbol{x}^* = \begin{pmatrix} 4/5 \\ -3/5 \end{pmatrix}$ è una soluzione. La (7.11) può essere scritta nella forma $\boldsymbol{x} = \Phi(\boldsymbol{x})$, con

$$\Phi(\boldsymbol{x}) = \begin{pmatrix} \sqrt{1 - x_2^2} \\ 1 - 2x_1 \end{pmatrix} \quad \text{e} \quad \Phi'(\boldsymbol{x}) = \begin{pmatrix} 0 & -\dfrac{x_2}{\sqrt{1 - x_2^2}} \\ -2 & 0 \end{pmatrix}.$$

Una semplice visualizzazione grafica delle due equazioni permette di stabilire che $\boldsymbol{x}^*$ è un punto interno al rettangolo

$$R = [\tfrac{3}{5}, 1] \times [-\tfrac{4}{5}, -\tfrac{1}{2}].$$

Per verificare se la condizione di Ostrowski ($\rho(\Phi'(\boldsymbol{x}^*)) < 1$) per gli iterati

$$\boldsymbol{x}^{(i+1)} = \Phi(\boldsymbol{x}^{(i)}), \quad \text{con } \boldsymbol{x}^* \in R,$$

è soddisfatta, calcoliamo il

$$\sup_{\boldsymbol{x} \in R} \rho(\Phi'(\boldsymbol{x})).$$

Fissato $\boldsymbol{x}$ in R, i due autovalori di $\Phi'(\boldsymbol{x})$ sono le due soluzioni dell'equazione

$$\lambda(\boldsymbol{x})^2 - \frac{2x_2}{\sqrt{1 - x_2^2}} = 0 \Longrightarrow \rho(\Phi(\boldsymbol{x})) = \sqrt{\frac{2|x_2|}{\sqrt{1 - x_2^2}}}.$$

Da tale espressione segue che

$$\rho(\Phi'(\boldsymbol{x})) > 1 \text{ per } x_2 \in [-\tfrac{4}{5}, -\tfrac{1}{2}],$$

essendo $4x_2^2 > 1 - x_2^2$. Questo significa che $\boldsymbol{x}^*$ potrebbe non essere un punto di attrazione per tali iterati. Come abbiamo già osservato, $\boldsymbol{x}^*$ è invece punto di attrazione per gli iterati del metodo di Newton

$$\boldsymbol{x}^{(i+1)} = \Phi(\boldsymbol{x}^{(i)}), \qquad i = 0, 1, 2, \ldots, \quad \boldsymbol{x}^{(0)} \in R,$$

nell'ipotesi che la matrice Jacobiana $\boldsymbol{F}'(\boldsymbol{x}^*)$ sia non singolare. Condizione certamente soddisfatta, in quanto

$$\boldsymbol{F}'(\boldsymbol{x}) = \begin{pmatrix} 2x_1 & 2x_2 \\ 2 & 1 \end{pmatrix} \Longrightarrow \det(\boldsymbol{F}'(\boldsymbol{x})) > 0 \text{ in } R.$$

La funzione di iterazione è pertanto

$$\Phi(\boldsymbol{x}) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \frac{1}{2(x_1 - 2x_2)} \begin{pmatrix} 1 & -2x_2 \\ -2 & 2x_1 \end{pmatrix} \begin{pmatrix} x_1^2 + x_2^2 - 1 \\ 2x_1 + x_2 - 1 \end{pmatrix}.$$

Esempio 7.5 L'esempio

$$\begin{cases} 5x_1 - x_2 + x_1^3 = 5, \\ -x_1 + 5x_2 + x_2^3 = 5, \end{cases}$$

può essere espresso nella forma $\boldsymbol{x} = \Phi(\boldsymbol{x})$, dove

$$\Phi(\boldsymbol{x}) = \frac{1}{5} \begin{pmatrix} x_2 - x_1^3 + 5 \\ x_1 - x_2^3 + 5 \end{pmatrix}.$$

Scelto un vettore di innesco $\boldsymbol{x}^{(0)}$, utilizzando la $\Phi(\boldsymbol{x})$ come funzione di iterazione si possono calcolare gli iterati

$$\boldsymbol{x}^{(i+1)} = \Phi(\boldsymbol{x}^{(i)}), \qquad i = 0, 1, 2, \ldots.$$

Osservato che $\boldsymbol{x}^* = \left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right)$ è una soluzione del sistema, è facile dimostrare che $\boldsymbol{x}^*$ è un punto di attrazione per gli iterati. La matrice Jacobiana della Φ è

$$\Phi'(\boldsymbol{x}) = \frac{1}{5} \begin{pmatrix} -3x_1^2 & 1 \\ 1 & -3x_2^2 \end{pmatrix}, \quad \text{da cui } \Phi'(\boldsymbol{x}^*) = \frac{1}{5} \begin{pmatrix} -3 & 1 \\ 1 & -3 \end{pmatrix},$$

i cui autovalori sono $\lambda_1 = -\frac{4}{5}$ e $\lambda_2 = -\frac{2}{5}$. Conseguentemente $\boldsymbol{x}^*$ è di attrazione per gli iterati, dato che $\rho(\Phi'(\boldsymbol{x}^*)) = \frac{4}{5}$.

7.2 Caso unidimensionale

In questo paragrafo verranno presentati risultati basilari sui metodi per la valutazione delle radici di un'equazione non lineare

$$f(x) = 0, \tag{7.12}$$

dove la f è derivabile e ha derivata prima continua. Approssimando il grafico della funzione $f(x)$ in un intorno del punto $(x_n, f(x_n))$ mediante la sua retta tangente $y - f(x_n) = f'(x_n)(x - x_n)$ e cercando il valore di $x = x_{n+1}$ per cui $(x_{n+1}, 0)$ è un punto di tale retta tangente, si ottiene

$$x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})}{f'(x^{(n)})}, \tag{7.13}$$

nell'ipotesi che la retta tangente in $(x_n, f(x_n))$ non sia orizzontale. L'iterazione del tipo (7.13) si chiama *metodo di Newton-Raphson* oppure, semplicemente, metodo di Newton. La sua funzione iterativa è $\varphi(x) = x - (f(x)/f'(x))$, che rappresenta il caso unidimensionale della (7.6).

Teorema 7.6 *Nel caso scalare, condizione sufficiente perché una funzione φ, derivabile con continuità in un intervallo $[a, b]$, sia contrattiva è che esista un numero $m \in (0, 1)$ tale che*

$$|\varphi'(x)| \leq m < 1, \qquad x \in [a, b].$$

Dimostrazione. Per il teorema del valor medio, applicato ad un intervallo $[x', x''] \subset [a, b]$,

$$\varphi(x') - \varphi(x'') = \varphi'(\theta)(x' - x''), \qquad \theta \in (x', x'').$$

Pertanto, essendo $|\varphi'(x)| \leq m < 1$ per ipotesi,

$$|\varphi(x') - \varphi(x'')| \leq m|x' - x''|$$

con $0 \leq m < 1$ per ogni $x', x'' \in [a, b]$, e quindi la φ è contrattive in $[a, b]$. □

Esempio 7.7 Per $a \in [-\frac{1}{2}, \frac{1}{2}]$, la funzione $\varphi = a \log(x + 1) - \frac{1}{3}x$ è contrattiva in $[0, 1]$. Infatti, essendo

$$\varphi'(x) = \frac{a}{x + 1} - \frac{1}{3},$$

basta osservare che per $a \in [-\frac{1}{2}, \frac{1}{2}]$

$$\left| \frac{a}{x + 1} - \frac{1}{3} \right| \leq \frac{5}{6} < 1, \qquad x \in [0, 1].$$

Esempio 7.8 Il metodo iterativo

$$x^{(i+1)} = \cos(x^{(i)})$$

è globalmente convergente. Infatti, indicato con ξ il solo punto fisso dell'equazione $x = \cos(x)$ ed osservato che $\xi \in (0, 1)$, per ogni $x^{(0)} \in \mathbb{R}$ possiamo scrivere

$$\begin{aligned} &x^{(1)} = \cos(x^{(0)}) \in [-1, 1] \\ &x^{(i+1)} = \cos(x^{(i)}) = \cos(\xi) - (x^{(i)} - \xi)\sin(\theta^{(i)}) \\ &\theta^{(i)} \in (x^{(i)}, \xi) \in [-1, 1], \quad i = 1, 2, \ldots. \end{aligned}$$

Pertanto, essendo per ogni i

$$|x^{(i+1)} - \xi| = |x^{(i+1)} - \cos(\xi)| = |\sin(\theta^{(i)})|\,|x^{(i)} - \xi| < (\sin(1))|x^{(i)} - \xi|,$$

il metodo è globalmente convergente e la sua convergenza è lineare (convergenza di ordine 1).

Se $f'(x) \neq 0$ per $x \in [a, b]$, il Teorema 7.6 può essere applicato al metodo di Newton, per il quale la funzione di iterazione è

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

Essendo $f'(x) \neq 0$ per $x \in [a, b]$,

$$\varphi'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2}.$$

Supponendo che ξ sia uno zero semplice della f (cioè, che $f(\xi) = 0$ e $f'(\xi) \neq 0$), esiste un intervallo contenente ξ al suo interno in cui $|\varphi'(x)| < 1$. Pertanto, qualunque sia il punto d'innesco appartenente a tale intervallo, gli iterati convergono a ξ.

Teorema 7.9 *Supponiamo che in un intervallo $[a, b]$ valgono le seguenti ipotesi:*

1) $f(a)f(b) < 0$;

2) la f è derivabile due volte con continuità in $[a, b]$;

3) $f'(x)$ e $f''(x)$ sono di segno costante in $[a, b]$.

Se inoltre, come punto iniziale $x^{(0)}$, si prende a qualora sia $f(a)f''(a) > 0$, oppure b qualora sia $f(b)f''(b) > 0$, la successione $\{x^{(n)}\}_{n=0}^{\infty}$, ottenuta con l'algoritmo (7.13)*, converge monotonamente alla sola soluzione ξ dell'equazione* (7.12).

Dimostrazione. Siano, per esempio, $f(a) < 0$, $f(b) > 0$, $f'(x) > 0$ e $f''(x) > 0$ per $a \leq x \leq b$ (la dimostrazione nelle altre possibili situazioni è del tutto analoga). Poiché $f(b)f''(b) > 0$, poniamo $x^{(0)} = b$. La 1) assicura l'esistenza di almeno uno zero ξ per la f in $[a, b]$. La non variazione di segno di $f'(x)$ in $[a, b]$ ne assicura l'unicità. Mostriamo ora, per induzione, che ogni iterato $x^{(n)} > b$. Supponendo che $x^{(m)} > \xi$ per $m = 0, 1, \ldots, i$, dobbiamo dunque dimostrare che $x^{(m+1)} > \xi$. Sviluppando $f(\xi)$ con la formula di Taylor con punto iniziale $x^{(0)}$ e osservato che

$$\xi = x^{(n)} + (\xi - x^{(n)}),$$

si ha

$$f(\xi) = 0 = f(x^{(n)}) + f'(x^{(n)})(\xi - x^{(n)}) + \tfrac{1}{2}f''(\theta^{(n)})(\xi - x^{(n)})^2$$

con $\xi < \theta^{(n)} < x^{(n)}$. Poiché $f''(x) > 0$, risulta

$$f(x^{(n)}) + f'(x^{(n)})(\xi - x^{(n)}) < 0$$

e, di conseguenza, essendo

$$\xi < x^{(n)} - \frac{f(x^{(n)})}{f'(x^{(n)})} = x^{(n+1)},$$

l'intera successione $\{x^{(n)}\}_{n=0}^{\infty}$ consiste di punti appartenenti all'intervallo semiaperto $(\xi, b]$.

Poiché $f'(x^{(n)}) > 0$ per ipotesi e $f(x^{(n)}) > 0$, in quanto $x^{(n)} > \xi$ ed $f(\xi) = 0$, la successione $\{x^{(n)}\}_{n=0}^{\infty}$ è monotona decrescente e inferiormente limitata da ξ. Esiste dunque un limite finito $\overline{\xi} = \lim_{n\to\infty} x^{(n)}$. Passando al limite nell'uguaglianza (7.13) si ha

$$\overline{\xi} = \overline{\xi} - \frac{f(\overline{\xi})}{f'(\overline{\xi})}$$

da cui $f(\overline{\xi}) = 0$ ed infine, per l'unicità degli zeri della f in $[a, b]$, $f(\overline{\xi}) = \xi$. □

Talvolta, come è dimostrato nel seguente teorema, è possibile valutare la distanza dell'n-esimo iterato dalla radice dell'equazione.

Teorema 7.10 *Supponiamo che valgano le ipotesi del Teorema* 7.9. *Se esistono due numeri positivi* m_1 *ed* M_2 *tali che:*

1) $m_1 \leq |f'(x)|$ *per ogni* $x \in [a,b]$*;*

2) $|f''(x)| \leq M_2$ *per ogni* $x \in [a,b]$*;*

tra l'n-esimo iterato di Newton e la radice ξ *dell'equazione* $f(x) = 0$ *vale la seguente importante stima dell'errore*

$$|x^{(n)} - \xi| < \frac{M_2}{2m_1}(x^{(n)} - x^{(n-1)})^2. \tag{7.14}$$

Dimostrazione. Per il teorema del valor medio, applicato ad un generico intervallo $[\overline{x}, \xi] \subset [a,b]$ con $f(\xi) = 0$ e $f(\overline{x}) \neq 0$,

$$f(\overline{x}) - f(\xi) = (\overline{x} - \xi)f'(\theta), \qquad \theta \in (\overline{x}, \xi).$$

Pertanto, esistendo un numero positivo m_1 tale che $|f'(x)| \geq m_1$ per $x \in [a,b]$, risulta

$$|f(\overline{x})| = |f(\overline{x}) - f(\xi)| \geq m_1|\overline{x} - \xi|$$

e quindi anche

$$|\overline{x} - \xi| \leq \frac{|f(\overline{x})|}{m_1}. \tag{7.15}$$

Inoltre, dall'applicazione della formula di Taylor di ordine due,

$$f(x^{(n)}) = f[x^{(n-1)} + (x^{(n)} - x^{(n-1)})] = f(x^{(n-1)}) + f'(x^{(n-1)})(x^{(n)} - x^{(n-1)}) + \tfrac{1}{2}f''(\theta^{(n)})(x^{(n)} - x^{(n-1)})^2,$$

per un'opportuna $\theta^{(n)} \in (x^{(n-1)}, x^{(n)})$. Poiché, per costruzione, nel metodo di Newton

$$f[x^{(n-1)}) + f'(x^{(n-1)})(x^{(n)} - x^{(n-1)}) = 0,$$

dalla formula di Taylor di secondo ordine deriva che

$$|f(x^{(n)}| \leq \tfrac{1}{2}M_2(x^{(n)} - x^{(n-1)})^2, \tag{7.16}$$

dato che

$$|f''(x)| \leq M_2, \qquad x \in [a,b].$$

Infine, utilizzando la (7.15) e la (7.16), si ha la (7.14). □

Esempio 7.11 Valutare $\sqrt{2}$ con un errore dell'ordine di 10^{-5}, con il metodo di Newton.

Consideriamo, allo scopo, $f(x) = x^2 - 2$ nell'intervallo $[1,2]$. Il metodo di Newton è applicabile in quanto:

1) $f(a)f(b) = (-1)(2) < 0$;

2) $f'(x) = 2x$ è di segno costante in $[1, 2]$;

3) $f''(x) = 2 > 0$ in $[1, 2]$.

Poiché $f(2)f''(2) = 4 > 0$, come punto iniziale si prende 2. Pertanto

$$x^{(1)} = x^{(0)} - \frac{f(x^{(0)})}{f'(x^{(0)})} = 2 - \frac{2}{4} = \frac{3}{2},$$

$$x^{(2)} = \frac{17}{12}, x^{(3)} \simeq 1.41422.$$

Essendo, per la (7.14),

$$|x^{(3)} - \sqrt{2}| \leq \frac{M_2}{2m_1}(x^{(3)} - x^{(2)}) \leq \frac{2}{4}(1.41422 - 1.41667)^2 \simeq 3 \times 10^{-6},$$

$x^{(3)}$ rappresenta un'approssimazione valida di $\sqrt{2}$.

Teorema 7.12 *Se ξ è uno zero semplice per la f ($f(\xi) = 0$, $f'(\xi) \neq 0$), il metodo di Newton ha convergenza del secondo ordine. Se ξ è uno zero multiplo ($f(\xi) = \ldots = f^{(m-1)}(\xi) = 0$, $f^{(m)} \neq 0$, $m > 1$), il metodo di Newton ha convergenza lineare.*

Dimostrazione. Se ξ è uno zero semplice, essendo

$$\varphi'(\xi) = \frac{f(\xi)f''(\xi)}{(f'(\xi))^2} = 0,$$

per la formula di Taylor di ordine due,

$$x^{(n+1)} = \varphi(x^{(n)}) = \varphi(\xi) + \frac{(x^{(n)} - \xi)^2}{2}\varphi''(\theta^{(n)}), \qquad \theta^{(n)} \in (x^{(n)}, \xi),$$

da cui, posto $|\varphi''(x)| \leq M_2$ in $[a, b]$,

$$|x^{(n+1)} - \xi| \leq \tfrac{1}{2}M_2 \, |x^{(n)} - \xi|^2$$

e quindi il metodo ha convergenza quadratica.

Se invece ξ è uno zero m-plo, esiste una $g(x)$ con $g(\xi) \neq 0$ tale che

$$f(x) = (x - \xi)^m g(x).$$

Di conseguenza, se $g(x)$ è derivabile,

$$\varphi'(\xi) = 1 - \frac{1}{m} \neq 0.$$

Pertanto, per la formula del valor medio,

$$x^{(n+1)} = \varphi(x^{(n)}) = \varphi(\xi) + (x^{(n)} - \xi)\varphi'(\theta^{(n)}), \qquad \theta^{(n)} \in (x^{(n)}, \xi),$$

da cui, posto $|\varphi'(x)| \leq M_1$ in $[a, b]$,

$$|x^{(n+1)} - \xi| \leq M_1 |x^{(n)} - \xi|.$$

L'ultima disuguaglianza dimostra che, essendo almeno localmente $M_1 < 1$, il metodo è convergente con convergenza lineare. □

Il metodo della secante o della falsa posizione, viene utilizzato per risolvere in un'intervallo $[a, b]$ un'equazione $f(x) = 0$, nell'ipotesi che in $[a, b]$ essa abbia una sola radice. Algoritmicamente, esso si presenta in uno dei modi seguenti:

$$x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})}{f(x^{(n)}) - f(a)}(x^{(n)} - a), \tag{7.17}$$

con $x^{(0)} = b$ e $n = 0, 1, \ldots,$;

$$x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})}{f(b) - f(x^{(n)})}(b - x^{(n)}), \tag{7.18}$$

con $x^{(0)} = a$ e $n = 0, 1, \ldots,$. Geometricamente corrisponde alla sostituzione, nell'intervallo $[a, b]$, della curva $y = f(x)$ con la corda che unisce il punto $(x^{(n)}, f(x^{(n)})$ con $(a, f(a))$ o $(b, f(b))$ e assume $x^{(n+1)}$ come l'intercetta della corda con l'asse x.

Teorema 7.13 *Se in un intervallo $[a, b]$ sono soddisfatte le seguenti ipotesi:*

1) $f(a)f(b) < 0$;

2) $f(x)$ è derivabile due volte con continuità;

3) $f'(x)$ ed $f''(x)$ sono di segno costante;

il metodo della secante è convergente all'unica radice ξ dell'equazione $f(x) = 0$ in (a, b), purché si usi la formula (7.17) se $f(a)f''(a) > 0$, la formula (7.18) se $f(b)f''(b) > 0$.

Dimostrazione. Sia, per fissare le idee, $f''(x) > 0$ in $a \leq x \leq b$ (se questa ipotesi non fosse soddisfatta le considerazioni seguenti cambierebbero solo formalmente). La curva $y = f(x)$ è allora convessa e disposta al di sotto della corda che unisce i punti $(a, f(a))$, $(b, f(b))$. Due casi sono allora possibili:

1) $f(a) > 0$;

2) $f(b) > 0$.

Nel primo caso, l'estremità a è fissa e le approssimazioni successive:

$$x^{(0)} = b, \quad x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})}{f(x^{(n)}) - f(a)}(x^{(n)} - a), \quad n = 0, 1, \ldots,$$

formano una successione decrescente e limitata inferiormente da ξ.

Nel secondo caso, l'estremità b è fissa e le approssimazioni successive:

$$x^{(0)} = a, \quad x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})}{f(b) - f(x^{(n)})}(b - x^{(n)}), \quad n = 0, 1, \ldots,$$

formano una successione crescente e limitata da ξ. Poiché, in entrambi i casi, la successione $\{x^{(n)}\}_{n=0}^{\infty}$ è monotona e contenuta in (a, b), esiste il

$$\lim_{n\to\infty} x^{(n)} = \overline{\xi} \quad \text{con} \quad a < \overline{\xi} < b.$$

Pertanto, nel primo caso, passando al limite nella (7.17) si ha

$$\overline{\xi} = \overline{\xi} - \frac{f(\overline{\xi})}{f(\overline{\xi}) - f(a)}(\overline{\xi} - a)$$

da cui $f(\overline{\xi}) = 0$ e quindi $\overline{\xi} = \xi$, dato che, nelle ipotesi $1) - 2)$ del teorema, l'equazione $f(x) = 0$ possiede una sola radice in (a, b). Procedendo allo stesso modo si trova che, anche nel secondo caso, la successione degli iterati $\{x^{(n)}\}_{n=0}^{\infty}$ converge all'unica radice ξ dell'equazione $f(x) = 0$ in (a, b). □

Il seguente teorema consente, sotto opportune ipotesi, di valutare la distanza tra l'n-esimo iterato e la radice in (a, b) dell'equazione $f(x) = 0$.

Teorema 7.14 *Se esistono due numeri m_1 e M_1 tali che*

$$0 < m_1 \leq |f'(x)| \leq M_1 < \infty$$

per

$$\begin{cases} a \leq x \leq x^{(n)}, & \textit{nel primo caso,} \\ x^{(n)} \leq x \leq b, & \textit{nel secondo caso,} \end{cases}$$

tra l'n-esimo iterato $x^{(n)}$ del metodo della secante e l'unica radice ξ dell'equazione $f(x) = 0$ in (a, b) vale la seguente relazione:

$$|x^{(n)} - \xi| \leq \frac{M_1 - m_1}{m_1}|x^{(n)} - x^{(n-1)}|. \tag{7.19}$$

Esempio 7.15 Valutare, a meno di $\frac{1}{100}$, la radice positiva dell'equazione $f(x) = x^3 - 0.2x^2 - 0.2x - 1.2 = 0$.

Poiché $f(1) = -0.6$ ed $f(2) = 5.6$, la radice cercata appartiene all'intervallo $(1, 2)$. Dato che l'intervallo di indeterminazione è grande, possiamo ridurlo per bisezione calcolando $f(3/2)$. Poiché $f(1/2) = 1.425$, la radice cercata appartiene all'intervallo $(1, 1.5)$. Applicando ora la relazione di ricorrenza (7.18) otteniamo:

$$
\begin{aligned}
x^{(1)} &= 1 + \frac{0.6}{1.425 + 0.6}(1.5 - 1) = 1.15, \\
x^{(2)} &= 1.15 + \frac{0.173}{1.425 + 0.173}(1.5 - 1.15) = 1.190, \\
x^{(3)} &= 1.190 + \frac{0.036}{1.425 + 0.036}(1.5 - 1.190) = 1.198.
\end{aligned}
$$

Poiché, nell'intervallo $[1.198, 1.5]$,

$$6.073(1.5)^2 - 0.4(1.198) - 0.2\, f'(x)3(2.298)^2 - 0.4(1.5) - 0.2 = 3.49$$

per la (7.19)

$$0 < \xi - x^{(3)} < \frac{2.58}{6.07}\, 0.008 < \frac{4}{1000}$$

e pertanto $x^{(3)}$ è accettabile come valutazione della radice. Il valore esatto della radice è 1.2.

Osservazione. Mentre il metodo di Newton, almeno quando la radice ξ dell'equazione $f(x) = 0$ è semplice, ha convergenza quadratica, il metodo della secante ha sempre convergenza lineare. Tuttavia l'importanza del metodo della secante non è trascurabile, soprattutto perché, mentre nel metodo di Newton il calcolo di un nuovo iterato $x^{(i)}$ richiede la valutazione sia della funzione che della sua derivata in $x^{(i-1)}$, il metodo della secante, tranne nel primo passo (nel quale ne sono richieste due) richiede soltanto la valutazione della funzione in $x^{(i-1)}$. Nel calcolo di radici o di logaritmi, ottenuti mediante la risoluzione di equazioni del tipo

$$
\begin{aligned}
&x^k - \alpha = 0, \qquad k \in \mathbb{N}, \qquad \alpha = 0, \\
&\log(x) - \alpha = 0, \qquad \alpha \in \mathbb{R},
\end{aligned}
$$

risulta tuttavia più efficiente il metodo di Newton, anche se si tiene conto del maggior numero di valutazioni funzionali richiesto.

7.3 Caso multidimensionale

Teorema 7.16 *Nel caso n-dimensionale,* **condizione sufficiente** *perché una funzione* Φ *sia contrattiva in un intervallo* $I \subset \mathbb{R}^n$ *è che esista un numero* $\lambda \in [0,1)$ *tale che*

$$\sum_{j=1}^{n} \left| \frac{\partial \varphi_i(\boldsymbol{x})}{\partial x_j} \right| \leq \lambda < 1, \qquad i = 1, 2, \ldots, n, \quad \boldsymbol{x} \in I. \tag{7.20}$$

Dimostrazione. Per la formula di Taylor di primo ordine, per $i = 1, 2, \ldots, n$ e per ogni coppia di punti $\boldsymbol{x}', \boldsymbol{x}'' \in I$,

$$\varphi_i(\boldsymbol{x}') - \varphi_i(\boldsymbol{x}'') = \sum_{j=1}^{n} \frac{\partial \varphi_i(\theta^{(i,j)})}{\partial x_j}(x_j' - x_j''),$$

essendo $\theta^{(i,j)}$ un punto interno all'intervallo di estremi $\boldsymbol{x}', \boldsymbol{x}''$. Di conseguenza, dato che $\theta^{(i,j)} \in I$, per la (7.20)

$$\begin{aligned} |\varphi_i(\boldsymbol{x}') - \varphi_i(\boldsymbol{x}'')| &\leq \sum_{j=1}^{n} \left| \frac{\partial \varphi_i(\theta^{(i,j)})}{\partial x_j} \right| |x_j' - x_j''| \\ &\leq \|\boldsymbol{x}' - \boldsymbol{x}''\|_\infty \sum_{j=1}^{n} \left| \frac{\partial \varphi_i(\theta^{(i,j)})}{\partial x_j} \right| \leq \lambda \|\boldsymbol{x}' - \boldsymbol{x}''\|_\infty. \end{aligned}$$

Poiché la suddetta diseguaglianza vale per $i = 1, \ldots, n$, risulta anche

$$\|\varphi(\boldsymbol{x}') - \varphi(\boldsymbol{x}'')\|_\infty \leq \lambda \|\boldsymbol{x}' - \boldsymbol{x}''\|_\infty$$

e pertanto la φ è contrattiva. Ricordando la definizione di $\|\cdot\|_\infty$ di una matrice, la (7.20), può esprimersi nella forma

$$\|\Phi'(\boldsymbol{x})\|_\infty \leq \lambda < 1.$$

□

Esempio 7.17 Dimostrare che la funzione

$$\Phi(\boldsymbol{x}) = \begin{cases} \varphi_1(x_1, x_2) = \sqrt{[x_1(x_2 + 5) - 1]/2}, \\ \varphi_2(x_1, x_2) = \sqrt{3 + \log_{10} x_1}, \end{cases}$$

è contrattiva nell'insieme

$$I = \{(x_1, x_2) \text{ con } |3.5 - x_1| \leq 0.1 \text{ e } |2.2 - x_2| \leq 0.1\}.$$

In tale intervallo

$$\left|\frac{\partial\varphi_1}{\partial x_1}\right| \leq \frac{2.3+5}{4\sqrt{[3.4(2.1+5)-1]/2}} < 0.54,$$
$$\left|\frac{\partial\varphi_1}{\partial x_2}\right| \leq \frac{3.6}{4\sqrt{[3.4(2.1+5)-1]/2}} < 0.27,$$
$$\left|\frac{\partial\varphi_2}{\partial x_1}\right| \leq \frac{1+\frac{3(0.43)}{3.4}}{2\sqrt{3.4+2\log_{10}3.4}} < 0.42,$$
$$\left|\frac{\partial\varphi_2}{\partial x_2}\right| = 0.$$

Di conseguenza, essendo

$$\left|\frac{\partial\varphi_1}{\partial x_1}\right| + \left|\frac{\partial\varphi_1}{\partial x_2}\right| < 0.81 < 1,$$
$$\left|\frac{\partial\varphi_2}{\partial x_1}\right| + \left|\frac{\partial\varphi_2}{\partial x_2}\right| < 0.42 < 1,$$

la funzione $\Phi(\boldsymbol{x})$ è contrattiva in I.

Teorema 7.18 *Supponiamo che in un intervallo chiuso $I = \{\boldsymbol{x} \in \mathbb{R}^n$ con $\boldsymbol{a}_i \leq \boldsymbol{x}_i \leq b_i$ per $i = 1, \ldots, n\}$ esista una sola soluzione ξ dell'equazione $\boldsymbol{F}(\boldsymbol{x}) = 0$ e quindi un solo punto fisso per una sua funzione di iterazione $\Phi(\boldsymbol{x})$. Allora, se:*

1) esistono continue in I le derivate parziali della Φ;

2) il punto iniziale $\boldsymbol{x}^{(0)}$ e tutti gli iterati successivi appartengono ad I;

3) per $i = 1, \ldots, n$ valgono in I le disuguaglianze

$$\left|\frac{\partial\varphi_i}{\partial x_1}\right| + \ldots + \left|\frac{\partial\varphi_i}{\partial x_n}\right| \leq m_i < 1,$$

essendo $\Phi = (\varphi_1, \ldots, \varphi_n)$,

la successione degli iterati converge a ξ.

La condizione 3) equivale a dire che esiste un $0 \leq m < 1$ tale che

$$\|\Phi'(\boldsymbol{x})\|_\infty \leq m < 1,$$

per ogni $\boldsymbol{x} \in I$.

Esempio 7.19 Valutare, a meno di $\frac{1}{1000}$, la soluzione positiva del sistema

$$\begin{cases} f_1(x,y) = 2x^2 - xy - 5x + 1 = 0, \\ f_2(x,y) = x + 3\log_{10} x - y^2 = 0. \end{cases}$$

Allo scopo di trovare un intervallo nel quale cade la soluzione positiva del sistema, costruiamo le curve $f_1(x,y) = 0$, $f_2(x,y) = 0$. L'approssimazione della soluzione positiva, ottenuta graficamente, è

$$x^{(0)} = 3.5, \qquad y^{(0)} = 2.2.$$

Per ottenere una funzione di iterazione Φ scriviamo il sistema nella forma seguente:

$$\begin{aligned} x = \varphi_1(x,y) &= \sqrt{\frac{x(y+5)-1}{2}}, \\ y = \varphi_2(x,y) &= \sqrt{3 + \log_{10} x}. \end{aligned}$$

Supponiamo inoltre che l'intervallo I sia così definito:

$$I = \{(x,y) \text{ con } |3.5 - x| \le 0.1 \text{ e } |2.2 - y| \le 0.1\}.$$

In tale intervallo

$$\begin{aligned} \left|\frac{\partial \varphi_1}{\partial x_1}\right| &\le \frac{2.3+5}{4\sqrt{[3.4(2.1+5)-1]/2}} < 0.54, \\ \left|\frac{\partial \varphi_1}{\partial x_2}\right| &\le \frac{3.6}{4\sqrt{[3.4(2.1+5)-1]/2}} < 0.27, \\ \left|\frac{\partial \varphi_2}{\partial x_1}\right| &\le \frac{1 + \frac{3(0.43)}{3.4}}{2\sqrt{3.4 + 2\log_{10} 3.4}} < 0.42, \\ \left|\frac{\partial \varphi_2}{\partial x_2}\right| &= 0. \end{aligned}$$

Di conseguenza, essendo

$$\begin{aligned} \left|\frac{\partial \varphi_1}{\partial x_1}\right| + \left|\frac{\partial \varphi_1}{\partial x_2}\right| &< 0.81 < 1, \\ \left|\frac{\partial \varphi_2}{\partial x_1}\right| + \left|\frac{\partial \varphi_2}{\partial x_2}\right| &< 0.42 < 1, \end{aligned}$$

se come punto iniziale prendiamo la soluzione grafica e se gli iterati successivi rimangono in I (ciò che deve essere verificato durante la esecuzione dei calcoli), il processo iterativo è convergente.

Calcolando gli iterati mediante le formule

$$x^{(i)} = \sqrt{\frac{x^{(i-1)}(y^{(i-1)}+5)-1}{2}},$$
$$y^{(i)} = \sqrt{3+\log_{10} x^{(i-1)}}, \qquad i = 1, 2, \ldots,$$

ed approssimando alla terza decimale, si costruisce la seguente tabella

$$\begin{aligned}
x^{(0)} &= 3.5, & y^{(0)} &= 2.2,\\
x^{(1)} &= 3.479, & y^{(1)} &= 2.259,\\
x^{(2)} &= 3.481, & y^{(2)} &= 2.260,\\
x^{(3)} &= 3.484, & y^{(3)} &= 2.261,\\
x^{(4)} &= 3.486, & y^{(4)} &= 2.261,\\
x^{(5)} &= 3.487, & y^{(5)} &= 2.262,\\
x^{(6)} &= 3.487, & y^{(6)} &= 2.262.
\end{aligned}$$

Avendo ottenuto $x^{(6)} = x^{(5)} = 3.487$ e $y^{(6)} = y^{(5)} = 2.262$, si può porre $\xi^{(1)} = 3.487$ e $\xi^{(2)} = 2.262$.

Osservazione. L'ipotesi 3) del Teorema 7.18, come dimostrato nel seguente Teorema 7.20, può essere sostituita dall'ipotesi meno restrittiva della contrattività della Φ.

Teorema 7.20 *Siano:*

- Φ *una funzione di iterazione definita in un insieme* Ω*;*
- $\boldsymbol{x}^{(0)}$ *un punto iniziale appartenente da* Ω*;*
- $I_{\boldsymbol{x}^{(0)},r}$ *un intorno circolare con centro* $\boldsymbol{x}^{(0)}$ *e raggio* r*.*

Allora, se esiste un numero $\kappa \in [0, 1)$ *tale che:*

1) $\|\Phi(\boldsymbol{x}') - \Phi(\boldsymbol{x}'')\| \leq \kappa \|\boldsymbol{x}' - \boldsymbol{x}''\|, \qquad \boldsymbol{x}', \boldsymbol{x}'' \in I_{\boldsymbol{x}^{(0)},r}$*;*

2) $\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\| = \|\Phi(\boldsymbol{x}^{(0)}) - \boldsymbol{x}^{(0)}\| \leq (1-\kappa)r,$

risultano valide le seguenti affermazioni:

1′) $\boldsymbol{x}^{(i)} = \Phi(\boldsymbol{x}^{(i-1)}) \in I_{\boldsymbol{x}^{(0)},r}$ *per* $i = 1, 2, \ldots$*;*

2′) Φ *possiede in* $I_{\boldsymbol{x}^{(0)},r}$ *un solo punto fisso* ξ*;*

3′) gli iterati convergono a ξ secondo la relazione seguente

$$\|\boldsymbol{x}^{(i)} - \xi\| \leq \frac{\kappa^i}{1-\kappa}\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\|. \tag{7.21}$$

Dimostrazione. La 1′) può essere verificata per induzione. Poiché $\boldsymbol{x}^{(1)} \in I_{\boldsymbol{x}^{(0)},r}$ per la 2), resta da dimostrare che se $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(i)} \in I_{\boldsymbol{x}^{(0)},r}$ anche $\boldsymbol{x}^{(i+1)} \in I_{\boldsymbol{x}^{(0)},r}$.

Effettivamente in tale ipotesi $\boldsymbol{x}^{(i+1)} \in I_{\boldsymbol{x}^{(0)},r}$ in quanto, essendo

$$\begin{aligned}\|\boldsymbol{x}^{(i+1)} - \boldsymbol{x}^{(i)}\| = \|\Phi(\boldsymbol{x}^{(i)}) - \Phi(\boldsymbol{x}^{(i-1)})\| &\leq \kappa\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(i-1)}\| \\ &\leq \kappa^i\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\|,\end{aligned} \tag{7.22}$$

per la disuguaglianza triangolare e per la 2),

$$\begin{aligned}\|\boldsymbol{x}^{(i+1)} - \boldsymbol{x}^{(0)}\| &\leq \|\boldsymbol{x}^{(i+1)} - \boldsymbol{x}^{(i)}\| + \|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(i-1)}\| + \ldots + \|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\| \\ &\leq (\kappa^i + \kappa^{i-1} + \ldots + 1)\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\| \leq (1 - \kappa^{i+1}) < r.\end{aligned}$$

Per verificare la 2′) dimostriamo dapprima che la successione $\{\boldsymbol{x}^{(i)}\}_{i=0}^{\infty}$ è di Cauchy. Dalla disuguaglianza (7.22) e dall'ipotesi 2) segue che, per $h = i + j$,

$$\begin{aligned}\|\boldsymbol{x}^{(h)} - \boldsymbol{x}^{(i)}\| &\leq \|\boldsymbol{x}^{(h)} - \boldsymbol{x}^{(h-1)}\| + \ldots + \|\boldsymbol{x}^{(i+1)} - \boldsymbol{x}^{(i)}\| \\ &\leq (\kappa^{(i+j-1)} + \kappa^{(i+j-2)} + \ldots + \kappa^i)\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\| \\ &\leq \kappa^i\frac{1-\kappa^j}{1-\kappa}\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\| < \frac{\kappa^i}{1-\kappa}\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\| < \kappa^i\, r.\end{aligned} \tag{7.23}$$

Pertanto, poiché $\{\kappa^i\, r\}_{i=0}^{\infty}$ è strettamente decrescente a zero, la successione $\{\boldsymbol{x}^{(i)}\}_{i=0}^{\infty}$ è di Cauchy. Inoltre, poiché in $\mathbb{R}^n$ le successioni di Cauchy sono convergenti, esiste un vettore $\xi \in \mathbb{R}^n$ al quale converge $\{\boldsymbol{x}^{(i)}\}_{i=0}^{\infty}$ ed inoltre, essendo ξ un punto di accumulazione per $I_{\boldsymbol{x}^{(0)},r}$, esso appartiene alla chiusura di $I_{\boldsymbol{x}^{(0)},r}$ ($\xi \in \overline{I}_{\boldsymbol{x}^{(0)},r}$). Infine ξ è un punto fisso per la funzione di interazione, essendo, per ogni $\varepsilon \in \mathbb{R}^+$ ed i sufficientemente elevato,

$$\begin{aligned}\|\Phi(\xi) - \xi\| &\leq \|\Phi(\xi) - \Phi(\boldsymbol{x}^{(i)})\| + \|\Phi(\boldsymbol{x}^{(i)}) - \xi\| \\ &\leq \kappa\|\xi - \boldsymbol{x}^{(i)}\| + \|\boldsymbol{x}^{(i+1)} - \xi\| < \varepsilon.\end{aligned}$$

per il fatto che $\boldsymbol{x}^{(i)} \to \xi$.

L'unicità del punto fisso segue immediatamente dall'osservazione che se ξ' fosse un altro punto fisso di avrebbe

$$\|\xi - \xi'\| = \|\Phi(\xi) - \Phi(\xi')\| \leq \kappa\|\xi - \xi'\|$$

e pertanto, dato che $\kappa \in [0,1)$, deve essere $\xi = \xi'$. Infine per la stima (7.21) dell'errore basta osservare che, in base alla (7.23)

$$\lim_{h\to\infty} \|\boldsymbol{x}^{(h)} - \boldsymbol{x}^{(i)}\| = \|\xi - \boldsymbol{x}^{(i)}\| \leq \frac{\kappa^i}{1-\kappa}\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\|.$$

□

Il teorema 7.20 è molto utile, in quanto: esso prescinde dalla conoscenza di un intervallo contenente il punto fisso ξ e fornisce una maggiorazione della distanza tra punto fisso e l'$i-$esimo iterato, computazionalmente utilizzabile come criterio di interruzione dei calcoli.

Analogie e differenze con il caso lineare. Nel caso lineare $\Phi(\boldsymbol{x}) = M\boldsymbol{x} - \boldsymbol{c}$, per cui la matrice Jacobiana $\Phi'(\boldsymbol{x}) = M$ coincide con la matrice stessa del sistema. Di conseguenza la condizione $\rho(\Phi'(\boldsymbol{x}^*)) = \rho(M) < 1$ del teorema di Ostrowski fornisce la nota condizione sufficiente sulla convergenza di un sistema lineare. Tuttavia, mentre nei sistemi lineari la convergenza è globale, ossia è indipendente dal punto iniziale, nei sistemi non lineari è soltanto locale, ossia la convergenza è assicurata soltanto se $\boldsymbol{x}^{(0)}$ è *abbastanza* vicino a $\boldsymbol{x}^*$. Da notare inoltre che, mentre nei sistemi lineari la condizione $\rho(\Phi(\boldsymbol{x}^*)) = \rho(M) < 1$ è necessaria per la convergenza, questa condizione non è necessaria per i sistemi non lineari. Il risultato non deve meravigliare dato che $\rho(M) < 1$ (nei sistemi lineari) è un vincolo necessario per la convergenza globale non per la semplice convergenza locale e, nei sistemi non lineari, la convergenza di cui si parla è semplicemente locale.

Supponiamo ora che la funzione di iterazione sia del tipo

$$\Phi(\boldsymbol{x}) = x - [\boldsymbol{C}(\boldsymbol{x})]^{-1}\boldsymbol{F}(\boldsymbol{x}), \tag{7.24}$$

dove $\boldsymbol{C}(\boldsymbol{x})$ è una matrice nonsingolare in $\boldsymbol{x}^*$. Sotto tali ipotesi, essendo $\boldsymbol{F}(\boldsymbol{x}^*) = \boldsymbol{0}$,

$$\Phi'(\boldsymbol{x}^*) = I - [\boldsymbol{C}(\boldsymbol{x}^*)]^{-1}\boldsymbol{F}'(\boldsymbol{x}^*). \tag{7.25}$$

Nel metodo di Newton $\boldsymbol{C}(\boldsymbol{x}) = \boldsymbol{F}'(\boldsymbol{x})$, ossia la funzione di iterazione è

$$\Phi(\boldsymbol{x}) = \boldsymbol{x} - [\boldsymbol{F}'(\boldsymbol{x})]^{-1}\boldsymbol{F}(\boldsymbol{x}).$$

Di conseguenza, nell'ipotesi che nel punto fisso $\boldsymbol{x}^*$ la matrice Jacobiana $\boldsymbol{F}'(\boldsymbol{x})$ sia nonsingolare, la (7.25) implica che $\Phi'(\boldsymbol{x}^*)$ è identicamente nulla, ossia che $\rho(\Phi'(\boldsymbol{x}^*)) = 0$. Sotto tale ipotesi il punto fisso $\boldsymbol{x}^*$ è evidentemente un punto di attrazione per gli iterati $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \boldsymbol{F}'(\boldsymbol{x}^{(k)})^{-1}\boldsymbol{F}(\boldsymbol{x}^{(k)})$, $k = 0, 1, \ldots$.

Metodo di Newton-Jacobi. Supponiamo ora di decomporre $\boldsymbol{F}'(\boldsymbol{x})$ nel modo seguente:

$$\boldsymbol{F}'(\boldsymbol{x}) = \boldsymbol{D}(\boldsymbol{x}) - [\boldsymbol{L}(\boldsymbol{x}) + \boldsymbol{U}(\boldsymbol{x})]$$

dove $\boldsymbol{D}(\boldsymbol{x})$ è la diagonale di $\boldsymbol{F}'(\boldsymbol{x})$, $\boldsymbol{L}(\boldsymbol{x})$ e $\boldsymbol{U}(\boldsymbol{x})$ rispettivamente i triangoli strettamente inferiore e superiore di $\boldsymbol{F}'(\boldsymbol{x})$. Il metodo di Newton-Jacobi è caratterizzato da una funzione di iterazione di tipo (7.24) con

$$\boldsymbol{C}(\boldsymbol{x}) = \boldsymbol{D}(\boldsymbol{x}) \Longrightarrow \Phi(\boldsymbol{x}) = \boldsymbol{x} - [\boldsymbol{D}(\boldsymbol{x})]^{-1}\boldsymbol{F}(\boldsymbol{x})$$

naturalmente nell'ipotesi che tutti gli elementi diagonali di $\boldsymbol{D}(\boldsymbol{x})$ siano strettamente non nulli.

Osserviamo ora che, nel caso lineare, dove $\boldsymbol{F}(\boldsymbol{x}) = A\boldsymbol{x} - \boldsymbol{b}$, il metodo di Newton-Jacobi coincide con il metodo di Jacobi. In questo caso, infatti, essendo $\boldsymbol{F}'(\boldsymbol{x}) = \boldsymbol{D} - (\boldsymbol{L} + \boldsymbol{U})$,

$$\begin{aligned}\Phi(\boldsymbol{x}) &= \boldsymbol{x} - \boldsymbol{D}^{-1}(A\boldsymbol{x} - \boldsymbol{b}) = \boldsymbol{x} - \boldsymbol{D}^{-1}\{[\boldsymbol{D} - (\boldsymbol{L} + \boldsymbol{U})]\boldsymbol{x} - \boldsymbol{b}\} \\ &= \boldsymbol{D}^{-1}[(\boldsymbol{L} + \boldsymbol{U})\boldsymbol{x} + \boldsymbol{b}] = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{c},\end{aligned}$$

essendo

$$\boldsymbol{H} = \boldsymbol{D}^{-1}(\boldsymbol{L} + \boldsymbol{U}), \quad \boldsymbol{c} = \boldsymbol{D}^{-1}\boldsymbol{b}.$$

Supponiamo ora che la $\boldsymbol{F}$ sia differenziabile in tutto un intorno di $\boldsymbol{x}^*$ nel quale $\boldsymbol{F}(\boldsymbol{x}^*) = \boldsymbol{0}$ e che la matrice Jacobiana $\boldsymbol{F}'$ sia continua in $\boldsymbol{x}^*$. Se $\boldsymbol{D}(\boldsymbol{x}^*)$ è nonsingolare e inoltre il raggio spettrale

$$\rho\{[\boldsymbol{D}(\boldsymbol{x}^*)]^{-1}[\boldsymbol{L}(\boldsymbol{x}^*) + \boldsymbol{U}(\boldsymbol{x}^*)]\} < 1, \tag{7.26}$$

la funzione di iterazione $\Phi(\boldsymbol{x}) = \boldsymbol{x} - [\boldsymbol{D}(\boldsymbol{x})]^{-1}\boldsymbol{F}(\boldsymbol{x})$ soddisfa, in $\boldsymbol{x}^*$, il teorema di Ostrowski e di conseguenza $\boldsymbol{x}^*$ è di attrazione per gli iterati $\boldsymbol{x}^{(k)} = \Phi(\boldsymbol{x}^{(k-1)})$. Per la dimostrazione è sufficiente osservare che per la (7.25),

$$\begin{aligned}\Phi'(\boldsymbol{x}^*) &= I - [\boldsymbol{D}(\boldsymbol{x}^*)]^{-1}\{\boldsymbol{D}(\boldsymbol{x}^*) - [\boldsymbol{L}(\boldsymbol{x}^*) + \boldsymbol{U}(\boldsymbol{x}^*)]\} \\ &= [\boldsymbol{D}(\boldsymbol{x}^*)]^{-1}[\boldsymbol{L}(\boldsymbol{x}^*) + \boldsymbol{U}(\boldsymbol{x}^*)].\end{aligned}$$

La verifica della (7.26) naturalmente non può essere diretta, dato che non conosciamo $\boldsymbol{x}^*$. La condizione è tuttavia verificata se si trova un intorno $I_{\boldsymbol{x}^*}$ di $\boldsymbol{x}^*$, per ogni punto $\boldsymbol{x}$ del quale risulti

$$\rho\{\boldsymbol{D}^{-1}(\boldsymbol{x})[\boldsymbol{L}(\boldsymbol{x}) + \boldsymbol{U}(\boldsymbol{x})]\} < 1.$$

Osservazione. Nel caso tale condizione sia verificata per ogni $\boldsymbol{x} \in \mathbb{R}^n$, la funzione di iterazione risulta globalmente contrattiva, con la conseguenza che la convergenza è globale, ossia gli iterati convergono a $\boldsymbol{x}^*$, qualunque sia $\boldsymbol{x}^{(0)}$.

Esempio 7.21 Applicare il metodo di Newton-Jacobi alla risoluzione del sistema ottenuto mediante la discretizzazione, con il metodo delle differenze centrali, del problema

$$y'' = g(y(x)), \quad a \le x \le b, \quad y(a) = \alpha, \quad y(b) = \beta,$$

nell'ipotesi che la g sia differenziabile con continuità su $[a, b]$ e che soddisfi la condizione $g'(t) > 0$, $t \in \mathbb{R}$.

Discretizzando con la usuale notazione si genera il sistema

$$\boldsymbol{F}(\boldsymbol{y}) = 0, \text{ con } f_i(\boldsymbol{y}) = -y_{i+1} + 2y_i - y_{i-1} + h^2 g(y_i) = 0, \tag{7.27}$$

con $i = 1, \ldots, n$, $y_0 = \alpha$ e $y_{n+1} = \beta$.

Da essa segue immediatamente che la matrice Jacobiana $\boldsymbol{F}'(\boldsymbol{y}) = A + \boldsymbol{G}'(\boldsymbol{y})$, essendo

$$A = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{pmatrix},$$

$$\boldsymbol{G}'(\boldsymbol{y}) = h^2 \operatorname{diag}(g'(y_1), g'(y_2), \ldots, g'(y_n)).$$

La condizione di nonnegatività della g implica che

$$(\boldsymbol{F}')_{i,i}(\boldsymbol{y}) = 2 + h^2 g(y_i) > |L_i(\boldsymbol{y}) + U_i(\boldsymbol{y})| = 2, \quad i = 1, \ldots, n,$$

$$\|\boldsymbol{D}^{-1}(\boldsymbol{y})(\boldsymbol{L}(\boldsymbol{y}) + \boldsymbol{U}(\boldsymbol{y}))\|_\infty < 1,$$

e pertanto anche

$$\rho\{\boldsymbol{D}^{-1}(\boldsymbol{y})(\boldsymbol{L}(\boldsymbol{y}) + \boldsymbol{U}(\boldsymbol{y}))\} < 1,$$

qualunque sia $\boldsymbol{y} \in \mathbb{R}^n$. Questo implica che la soluzione del sistema è un punto di attrazione globale per gli iterati del metodo di Newton-Jacobi.

Metodo di Newton-Gauss-Seidel. In questo caso la funzione di iterazione è

$$\Phi(\boldsymbol{x}) = \boldsymbol{x} - [\boldsymbol{D}(\boldsymbol{x}) - \boldsymbol{L}(\boldsymbol{x})]^{-1} \boldsymbol{F}(\boldsymbol{x})$$

dove $\boldsymbol{D}(\boldsymbol{x})$ e $\boldsymbol{L}(\boldsymbol{x})$ hanno il significato precedentemente specificato. Procedendo come nel caso di Newton-Jacobi, è facile dimostrare che, nel caso $\boldsymbol{D}(\boldsymbol{x}^*) - \boldsymbol{L}(\boldsymbol{x}^*)$ sia nonsingolare, il punto fisso $\boldsymbol{x}^*$ è di attrazione per gli iterati del metodo di Newton-Gauss-Seidel se

$$\rho\{[\boldsymbol{D}(\boldsymbol{x}^*) - \boldsymbol{L}(\boldsymbol{x}^*)]^{-1} \boldsymbol{U}(\boldsymbol{x}^*)\} < 1. \tag{7.28}$$

Per quanto concerne la verifica della diseguaglianza (7.28), valgono lo stesso tipo di considerazioni fatte per il metodo di Newton-Jacobi.

Per approfondimenti sulla convergenza locale e globale del metodo di Newton si rinvia al Cap. 5 del libro di Stoer e Burlish [28] e ai libri di Ortega [20], di Ortega e Rheinboldt [21] e di Kelley [16].

Capitolo 8

METODO AGLI ELEMENTI FINITI

8.1 Introduzione

Il metodo agli elementi finiti (Finite Element Method, FEM) è, attualmente, la tecnica numerica più utilizzata nella risoluzione delle PDEs con assegnate condizioni al bordo (Boundary Value Problems, BVPs). Questo non esclude, naturalmente, l'esistenza di settori specifici nei quali altri metodi siano più utilizzati. Nel caso di domini regolari, ad esempio, i metodi tuttora più utilizzati sono quelli alle differenze finite. L'utilizzo degli elementi finiti è essenziale nel caso di domini complessi (telaio di un'automobile, motore di un aereo, ecc.).

Nel settore esistono molti libri eccelenti, come i libri di Brennes e Scott [2], Strang e Fix [30], Ciarlet [5] e Raviart e Thomas [23]. Essi tuttavia non riservano abbastanza spazio ai dettagli pratici sugli algoritmi e sulla programmazione. Per questo motivo, consideriamo più adatti alle esigenze didattiche degli studenti di Ingegneria quelli di Gockenbach [9] e di Quarteroni [22].

Dal punto di vista teorico un BVP lineare può essere rappresentato nella forma

$$Af = g, \tag{8.1}$$

dove $A : X \to Y$ è un operatore lineare da uno spazio di Hilbert X ad un altro spazio di Hilbert Y. Nelle applicazioni, spesso, gli spazi X e Y sono coincidenti.

Nei BVPs, A è identificato da un operatore a derivate parziali lineare e dalle condizioni assegnate al bordo. Nei problemi di tipo ellittico, l'operatore differenziale in n variabili è del tipo

$$\hat{A} = -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}(\boldsymbol{x})\frac{\partial}{\partial x_j}\right) + a_0(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega \subset \mathbb{R}^n, \tag{8.2}$$

con X e Y opportuni spazi di Hilbert.

Nel caso particolare $X = Y$, generalmente verificato nelle applicazioni, condizione necessaria e sufficiente perchè l'equazione funzionale (8.1) sia univocamente risolubile in X, qualunque sia $g \in X$, è che esista limitato l'operatore inverso $A^{-1} : X \to X$.

Allo scopo di dare condizioni sufficienti per stabilire tale proprietà, nel caso degli operatori ellittici, si fa riferimento al funzionale bilineare

$$a(u,v) = -\sum_{i,j=1}^{n} \int_\Omega \frac{\partial}{\partial x_i}\left(a_{ij}(\boldsymbol{x})\frac{\partial u}{\partial x_j}\right) v(\boldsymbol{x})\, d\boldsymbol{x} + \int_\Omega a_0(\boldsymbol{x}) u(\boldsymbol{x}) v(\boldsymbol{x})\, d\boldsymbol{x} \quad (8.3)$$

con u e v appartenenti ad X. Vale infatti il seguente importante

Lemma 8.1 (Lax-Milgram) *Condizione sufficiente perchè l'operatore* (*differenziale*) *lineare $A : X \to X$ abbia inverso limitato è che esso sia* **coercivo**, *ossia che esiste una costante $c > 0$ tale che*

$$a(v,v) \geq c\|v\|^2, \quad \textit{qualunque sia } v \in X, \quad (8.4)$$

dove il simbolo $\|\cdot\|$ indica la norma nello spazio di Hilbert X.

Nel caso dei modelli differenziali di tipo ellittico, parabolico e iperbolico considerati in questo libro, la condizione di coercività, introdotta nel Lemma di Lax-Milgram, è equivalente a quella di ellitticità introdotta nella Sezione 1.2 del libro.

Il primo passo nella risoluzione di un BVP consiste nell'associare al problema iniziale una *formulazione variazionale*

$$a(u,v) = L(v), \quad \text{per ogni } v \in X, \quad (8.5)$$

dove $a(u,v)$ è il funzionale bilineare (8.3) e $L(v)$ è un operatore lineare limitato. Nell'ipotesi di coercività dell'operatore A, il Lemma di Lax-Milgram implica che l'equazione (8.5) possiede una e una sola soluzione $u \in X$.

Allo spazio infinito dimensionale X viene associata una successione di spazi finito dimensionali $\{X_n\}_{n=1}^{\infty}$ soddisfacenti la seguente proprietà:

$$X_n \subset X_{n+1} \quad \text{con} \quad \overline{\bigcup_{n=1}^{\infty} X_n} = X,$$

dove il soprassegno indica la chiusura dell'unione degli spazi introdotti.

Al problema infinito dimensionale (8.5) viene quindi associato un problema finito dimensionale

$$a(u_n, v) = L(v) \quad \text{per ogni } v \in X_n. \quad (8.6)$$

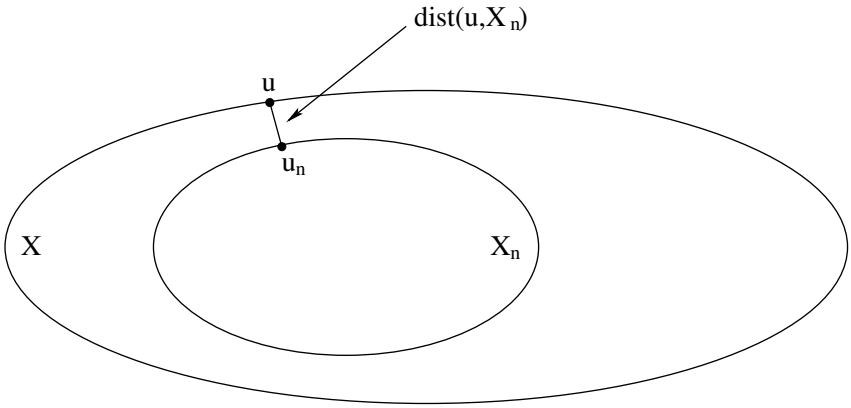


Figura 8.1: Distanza $\text{dist}(u, X_n) = \min_{u_n \in X_n} \|u - u_n\|$ tra la soluzione u in X e la soluzione u_n nello spazio finito dimensionale X_n.

Il Lemma di Lax-Milgram garantisce che, sotto le ipotesi di coercività, anche il problema finito dimensionale (8.6) possiede esattamente una soluzione in X_n [Fig. 8.1]. Al tendere di $n \to \infty$ la $\|u - u_n\| \to 0$, come stabilito dal seguente importante

Lemma 8.2 (di Céa) *[5] Sotto le ipotesi di coercività dell'operatore differenziale e di approssimazione della successione $\{X_n\}_{n=1}^{\infty}$, vale la seguente diseguaglianza*

$$\|u - u_n\| \leq c\, dist(u, X_n), \tag{8.7}$$

essendo c una costante indipendente da n.

Il risultato è molto importante in quanto consente di utilizzare, in tale ambiente, i numerosi risultati della teoria dell'approssimazione sulla stima della distanza tra una funzione in uno spazio di Hilbert X e la sua approssimazione ottimale in molteplici sottospazi finito dimensionali X_n.

8.2 Formulazione variazionale

8.2 a Formulazione variazionale di una ODE con valori agli estremi assegnati

Consideriamo il seguente problema differenziale (con condizioni di omogeneità agli estremi):

$$\begin{cases} -\dfrac{d}{dx}\left(p(x)\dfrac{dy}{dx}\right) + q(x)y = f(x), \quad x \in [a,b] \subset \mathbb{R}, \\ y(a) = \alpha, \quad y(b) = \beta, \end{cases} \tag{8.8a}$$

dove $p(x) > 0$ in (a,b), essendo $p \in C^1[a,b]$, con $q, f \in C^0[a,b]$, avendo indicato con $C^0[a,b]$ lo spazio delle funzioni continue in $[a,b]$ e con $C^k[a,b]$, $k = 1, 2, \ldots$, lo spazio delle funzioni continue in $[a,b]$ assieme alle sue derivate fino all'ordine k. Nel caso $\alpha = \beta = 0$, si dice che il problema (8.8a) soddisfa condizioni di omogeneità agli estremi. Diversamente si dice che le condizioni agli estremi sono di inomogeneità. Per ragioni di semplicità, nell'associare al problema (8.8a) la sua formulazione variazionale, si ipotizzano condizioni di omogeneità agli estremi. Tale ipotesi non è limitativa in quanto, con una semplice trasformazione lineare, è sempre possibile ricondursi a tale situazione. Posto infatti $y = z + \varphi$ con $\varphi(x) = \alpha + \frac{x-a}{b-a}(\beta - \alpha)$, dove φ è l'interpolante lineare tra (a,α) e (b,β), l'equazione (8.8a) si trasforma nella seguente:

$$\begin{cases} -\dfrac{d}{dx}\left(p(x)\dfrac{dz}{dx}\right) + q(x)z = g(x), \quad x \in [a,b] \subset \mathbb{R}, \\ z(a) = z(b) = 0, \end{cases} \tag{8.8b}$$

dove $g(x)$ è la funzione $g(x) = f(x) + \frac{d}{dx}\left[p(x)\varphi(x)\right] - q(x)\varphi(x)$, continua in $[a,b]$. Per questo motivo nel seguito supporremo, per semplicità, che nel problemi (8.8a) le condizioni agli estremi siano di tipo omogeneo. L'esistenza puntuale della (8.8a) presuppone che $p \in C^1[a,b]$ e che $q, f \in C^0[a,b]$. Questo implica che la sua soluzione puntuale

$$y \in C_0^2[a,b] = \left\{ y; y, y', y'' \in C^0[a,b], \text{ con } y(a) = y(b) = 0 \right\}.$$

Per generarla introduciamo, come spazio delle "funzioni test", lo spazio $C_0^2[a,b]$. Indicata con v la generica funzione test, moltiplicando primo e secondo membro delle (8.8) per v e integrando su $[a,b]$ otteniamo l'equazione

$$-\int_a^b \frac{d}{dx}\left(p(x)\frac{dy}{dx}\right) v(x)\,dx + \int_a^b q(x)y(x)v(x)\,dx = \int_a^b f(x)v(x)\,dx.$$

Da essa, integrando per parti il primo termine e tenendo presente che (per ipotesi) $v(a) = v(b) = 0$, otteniamo che, per ogni funzione test v, vale l'equazione

$$\int_a^b p(x)y'(x)v'(x)\,dx + \int_a^b q(x)y(x)v(x)\,dx = \int_a^b f(x)v(x)\,dx. \tag{8.9}$$

La (8.9) viene definita *formulazione variazionale* delle (8.8). È importante osservare che le due formulazioni sono equivalenti, nel senso che: (a) ogni soluzione y della (8.8) soddisfa la (8.9), qualunque sia $v \in C_0^2[a,b]$; (b) ogni funzione y che soddisfa la (8.9), qualunque sia $v \in C_0^2[a,b]$, soddisfa la (8.8). Essendo la (a) ovvia, ci limitiamo a dimostrare la (b). Procedendo per assurdo

supponiamo che, pur essendo la (8.9) soddisfatta (qualunque sia $v \in C_0^2[a,b]$) esista un punto $x_0 \in (a,b)$ nel quale risulti

$$-\frac{d}{dx}\left(p(x)\frac{dy}{dx}\right) + q(x)y(x) - f(x) > 0. \tag{8.10}$$

Sotto tale ipotesi, per continuità esiste un intorno $I_{x_0,\delta}$ di centro x_0 e raggio δ, in ogni punto del quale la (8.10) risulta soddisfatta. Questo implica che possiamo costruire una $v_0 \in C_0^2[a,b]$, positiva in $I_{x_0,\delta}$ e nulla altrove. Ma allora

$$I_0 = \int_a^b \left[-\frac{d}{dx}\left(p(x)\frac{dy}{dx}\right) + q(x)y(x) - f(x)\right] v_0(x)\,dx > 0,$$

contro l'ipotesi che

$$\int_a^b [p(x)y'(x)v'(x) + q(x)y(x)v(x) - f(x)v(x)] = 0,$$

qualunque sia $v \in C_0^2[a,b]$.

Per estendere tale tipo di formulazione ai modelli differenziali alle derivate parziali (BVPs per PDEs) è necessario premettere alcuni risultati di calcolo vettoriale differenziale.

8.2 b Richiami di calcolo vettoriale-differenziale

Gradiente di una funzione. Indicata con $u(\boldsymbol{x}) = u(x_1, x_2, \ldots, x_n)$ una funzione definita in un dominio Ω e ivi differenziabile, per gradiente della u in Ω, in simboli, $\nabla u = \operatorname{grad} u$ (nabla u, gradiente di u), si intende il vettore

$$\nabla u = \left(\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, \ldots, \frac{\partial u}{\partial x_n}\right)^T.$$

Esempio 8.3 Il gradiente della funzione $u(x_1, x_2, x_3) = x_1^2 + x_2^2 + \cos(x_1 x_2 x_3)$ è il vettore

$$\nabla u = \begin{pmatrix} 2x_1 - x_2 x_3 \sin(x_1 x_2 x_3) \\ 2x_2 - x_1 x_3 \sin(x_1 x_2 x_3) \\ -x_1 x_2 \sin(x_1 x_2 x_3) \end{pmatrix}.$$

Derivata direzionale-gradiente. Supponiamo che $u(x_1, x_2, \ldots, x_n)$ e le sue derivate parziali siano continue in una sfera con centro $\boldsymbol{x}^{(0)}$ e che $\nabla u(\boldsymbol{x}^{(0)}) \neq \boldsymbol{0}$. Per derivata direzionale della u in $\boldsymbol{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})$, lungo la direzione $\boldsymbol{d} = (d_1, d_2, \ldots, d_n)^T$, si intende il prodotto vettoriale

$$\partial_{\boldsymbol{d}} u(\boldsymbol{x}^{(0)}) \cdot \boldsymbol{d} = \frac{\partial u(\boldsymbol{x}^{(0)})}{\partial x_1} d_1 + \ldots + \frac{\partial u(\boldsymbol{x}^{(0)})}{\partial x_n} d_n,$$

dove, come mostra la definizione e come usuale, il punto indica il prodotto interno (prodotto scalare) tra i vettori $\nabla u(\boldsymbol{x}^{(0)})$ e $\boldsymbol{d}$. Dalle quale, ricordando che (per definizione di prodotto tra vettori)

$$\partial_{\boldsymbol{d}} u(\boldsymbol{x}^{(0)}) \cdot \boldsymbol{d} = \left\|\nabla u(\boldsymbol{x}^{(0)})\right\| \, \|\boldsymbol{d}\| \cos\theta,$$

dove θ è l'angolo tra i vettori $\nabla u(\boldsymbol{x}^{(0)})$ e $\boldsymbol{d}$, seguono le seguenti osservazioni:

a) la direzione di massima crescita, ossia la direzione $\boldsymbol{d}$, di lunghezza unitaria, lungo la quale, a partire da $\boldsymbol{x}^{(0)}$ la funzione u cresce maggiormente è rappresentata dal vettore unitario (versore)

$$\boldsymbol{d} = \frac{\nabla u(\boldsymbol{x}^{(0)})}{\|\nabla u(\boldsymbol{x}^{(0)})\|}.$$

b) Per lo stesso tipo di considerazioni, il versore

$$\boldsymbol{d} = -\frac{\nabla u(\boldsymbol{x}^{(0)})}{\|\nabla u(\boldsymbol{x}^{(0)})\|}$$

indica la direzione di massima decrescita, ossia la direzione lungo la quale, a partire da $\boldsymbol{x}^{(0)}$, la u presenta la più ripida discesa.

Al fine di interpretare geometricamente l'osservazione b), consideriamo la famiglia delle curve di livello

$$u(x_1, x_2) = c, \qquad c \in \mathbb{R},$$

associata ad una funzione differenziabile u in un dominio bidimensionale Ω. Considerato un punto $\boldsymbol{x}^{(0)}$, la direzione di più ripida discesa, come indicato nella Fig. 8.2, rappresenta il versore normale alla tangente di una curva di livello, in $\boldsymbol{x}^{(0)}$, rivolto verso l'interno, ossia nel verso delle curve di livello con c decrescenti.

Esempio 8.4 Se $u(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2$, il versore normale di più ripida discesa in $\boldsymbol{x}^{(0)} = (1, 1, 1)^T$ è $\boldsymbol{d} = -\frac{1}{\sqrt{3}}(1, 1, 1)^T$. Esso indica infatti che, qualunque sia la superficie sferica di equazione $x_1^2 + x_2^2 + x_3^2 = c$, $c > 0$, il versore $\boldsymbol{d}$, partendo da $\boldsymbol{x}^{(0)}$, punta verso l'origine, ossia al centro della sfera caratterizzata da $c = 0$.

Versore normale esterno. Per versore normale esterno (vettore normale esterno di norma 1) ad una superficie, dotata di piano tangente in un punto

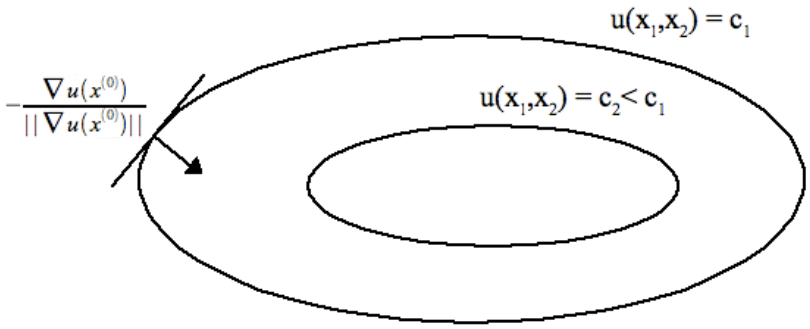


Figura 8.2: La figura mostra due curve di livello della funzione $u(x_1, x_2)$ e il suo (versore) gradiente.

$\boldsymbol{x}^{(0)}$, si intende il versore normale al piano tangente, orientato nel verso delle curve di livello crescenti della $u(\boldsymbol{x})$. Generalmente indicato con

$$\boldsymbol{n} = \frac{\nabla u(\boldsymbol{x}^{(0)})}{\|\nabla u(\boldsymbol{x}^{(0)})\|},$$

esso rappresenta, nel punto, la direzione di più rapida crescita.

Esempio 8.5 Consideriamo la superficie rappresentata dall'equazione

$$x_3 = \sqrt{x_1^2 + x_2^2}\,,$$

in cui x_1 e x_2 sono numeri reali qualsiasi. Il versore normale esterno alla superficie $x_3 - \sqrt{x_1^2 + x_2^2} = 0$ in (x_1, x_2, x_3), è

$$\boldsymbol{n} = \frac{1}{\sqrt{2}} \left(\frac{x_1}{\sqrt{x_1^2 + x_2^2}}, \frac{x_2}{\sqrt{x_1^2 + x_2^2}}, -1 \right)^T = \frac{1}{\sqrt{2}} \left(\frac{x_1}{x_3}, \frac{x_2}{x_3}, -1 \right)^T.$$

Il versore normale alla suddetta superficie conica, in $(1, 1, \sqrt{2})$ è pertanto

$$\boldsymbol{n} = \frac{1}{\sqrt{2}} \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, -1 \right)^T.$$

Divergenza di un campo vettoriale. Indicato con

$$\boldsymbol{v}(\boldsymbol{x}) = (v_1(\boldsymbol{x}), v_2(\boldsymbol{x}), \ldots, v_n(\boldsymbol{x}))^T, \quad \boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n,$$

un vettore n-dimensionale differenziabile in un punto $\boldsymbol{x} \in \Omega$, per divergenza di v in $\boldsymbol{x}$, in simboli

$$(\nabla \cdot v)(\boldsymbol{x}) = (\operatorname{div} v)(\boldsymbol{x}) \qquad (\text{nabla} \cdot v \text{ in } \boldsymbol{x},\ \text{divergenza di } v \text{ in } \boldsymbol{x}),$$

si intende la funzione

$$(\nabla \cdot v)(\boldsymbol{x}) = \frac{\partial v_1(\boldsymbol{x})}{\partial x_1} + \frac{\partial v_2(\boldsymbol{x})}{\partial x_2} + \ldots + \frac{\partial v_n(\boldsymbol{x})}{\partial x_n}.$$

Esempio 8.6 Se $v(\boldsymbol{x}) = (x_1^2 + x_2^2, x_1 x_2 x_3, x_3 \sin(x_1 x_2))^T$, la divergenza di v in $\boldsymbol{x}$ è la funzione

$$(\nabla \cdot v)(\boldsymbol{x}) = 2x_1 + x_1 x_3 + \sin(x_1 x_2).$$

Se, in particolare, v è differenziabile due volte in Ω,

$$\begin{aligned}(\nabla \cdot \nabla v)(\boldsymbol{x}) &= \frac{\partial}{\partial x_1}\frac{\partial v(\boldsymbol{x})}{\partial x_1} + \frac{\partial}{\partial x_2}\frac{\partial v(\boldsymbol{x})}{\partial x_2} + \ldots + \frac{\partial}{\partial x_n}\frac{\partial v(\boldsymbol{x})}{\partial x_n} \\ &= \frac{\partial^2 v(\boldsymbol{x})}{\partial x_1^2} + \frac{\partial^2 v(\boldsymbol{x})}{\partial x_2^2} + \ldots + \frac{\partial^2 v(\boldsymbol{x})}{\partial x_n^2}\end{aligned}$$

che, come noto, indica il Laplaciano della v. In simboli

$$(\nabla \cdot \nabla v)(\boldsymbol{x}) = \Delta v(\boldsymbol{x}) = \nabla^2 v(\boldsymbol{x}).$$

Nell'esempio 8.4 ($u(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2$)

$$(\nabla \cdot \nabla v)(\boldsymbol{x}) = (\nabla \cdot \nabla u)(\boldsymbol{x}) = 6.$$

Il risultato del calcolo vettoriale più importante nella trattazione della forma debole di un problema alle derivate parziali con assegnate condizioni al contorno (BVPs per PDEs) è rappresentato dal teorema della divergenza. In un certo senso, esso rappresenta l'analogo multidimensionale del teorema fondamentale del calcolo integrale.

Anche se i risultati, per semplicità, sono illustrati in due dimensioni, grazie alla generalità delle notazioni usate, essi si estendono facilmente al caso n-dimensionale ($n = 3, 4, \ldots$). Se Ω è un dominio di $\mathbb{R}^2$ con contorno $\partial\Omega$ regolare a tratti ed $\boldsymbol{f}$ è un campo vettoriale definito su $\overline{\Omega}$ (chiusura di Ω, $\overline{\Omega} = \Omega \cup \partial\Omega$), il *teorema della divergenza* stabilisce che

$$\int_\Omega \nabla \cdot \boldsymbol{f}\, d\boldsymbol{x} = \int_{\partial\Omega} \boldsymbol{f} \cdot \boldsymbol{n}\, d\sigma, \tag{8.11}$$

dove $\boldsymbol{n}$ è il versore normale a $\partial\Omega$ (di lunghezza unitaria) che è variabile in $\partial\Omega$. Il teorema della divergenza, per la sua applicabilità, richiede ovviamente che il campo vettoriale f sia sufficiemente regolare. Sotto tale ipotesi di regolarità esso stabilisce una relazione tra una quantità definita in un dominio Ω e un'altra definita sulla frontiera $\partial\Omega$. Esso permette di stabilire, in particolare, che per un problema del calore, in condizioni di stazionarità, nel dominio Ω vale l'equazione

$$-\nabla \cdot (k\, \nabla u) = f, \tag{8.12}$$

dove k indica una costante fisica del dominio e f una sorgente termica.

Indicato infatti con ω un subdominio di Ω, la quantità di calore che entra in Ω può essere descritto sia in termini di una sorgente f, sia in termini del calore che attraversa la frontiera, rappresentato dal flusso termico $-k\,\nabla u$. In condizioni di equilibrio termico i due flussi si compensano, ossia risulta che

$$\int_{\partial\omega} k\,\nabla u\cdot\boldsymbol{n}\,d\sigma+\int_{\omega} f\,d\boldsymbol{x}=0.$$

Per il teorema della divergenza (applicato al primo integrale), qualunque sia il subdominio ω di Ω

$$\int_{\omega}\left[\nabla\cdot(k\,\nabla u)+f\right]d\boldsymbol{x}=0,$$

da cui (per l'arbitrarietà di ω in Ω) segue che, in condizioni di stabilità termica, vale la (8.12). Il teorema della divergenza permette anche di stabilire facilmente le "condizioni di compatibilità" per il seguente problema di Neumann:

$$\begin{cases}-\nabla\cdot(k\,\nabla u)=f \quad \text{in} \quad \Omega,\\ k\,\dfrac{\partial u}{\partial u}=h \quad \text{in} \quad \partial\Omega.\end{cases}\tag{8.13a}$$

È intuitivo pensare che per l'esistenza e l'unicità della soluzione del problema (8.13a), le funzioni f ed h non possono essere del tutto indipendenti tra loro. È infatti naturale attendersi che esista tra loro un problema di compatibilità. Tale condizione la si può ricavare, in modo molto semplice, mediante il teorema della divergenza.

Basta infatti osservare che, indicata con $\frac{\partial u}{\partial n}$ la derivata normale di u nel punto (x,y) di $\partial\Omega$ e ricordando che $\frac{\partial u}{\partial n}=\nabla u\cdot\boldsymbol{n}$ (essendo $\boldsymbol{n}$ il versore normale della u in (x,y) orientato verso l'esterno), per il teorema della divergenza

$$\int_{\Omega} f\,d\boldsymbol{x}=-\int_{\Omega}\nabla\cdot(k\,\nabla u)\,d\boldsymbol{x}=-\int_{\partial\Omega} k\,\frac{\partial u}{\partial n}\,d\sigma=-\int_{\partial\Omega} h\,d\sigma.$$

Le condizioni di compatibilità tra le funzioni f ed h sono dunque espresse dalla condizione

$$\int_{\Omega} f\,d\boldsymbol{x}+\int_{\partial\Omega} h\,d\sigma=0.\tag{8.13b}$$

Per il seguito è *utile osservare* che, nel caso $f(x,y)=\binom{f_1(x,y)}{0}$, il teorema della divergenza si semplifica nel modo seguente:

$$\int_{\Omega}\frac{\partial f}{\partial x}\,dxdy=\int_{\partial\Omega} f_1 n_1\,d\sigma,\tag{8.14a}$$

essendo n_1 la prima componente del vettore $\boldsymbol{n}$ orientato verso l'esterno. Per analogia è evidente che, qualora sia nulla la prima componente del campo vettoriale f, risulta

$$\int_\Omega \frac{\partial f_2}{\partial y}\, dxdy = \int_{\partial\Omega} f_2 n_2\, d\sigma, \tag{8.14b}$$

essendo f_2 la seconda componente di f e n_2 la seconda componente di $\boldsymbol{n}$.

Prima identità di Green. La derivazione della formulazione variazionale (debole) di un BVP richiede l'utilizzo della prima identità di Green, considerato l'analogo multidimensionale della regola di integrazione per parti. Dalla regola di derivazione di un prodotto sappiamo che

$$\frac{d}{dx}(uv) = \frac{du}{dx}v + u\frac{du}{dx}, \tag{8.15a}$$

come anche

$$\int_a^b \frac{du}{dx}v\, dx + \int_a^b u\frac{dv}{dx}\, dx = [uv]_a^b\,,$$

da cui segue la regola

$$\int_a^b u\frac{dv}{dx}\, dx = [uv]_a^b - \int_a^b \frac{du}{dx}v\, dx.$$

La prima identità di Green segue dall'applicazione del teorema della divergenza alla seguente regola di prodotto multidimensionale

$$\nabla\cdot(v\,\nabla u) = \nabla v\cdot\nabla u + v\,\Delta u, \tag{8.15b}$$

dove $\Delta u = \nabla\cdot\nabla u$ è il cosiddetto Laplaciano di u. (La sua derivazione segue dalla applicazione della regola del prodotto al primo membro, esplicitamente espressa rispetto alle sue coordinate). Integrando in Ω ambedue i membri della (8.15b), otteniamo

$$\int_\Omega \nabla\cdot(v\,\nabla u)\, dxdy = \int_\Omega \nabla u\cdot\nabla v\, dxdy + \int_\Omega v\,\Delta u\, dxdy.$$

Da cui, applicando il teorema della divergenza, segue che

$$\int_{\partial\Omega} v(\nabla u\cdot\boldsymbol{n})\, d\sigma = \int_\Omega \nabla u\cdot\nabla v\, dxdy + \int_\Omega v\,\Delta u\, dxdy.$$

Da essa, ricordando che $\nabla u\cdot\boldsymbol{n} = \frac{\partial u}{\partial n}$, segue la *prima identità di Green*

$$\int_\Omega v\,\Delta u\, dxdy = \int_{\partial\Omega} v\,\frac{\partial u}{\partial n}\, d\sigma - \int_\Omega \nabla u\cdot\nabla v\, dxdy. \tag{8.16a}$$

Una forma alternativa alla (8.16a), molto utile nello studio dei BVPs, è la seguente:

$$-\int_\Omega v\,\nabla\cdot(k\,\nabla u)\,dxdy=\int_\Omega k\,\nabla u\cdot\nabla v\,dxdy-\int_{\partial\Omega}kv\,\frac{\partial u}{\partial n}\,d\sigma. \tag{8.16b}$$

Forma ottenibile dalla seguente applicazione della regola di differenziazione del prodotto:

$$\nabla\cdot(v(k\,\nabla u))=k\,\nabla u\cdot\nabla v+v\,\nabla\cdot(k\,\nabla u)$$

e del teorema della divergenza. Il teorema della divergenza implica infatti che

$$\begin{aligned}\int_\Omega\nabla\cdot(v(k\,\nabla u))\,dxdy&=\int_{\partial\Omega}kv(\nabla\cdot\boldsymbol{n})\,d\sigma=\int_{\partial\Omega}kv\,\frac{\partial u}{\partial n}\,d\sigma\\&=\int_\Omega k\,\nabla u\cdot\nabla v\,dxdy+\int_\Omega v\,\nabla\cdot(k\,\nabla u)\,dxdy,\end{aligned}$$

da cui

$$-\int_\Omega(\nabla\cdot k\,\nabla u)\,dxdy=-\int_{\partial\Omega}kv\,\frac{\partial u}{\partial n}\,d\sigma+\int_\Omega k\,\nabla u\cdot\nabla v\,dxdy.$$

La prima identità di Green e il teorema della divergenza (più che per l'effettivo calcolo di specifici integrali) sono molto importanti per ricavare altre formule di pratica utilità.

Vediamo ora alcune forme particolari del teorema della divergenza e della prima identità di Green, molto utilizzate nel campo degli elementi finiti. Dalla (8.10), ponendo $f=\left(\begin{smallmatrix}uv\\0\end{smallmatrix}\right)$ e $f=\left(\begin{smallmatrix}0\\uv\end{smallmatrix}\right)$ e indicati con n_1 e n_2 la prima e la seconda componente di $\boldsymbol{n}$, possiamo ricavare le seguenti utili formule:

$$\begin{cases}\displaystyle\int_\Omega u\frac{\partial v}{\partial x}\,dxdy+\int_\Omega v\,\frac{\partial u}{\partial x}\,dxdy=\int_\Omega uvn_1\,dxdy,\\[2ex]\displaystyle\int_\Omega u\frac{\partial v}{\partial y}\,dxdy+\int_\Omega v\,\frac{\partial u}{\partial y}\,dxdy=\int_\Omega uvn_2\,dxdy.\end{cases} \tag{8.17}$$

Altra formula molto utilizzata e facilmente ricavabile dalla regola di derivazione di un prodotto e dal teorema della divergenza, è la seguente:

$$\int_\Omega v\,\nabla\cdot(k\,\nabla u)\,dxdy=\int_{\partial\Omega}vk\,\frac{\partial u}{\partial n}\,d\sigma-\int_\Omega k\,\nabla u\cdot\nabla v\,dxdy. \tag{8.18}$$

8.2 c Forma variazionale di tipici BVPs

1. Come primo esempio consideriamo il classico problema di Dirichlet con condizioni di omogeneità sul bordo:

$$\begin{cases}-\nabla\cdot(k\,\nabla u)=f\quad\text{in}\quad\Omega,\\u|_{\partial\Omega}=0,\end{cases} \tag{8.19}$$

dove $k(x,y) > 0$ in Ω e $u|_{\partial\Omega} = 0$, ossia $u(x,y) = 0$ in ogni punto della frontiera di Ω.

Se la f è continua e u è una soluzione (in senso classico) della (8.19), la u e le sue derivate parziali prime e seconde sono ovunque continue in Ω, con $u(x,y) = 0$ in ogni punto di $\partial\Omega$. Possiamo dunque affermare che u appartiene allo spazio

$$C_0^2(\Omega) = \{v \in C^2(\overline{\Omega}) : v = 0 \text{ su } \partial\Omega\},$$

avendo indicato con $\overline{\Omega}$ la chiusura di Ω e con $C^2(\overline{\Omega})$ lo spazio delle funzioni continue, assieme alle sue derivate parziali prime e seconde, su $\overline{\Omega}$. Se u è soluzione della (8.19), è evidente che, qualunque sia $v \in C_0^2(\Omega)$,

$$\int_{\partial\Omega} [\nabla \cdot (k\,\nabla u) + f]\, v\, d\sigma = 0. \tag{8.20}$$

Equazione che, in conseguenza della (8.18) e dell'ipotesi che v sia nulla su $\partial\Omega$, può essere scritta nella seguente forma variazionale:

$$\int_\Omega k\,\nabla u \cdot \nabla v\, dxdy = \int_\Omega f v\, dxdy, \quad \text{qualunque sia } v \in C_0^2(\Omega). \tag{8.21}$$

È importante osservare che vale anche l'inverso, nel senso che se la (8.21) è verificata qualunque sia $v \in C_0^2(\Omega)$, è verificato anche il problema di Dirichlet (8.19). Per dimostrarlo basta osservare che se esiste un punto $(x_0, y_0) \in \Omega$ nel quale risulta

$$\int_\Omega (\nabla \cdot (k\,\nabla u) + f)\, dxdy > 0, \tag{8.22}$$

per continuità esiste un intorno $I_{(x_0,y_0)} \subset \Omega$ in ogni punto del quale la (8.22) risulta soddisfatta.

Ma allora è possibile costruire una funzione $v \in C_0^2(\Omega)$, positiva in ogni punto interno di $I_{(x_0,y_0)}$ e nulla altrove in $\overline{\Omega}$, per la quale risulta

$$\int_\Omega [\nabla \cdot (k\,\nabla u) + f]\, v\, dxdy > 0,$$

contro l'ipotesi che tale integrale sia nullo, qualunque sia $v \in C_0^2(\Omega)$. Lo stesso tipo di considerazioni vale (naturalmente) nel caso che esiste un punto di Ω nel quale sia

$$\nabla \cdot (k\,\nabla u) + f < 0.$$

Si può pertanto affermare che se una funzione u verifica la (8.20), qualunque sia $v \in C_0^2(\Omega)$, la u è anche soluzione del problema di Dirichlet (8.19). Risultato pertanto dimostrato che risolvere il BVP (8.19) equivale a determinare la funzione $u \in C_0^2(\Omega)$ che soddisfa la (8.21), qualunque sia $v \in C_0^2(\Omega)$.

2. Come secondo esempio consideriamo il classico problema di Neumann, con condizioni di omogeneità al bordo:

$$\begin{cases} -\nabla \cdot (k\,\nabla u) = f, \quad (x,y) \in \Omega, \\ k\,\dfrac{\partial u}{\partial n} = 0, \quad \text{su } \partial\Omega. \end{cases} \tag{8.23}$$

La continuità della f in Ω implica che, come nel precedente problema di Dirichlet, u è continua in Ω assieme alle sue derivate parziali prime e seconde. La differenza, rispetto al caso precedente, dipende unicamente dalle diverse condizioni al bordo. In questo caso supponiamo che l'insieme delle funzioni test sia $C^2(\overline{\Omega})$. La (8.23) implica che, qualunque sia la funzione test $v \in C^2(\overline{\Omega})$,

$$-\int_\Omega \nabla \cdot (k\,\nabla u) v\,dxdy = \int_\Omega f v\,dxdy,$$

da cui, applicando la prima identità di Green, segue che (per ogni $v \in C^2(\overline{\Omega})$)

$$-\int_{\partial\Omega} vk\,\frac{\partial u}{\partial n}\,d\sigma + \int_{\partial\Omega} k\,\nabla u \cdot \nabla v\,d\sigma = \int_\Omega f v\,dxdy, \quad \text{per ogni } v \in C^2(\overline{\Omega}). \tag{8.24}$$

Da quest'ultima relazione, essendo $k\,\frac{\partial u}{\partial n} = 0$ su $\partial\Omega$, segue infine che

$$\int_\Omega k\,\nabla u \cdot \nabla v\,dxdy = \int_\Omega f v\,dxdy, \quad \text{per ogni } v \in C^2(\overline{\Omega}). \tag{8.25}$$

La condizione di omogeneità sulla derivata normale prevista dal problema di Neumann non richiede alcun vincolo della funzione test, a differenza di quanto richiesto dal problema di Dirichlet.

La (8.25) rappresenta la formulazione variazionale del problema (8.23). Mentre è evidente che la (8.23) implica (8.24) e conseguentemente la (8.25), l'inverso non lo è. Non è evidente, in particolare, per quale motivo la u deve soddisfare la condizione al bordo $k\,\frac{\partial u}{\partial n} = 0$. Per dimostrarlo, osserviamo che l'applicazione della prima identità di Green alla (8.25) implica che

$$-\int_\Omega \nabla \cdot (k\,\nabla u) v\,dxdy + \int_{\partial\Omega} k\,\frac{\partial u}{\partial n} v\,d\sigma = \int_\Omega f v\,dxdy, \text{ qualunque sia } v \in C^2(\overline{\Omega}).$$

Da tale relazione, essendo $-\nabla \cdot (k\,\nabla u) = f$, segue ovviamente che

$$\int_\Omega \left(k\,\frac{\partial u}{\partial n} \right) v\,dxdy = 0, \quad \text{per ogni } v \in C^2(\overline{\Omega}). \tag{8.26}$$

Di conseguenza, dall'arbitrarietà di v in $C^2(\overline{\Omega})$ segue che

$$k\,\frac{\partial u}{\partial n} = 0.$$

Per questo motivo la (8.25) viene considerata equivalente alla (8.23).

3. Come terzo esempio consideriamo il seguente problema con condizioni alla frontiera di tipo misto (problema misto):

$$\begin{cases} -\nabla \cdot (k\,\nabla u) = f, \quad (x,y) \in \Omega, \\ u = 0 \quad \text{su } \Gamma_1, \\ k\,\dfrac{\partial u}{\partial n} = 0 \quad \text{su } \Gamma_2, \end{cases} \tag{8.27}$$

dove $\Gamma_1 \cup \Gamma_2 = \partial\Omega$, con $\Gamma_1 \cap \Gamma_2 = \emptyset$. Come nel caso Dirichlet la condizione $u = 0$ su Γ_1 è considerata *essenziale*, nel senso che deve essere presente anche nello spazio delle funzioni test. La condizione $k\,\frac{\partial u}{\partial n} = 0$ su Γ_2 è invece considerata di tipo *naturale*, nel senso che contribuisce all'annullarsi dell'integrale su $\partial\Omega$, senza implicare alcun vincolo sulle funzioni test. Lo spazio delle funzioni test è infatti

$$C_1^2(\Omega) = \{v \in C^2(\overline{\Omega}) : v = 0 \text{ su } \Gamma_1\}.$$

Moltiplicando primo e secondo membro dell'equazione per v e integrando in Ω otteniamo l'equazione

$$-\int_\Omega \nabla \cdot (k\,\nabla u) v\,dxdy = \int_\Omega fv\,dxdy, \quad \text{qualunque sia } v \in C_1^2(\Omega).$$

Utilizzando quindi la prima identità di Green otteniamo l'equazione

$$\int_\Omega k\,\nabla u \cdot \nabla v\,dxdy - \int_{\partial\Omega} kv\,\frac{\partial u}{\partial n}\,d\sigma = \int_\Omega fv\,dxdy, \quad \text{per ogni } v \in C_1^2(\Omega).$$

Tenendo conto delle due condizioni al bordo, possiamo scrivere che

$$\int_{\partial\Omega} kv\,\frac{\partial u}{\partial n}\,d\sigma = \int_{\Gamma_1} kv\,\frac{\partial u}{\partial n}\,d\sigma + \int_{\Gamma_2} kv\,\frac{\partial u}{\partial n}\,d\sigma,$$

da cui segue che l'integrale al bordo è nullo, dato che il primo integrale è nullo perchè $v = 0$ su Γ_1 e il secondo lo è in quanto $k\,\frac{\partial u}{\partial n} = 0$ su Γ_2. Di conseguenza possiamo affermare che

$$\int_\Omega k\,\nabla u \cdot \nabla v\,dxdy = \int_\Omega fv\,dxdy, \quad \text{qualunque sia } v \in C_1^2(\Omega). \tag{8.28}$$

La (8.28) rappresenta la formulazione variazionale della (8.27).

Con un procedimento analogo ai due usati in precedenza, si può dimostrare che le due formulazioni sono equivalenti nel senso che la (8.27) implica la (8.28) e viceversa la validità della (8.28), qualunque sia $v \in C_1^2(\Omega)$, implica la validità della (8.27).

8.2 d Differenza fondamentale tra la formulazione classica e quella variazionale

Problema (8.8). La formulazione classica, nell'ipotesi che la f sia continua in $[a, b]$, richiede che la soluzione y sia due volte derivabile con continuità in $[a, b]$ ($y \in C^2[a, b]$). La sua formulazione variazionale (8.9) richiede invece la semplice derivabilità della y. Da notare anche che, mentre la (8.8) richiede la continuità della f, la (8.9) richiede semplicemente la sua integrabilità in $[a, b]$. Questo suggerisce di ampliare notevolmente lo spazio nel quale ricercare la soluzione del problema differenziale con valori agli estremi assegnati. Vedremo che tale spazio è lo spazio di Sobolev $H_0^1[a, b]$, definito nel seguito.

Problema (8.19). Il problema (8.19), nella sua formulazione classica, richiede che la f sia continua in $\overline{\Omega} = \Omega \cup \partial\Omega$ e che la u sia due volte differenziabile con continuità in $\overline{\Omega}$. La sua formulazione variazionale (8.21) richiede, invece, la semplice differenziabilità della u in Ω e la semplice integrabilità in Ω della f. L'interesse a risolvere il problema (8.19) nella sua formulazione variazionale deriva dalla possibilità che questa offre di ricercare la soluzione in spazi molto più ampi rispetto a $C_0^2(\overline{\Omega})$. Considerazioni del tutto analoghe valgono per i BVPs (8.23) e (8.27). È questa la ragione che ha portato all'introduzione degli spazi di Sobolev, che sono gli spazi nei quali si cerca la soluzione variazionale degli assgnati BVPs. Le loro soluzioni, per contrasto con quelle classiche, definite *forti*, sono definite *deboli*.

8.2 e Spazi di Sobolev

Per la loro introduzione è necessario premettere alcune definizioni. In primo luogo occorre precisare che per *supporto* di una funzione $u(x, y)$ intendiamo la chiusura dell'insieme dei punti di $\mathbb{R}^2$ nei quali la u è non nulla. In simboli

$$\text{supp}(u) = \overline{\{(x, y) \in \mathbb{R}^2 : u(x, y) \neq 0\}}.$$

Se u è definita su Ω e il supp(u) è un compatto (insieme chiuso e limitato), per brevità diciamo che u è a supporto compatto su Ω. Da notare che una funzione continua a supporto compatto su Ω è zero sulla frontiera di Ω e nelle sue immediate vicinanze. Lo spazio $C_0^\infty(\overline{\Omega})$, al quale spesso si ricorre nei metodi agli elementi finiti, è lo spazio delle funzioni indefinitamente differenziabili in Ω, aventi supporto compatto in Ω.

Esempio 8.7 (di una funzione in $C_0^\infty(\overline{\Omega})$) Indicati con $\boldsymbol{x}$ e $\boldsymbol{x}_0$ due punti di $\Omega \subset \mathbb{R}^n$, sia $r > 0$ un numero tale che la sfera di $\mathbb{R}^n$ con centro $\boldsymbol{x}_0$ e raggio

r sia interna ad Ω. In tale ipotesi la funzione

$$v(\boldsymbol{x}) = \begin{cases} e^{-1/(r^2 - \|\boldsymbol{x}-\boldsymbol{x}_0\|^2)}, & \text{per } \|\boldsymbol{x}-\boldsymbol{x}_0\| \le r, \\ 0, & \text{altrove in } \overline{\Omega}, \end{cases}$$

dove il simbolo $\|\cdot\|$ indica la norma Euclidea in $\mathbb{R}^n$, appartiene allo spazio $C_0^\infty(\overline{\Omega})$. Per rendersene conto basta osservare che v è indefinitivamente differenziabile e che $1/(r^2 - \|\boldsymbol{x}-\boldsymbol{x}_0\|^2) \to +\infty$ per $\|\boldsymbol{x}-\boldsymbol{x}_0\| \to r^-$.

Derivazione in senso debole (**nel senso delle distribuzioni**) Per semplicità, iniziamo con la *derivazione distribuzionale* (*in senso debole*) di una funzione definita in un intervallo $[a,b]$. A tal fine indichiamo con $C_0^\infty[a,b]$ l'insieme delle funzioni indefinitamente derivabili (in senso ordinario) in (a,b) e nulle in a e b. Questo implica che, per continuità, esistono un intorno destro di a e uno sinistro di b nei quali tali funzioni sono nulle. In sintesi, $C_0^\infty[a,b]$ è lo spazio delle funzioni indefinitamente derivabili in (a,b), aventi supporto compatto in $[a,b]$.

Per definire la derivazione, in senso debole (distribuzionale), della y in $[a,b]$, ricordiamo che se $y' \in C^1[a,b]$, per la regola di integrazione per parti, per ogni z sufficientemente regolare in $[a,b]$,

$$\int_a^b \frac{dy}{dx} z\, dx = [yz]_a^b - \int_a^b y \frac{dz}{dx}\, dx.$$

Nel caso $z \in C_0^\infty[a,b]$, essendo $z(a) = z(b) = 0$, risulta più semplicemente

$$\int_a^b \frac{dy}{dx} z\, dx = -\int_a^b y \frac{dz}{dx}\, dx.$$

Questo equivale ad affermare che $\frac{dy}{dx}$ è quella funzione u per cui

$$\int_a^b uz\, dx = -\int_a^b y \frac{dz}{dx}\, dx, \quad \text{qualunque sia } z \in C_0^\infty[a,b]. \tag{8.29}$$

Come è naturale aspettarsi, si può dimostrare che per ogni funzione $y \in C_0^1[a,b]$, esiste una e una sola funzione u che soddisfa la (8.29). Per questo motivo essa, per contrasto con la definizione di derivata ordinaria, viene definita derivata in senso debole (distribuzionale) della y.

Prima di passare alla introduzione della derivazione parziale in senso debole consideriamo due esempi.

Esempio 8.8 Consideriamo la seguente funzione continua in $[0,1]$, con un punto angoloso in $x = \frac{1}{2}$:

$$f(x) = \begin{cases} x, & 0 \le x < \frac{1}{2}, \\ 1-x, & \frac{1}{2} \le x < 1. \end{cases}$$

Tale funzione, anche se non è derivabile (in senso ordinario) in $x = \frac{1}{2}$, è derivabile in senso distribuzionale (debole) e la sua derivata è

$$f'(x) = \begin{cases} 1, & 0 \le x < \frac{1}{2}, \\ -1, & \frac{1}{2} < x < 1. \end{cases}$$

Per la sua dimostrazione basta osservare che, qualunque sia $g \in C_0^\infty[0,1]$,

$$\begin{aligned} \int_0^1 f(x)g'(x)\,dx &= \int_0^{\frac{1}{2}} xg'(x)\,dx + \int_{\frac{1}{2}}^1 (1-x)g'(x)\,dx \\ &= xg(x)|_0^{\frac{1}{2}} - \int_0^{\frac{1}{2}} g(x)\,dx + (1-x)g(x)|_{\frac{1}{2}}^1 + \int_{\frac{1}{2}}^1 g(x)\,dx \\ &= -\int_0^{\frac{1}{2}} g(x)\,dx + \int_{\frac{1}{2}}^1 g(x)\,dx = -\int_0^1 f'(x)g(x)\,dx. \end{aligned}$$

Esempio 8.9 Consideriamo ora la seguente funzione di Heaviside:

$$H(x) = \begin{cases} 0, & -1 \le x < 0, \\ 1, & 0 \le x \le 1. \end{cases}$$

La funzione H, anche se discontinua in $x = 0$, è al quadrato integrabile in $[-1,1]$, ossia $H \in L^2[-1,1]$. Poiché il suo supporto è $[0,1]$, la funzione w è la derivata (in senso debole) di H se

$$\int_{-1}^1 wv\,dx = -\int_{-1}^1 Hv'\,dx = -\int_0^1 v'\,dx = v(0), \text{ qualunque sia } v \in C_0^\infty[-1,1].$$

Questo implica che w è la restrizione della δ di Dirac su $[-1,1]$ in quanto, per definizione di δ di Dirac,

$$\int_{-1}^1 \delta(x)v(x)\,dx = v(0), \quad \text{qualunque sia } v \in C_0^\infty[-1,1].$$

Supponendo ora che $u \in C^1(\overline{\Omega})$, integrando per parti, come indicato nella (8.17), abbiamo che

$$\int_\Omega \frac{\partial u}{\partial x} v \, dxdy = \int_{\partial\Omega} uvn_1 \, d\sigma - \int_\Omega u \frac{\partial v}{\partial x} \, dxdy,$$

essendo n_1 la prima componente della normale $\boldsymbol{n}$, qualunque sia la funzione test v sufficientemente regolare. Nel caso $v \in C_0^\infty(\Omega)$ l'integrale su $\partial\Omega$ è nullo e pertanto

$$\int_\Omega \frac{\partial u}{\partial x} v \, dxdy = -\int_\Omega u \frac{\partial v}{\partial x} \, dxdy, \quad \text{qualunque sia } v \in C_0^\infty(\Omega).$$

Questo significa che $\frac{\partial u}{\partial x}$ è la funzione w per la quale

$$\int_\Omega wv \, dxdy = -\int_\Omega u \frac{\partial v}{\partial x} \, dxdy, \quad \text{qualunque sia } v \in C_0^\infty(\Omega). \tag{8.30a}$$

Come c'è da aspettarsi si può dimostrare che ad ogni $u \in C_0^1(\overline{\Omega})$, corrisponde una e una sola funzione w per la quale vale la (8.30a) per ogni $v \in C_0^\infty(\Omega)$. Per questo motivo, tale funzione viene definita la derivata parziale in senso debole (distribuzionale) della u. Utilizzando lo stesso procedimento si definisce la derivata in senso debole (distribuzionale) $\frac{\partial u}{\partial y}$ che, evidentemente, è la unica funzione z per la quale

$$\int_\Omega zv \, dxdy = -\int_\Omega u \frac{\partial v}{\partial y} \, dxdy, \quad \text{qualunque sia } v \in C_0^\infty(\Omega). \tag{8.30b}$$

Esempio 8.10 Indicato con Ω il quadrato, $\Omega = (0,1) \times (0,1)$, consideriamo la funzione

$$u(x,y) = \begin{cases} xy, & 0 < x < \frac{1}{2},\ 0 < y < 1, \\ (1-x)y, & \frac{1}{2} < x < 1,\ 0 < y < 1. \end{cases} \tag{8.31}$$

Tale funzione, ovviamente non derivabile in senso forte, lo è in senso debole. È facile infatti dimostrare che (in senso debole)

$$\frac{\partial u}{\partial x}(x,y) = w(x,y) = \begin{cases} y, & 0 < x < \frac{1}{2},\ 0 < y < 1, \\ -y, & \frac{1}{2} < x < 1,\ 0 < y < 1. \end{cases} \tag{8.32}$$

Per verificarlo osserviamo dapprima che, per ogni funzione test $v \in C_0^\infty(\Omega)$,

$$\begin{aligned} \int_\Omega wv \, dxdy &= \int_0^1 \int_0^1 w(x,y)v(x,y) \, dxdy \\ &= \int_0^1 y \int_0^{\frac{1}{2}} v(x,y) \, dxdy - \int_0^1 y \int_{\frac{1}{2}}^1 v(x,y) \, dxdy. \end{aligned}$$

D'altro canto

$$
\begin{aligned}
-\int_\Omega u\frac{\partial v}{\partial x}\,dxdy &= -\int_0^1\int_0^1 u(x,y)\frac{\partial v}{\partial x}(x,y)\,dxdy = -\int_0^1 y\int_0^{\frac12} x\frac{\partial v}{\partial x}\,dxdy \\
&\quad -\int_0^1 y\int_{\frac12}^1 (1-x)\frac{\partial v}{\partial x}\,dxdy = -\int_0^1 y\left\{[xv(x,y)]_0^{\frac12} - \int_0^{\frac12} v(x,y)\,dx\right\}dy \\
&\quad -\int_0^1 y\left\{[(1-x)v(x,y)]_{\frac12}^1 + \int_{\frac12}^1 v(x,y)\,dx\right\}dy = -\int_0^1 y\left\{\tfrac12 v(\tfrac12,y)\right. \\
&\quad \left.-\int_0^{\frac12} v(x,y)\,dx\right\}dy - \int_0^1 y\left\{-\tfrac12 v(\tfrac12,y) + \int_{\frac12}^1 v(x,y)\,dx\right\}dy \\
&\quad = \int_0^1 y\int_0^{\frac12} v(x,y)\,dxdy - \int_0^1 y\int_{\frac12}^1 v(x,y)\,dxdy = \int_\Omega w(x,y)v(x,y)\,dxdy.
\end{aligned}
$$

Risulta pertanto verificato che la funzione w definita nella (8.32) è la derivata $\frac{\partial u}{\partial x}$ in senso debole della (8.31).

Osservazione. Nelle forme variazionali considerate le uniche derivate parziali che compaiono sono del primo ordine. Per questo motivo, pur non essendoci difficoltà particolari ad estendere la definizione alle derivate parziali di ordine superiore in senso debole, ci limitiamo a quelle del primo ordine. Occorre comunque tenere presente che le formulazioni variazionali considerate richiedono che i prodotti $\nabla u \cdot \nabla v$ siano integrabili, ossia che siano integrabili in Ω i prodotti

$$\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} \quad \text{e} \quad \frac{\partial u}{\partial y}\frac{\partial v}{\partial y}.$$

8.3 Proprietà basilari degli spazi di Sobolev

Gli spazi di Sobolev sono gli spazi di Hilbert nei quali vengono risolte le equazioni differenziali (ordinarie e alle derivate parziali) mediante il metodo degli elementi finiti. Per la loro introduzione è preliminare osservare che gli integrali utilizzati sono definiti nel senso di Lebesgue, non nel senso di Riemann. Questi ultimi, essendo più semplici da definire, vengono introdotti nei corsi introduttivi di Analisi Matematica. Essi non vengono però utilizzati negli elementi finiti per varie ragioni. La più importante è questa: una successione di funzioni integrabili nel senso di Riemann non sempre converge a una funzione integrabile nel senso di Riemann. È importante comunque tener presente che ogni funzione integrabile nel senso di Riemann (in un dominio limitato)

lo è anche nel senso di Lebesgue e i due integrali coincidono. L'insieme delle funzioni integrabili nel senso di Lebesgue è più ampio, rispetto a quello delle funzioni integrabili nel senso di Riemann, come richiesto nella teoria degli elementi finiti. Per questo motivo, negli elementi finiti, si definiscono misurabili le funzioni in valore assoluto integrabili nel senso di Lebesgue. Gli integrali di Lebesgue sono basati sulla misura degli insiemi, nel senso che la misura di Lebesgue indica l'area di un insieme che può essere anche parecchio complesso. Per approfondimenti sulle loro proprietà si rinvia al libro di Royden [25].

Lo spazio di Sobolev più utilizzato è

$$L^2(\Omega) = \left\{ f : \Omega \to \mathbb{R} : \int_\Omega |f|^2 \, d\boldsymbol{x} < \infty \right\}. \tag{8.33}$$

dove l'integrale è nel senso di Lebesgue. Dalla (8.33) segue immediatamente che due funzioni f e g in $L^2(\Omega)$ per le quali

$$\int_\Omega |f - g|^2 \, d\boldsymbol{x} = 0$$

sono da considerarsi uguali, in quanto differiscono tra loro unicamente in un sottoinsieme di Ω di misura nulla. Da questo deriva che in $L^2(\Omega)$: (a) una funzione può restare indefinita in un insieme di misura nulla; (b) non può essere definita unicamente in un insieme di misura nulla. Di conseguenza non può essere definita univocamente nella frontiera $\partial\Omega$ di un insieme limitato $\Omega \subset \mathbb{R}^n$, dato che essa è di misura nulla.

Completezza di $L^2(\Omega)$. Lo spazio $C(\overline{\Omega})$, ossia lo spazio delle funzioni continue nella chiusura $\overline{\Omega} = \Omega \cup \partial\Omega$ di Ω è denso in $L^2(\Omega)$, ma non è completo. Questo significa che: (a) qualunque funzione di $L^2(\Omega)$ può essere approssimata, in norma $L^2(\Omega)$, con la precisione che si vuole, con una funzione continua in $\overline{\Omega}$; (b) una successione di funzioni di $C(\overline{\Omega})$, di Cauchy rispetto alla norma di $L^2(\Omega)$, non necessariamente converge a una funzione di $C(\overline{\Omega})$. Si dimostra invece che ogni successione di funzioni di $C(\overline{\Omega})$ di Cauchy, rispetto alla norma di $L^2(\Omega)$ converge ad una funzione di $L^2(\Omega)$. Per questo motivo si dice che $L^2(\Omega)$ è il completamento di $C(\overline{\Omega})$, rispetto alla norma di $L^2(\Omega)$. Esso è uno spazio di Hilbert nel quale il prodotto interno, relativo a una qualsiasi coppia di funzioni $f, g \in L^2(\Omega)$, è così definito:

$$< f, g >_{L^2(\Omega)} = \int_\Omega f g \, d\boldsymbol{x}.$$

La norma di una qualsiasi $f \in L^2(\Omega)$ è pertanto

$$\|f\|_{L^2(\Omega)} = \sqrt{\int_\Omega |f|^2 \, d\boldsymbol{x}}.$$

Spazi di Sobolev $H^1(\Omega)$ e $H_0^1(\Omega)$. Sia $u \in L^2(\Omega)$. Indicate con $\frac{\partial u}{\partial x}$ e $\frac{\partial u}{\partial y}$ le derivate parziali (in senso debole) della u in Ω, lo spazio $H^1(\Omega)$ è così definito:

$$H^1(\Omega) = \left\{ u \in L^2(\Omega) : \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y} \in L^2(\Omega) \right\}.$$

Esso rappresenta uno spazio di Hilbert rispetto al prodotto interno

$$< u, v >_{H^1(\Omega)} = \int_\Omega \left(uv + \frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y} \right) dxdy = \int_\Omega \left(uv + \nabla u \cdot \nabla v \right) dxdy,$$

dove u e v sono due generiche funzioni di $H^1(\Omega)$ le cui derivate parziali sono definite in senso debole. Di conseguenza

$$\|u\|^2_{H^1(\Omega)} = \int_\Omega \left(|u|^2 + \left|\frac{\partial u}{\partial x}\right|^2 + \left|\frac{\partial u}{\partial y}\right|^2 \right) dxdy = \int_\Omega \left(|u|^2 + \|\nabla u\|^2 \right) dxdy.$$

Per evidenziare una importante proprietà di $H^1(\Omega)$, è bene far riferimento allo spazio

$$C^1(\overline{\Omega}) = \left\{ f \in C^0(\overline{\Omega}) : \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y} \in C^0(\overline{\Omega}) \right\}.$$

Tra $H^1(\Omega)$ e $C^1(\overline{\Omega})$ esiste lo stesso tipo di relazione esistente tra $H^0(\Omega)$ e $C^0(\overline{\Omega})$. Si può dimostrare infatti che:

> $C^1(\overline{\Omega})$ è un sottospazio denso di $H^1(\Omega)$. Per questo motivo $H^1(\Omega)$ viene considerato il completamento di $C^1(\overline{\Omega})$ rispetto alla norma di $H^1(\Omega)$.

Con riferimento agli elementi finiti è importante notare che, indicato con $C_0^\infty(\Omega)$ l'insieme delle funzioni con supporto limitato e ivi indefinitamente derivabili, lo spazio $C_0^\infty(\Omega)$ è denso in $L^2(\Omega)$, rispetto alla norma di $L^2(\Omega)$.

Altro spazio di Sobolev, fondamentale negli elementi finiti, è lo spazio

$$H_0^1(\Omega) = \left\{ u \in H^1(\Omega) : u|_{\partial\Omega} = 0 \right\}.$$

Indicato con

$$C_0^1(\overline{\Omega}) = \left\{ u \in C^1(\overline{\Omega}) : u|_{\partial\Omega} = 0 \right\},$$

si può dimostrare che $C_0^1(\overline{\Omega})$ è denso in $H_0^1(\Omega)$. Lo spazio $H_0^1(\Omega)$ può, pertanto, essere considerato come il completamento di $C_0^1(\overline{\Omega})$ rispetto alla norma di $H^1(\Omega)$.

Per i BVPs di tipo ellittico considerati, con condizioni di omogeneità al bordo ($u|_{\partial\Omega} = 0$). $H_0^1(\Omega)$ è lo spazio di Sobolev di riferimento. Nel caso di non omogeneità al bordo, lo spazio di riferimento è $H^1(\Omega)$.

8.4 Risoluzione di BVPs con il metodo degli elementi finiti

Iniziamo con la risoluzione della forma variazionale (8.9) del problema (8.8), nell'ipotesi $u(a) = u(b) = 0$. In questo caso, essendo il problema omogeneo, relativamente alle condizioni agli estremi, lo spazio di Sobolev di riferimento è $H_0^1[a, b]$. Il primo passo consiste nel costruire una successione di spazi $\{X_n\}_{n=1}^{\infty}$ di $H_0^1[a, b]$, con

$$X_n \subset H_{n+1}, \qquad \bigcup_{n=1}^{\infty} X_n = X,$$

essendo X un sottospazio denso di $H_0^1[a, b]$ per la norma di $H_0^1[a, b]$.

A tale scopo si decompone l'intervallo $[a, b]$ con punti nodali equidistanziati (per semplice comodità)

$$x_i = a + ih, \quad i = 0, 1, \ldots, n + 1, \qquad h = \frac{b - a}{n + 1}.$$

Si introduce quindi uno spazio n-dimensionale X_n, tipicamente rappresentato da n polinomi a tratti opportunamente raccordati (spline polinomiali). Nel caso più semplice, che è anche quello più usato, tale base è formata da n splines lineari

$$\{H_i(x)\}_{i=1}^{n},$$

ciascuna delle quali, per soddisfare le condizioni di omgeneità agli estremi, richieste della loro appartenenza a $H_0^1[a, b]$, è nulla in $x_0 = a$ e $x_{n+1} = b$. Il loro tipico andamento è quello rappresentato nella Fig. 8.3. Analiticamente la

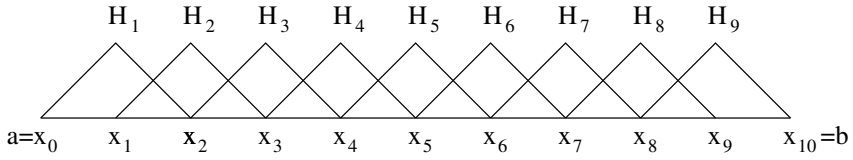


Figura 8.3: Esempio di spline lineari per $n = 9$.

i-esima spline H_i, $i = 1, 2, \ldots, n$, il cui supporto è l'intervallo $[x_{i-1}, x_{i+1}]$, di ampiezza $2h$, è così definita

$$H_i(x) = \begin{cases} 0, & a \le x \le x_{i-1}, \\ \dfrac{x - x_{i-1}}{h}, & x_{i-1} \le x \le x_i, \\ \dfrac{x_{i+1} - x}{h}, & x_i \le x \le x_{i+1}, \\ 0, & x_{i+1} \le x \le b. \end{cases} \tag{8.34}$$

L'approssimazione n-dimensional $y_n(x)$ di una funzione $y \in H_0^1[a,b]$ è data dall'interpolante

$$y_n(x) = \sum_{j=1}^{n} y_n(x_j) H_j(x), \tag{8.35}$$

nella quale $y_n(x_i)$ è il coefficiente di $H_i(x)$, dato che $H_i(x_j) = \delta_{ij}$ (1 per $j = i$ e 0 per $j \neq i$), e $y_n(a) = y_n(b) = 0$. Lo spazio delle approssimanti (8.35), al variare di $y \in C[a,b]$, è indicato con $S_0^1 \subset H_0^1[a,b]$. Ogni spline $H_i(x)$, pur avendo un punto angoloso in x_i, è derivabile in senso debole, e la sua derivata è:

$$H_i'(x) = \begin{cases} 0, & a \leq x < x_{i-1}, \\ \dfrac{1}{h}, & x_{i-1} < x < x_i, \\ -\dfrac{1}{h}, & x_i < x < x_{i+1}, \\ 0, & x_{i+1} < x \leq b. \end{cases} \tag{8.36}$$

Sostituendo nella (8.9) la y con la y_n e la funzione test v con H_i, $i = 1, 2, \ldots, n$, si ottiene il sistema

$$\sum_{j=1}^{n} \left\{ \int_a^b p(x) H_j'(x) H_i'(x)\, dx + \int_a^b q(x) H_j(x) H_i(x)\, dx \right\} y_{nj} = \int_a^b f(x) H_i(x)\, dx, \tag{8.37}$$

dove $y_{nj} = y_n(x_j)$ $(i = 1, 2, \ldots, n)$. In tale sistema il vettore

$$\boldsymbol{y}_n = (y_{n1}, y_{n2}, \ldots, y_{nn})^T$$

fornisce i valori nodali dell'approssimante y_n della soluzione y. La sua valutazione permette di ottenere, mediante la (8.35), una approssimazione della soluzione y del problema iniziale (8.8).

Esprimendo il sistema (8.37) nella forma matriciale

$$A\, \boldsymbol{y}_n = \boldsymbol{b}_n, \tag{8.38}$$

è immediato osservare che la matrice A è simmetrica e tridiagonale, dato che

$$a_{ij} = a_{ji} \quad \text{e} \quad a_{ij} = 0 \quad \text{per} \quad |i - j| \geq 2,$$

essendo $[x_{i-1}, x_{i+1}]$ il supporto di H_i e $[x_{j-1}, x_{j+1}]$ quello di H_j. Il motivo per cui le splines lineari sono molto utilizzate negli elementi finiti è fondamentalmente dovuto a tali caratteristiche (simmetria e sparsità della matrice). Grazie

al supporto minimo delle splines lineari considerate, risulta

$$
\begin{aligned}
a_{ij} &= 0 \quad \text{per} \quad |i-j| \geq 2, \\
a_{ii} &= \frac{1}{h^2}\left[\int_{x_{i-1}}^{x_{i+1}} p(x)\,dx + \int_{x_{i-1}}^{x_i} q(x)(x-x_{i-1})^2\,dx + \int_{x_i}^{x_{i+1}} q(x)(x_{i+1}-x)^2\,dx\right], \\
a_{i,i-1} &= a_{i-1.i} = -\frac{1}{h^2}\left[\int_{x_{i-1}}^{x_i} p(x)\,dx + \int_{x_{i-1}}^{x_i} q(x)(x-x_i)(x-x_{i-1})\,dx\right], \\
a_{i,i+1} &= a_{i+1,i} = -\frac{1}{h^2}\left[\int_{x_i}^{x_{i+1}} p(x)\,dx + \int_{x_i}^{x_{i+1}} q(x)(x-x_i)(x-x_{i+1})\,dx\right], \\
b_{ni} &= \frac{1}{h}\left[\int_{x_{i-1}}^{x_i} f(x)(x-x_{i-1})\,dx + \int_{x_i}^{x_{i+1}} f(x)(x_{i+1}-x)\,dx\right],
\end{aligned}
$$

con $i, j = 1, 2, \ldots, n$. È immediato osservare che per $q(x) \equiv 1$ la matrice A è anche diagonalmente dominante. Il calcolo degli integrali è tecnicamente possibile solo in casi particolari. Esso è chiaramente immediato nel caso $p(x) = q(x) = 1$. Nel caso generale si ricorre a formule di quadratura, scelte in funzione delle caratteristiche analitiche di $p(x)$ e $q(x)$ e, per quanto riguarda $\boldsymbol{b}_n$, di quelle di $f(x)$.

Stima dell'errore. Stime dell'errore di approssimazione della soluzione, con il metodo descritto, si possono facilmente ottenere nell'ipotesi che la soluzione y sia abbastanza regolare, ossia che

$$
y \in H^2[a,b] = \left\{ y \in L^2[a,b] : y', y'' \text{ anch'esse in } L^2[a,b] \right\},
$$

essendo y' e y'' derivate in senso debole. Ipotesi certamente verificate nel caso che la y sia la soluzione in senso ordinario della (8.8). Tali stime discendono dal Lemma di Céa e dalle proprietà di approssimazione di una funzione "abbastanza regolare" mediante le splines lineari [1]. Più precisamente, indicata con il simbolo $|\cdot|_{H^2[a,b]}$ la seminorma in $H^2[a,b]$, così definita

$$
|y|^2_{H^2[a,b]} = \int_a^b |y''(x)|^2\,dx,
$$

si dimostra la validità delle seguenti stime dell'errore di approssimazione (della soluzione y mediante la spline lineare y_n):

$$
\|y - y_n\|_{L^2[a,b]} \leq c h^2 |y|_{H^2[a,b]}, \tag{8.39a}
$$

$$
\|y - y_n\|_{H^1[a,b]} \leq \hat{c} h |y|_{H^2[a,b]}, \tag{8.39b}
$$

dove c e $\hat{c}$ sono costanti indipendenti da $h = \frac{b-a}{n+1}$. Tali stime evidenziano la convergenza di y_n a y per $n \to +\infty$ ($h \to 0^+$). Aumentare n significa incrementare l'ordine della matrice del sistema e conseguentemente la complessità

computazionale per la risoluzione del sistema (8.38). In questo caso, fortunatamente, l'incremento della complessità è molto limitato, in quanto la matrice del sistema è tridiagonale. Esistono infatti metodi appositi che risolvono tali tipi di sistemi per i quali la complessità è di $O(n)$, ossia la complessità cresce linearmente con n.

Osservazione. Relativamente alla completezza di importanti spazi di Hilbert, valgono le seguenti affermazioni:

(1) $C(\overline{\Omega})$ è denso in $L^2(\Omega) \iff L^2(\Omega)$ è il completamento di $C(\overline{\Omega})$ nella norma di $L^2(\Omega)$;

(2) $C^1(\overline{\Omega})$ è denso in $H^1(\Omega) \iff H^1(\Omega)$ è il completamento di $C^1(\overline{\Omega})$ nella norma di $H^1(\Omega)$;

(3) $C_0^1(\overline{\Omega})$ è denso in $H_0^1(\Omega) \iff H_0^1(\Omega)$ è il completamento di $C_0^1(\overline{\Omega})$ nella norma di $H^1(\Omega)$.

Altri spazi densi in $L^2(\Omega)$ sono $C_0^\infty(\Omega)$ e $C_0^\infty(\Omega) \cap L^2(\Omega)$. Lo spazio $C^\infty(\Omega) \cap H^1(\Omega)$ è denso in $H^1(\Omega)$ e lo spazio $C_0^\infty(\Omega)$ è denso in $H_0^1(\Omega)$.

Consideriamo ora la risoluzione del BVP di tipo ellittico (8.19), la cui formulazione variazionale è data dalla (8.21). Per descrivere l'analogo di quanto visto nel caso di una ODE, è necessario premettere qualche definizione e qualche considerazione. La prima operazione consiste nell'associare al dominio Ω un reticolo (mesh) formato da un insieme di triangoli come indicato nella Fig. 8.4. Come si vede nella Fig. 8.4, due triangoli possono avere in comune

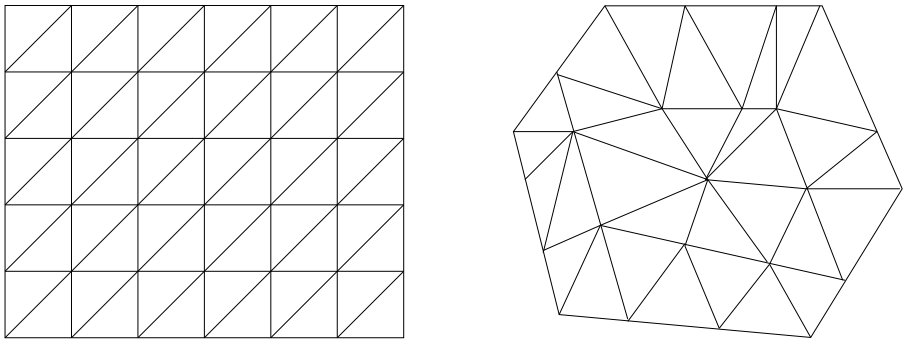


Figura 8.4: Reticolazione di due domini.

un lato, oppure una vertice. È evidente che, indicato con h il diametro dei triangoli considerati, il reticolo associato Ω_h deve riempire il più possibile Ω e, per $h \to 0^+$, $\Omega_h \to \Omega$. Per l'applicazione del metodo, i triangoli e i vertici

debbono essere numerati. Supponendo che la triangolazione abbia generato N_t triangoli, possiamo indicarli con

$$T_1, T_2, \ldots, T_{N_t}.$$

I vertici dei triangoli li indichiamo con

$$V_1, V_2, \ldots, V_{N_v},$$

dove, indicato con (a_i, y_j) le coordinate locali del j-esimo vertice, $z_j = (x_j, y_j)$, $j = 1, 2, \ldots, N_v$. I vertici del triangolo T_i, identificati con gli indici $x_{i,1}$, $x_{i,2}$ e $x_{i,3}$, vengono pertanto indicati con

$$V_{n_i,1}, \quad V_{n_i,2} \quad \text{e} \quad V_{n_i,3}.$$

È altresì importante identificare una trasformazione $n_{ij} = n(i, j)$ che, a ciascuno dei tre vertici del triangolo T_i, associ il valore che gli corrisponde nella numerazione globale $\{n_{ij}\}$. Nel caso rettangolare $\Omega = [a, b] \times [c, d]$, introdotte le coordinate locali

$$x_i = a + ih, \quad i = 0, 1, \ldots, n+1 \quad \text{e} \quad y_j = c + jk, \quad j = 0, 1, \ldots, m,$$

la numerazione globale (sulla base dell'ordinamento lessicografico) è ottenuta (Fig. 8.5) con la regola

$$n_{ij} = i + 1 + (n+1)j, \qquad i = 0, 1, \ldots, n+1 \quad \text{e} \quad j = 0, 1, \ldots, n+1.$$

Altro passo importante è rappresentato dalla identificazione dei nodi liberi ossia dei nodi (della triangolazione del dominio) nei quali il valore della soluzione del problema è incognita. Nel caso del problema ellittico (8.19), essendo prefissati i valori della soluzione al bordo ($u(x, y) = 0$ per $(x, y) \in \partial\Omega$), i nodi liberi coincidono con i nodi interni. Indicato con N_v il numero complessivo det vertici dei triangoli di Ω_h, indichiamo con n_v il numero di vertici interni che, per quanto osservato, coincide con il numero dei nodi liberi. Per la risoluzione (con gli elementi finiti) del problema variazionale (8.21), si inizia con l'associare ad ogni punto nodale libero (vertice interno) z_j, una box-spline ϕ_j che assume il valore 1 in z_j e zero in ogni altro punto nodale (sia interno sia di frontiera) della mesh Ω_h. Come notato nella Fig. 8.6, la ϕ_j rappresenta una piramide a facce piane e base poligonale (esagonale nel caso specifico) che assume il valore 1 nel punto z_j, relativo al vertice della piramide e zero in tutti gli altrinodi della mesh. La soluzione del problema variazionale (8.21), relativo alla reticolazione Ω_h di Ω, viene quindi approssimata con la seguente combinazione lineare delle box-splines $\{\phi_j\}$:

$$u_h(x, y) = \sum_{j=1}^{n_v} u_{h,j}\phi_j(x, y), \tag{8.40}$$

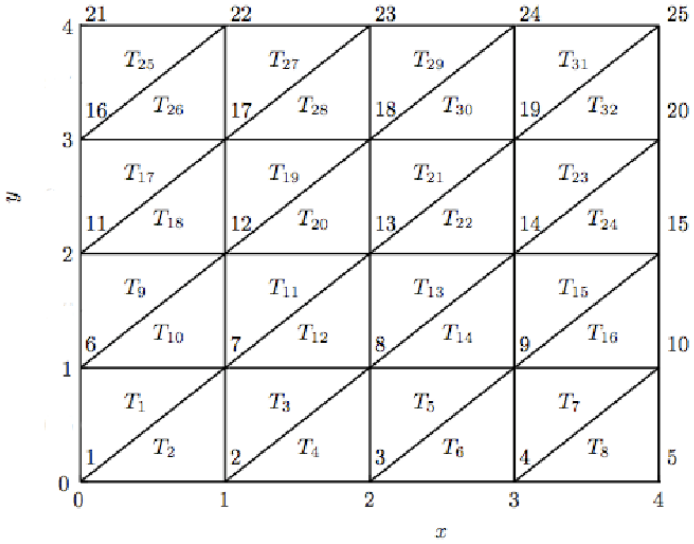


Figura 8.5: La figura rappresenta un esempio di mesh. Il dominio scelto è un quadrato, la mesh è regolare e rappresentata da 32 triangoli e 25 nodi. Dei 25 nodi, 9 sono interni e 16 sono di frontiera.

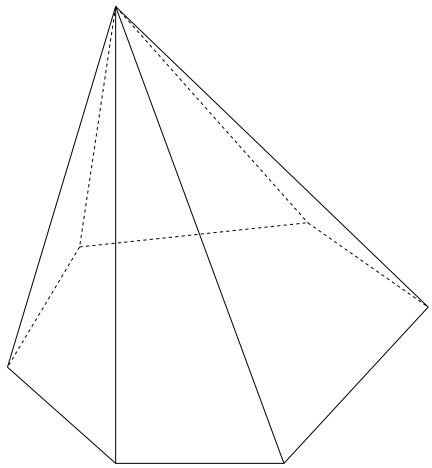


Figura 8.6: Rappresentazione grafica di $\phi_j(x, y)$.

dove, per la cardinalità delle box-splines, $u_{h,j}$ rappresenta il valore della u_h nel vertice z_j, ossia $u_{h,j} = u_h(z_j)$, $z_j = (x_j, y_j)$, $j = 1, 2, \ldots, n_v$, essendo n_v il numero dei vertici liberi. Poichè $\phi_j \in H_0^1(\Omega_h)$, come è immediato osservare, anche $u_h \in H_0^1(\Omega_h)$. Questo implica che u_h soddisfa le condizioni al bordo, a prescindere dei valori dei coefficienti $\{u_{h,j}\}_{j=1}^{n_v}$. Assunto, pertanto, $H_0^1(\Omega_h)$ come spazio delle funzioni test, assumiamo $\phi_i(x, y)$, $i = 1, 2, \ldots, n_v$, come i-esima funzione test. Di conseguenza il calcolo del vettore dei coefficienti $\{u_{h,j}\}$

della approssimante u_h, viene ricondotto alla risoluzione del sistema

$$\int_{\Omega_h} k(x,y)\nabla u_h \cdot \nabla \phi_i \, dxdy = \int_{\Omega_h} f(x,y)\phi_i(x,y)\, dxdy, \qquad i = 1,2,\ldots,n_v. \tag{8.41}$$

Sostituendo quindi la rappresentazione di u_h nella (8.41), otteniamo il sistema determinato

$$\sum_{j=1}^{n_v} u_{h,j} \int_{\Omega_h} k(x,y)\nabla\phi_j \cdot \nabla\phi_i \, dxdy = \int_{\Omega_h} f(x,y)\phi_i(x,y)\, dxdy, \quad i = 1,2,\ldots,n_v, \tag{8.42}$$

dove le derivate parziali di ϕ_i e di ϕ_j sono, ovviamente, da intendersi in senso debole. Il sistema (8.42), in forma matriciale, è dunque del tipo

$$\boldsymbol{K}_h \boldsymbol{u}_h = \boldsymbol{f}_h, \tag{8.43}$$

dove $\boldsymbol{K}_h$ è la cosiddetta stiffness matrix e $\boldsymbol{f}_h$ il load vector. Da quanto detto è evidente che

$$(\boldsymbol{K}_h)_{ij} = \int_{\Omega_h} k(x,y)\nabla\phi_i \cdot \nabla\phi_j \, dxdy, \qquad i,j = 1,2,\ldots,n_v,$$
$$f_{n,i} = \int_{\Omega_h} f(x,y)\phi_i(x,y)\, dxdy, \qquad i = 1,2,\ldots,n_v.$$

La matrice $\boldsymbol{K}_h$ è chiaramente sparsa, dato che $(\boldsymbol{K}_h)_{ij} \neq 0$ solo nel caso in cui i supporti delle box-splines ϕ_i e ϕ_j abbiano almeno un triangolo in comune.

Il calcolo degli elementi della matrice $\boldsymbol{K}_h$ e del vettore f_h può essere effettuato analiticamente solo se $k(x,y)$ e $f(x,y)$ sono molto semplici. Nella generalità dei casi deve essere effettuato numericamente utilizzando formule di risoluzione numerica che tengono conto delle particolarità di k e f in Ω. Il loro calcolo è, in generale, molto complesso in conseguenza dell'elevato numero di triangoli che compaiono in Ω_h, ma anche delle loro molteplici configurazioni rispetto alle coordinate. Questa seconda difficoltà, come vedremo nel seguito, può essere superata con l'introduzione del cosiddetto triangolo di riferimento.

8.4 a Calcolo delle funzioni di base $\phi_l(x,y)$

Per trovare la rappresentazione analitica di ogni elemento della base, occorre tenere presente che, nella frontiera del dominio Ω, gli elementi della base si devono annullare e che in ogni triangolo T_i della mesh (facente parte del supporto di una funzione di base), il piano che la rappresenta ha valore uno nel comune vertice e zero nei restanti vertici (Fig. 8.7). Pertanto è sempre possibile trovare

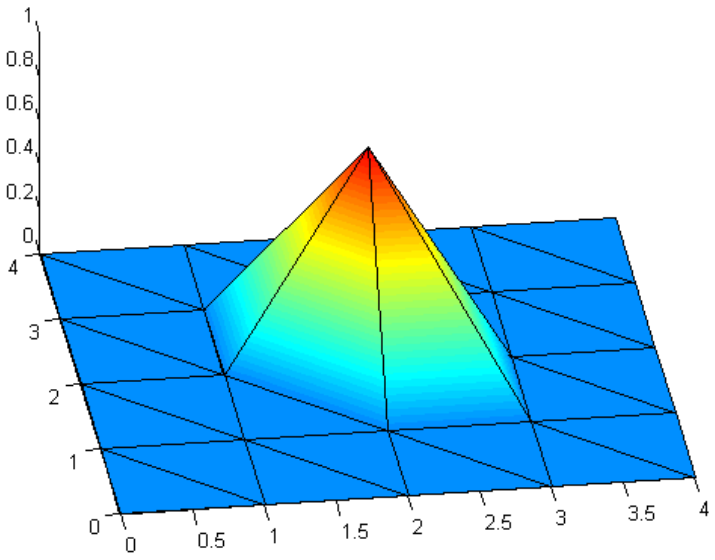


Figura 8.7: La figura rappresenta un esempio di box-spline lineare in un dominio esagonale. Da notare, in particolare, come l'unione dei sei piani aventi valore uno nel vertice comune, formano una piramide a base esagonale.

l'equazione di ogni generico piano su un generico triangolo, visto che per tre punti non allineati dello spazio passa uno e un solo piano.

Calcolo di $\phi_1(x, y)$. Gli estremi del triangolo T_1 (Fig. 8.5) sono dati dai punti

$$\begin{cases} x_1 = 0, \\ y_1 = 0, \end{cases} \qquad \begin{cases} x_7 = h, \\ y_7 = h, \end{cases} \qquad \begin{cases} x_6 = 0, \\ y_6 = h, \end{cases}$$

per cui il piano $z(x, y) = a + bx + cy$ che la caratterizza soddisfa le condizioni:

$$\begin{cases} z(x_1, y_1) = 0, \\ z(x_7, y_7) = 1, \\ z(x_6, y_6) = 0, \end{cases} \implies \begin{cases} a = 0, \\ a + hb + hc = 1, \\ a + hc = 0, \end{cases} \implies \begin{cases} a = 0, \\ b = \dfrac{1}{h}, \\ c = 0. \end{cases}$$

Di conseguenza l'equazione del piano cercato è $z = \frac{x}{h}$.

Piano con supporto il triangolo T_2:

$$\begin{cases} x_1 = 0, \\ y_1 = 0, \end{cases} \qquad \begin{cases} x_2 = h, \\ y_2 = 0, \end{cases} \qquad \begin{cases} x_7 = h, \\ y_7 = h, \end{cases}$$

$$\begin{cases} z(x_1,y_1)=0, \\ z(x_2,y_2)=0, \\ z(x_7,y_7)=1, \end{cases} \Rightarrow \begin{cases} z(0,0)=0, \\ z(h,0)=0, \\ z(h,h)=1, \end{cases} \Rightarrow \begin{cases} a=0, \\ a+hb=0, \\ a+hb+hc=1, \end{cases} \Rightarrow \begin{cases} a=0, \\ b=0, \\ c=\dfrac{1}{h}. \end{cases}$$

Di conseguenza l'equazione del piano cercato è $z=\frac{y}{h}$.

Piano con supporto il triangolo T_3:

$$\begin{cases} x_2=h, \\ y_2=0, \end{cases} \qquad \begin{cases} x_8=2h, \\ y_8=h, \end{cases} \qquad \begin{cases} x_7=h, \\ y_7=h, \end{cases}$$

$$\begin{cases} z(x_2,y_2)=0, \\ z(x_8,y_8)=0, \\ z(x_7,y_7)=1, \end{cases} \Rightarrow \begin{cases} z(h,0)=0, \\ z(2h,h)=0, \\ z(h,h)=1, \end{cases} \Rightarrow \begin{cases} a+hb=0, \\ a+2hb+hc=0, \\ a+hb+hc=1, \end{cases}$$

$$\begin{pmatrix} 1 & h & 0 \\ 1 & 2h & h \\ 1 & h & h \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \Rightarrow \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 \\ 0 & \frac{1}{h} & -\frac{1}{h} \\ -\frac{1}{h} & 0 & \frac{1}{h} \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Di conseguenza l'equazione del piano cercato è $z=1-\frac{x}{h}+\frac{y}{h}$.

Piano con supporto il triangolo T_{12}:

$$\begin{cases} x_7=h, \\ y_7=h, \end{cases} \qquad \begin{cases} x_8=2h, \\ y_8=h, \end{cases} \qquad \begin{cases} x_{13}=2h, \\ y_{13}=2h, \end{cases}$$

$$\begin{cases} z(x_7,y_7)=1, \\ z(x_8,y_8)=0, \\ z(x_{13},y_{13})=0, \end{cases} \Rightarrow \begin{cases} z(h,h)=1, \\ z(2h,h)=0, \\ z(2h,2h)=0, \end{cases} \Rightarrow \begin{cases} a+hb+hc=1, \\ a+2hb+hc=0, \\ a+2hb+2hc=0, \end{cases}$$

$$\begin{pmatrix} 1 & h & h \\ 1 & 2h & h \\ 1 & 2h & 2h \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \Rightarrow \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 2 & 0 & -1 \\ -\frac{1}{h} & \frac{1}{h} & 0 \\ 0 & -\frac{1}{h} & \frac{1}{h} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

Di conseguenza l'equazione del piano cercato è $z=2-\frac{x}{h}$.

Piano con supporto il triangolo T_{11}:

$$\begin{cases} x_7=h, \\ y_7=h, \end{cases} \qquad \begin{cases} x_{13}=2h, \\ y_{13}=2h, \end{cases} \qquad \begin{cases} x_{12}=h, \\ y_{12}=2h, \end{cases}$$

$$\begin{cases} z(x_7,y_7)=1, \\ z(x_{13},y_{13})=0, \\ z(x_{12},y_{12})=0, \end{cases} \Rightarrow \begin{cases} z(h,h)=1, \\ z(2h,2h)=0, \\ z(h,2h)=0, \end{cases} \Rightarrow \begin{cases} a+hb+hc=1, \\ a+2hb+2hc=0, \\ a+hb+2hc=0, \end{cases}$$

$$\begin{pmatrix} 1 & h & h \\ 1 & 2h & 2h \\ 1 & h & 2h \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \Rightarrow \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 2 & -1 & 0 \\ 0 & \frac{1}{h} & -\frac{1}{h} \\ -\frac{1}{h} & 0 & \frac{1}{h} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

Di conseguenza l'equazione del piano cercato è $z = 2 - \frac{y}{h}$.

Piano con supporto il triangolo T_{10}:

$$\begin{cases} x_6 = 0, \\ y_6 = h, \end{cases} \qquad \begin{cases} x_7 = h, \\ y_7 = h, \end{cases} \qquad \begin{cases} x_{12} = h, \\ y_{12} = 2h, \end{cases}$$

$$\begin{cases} z(x_6, y_6) = 0, \\ z(x_7, y_7) = 1, \\ z(x_{12}, y_{12}) = 0, \end{cases} \Rightarrow \begin{cases} z(0, h) = 0, \\ z(h, h) = 1, \\ z(h, 2h) = 0, \end{cases} \Rightarrow \begin{cases} a + hc = 0, \\ a + hb + hc = 1, \\ a + hb + 2hc = 0, \end{cases}$$

$$\begin{pmatrix} 1 & 0 & h \\ 1 & h & h \\ 1 & h & 2h \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \Rightarrow \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 1 & 1 & -1 \\ -\frac{1}{h} & \frac{1}{h} & 0 \\ 0 & -\frac{1}{h} & \frac{1}{h} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}.$$

Di conseguenza l'equazione del piano cercato è $z = 1 + \frac{x}{h} - \frac{y}{h}$. Pertanto la

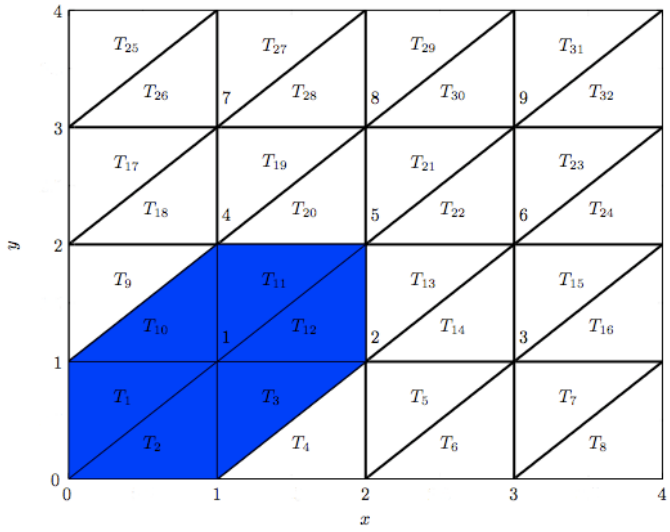


Figura 8.8: La figura rappresenta la mesh nella quale sono stati rappresentati solo i nodi interni e dove il supporto della funzione base $\phi_1(x, y)$ è evidenziato in blu.

funzione di base $\phi_1(x,y)$ è data da (Fig. 8.8):

$$\phi_1(x,y) = \begin{cases} \frac{x}{h}, & (x,y) \in T_1, \\ \frac{y}{h}, & (x,y) \in T_2, \\ 1 - \frac{x}{h} + \frac{y}{h}, & (x,y) \in T_3, \\ 2 - \frac{x}{h}, & (x,y) \in T_{12}, \\ 2 - \frac{y}{h}, & (x,y) \in T_{11}, \\ 1 + \frac{x}{h} - \frac{y}{h}, & (x,y) \in T_{10}. \end{cases} \tag{8.44}$$

Calcolo di $\phi_2(x,y)$.

- *Piano con supporto il triangolo* T_3. Tale piano ha la stessa inclinazione del piano avente come supporto il triangolo T_1 e assume valore uno in $(2h,h)$. Pertanto la sua equazione è data da (si veda la (8.44)) $z = \frac{x}{h} - 1$.
- *Piano con supporto il triangolo* T_4. Tale piano ha la stessa inclinazione del piano avente come supporto il triangolo T_2 e assume valore uno in $(2h,h)$. Pertanto la sua equazione è data da (si veda la (8.44)) $z = yh$.
- *Piano con supporto il triangolo* T_5. Tale piano ha la stessa inclinazione del piano avente come supporto il triangolo T_3 e assume valore uno in $(2h,h)$. Pertanto la sua equazione è data da (si veda la (8.44)) $z = 2 - \frac{x}{h} + \frac{y}{h}$.
- *Piano con supporto il triangolo* T_{14}. Tale piano ha la stessa inclinazione del piano avente come supporto il triangolo T_{12} e assume valore uno in $(2h,h)$. Pertanto la sua equazione è data da (si veda la (8.44)) $z = 3 - \frac{x}{h}$.
- *Piano con supporto il triangolo* T_{13}. Tale piano ha la stessa inclinazione del piano avente come supporto il triangolo T_{11} e assume valore uno in $(2h,h)$ e pertanto (si veda la (8.44)) si ha: $z = 2 - \frac{y}{h}$.
- *Piano con supporto il triangolo* T_{12}. Tale piano ha la stessa inclinazione del piano avente come supporto il triangolo T_{10} e pertanto la sua equazione è data da (si veda la (8.44)) $z = \frac{x}{h} - \frac{y}{h}$.

Pertanto la funzione di base $\phi_2(x,y)$ è data da:

$$\phi_2(x,y) = \begin{cases} \frac{x}{h} - 1, & (x,y) \in T_3, \\ \frac{y}{h}, & (x,y) \in T_4, \\ 2 - \frac{x}{h} + \frac{y}{h}, & (x,y) \in T_5, \\ 3 - \frac{x}{h}, & (x,y) \in T_{14}, \\ 2 - \frac{y}{h}, & (x,y) \in T_{13}, \\ \frac{x}{h} - \frac{y}{h}, & (x,y) \in T_{12}. \end{cases} \tag{8.45}$$

Calcolo delle restanti funzioni base. Sfruttando la simmetria della mesh in Fig. 8.8 e la linearità delle funzioni base, si trova facilmente che

$$\phi_3(x,y) = \begin{cases} -2 + \frac{x}{h}, & (x,y) \in T_5, \\ \frac{y}{h}, & (x,y) \in T_6, \\ 3 - \frac{x}{h} + \frac{y}{h}, & (x,y) \in T_7, \\ 4 - \frac{x}{h}, & (x,y) \in T_{16}, \\ 2 - \frac{y}{h}, & (x,y) \in T_{15}, \\ -1 + \frac{x}{h} - \frac{y}{h}, & (x,y) \in T_{14}, \end{cases} \tag{8.46}$$

$$\phi_4(x,y) = \begin{cases} \frac{x}{h}, & (x,y) \in T_9, \\ -1 + \frac{y}{h}, & (x,y) \in T_{10}, \\ -\frac{x}{h} + \frac{y}{h}, & (x,y) \in T_{11}, \\ 2 - \frac{x}{h}, & (x,y) \in T_{20}, \\ 3 - \frac{y}{h}, & (x,y) \in T_{19}, \\ 2 + \frac{x}{h} - \frac{y}{h}, & (x,y) \in T_{18}, \end{cases} \tag{8.47}$$

$$\phi_5(x,y) = \begin{cases} -1 + \frac{x}{h}, & (x,y) \in T_{11}, \\ -1 + \frac{y}{h}, & (x,y) \in T_{12}, \\ 1 - \frac{x}{h} + \frac{y}{h}, & (x,y) \in T_{13}, \\ 3 - \frac{x}{h}, & (x,y) \in T_{22}, \\ 3 - \frac{y}{h}, & (x,y) \in T_{21}, \\ 1 + \frac{x}{h} - \frac{y}{h}, & (x,y) \in T_{20}, \end{cases} \tag{8.48}$$

$$\phi_6(x,y) = \begin{cases} -2 + \frac{x}{h}, & (x,y) \in T_{13}, \\ -1 + \frac{y}{h}, & (x,y) \in T_{14}, \\ 2 - \frac{x}{h} + \frac{y}{h}, & (x,y) \in T_{15}, \\ 4 - \frac{x}{h}, & (x,y) \in T_{24}, \\ 3 - \frac{y}{h}, & (x,y) \in T_{23}, \\ \frac{x}{h} - \frac{y}{h}, & (x,y) \in T_{22}, \end{cases} \tag{8.49}$$

$$\phi_7(x,y) = \begin{cases} \frac{x}{h}, & (x,y) \in T_{17}, \\ -2 + \frac{y}{h}, & (x,y) \in T_{18}, \\ -1 - \frac{x}{h} + \frac{y}{h}, & (x,y) \in T_{19}, \\ 2 - \frac{x}{h}, & (x,y) \in T_{28}, \\ 4 - \frac{y}{h}, & (x,y) \in T_{27}, \\ 3 + \frac{x}{h} - \frac{y}{h}, & (x,y) \in T_{26}, \end{cases} \tag{8.50}$$

$$\phi_8(x,y) = \begin{cases} -1 + \frac{x}{h}, & (x,y) \in T_{19}, \\ -2 + \frac{y}{h}, & (x,y) \in T_{20}, \\ -\frac{x}{h} + \frac{y}{h}, & (x,y) \in T_{21}, \\ 3 - \frac{x}{h}, & (x,y) \in T_{30}, \\ 4 - \frac{y}{h}, & (x,y) \in T_{29}, \\ 2 + \frac{x}{h} - \frac{y}{h}, & (x,y) \in T_{28}, \end{cases} \tag{8.51}$$

$$\phi_9(x,y) = \begin{cases} -2 + \frac{x}{h}, & (x,y) \in T_{21}, \\ -2 + \frac{y}{h}, & (x,y) \in T_{22}, \\ 1 - \frac{x}{h} + \frac{y}{h}, & (x,y) \in T_{23}, \\ 4 - \frac{x}{h}, & (x,y) \in T_{32}, \\ 4 - \frac{y}{h}, & (x,y) \in T_{31}, \\ 1 + \frac{x}{h} - \frac{y}{h}, & (x,y) \in T_{30}. \end{cases} \tag{8.52}$$

8.4 b Calcolo della stiffness matrix e del load vector

Per calcolare la stiffness matrix $\boldsymbol{K}$ occorre fare uso della seguente relazione

$$K_{ij} = a(\phi_i, \phi_j) = \int_\Omega k(x,y) \nabla\phi_i(x,y) \cdot \nabla\phi_j(x,y)\, dxdy. \tag{8.53}$$

Supponendo che il mezzo sia omogeneo[1] ($k(x,y) = 1$), si ha

$$\begin{aligned} K_{11} = a(\phi_1, \phi_1) &= \int_\Omega \nabla\phi_1(x,y) \cdot \nabla\phi_1(x,y)\, dxdy \\ &= \int_{T_1} \nabla\phi_1(x,y) \cdot \nabla\phi_1(x,y)\, dxdy + \int_{T_2} \nabla\phi_1(x,y) \cdot \nabla\phi_1(x,y)\, dxdy \\ &+ \int_{T_3} \nabla\phi_1(x,y) \cdot \nabla\phi_1(x,y)\, dxdy + \int_{T_{10}} \nabla\phi_1(x,y) \cdot \nabla\phi_1(x,y)\, dxdy \\ &+ \int_{T_{11}} \nabla\phi_1(x,y) \cdot \nabla\phi_1(x,y)\, dxdy + \int_{T_{12}} \nabla\phi_1(x,y) \cdot \nabla\phi_1(x,y)\, dxdy, \end{aligned}$$

in quanto il supporto della funzione base $\phi_1(x,y)$ è $T_1 \cup T_2 \cup T_3 \cup T_{10} \cup T_{11} \cup T_{12}$.

Dal momento che $\nabla\phi_1(x,y) = (\frac{1}{h}, 0)^T$ quando $(x,y) \in T_1$, si ha

$$\int_{T_1} \frac{1}{h^2}\, dxdy = \frac{1}{h^2}\,\frac{h^2}{2} = \frac{1}{2}.$$

[1] Nel caso disomogeneo ($k(x,y)$ non costante) il procedimento è del tutto analogo, con la sola differenza che il calcolo degli elementi K_{ij} potrebbe non essere fattibile per via analitica. In tal caso viene effettuato con formule di integrazione numerica che tengono conto della espressione esplicita di $k(x,y)$.

Analogamente

$$(x,y) \in T_2 \Longrightarrow \nabla\phi_1(x,y) = \begin{pmatrix} 0 \\ \frac{1}{h} \end{pmatrix} \Longrightarrow \int_{T_2} \frac{1}{h^2}\, dxdy = \frac{1}{h^2}\,\frac{h^2}{2} = \frac{1}{2},$$
$$(x,y) \in T_3 \Longrightarrow \nabla\phi_1(x,y) = \begin{pmatrix} -\frac{1}{h} \\ \frac{1}{h} \end{pmatrix} \Longrightarrow \int_{T_3} \frac{2}{h^2}\, dxdy = \frac{2}{h^2}\,\frac{h^2}{2} = 1.$$

Per simmetria si vede facilmente che

$$\int_{T_{10}} \nabla\phi_1(x,y)\cdot\nabla\phi_1(x,y)\, dxdy = \int_{T_3} \nabla\phi_1(x,y)\cdot\nabla\phi_1(x,y)\, dxdy = 1, \quad (8.54a)$$
$$\int_{T_{11}} \nabla\phi_1(x,y)\cdot\nabla\phi_1(x,y)\, dxdy = \int_{T_2} \nabla\phi_1(x,y)\cdot\nabla\phi_1(x,y)\, dxdy = 1, \quad (8.54b)$$
$$\int_{T_{12}} \nabla\phi_1(x,y)\cdot\nabla\phi_1(x,y)\, dxdy = \int_{T_1} \nabla\phi_1(x,y)\cdot\nabla\phi_1(x,y)\, dxdy = 1, \quad (8.54c)$$

e quindi

$$K_{11} = 2\left(1 + \tfrac{1}{2} + \tfrac{1}{2}\right) = 4. \quad (8.55)$$

Per il calcolo dell'entrata K_{12} notiamo che i supporti delle funzioni base $\phi_1(x,y)$ e $\phi_2(x,y)$ si intersecano nei triangoli T_3 e T_{12}:

$$\operatorname{supp}\phi_1(x,y) \cap \operatorname{supp}\phi_2(x,y) = T_3 \cup T_{12}.$$

Pertanto si ha

$$(x,y) \in T_3 \Longrightarrow \begin{cases} \nabla\phi_1(x,y) = \begin{pmatrix} -1/h \\ 1/h \end{pmatrix}, \\ \nabla\phi_2(x,y) = \begin{pmatrix} 1/h \\ 0 \end{pmatrix}, \end{cases} \Longrightarrow \int_{T_3} \frac{-1}{h^2}\, dxdy = -\tfrac{1}{2},$$
$$(x,y) \in T_{12} \Longrightarrow \begin{cases} \nabla\phi_1(x,y) = \begin{pmatrix} -1/h \\ 0 \end{pmatrix}, \\ \nabla\phi_2(x,y) = \begin{pmatrix} 1/h \\ -1/h \end{pmatrix}, \end{cases} \Longrightarrow \int_{T_{12}} \frac{-1}{h^2}\, dxdy = -\tfrac{1}{2},$$

e quindi

$$K_{12} = -1. \quad (8.56)$$

L'elemento $K_{13} = 0$ in quanto

$$\operatorname{supp}\phi_1(x,y) \cap \operatorname{supp}\phi_3(x,y) = \emptyset,$$

mentre gli elementi K_{14} e K_{15} possono essere diversi da zero in quanto

$$\operatorname{supp}\phi_1(x,y) \cap \operatorname{supp}\phi_3(x,y) = T_{10} \cup T_{11},$$
$$\operatorname{supp}\phi_1(x,y) \cap \operatorname{supp}\phi_3(x,y) = T_{11} \cup T_{12}.$$

Tenendo conto delle funzioni base e delle loro relazioni di simmetria, si trova facilmente che la stiffness matrix è una matrice sparsa avente la seguente forma

$$\boldsymbol{K} = \begin{pmatrix} 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 4 & 0 & 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & -1 & 4 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 4 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 \end{pmatrix}. \tag{8.57}$$

Il calcolo del load vector $\boldsymbol{f} = ((f, \phi_1), \ldots, (f, \phi_9))^T$ dipende fortemente dalla forma del termine noto della PDE (8.19). In ogni caso, si procede in modo del tutto analogo a quanto visto per la stiffness matrix. Per esempio, per ottenere la prima componente di $\boldsymbol{f}$, si procede nel seguente modo

$$\begin{aligned} f_1 = (f, \phi_1) &= \int_\Omega f(x,y)\phi_1(x,y)\,dxdy \\ &= \int_{T_1} f(x,y)\frac{x}{h}\,dxdy + \int_{T_2} f(x,y)\frac{y}{h}\,dxdy \\ &+ \int_{T_3} f(x,y)\left(1 - \frac{x}{h} + \frac{y}{h}\right)dxdy + \int_{T_{10}} f(x,y)\left(1 + \frac{x}{h} - \frac{y}{h}\right)dxdy \\ &+ \int_{T_{11}} f(x,y)\left(2 - \frac{y}{h}\right)dxdy + \int_{T_{12}} f(x,y)\left(2 - \frac{x}{h}\right)dxdy, \end{aligned}$$

e facendo uso delle restanti funzioni base $\phi_i(x,y)$ si trovano formule analoghe per il calcolo delle rimanenti entrate del load vector.

Osservazione. Supponiamo ora che l'equazione ellittica nella (8.19) sia la seguente:

$$-\nabla \cdot (k\,\nabla u) + a_0 u = f, \quad \text{in} \quad \Omega. \tag{8.58}$$

La presenza del termine additivo a_0 non comporta alcuna modifica di tipo metodologico. In questo caso si ottiene la seguente forma variazionale:

$$\int_\Omega k\,\nabla u \cdot \nabla v\,dxdy + \int_\Omega a_0 uv\,dxdy = \int_\Omega fv\,dxdy, \quad \text{qualunque sia } v \in H_0^1(\Omega),$$

che differisce dalla precedente per la presenza del secondo integrale che non figura nella (8.21). Per quanto riguarda la sua risoluzione numerica, l'unica modifica significativa riguarda la sostituzione del sistema (8.43) con il seguente:

$$\boldsymbol{K}_h \boldsymbol{u}_h + \boldsymbol{B}_h \boldsymbol{u}_h = \boldsymbol{f}_h, \tag{8.59}$$

dove $(\boldsymbol{B}_h)_{ij} = \int_{\Omega_h} a_0\phi_i\phi_j\,dxdy$. La sua presenza non altera in modo rilevante la complessità di calcolo nella risoluzione del sistema, dato che la matrice $\boldsymbol{K}_h + \boldsymbol{B}_h$ presenta la stessa sparsità di $\boldsymbol{K}_h$. L'unico aggravio computazionale deriva dal calcolo del secondo integrale, la cui complessità dipende da a_0.

8.4 c Triangolo di riferimento e suo utilizzo nella costruzione del sistema (8.43)

Nella pratica gli elementi della matrice $\boldsymbol{K}_h$ e del vettore $\boldsymbol{f}_h$ non si calcolano (quasi mai) utilizzando la forma esplicita delle ϕ_j (rispetto alle variabili x e y) per diverse ragioni. La prima è che, nei problemi applicativi, il reticolo Ω_h è generalmente formato da decine di migliaia di triangoli. La seconda è che, per le più svariate forme che i triangoli possono presentare, rispetto agli assi coordinati, l'integrazione su di essi potrebbe comportare un numero molto rilevante di adattamenti tra loro diversi. Per ottimizzare la tecnica di calcolo si ricorre (di norma) ad un triangolo di riferimento T_R avante vertici $(0,0)$, $(1,0)$ e $(0,1)$ in un sistema di riferimento cartesiano (s,t). A tale scopo si definisce una trasformazione lineare che stabilisce una corrispondenza biunivoca tra il generico triangolo T_i, $i = 1, 2, \ldots, n_v$, della mesh, con il triangolo di riferimento T_R. Indichiamo con (x_{i1}, y_{i1}), (x_{i2}, y_{i2}) e (x_{i3}, y_{i3}) i tre vertici di T_i. Tra questi e quelli di T_R si stabilisce una corrispondenza, mediante una connessione di

$$(x_{i1}, y_{i1}) \text{ con } (0,0), \qquad (x_{i2}, y_{i2}) \text{ con } (1,0), \qquad (x_{i3}, y_{i3}) \text{ con } (0,1). \tag{8.60}$$

Come è immediato verificare, essa può essere stabilita mediante la trasformazione lineare

$$\begin{cases} x = x_{i1} + (x_{i2} - x_{i1})s + (x_{i3} - x_{i3})t, \\ y = y_{i1} + (y_{i2} - y_{i1})s + (y_{i3} - y_{i3})t. \end{cases} \tag{8.61}$$

Tale corrispondenza è geometricamente visualizzata nella Fig. 8.9 La trasformazione (8.61) permette di associare a una generica funzione $g(x,y)$ definita nel triangolo T_i della mesh una funzione $h(s,t)$, avente supporto T_R, così definita:

$$h(s,t) = g(x_{i1} + (x_{i2} - x_{i1})s + (x_{i3} - x_{i3})t, y_{i1} + (y_{i2} - y_{i1})s + (y_{i3} - y_{i3})t),$$

dove $0 \le s, t \le 1$. Questo equivale a dire che tra i vertici di T_i, relativi agli assi (x,y) e quelli di T_R, relativi agli assi (s,t), vale la corrispondenza (8.60), essendo

$$h(0,0) = g(x_{i1}, y_{i1}), \qquad h(1,0) = g(x_{i2}, y_{i2}), \qquad h(0,1) = g(x_{i3}, y_{i3}).$$

Nella costruzione delle box-splines, ad ogni triangolo T_i vengono quindi associati tre funzioni lineari: la $\phi_{i1} = a_1x + b_1y + c_1$ che assume il valore uno in

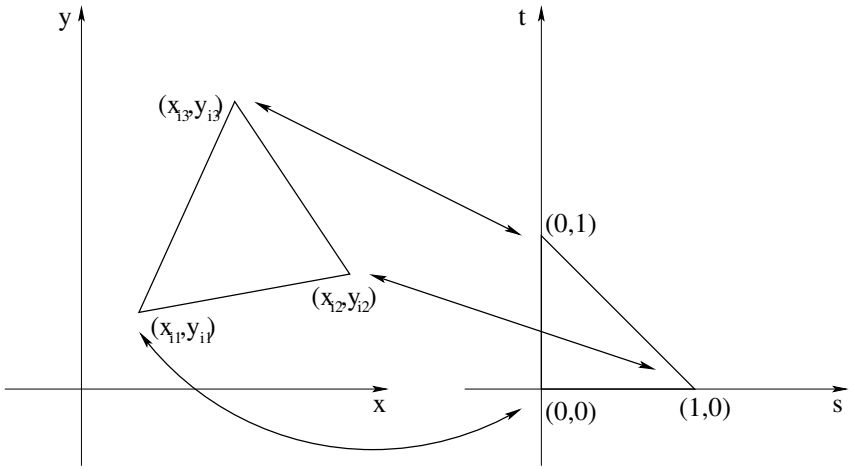


Figura 8.9: Corrispondenza biunivoca tra T_i e T_R.

(x_{i1}, y_{i1}) e zero in (x_{i2}, y_{i2}) e (x_{i3}, y_{i3}), la $\phi_{i2} = a_2x + b_2y + c_2$ che assume il valore uno in (x_{i2}, y_{i2}) e zero in (x_{i1}, y_{i1}) e (x_{i3}, y_{i3}), e la $\phi_{i3} = a_3x + b_3y + c_3$ che assume il valore uno in (x_{i3}, y_{i3}) e zero in (x_{i1}, y_{i1}) e (x_{i2}, y_{i2}). Per analogia al triangolo T_R vengono associate le seguenti tre funzioni lineari:

$$\begin{cases} \gamma_1(s,t) = 1 - s - t, \\ \gamma_2(s,t) = s, \\ \gamma_3(s,t) = t. \end{cases} \tag{8.62}$$

Le (8.61) e (8.62) implicono che la $\phi_{i1}(x, y)$ è in corrispondenza biunivoca con la $\gamma_1(s, t)$, la $\phi_{i2}(x, y)$ con la $\gamma_2(s, t)$, e la $\gamma_{i3}(x, y)$ con la $\gamma_3(s, t)$.

Calcolo del vettore $\boldsymbol{f}_h$. Ricorrendo al triangolo di riferimento, per il calcolo delle componenti $f_{h,i}$, $i = 1, 2, \ldots, n_v$, si procede in questo modo:

(a) identificati i triangoli T_{il} che formano il supporto della box-spline ϕ_i, si osserva che

$$f_{h,i} = \sum_{l=1}^{n_i} \int_{T_{il}} f(x,y)\varphi_i(x,y)\, dxdy,$$

dove T_{il} indica l'l-esimo triangolo del supporto della φ_i e n_i il numero dei triangoli che formano il suo supporto;

(b) il triangolo T_{il} viene quindi posto in corrispondenza biunivoca con il triangolo di riferimento T_R, associando alle φ_i la γ_1, qualora essa assume il valore 1 nel primo vertice di T_{il}, la γ_2, qualora essa assume il valore 1 nel secondo vertice,e la γ_3, nel caso assume il valore 1 nel terzo vertice;

(c) si procede quindi al calcolo dei vari integrali, osservando che

$$\int_{T_{il}} f(x,y)\varphi_i(x,y)\,dxdy = \int_{T_R} g(s,t)\gamma_l(s,t)|\det J_{il}|\,dsdt$$
$$= |\det J_{il}| \int_0^1 \int_0^{1-s} g(s,t)\gamma_l(s,t)\,dtds,$$

dove $g(s,t) = f(x(s,t), y(s,t))$ è la funzione γ che corrisponde alla φ_i su T_l e J_{il} è la matrice Jacobiana della trasformazione che al triangolo T_{il} nelle coordinate (x,y) associa T_R nelle coordinate (s,t).

Calcolo della matrice $\boldsymbol{K}_h$. Il calcolo degli elementi di $\boldsymbol{K}_h$ è più complesso, in quanto richiede quello dei gradienti $\nabla\varphi_i$ e $\nabla\varphi_j$ e del loro prodotto interno nell'intersezione dei rispettivi supporti. Indicato con T il generico triangolo dell'intersezione dei supporti di φ_i e φ_j, indichiamo con $\{(x_1,y_1),(x_2,y_2),(x_3,y_3)\}$ i suoi vertici e con φ la restrizione di una generica box-spline su T. La corrispondenza tra T e T_R è data dalle relazioni

$$\begin{cases} x = x_1 + (x_2 - x_1)s + (x_3 - x_1)t, \\ y = x_1 + (x_2 - x_1)s + (x_3 - x_1)t, \end{cases} \tag{8.63}$$

che associano (x_1,y_1) con $(0,0)$, (x_2,y_2) con $(1,0)$ e (x_3,y_3) con $(0,1)$ Di conseguenza alla $\varphi|_T$ corrisponde $\gamma_1(s,t) = 1 - s - t$ se $\varphi(x_1,y_1) = 1$, $\gamma_2(s,t) = s$ se $\varphi(x_2,y_2) = 1$ e $\gamma_3(s,t) = t$ se $\varphi(x_3,y_3) = 1$. Essendo

$$J = \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix}$$

la matrice Jacobiana della trasformazione, la (8.63) può essere scritta nella forma

$$\begin{pmatrix} x - x_1 \\ y - y_1 \end{pmatrix} = J \begin{pmatrix} s \\ t \end{pmatrix},$$

da cui, essendo J non singolare (in quanto i tre punti (x_j,y_j) sono distinti),

$$\begin{pmatrix} s \\ t \end{pmatrix} = J^{-1} \begin{pmatrix} x - x_1 \\ y - y_1 \end{pmatrix} = \frac{1}{\det J} \begin{pmatrix} y_3 - y_1 & x_1 - x_3 \\ y_1 - y_2 & x_2 - x_1 \end{pmatrix} \begin{pmatrix} x - x_1 \\ y - y_1 \end{pmatrix}.$$

Pertanto, essendo

$$\begin{pmatrix} s \\ t \end{pmatrix} = \begin{cases} \dfrac{1}{\det J}\left[(y_3 - y_1)(x - x_1) + (x_1 - x_3)(y - y_1)\right], \\ \dfrac{1}{\det J}\left[(y_1 - y_2)(x - x_1) + (x_2 - x_1)(y - y_1)\right], \end{cases}$$

si ha:

$$\begin{cases} \dfrac{\partial s}{\partial x} = \dfrac{y_3 - y_1}{\det J} & \text{e} \quad \dfrac{\partial s}{\partial y} = \dfrac{x_1 - x_3}{\det J}, \\ \dfrac{\partial t}{\partial x} = \dfrac{y_1 - y_2}{\det J} & \text{e} \quad \dfrac{\partial t}{\partial y} = \dfrac{x_2 - x_1}{\det J}. \end{cases}$$

Note le derivate parziali di s e t rispetto a x e y, otteniamo immediatamente la relazione tra $\nabla\gamma$ e $\nabla\gamma$. Basta infatti notare che (per regola di derivazione composta)

$$\begin{cases} \dfrac{\partial \varphi}{\partial x} = \dfrac{\partial \gamma}{\partial s}\dfrac{\partial s}{\partial x} + \dfrac{\partial \gamma}{\partial t}\dfrac{\partial t}{\partial x}, \\ \dfrac{\partial \varphi}{\partial y} = \dfrac{\partial \gamma}{\partial s}\dfrac{\partial s}{\partial y} + \dfrac{\partial \gamma}{\partial t}\dfrac{\partial t}{\partial y}, \end{cases} \quad \Longrightarrow \nabla\phi = \begin{pmatrix} \dfrac{\partial s}{\partial x} & \dfrac{\partial t}{\partial x} \\ \dfrac{\partial s}{\partial y} & \dfrac{\partial t}{\partial y} \end{pmatrix} \begin{pmatrix} \dfrac{\partial \gamma}{\partial s} \\ \dfrac{\partial \gamma}{\partial t} \end{pmatrix} = J^{-T}\,\nabla\gamma,$$

dove il simbolo T indica la trasposizione tra righe e colonne di una matrice e $J^{-T} = (J^{-1})^T = (J^T)^{-1}$. Ricordando infine che

$$dxdy = |\det J|\, dsdt,$$

il calcolo di $(\boldsymbol{K}_h)_{ij}$, su un generico triangolo T, può essere ricondotto ad una integrazione su T_R mediante la trasformazione

$$\int_T k(x,y)\nabla\varphi_j \cdot \nabla\varphi_i\, dxdy = \int_{T_R} \tilde{k}(s,t)(J^{-T}\nabla\gamma_j)\cdot(J^{-T}\nabla\gamma_i)\,|\det J|\, dsdt, \tag{8.64}$$

avendo indicato con $\tilde{k}$ la rappresentazione di k su T_R e con γ_j e γ_i le funzioni $\gamma(s,t)$ corrispondenti su T_R a φ_j e φ_i definite su T_R.

Esempio 8.11 Indicate con φ e ψ le box-splines aventi come rispettivi supporti i due esagoni della Fig. 8.10 e che assumono il valore 1 rispettivamente in $(2,1)$ e $(2+\frac{3}{2}h, 1+\frac{1}{3}k)$, illustrare il procedimento per il calcolo dell'integrale

$$I = \int_T (xy)\,\nabla\varphi\cdot\nabla\psi\, dxdy,$$

essendo T il triangolo con vertici $\{(2,1),(2+h,1-\frac{2}{3}k),(2+\frac{3}{2}h,1+\frac{1}{3}k)\}$. Iniziamo con la relazione tra T e il triangolo di riferimento T_R che, in coordinate (s,t) è caratterizzato dai vertici $\{(0,0),(1,0),(0,1)\}$. Tra T e T_R vale la seguente relazione lineare:

$$\begin{cases} x = 2 + hs + \frac{3}{2}ht, \\ y = 1 - \frac{2}{3}ks + \frac{1}{3}kt. \end{cases}$$

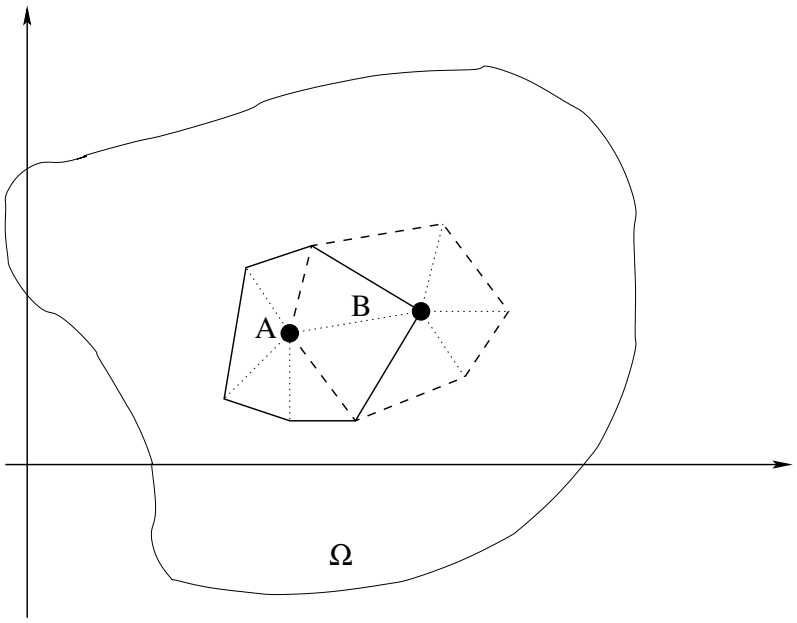


Figura 8.10: Supporti di φ e ψ, dove $A = (2,1)$ e $B = (2 + \frac{3}{2}h, 1 + \frac{1}{3}k)$.

Da essa risulta evidente che il vertice $(2,1)$, nel quale la φ assume il valore 1, è in corrispondenza biunivoca con $(0,0)$ e il vertice $(2 + \frac{3}{2}h, 1 + \frac{1}{3}k)$ con $(0,1)$. Questo implica che alla rappresentazione della φ in T corrisponde la $\gamma_1(s,t) = 1 - s - t$ in T_R e che alla ψ corrispondende la $\gamma_2(s,t) = t$. La matrice di Jordan della trasformazione è

$$J = \begin{pmatrix} h & \frac{3}{2}h \\ -\frac{2}{3}k & \frac{1}{3}k \end{pmatrix}, \quad \text{con} \quad J^{-1} = \frac{3}{4hk} \begin{pmatrix} \frac{1}{3}k & -\frac{3}{2}h \\ \frac{2}{3}k & h \end{pmatrix},$$

con $\det J = \frac{4}{3}hk$. Essendo inoltre $\nabla\gamma_1 = \left(\begin{smallmatrix} -1 \\ -1 \end{smallmatrix}\right)$ e $\nabla\gamma_3 = \left(\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right)$,

$$J^{-T}\nabla\gamma_1 = \begin{pmatrix} -\frac{3}{4h} \\ \frac{3}{8k} \end{pmatrix} = \frac{3}{8}\begin{pmatrix} -\frac{2}{h} \\ \frac{1}{k} \end{pmatrix}, \quad J^{-T}\nabla\gamma_3 = \frac{3}{8}\begin{pmatrix} \frac{3}{h} \\ \frac{2}{k} \end{pmatrix}.$$

Di conseguenza, tenuto conto della (8.64), l'integrale I può essere calcolato nel modo seguente:

$$\begin{aligned} I &= \int_{T_R} (2 + hs + \tfrac{3}{2}ht)(1 - \tfrac{2}{3}ks + \tfrac{1}{3}kt)\, \frac{3}{8}\begin{pmatrix} -\frac{2}{h} \\ \frac{1}{k} \end{pmatrix} \cdot \frac{3}{8}\begin{pmatrix} \frac{3}{h} \\ \frac{2}{k} \end{pmatrix} \frac{4}{3}hk\, dsdt \\ &= \frac{3}{16}hk\left(-\frac{6}{h^2} + \frac{2}{k^2}\right) \int_0^1 \int_0^{1-s} (2 + hs + \tfrac{3}{2}ht)(1 - \tfrac{2}{3}ks + \tfrac{1}{3}kt)\, dtds. \end{aligned}$$

8.5 BVPs con condizioni inomogenee al bordo

Condizioni di inomogeneità per il problema di Neumann. Consideriamo ora in luogo del problema (8.23), nel quale le condizioni sono di tipo omogeneo, il seguente problema:

$$\begin{cases} -\nabla \cdot (k\,\nabla u) = f, & (x,y) \in \Omega, \\ k\,\dfrac{\partial u}{\partial n} = g, & (x,y) \in \partial\Omega, \end{cases} \tag{8.65}$$

essendo g una funzione "abbastanza regolare" definita su $\partial\Omega$. La derivazione della forma debole associata viene ottenuta procedendo come nel caso omogeneo. Moltiplicando primo e secondo membro nella equazione per $v \in H^1(\Omega)$, richiediamo che

$$-\int_\Omega \nabla \cdot (k\,\nabla u) v\,dxdy = \int_\Omega f v\,dxdy, \quad \text{per qualsiasi } v \in H^1(\Omega).$$

Da tale relazione, applicando la prima identità di Green, segue la seguente

$$\int_\Omega k\,\nabla u \cdot \nabla v\,dxdy - \int_{\partial\Omega} \left(k\,\frac{\partial u}{\partial n}\right) v\,d\sigma = \int_\Omega f v\,dxdy,$$

da cui, tenendo conto delle condizioni al bordo, segue che

$$\int_\Omega k\,\nabla u \cdot \nabla v\,dxdy = \int_\Omega f v\,dxdy + \int_{\partial\Omega} g v\,d\sigma, \quad \text{qualunque sia } v \in H^1(\Omega). \tag{8.66}$$

La forma debole (variazionale), in presenza di condizioni di inomogeneità, differisce dunque da quella relativa al caso omogeneo per la presenza dell'integrale su $\partial\Omega$.

Problema di Dirichlet con condizioni di inomogeneità. Consideriamo ora in luogo del problema (8.19), nel quale le condizioni al bordo sono di tipo omogeneo, il seguente problema:

$$\begin{cases} -\nabla \cdot (k\,\nabla u) = f, & (x,y) \in \Omega, \\ u = h, & (x,y) \in \partial\Omega, \end{cases} \tag{8.67}$$

dove h è una funzione "abbastanza regolare" definita su $\partial\Omega$. Consideriamo inizialmente una funzione $H \in H^1(\Omega)$ con $H = h$ su $\partial\Omega$. La costruzione analitica di H, possibile nel caso di dominio rettangolare (come visto in precedenza), è molto complicata in generale. Essa [vedi la prossima Osservazione] è fortunatamente agevole nel contesto degli elementi finiti. Poichè H è nota e

$w = u - H \in H_0^1(\Omega)$, possiamo procedere al calcolo di w e richiedere infine che $u = w + H$.

Per ottenere la forma debole (distribuzionale) della (8.67), moltiplichiamo primo e secondo membro dell'equazione per $v \in H_0^1(\Omega)$, dopo aver posto $u = w + H$. Otteniamo ora l'equazione

$$-\int_\Omega [\nabla \cdot (k\,\nabla w + \nabla H)v]\,dxdy = \int_\Omega f v\,dxdy, \quad v \in H_0^1(\Omega).$$

Da essa, utilizzando la prima identità di Green, segue la formulazione debole della (8.67), relativa alla funzione incognita $w \in H_0^1(\Omega)$, che è la seguente: determinare la funzione $u = w + H$, con $w \in H_0^1(\Omega)$, soddisfacente l'equazione

$$\int_\Omega k\,\nabla w \cdot \nabla v\,dxdy = \int_\Omega f v\,dxdy - \int_\Omega k\,\nabla H \cdot \nabla v\,dxdy, \tag{8.68}$$

qualunque sia $v \in H_0^1(\Omega)$. La formulazione debole, nel caso inomogeneo, differisce dunque da quella del caso omogeneo, per la presenza dell'ultimo integrale.

Osservazione. La funzione H può essere costruita mediante un semplice procedimento di interpolazione lineare. Costruito il reticolo Ω_h relativo alla

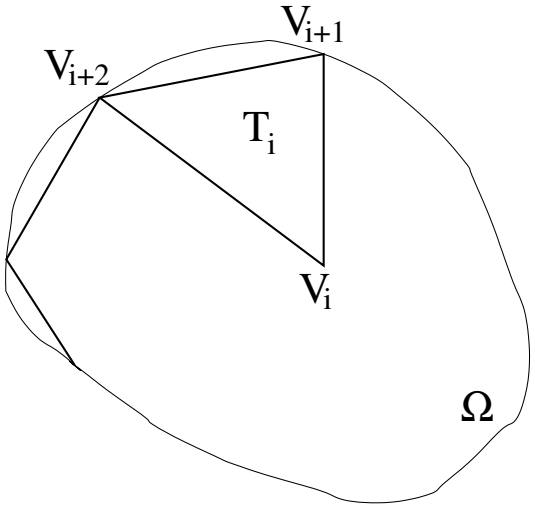


Figura 8.11: Il triangolo T_i inserito nel dominio Ω.

Ω, si considera l'interpolante continua in Ω_h e lineare a tratti che soddisfa le condizioni:

$$H(z_i) = \begin{cases} 0, & \text{se } z_i \text{ è un punto nodale interno di } \Omega_h, \\ h(z_i), & \text{se } z_i \text{ è un punto nodale di } \Gamma \text{ appartenente a } \partial\Omega_h. \end{cases}$$

Nel triangolo T_i riportato nella Fig. 8.11, ad esempio, $H(x,y)$ rappresenta l'equazione del piano identificato di interpolazione $H(z_i) = 0$, $H(z_{i+1}) = h(z_{i+1})$ e $H(z_{i+2}) = h(z_{i+2})$.

Condizioni di inomogeneità nel problema di Neumann. In luogo del problema (8.23) consideriamo ora il seguente problema di Neumann con condizioni di inomogeneità al bordo:

$$\begin{cases} -\nabla \cdot (k\,\nabla u) = f, & (x,y) \in \Omega, \\ k\,\dfrac{\partial u}{\partial n} = h, & (x,y) \in \partial\Omega, \end{cases} \tag{8.69}$$

dove h è una funzione nota al bordo.

In questo caso la forma debole (variazionale) dell'equazione relativa alla (8.69) procede nel modo seguente:

(a) moltiplicando primo e secondo membro dell'equazione per una generica $v \in H^1(\Omega)$ e integrando in Ω si scrive l'equazione

$$-\int_\Omega (\nabla \cdot k\,\nabla u) v\,dxdy = \int_\Omega f v\,dxdy, \quad \text{qualunque sia } v \in H^1_1(\Omega);$$

(b) applicando quindi l'identità di Green il primo integrale diventa

$$\int_\Omega k\,\nabla u \cdot \nabla v\,d\sigma - \int_{\partial\Omega} \left(k\,\frac{\partial u}{\partial n}\right) v\,dxdy, \quad \text{qualunque sia } v \in H^1(\Omega);$$

da cui, ricordando che $k\,\frac{\partial u}{\partial n} = h$, con h funzione nota, si ottiene la seguente forma variazionale della (8.69):

$$\int_\Omega k\,\nabla u \cdot \nabla v\,dxdy = \int_\Omega f v\,dxdy + \int_{\partial\Omega} h v\,d\sigma, \quad \text{qualunque sia } v \in H^1(\Omega). \tag{8.70}$$

In questo caso la differenza tra la forma variazionale associata al problema, con condizioni non omogenee al bordo, differisce dal caso omogeneo per la presenza dell'ultimo integrale.

Condizioni di inomogeneità nel caso misto. Consideriamo ora il caso misto, in condizioni di non omogeneità:

$$\begin{cases} -\nabla \cdot (k\,\nabla u) = f, & (x,y) \in \Omega, \\ u = g, & (x,y) \in \Gamma_1, \\ k\,\dfrac{\partial u}{\partial n} = h, & (x,y) \in \Gamma_2. \end{cases}$$

Costruito il reticolo Ω_h di Ω, come primo passo si costruisce una funzione F, continua e lineare a tratti in Ω_h, che interpola la g in Γ_1, imponendo le condizioni:

$$F(z_i) = \begin{cases} f(z_i), & \text{per } z_i \in \Gamma_1, \\ 0, & \text{per } z_i \notin \Gamma_1, \end{cases}$$

dove z_i indica un qualsiasi punto nodale di $\overline{\Omega_h}$. Posto $u = w+F$, moltiplicando primo e secondo membro dell'equazione per $v \in H^1(\Omega)$ con $v|_{\Gamma_1} = 0$, si ottiene l'equazione in w

$$-\int_\Omega \left[\nabla \cdot k(\nabla w + \nabla F)\right] v \, dxdy = \int_\Omega fv \, dxdy, \qquad v \in H^1(\Omega) \quad \text{con} \quad v|_{\Gamma_1} = 0,$$

dalla quale applicando la prima identità di Green

$$\int_\Omega k(\nabla w + \nabla F) \cdot \nabla v \, dxdy - \int_{\Gamma_1 \cup \Gamma_2} k \left(\frac{\partial w}{\partial n} + \frac{\partial F}{\partial n} \right) v \, d\sigma = \int_\Omega gv \, dxdy.$$

Da essa, ricordando che $k \frac{\partial u}{\partial n} = h$ su Γ_2 e che $v|_{\Gamma_1} = 0$, deriva l'equazione

$$\int_\Omega k \nabla w \cdot \nabla v \, dxdy + \int_\Omega k \nabla F \cdot \nabla v \, dxdy = \int_{\Gamma_2} hv \, d\sigma + \int_\Omega fv \, dxdy,$$

qualunque sia $v \in H^1(\Omega)$, con $v|_{\Gamma_1} = 0$. La forma debole, nel caso inomogeneo, differisce pertanto da quella ottenuta nel caso omogeneo per la presenza del secondo integrale in Ω e dell'integrale in Γ_2.

Osservazione. Tutte le considerazioni precedenti, incluse quelle relative alla metodologia di calcolo, restano sostanzialmente valide se l'equazione ellittica precedentemente considerata (nei vari BVPs)

$$-\nabla \cdot (k \nabla u) = f, \qquad (x, y) \in \Omega,$$

viene sostituita con la seguente (anch'essa ellittica)

$$-\nabla \cdot (k \nabla u) + a_0 u = f, \qquad (x, y) \in \Omega.$$

La forma debole associata al relativo problema di Dirichlet (con condizioni al bordo di tipo omogeneo) assume la seguente forma

$$\int_\Omega k \nabla u \cdot \nabla v \, dxdy + \int_\Omega a_0 uv \, dxdy = \int_\Omega fv \, dxdy, \quad \text{qualunque sia } v \in H_0^1(\Omega),$$

che differisce dalla precedente per la presenza del secondo integrale.

Per quanto riguarda la sua risoluzione numerica, l'unica modifica significativa è la sostituzione del sistema (8.43) con il seguente:

$$\boldsymbol{K}_h \boldsymbol{u}_h + \boldsymbol{B}_h \boldsymbol{u}_h = \boldsymbol{f}_h, \tag{8.71}$$

essendo $(\boldsymbol{B}_h)_{ij} = \int_{\Omega_h} a_0 \phi_i \phi_j \, dxdy$ per $i, j = 1, \ldots, n_h$. In essa $\boldsymbol{K}_h$ è la "stiffness matrix", $\boldsymbol{B}_h$ la "mass matrix" e $\boldsymbol{f}_h$ il "load vector".

8.6 Convergenza del metodo degli elementi finiti

Nonostante i dettagli sulla convergenza del metodo degli elementi finiti considerato (metodo di Galerkin) siano piuttosto complessi, la sua illustrazione è abbastanza semplice. Per capire quanto rapidamente l'errore d'approssimazione della soluzione tende a zero in norma, è necessario premettere qualche definizione. In primo luogo precisiamo che, indicata con Ω_h la mesh relativa ad Ω, indichiamo con u_h l'approssimazione della soluzione esatta u del problema in studio. L'errore di approssimazione dipende naturalmente dal parametro h della reticolazione Ω_h di Ω e dalla norma dello spazio di Sobolev considerato. D'altro canto l'approssimazione di una funzione "abbastanza regolare" nello spazio delle funzioni approssimanti scelte (box-splines nel nostro caso) dipende dalla finezza della mesh, ossia da h, e dalla norma adottata per stimare l'errore di approssimazione.

Indicata con f una funzione "abbastanza regolare" da approssimare in uno spazio X_h di funzioni prefissate (box-splines), rispetto alla norma di un prefissato spazio di Hilbert X, con $X_h \subset X$, interessa stimare la

$$\text{dist}(f, X_n) = \min_{f_h \in X_h} \|f - f_h\|.$$

Per quanto concerne la stima dell'errore, quando sia utile precisarlo, vengono indicate con i simboli $\|\cdot\|_{L^2(\Omega)}$ e $\|\cdot\|_{H^1(\Omega)}$ le norme utilizzate. In questo contesto, per funzione "abbastanza regolare" intendiamo funzioni appartenenti allo spazio di Sobolev

$$H^2(\Omega) = \left\{ f \in L^2(\Omega) \text{ con derivate parziali seconde appartenenti ad } L^2(\Omega) \right\},$$

al quale appartengono le "soluzioni ordinarie" dei BVPs considerati. Per la stima dell'errore di approssimazione si fa spesso riferimento alla seguente definizione di seminorma in $H^2(\Omega)$:

$$|u|_{H^2(\Omega)} = \sqrt{\int_\Omega \left[\left(\frac{\partial^2 u}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 u}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 u}{\partial y^2} \right)^2 \right] dxdy}.$$

Essa rappresenta una seminorma, non una norma, in quanto esistono funzioni non nulle $u \in H^2(\Omega)$ con $|u|_{H^2(\Omega)} = 0$.[2] Nel caso di funzioni "abbastanza regolari", le stime più importanti sugli errori di approssimazione sono le seguenti:

[2] Nel caso univariato, nel quale $\Omega = [a, b]$, tale seminorma coincide con quella introdotta nell'Appendice A per la stima dell'errore nella risoluzione del problema agli estremi (8.8).

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^2 |u|_{H^2(\Omega)}, \tag{8.72a}$$

$$\|u - u_h\|_{H^1(\Omega)} \leq \hat{C}h |u|_{H^2(\Omega)}, \tag{8.72b}$$

dove u_h è l'approssimante di u in X_h e C e $\hat{C}$ sono costanti non dipendenti da h. Questo significa che se u è una funzione abbastanza regolare in Ω, la sua approssimante u_h, ottenuta mediante combinazione lineare di box-splines costruita nel reticolo Ω_h, converge alla u con un errore d'ordine di h^2 rispetto alla norma in $L^2(\Omega)$ e con un errore dell'ordine di h rispetto alla norma in $H^1(\Omega)$.

Le stime (8.72), sull'errore di approssimazione della soluzione con gli elementi finiti, sono una immediata conseguenza della diseguaglianza (8.7) [Lemma di Céa] e dei risultati sull'approssimazione di funzioni "abbastanza regolari" con le box-splines. Risultati dello stesso tipo si ottengono quando, in luogo delle box-splines, vengano utilizzati altri tipi di approssimanti (splines di ordine superiore, polinomi trigonometrici, ecc.).

8.7 Problema spettrale di Helmholtz

In questo paragrafo consideriamo la risoluzione numerica del problema spettrale di Helmholtz con condizioni di periodicità [7]:

$$-\nabla \cdot \left(\frac{1}{\varepsilon(x,y)} \nabla u(x,y) \right) = \eta\, u(x,y), \tag{8.73}$$

con $(x,y) \in A = [0,a] \times [0,b]$ e $\varepsilon(x,y)$ funzione periodica in A, tale cioè da soddisfare le condizioni

$$\begin{cases} \varepsilon(0,y) = \varepsilon(a,y), \\ \varepsilon_x(0,y) = \varepsilon_x(a,y), \end{cases} \quad y \in [0,b]; \quad \begin{cases} \varepsilon(x,0) = \varepsilon(x,b), \\ \varepsilon_y(x,0) = \varepsilon_y(x,b), \end{cases} \quad x \in [0,a], \tag{8.74}$$

essendo η un parametro reale che rappresenta il generico autovalore. Vogliamo dunque studiare lo stesso problema spettrale precedentemente studiato mediante le differenze finite. Questo allo scopo di mostrare come, nel caso di domini regolari, le due diverse tecniche conducono essenzialmente agli stessi risultati. Da notare che, nel caso il doninio non sia sufficientemente regolare, il problema può essere risolto agevolmente con gli elementi finiti ma non con le differenze finite.

Come già osservato nel (8.73), la periodicità della funzione ε induce quella della soluzione u, così che

$$u(0,y) = u(a,y) \quad \text{e} \quad u_x(0,y) = u_x(a,y), \qquad 0 \leq y \leq b, \tag{8.75a}$$

$$u(x,0) = u(x,b) \quad \text{e} \quad u_y(x,0) = u_y(x,b), \qquad 0 \leq x \leq a. \tag{8.75b}$$

Introduciamo ora lo spazio H^1_{per} delle funzioni reali ϕ e periodiche (nel senso precisato in (8.75)) e limitate in A rispetto alla norma seguente

$$\|\phi\|^2_{H^1_{\text{per}}} = \iint_A \left(|\phi(x,y)|^2 + \|\nabla\phi(x,y)\|^2\right) dxdy.$$

Da quanto già visto sugli elementi finiti segue che la soluzione debole in H^1_{per} del problema (8.73) è rappresentata dalla soluzione del problema infinito-dimensionale

$$-\iint \left(\nabla \cdot \frac{1}{\varepsilon}\nabla u\right) \varphi \, dxdy = \eta \iint_A u\varphi \, dxdy,$$

essendo φ una qualunque funzione in H^1_{per}. Come conseguenza della prima identità di Green e delle condizioni di periodicità della ε in A, la precedente equazione può essere espressa nella forma

$$\iint_A \frac{1}{\varepsilon} \nabla u \cdot \nabla\varphi \, dxdy = \eta \iint_A u\varphi \, dxdy. \tag{8.76}$$

Per la sua risoluzione numerica dobbiamo associare al problema infinito dimensionale (8.76) un sistema lineare 9finito-dimensionale0 la cui soluzione fornisce i coefficienti delle funzioni di base che compaiono nell'approssimazione finito dimensionale delle u in H^1_{per}. A tale scopo iniziamo con l'introdurre una griglia (per comodità) regolare di nodi del dominio A, ponendo

$$x_i = ih_x, \quad i = 0, 1, \ldots, n, \quad h_x = \frac{a}{n},$$

e

$$y_j = jh_y, \quad i = 0, 1, \ldots, m, \quad h_y = \frac{b}{m}.$$

Per tenere conto delle periodicità, tale griglia viene estesa a tutto $\mathbb{R}^2$ mediante i nodi

$$(x_{i+nr}, y_{j+ms}), \qquad t = 0, \pm1, \pm2, \ldots, \quad s = 0, \pm1, \pm2, \ldots.$$

A questo punto costruiamo una famiglia infinita di funzioni periodiche e interpolanti (in senso Lagrangian) nella griglia estesa. Questo significa che ad ogni punto nodale della griglia estesa associamo una funzione periodica che vale uno nel punto modale preso come riferimento e come zero in tutti gli altri. A tale scopo, al generico punto nodale (x_j, y_l) associamo la funzione

$$\varphi_{j,l}(x,y) = \sum_{r=-\infty}^{\infty} \sum_{s=-\infty}^{\infty} \varphi\left(\frac{x - x_{j+nr}}{h_x}\right) \varphi\left(\frac{y - y_{l+ms}}{h_x}\right),$$

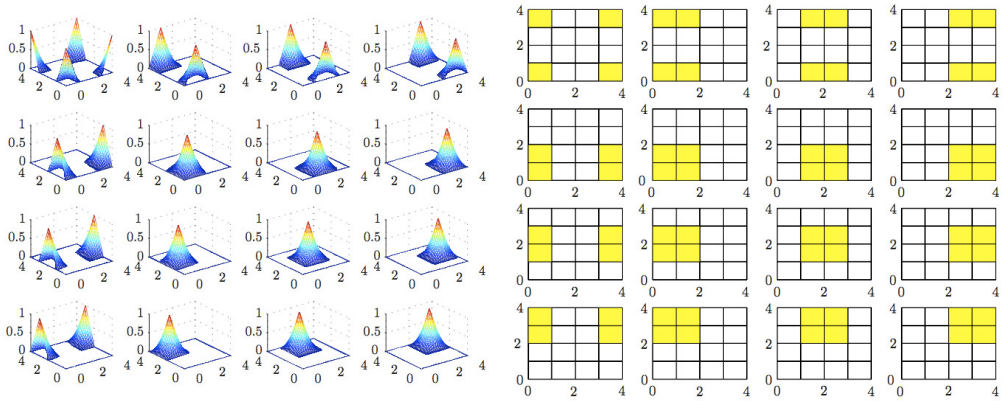


Figura 8.12: La figura a sinistra mostra alcune traslate della φ e i relativi supporti sono mostrati nella figura a destra nel caso di una griglia regolare 4×4.

essendo $\varphi(x)$ la spline lineare

$$\varphi(x) = \begin{cases} 1 - |x|, & -1 \le x \le 1, \\ 0, & x \le -1 \text{ o } x \ge 1. \end{cases}$$

Per facilitarne la visualizzazione, abbiamo riportato nella Fig. 8.12 la rappresentazione geometrica di alcune traslate della φ e dei relativi supporti, nell'ipotesi semplificativa di una griglia 4×4. Dalla costruzione segue immediatamente che ogni $\varphi_{j,l} \in H^1_{\text{per}}$ e soddisfa le condizioni di periodicità

$$\varphi_{j+nr,l}(x,y) = \varphi_{j,l+ms}(x,y) = \varphi_{j,l}(x,y), \qquad j,l = 0, \pm 1, \pm 2, \ldots.$$

È ugualmente immediato osservare che soddisfano la seguente condizione di *partizione dell'unità*

$$\sum_{j=0}^{n-1} \sum_{l=0}^{m-1} \varphi_{j,l}(x,y) = 1,$$

per ogni $(x,y) \in \mathbb{R}^2$, molto importante ai fini della valutazione dell'errore di approssimazione della soluzione effettiva con la tecniva degli elementivi finiti. Indicato ora con $S_{n,m}$ il sottospazio nm-dimensionale di H^1_{per}, espanso dalle nm splines lineari $\varphi_{j',l'}$, $j' = 0, 1, \ldots, n-1$, $l' = 0, 1, \ldots, m-1$, una generica funzione $v \in \mathcal{S}_{n,m}$ può essere rappresenta ponendo

$$v(x,y) = \sum_{j'=0}^{n-1} \sum_{l'=0}^{m-1} v_{j',l'} \varphi_{j',l'}(x,y),$$

essendo $v(x_{j'}, y_{l'}) = v_{j',l'}$, $j' = 0, 1, \ldots, n-1$; $l' = 0, 1, 2, \ldots, m-1$. La distribuzione in $\mathcal{S}_{n,m}$ del problema (8.76) fornisce un'approssimazione nm-dimensionale della soluzione infinito dimensionale, ossia della soluzione debole del problema iniziale. Di conseguenza lo spettro del problema nm-dimensionale così tende a rappresentare quello iniziale per $n, m \to\to \infty$. Tale problema spettrale, posto

$$u(x,y) = \sum_{j'=0}^{n-1} \sum_{l'=0}^{m-1} u_{j',l'} \varphi_{j',l'}(x,y), \tag{8.77}$$

si traduce nel seguente problema ad autovalori e autovalor:

$$\begin{aligned} &\sum_{j'=0}^{n-1} \sum_{l'=0}^{m-1} u_{j',l'} \int_0^a \int_0^b \frac{1}{\varepsilon(x,y)} \left(\nabla \varphi_{j',l'} \cdot \nabla \varphi_{j,l}\right) dxdy \\ &= \sum_{j'=0}^{n-1} \sum_{l'=0}^{m-1} u_{j',l'} \int_0^a \int_0^b \varphi_{j',l'}(x,y) \varphi_{j,l}(x,y)\, dxdy, \end{aligned} \tag{8.78}$$

$j = 0, 1, \ldots, n-1$ e $l = 0, 1, \ldots, m-1$.

Il sistema così ottenuto è della forma

$$A\boldsymbol{u} = \eta\, B\boldsymbol{u} \tag{8.79}$$

che rappresenta un problema autovalori/autovettori generalizzato, nel quale

$$\begin{aligned} A_{(j,l),(j',l')} &= \int_0^a \int_0^b \frac{1}{\varepsilon(x,y)} \left[\nabla \varphi_{j'\ldots l'} \cdot \nabla \varphi_{j,l}\right] dxdy, \\ B_{(j,l),(j',l')} &= \int_0^a \int_0^b \varphi_{j'\ldots l'} \varphi_{j,l}\, dxdy, \\ \boldsymbol{u} &= (u_{(j,l),(j',l')} = u_{(j,l),(j',l')}, \end{aligned}$$

and $j, j' = 0, 1, \ldots, n-1$ e $l' = 0, 1, \ldots, m-1$.

Da notare che la matrice A è simmetrica e semi-definita positiva (autovalori non negativi) e la matrice B è simmetrica e definita possitiva (autovalori tutti positivi), in quanto matrice di Gram di funzioni linearmente indipendenti. Per questo motivo è facile ricondurre il priblema ad autovalori/autofunzioni generalizzato ad un problema autovalori/autofunzioni classico. A tale scopo, decomposto B mediante la fattorizzazione di Cholesky, ossia posto $B = HH^T$, il sistema può essere scritto nella forma

$$A(H^T)^{-1}(H^T\boldsymbol{u}) = \eta\, H(H^T\boldsymbol{u}),$$

da cui, posto $\boldsymbol{v} = H^T\boldsymbol{u}$ e premoltiplicando il sistema per H^{-1}, si ottiene il classico problema ad autovalori/autofunzioni

$$H^{-1}A(H^T)^{-1}\boldsymbol{v} = \eta\, \boldsymbol{v} \iff \tilde{A}\boldsymbol{v} = \eta\, \boldsymbol{v}, \tag{8.80}$$

dove $\tilde{A} = H^{-1}A(H^T)^{-1} = \tilde{A}^T$.

8.8 Problemi parabolici e iperbolici

In questa sezione illustriamo sinteticamente la tecnica di risoluzione dei BVPs relativi alle PDEs di tipo parabolico e iperbolico. Per l'analisi delle condizioni di esistenza e unicità e delle proprietà della soluzione rinviamo a libri più specifici [9, 30, 8].

Caso parabolico. Per motivi di chiarezza, consideriamo inizialmente equazioni contenenti una sola variabile spaziale, ossia equazioni del tipo

$$\begin{cases} \dfrac{\partial u}{\partial t} = \dfrac{\partial}{\partial x}\left(k(x,t)\dfrac{\partial u}{\partial x}\right) - a_0(x,t)u + f(x,t), \quad a \le x \le b,\ 0 \le t \le T, \\ u(a,t) = f_1(t), \quad u(b,t) = f_2(t), \quad 0 \le t \le T, \\ u(x,0) = g(x), \quad a \le x \le b, \end{cases} \tag{8.81}$$

dove $k(x,t) \ge 0$ su $[a,b] \times [0,T]$.

La prima operazione consiste nel trasformare il problema iniziale in uno "equivalente" avente condizioni di omogeneità per $x = a$ e $x = b$. A tale scopo è sufficiente adottare la trasformazione

$$u(x,t) = v(x,t) + \varphi(x,t),$$

dove φ è la seguente interpolante lineare in x di $f_1(t)$ e $f_2(t)$ in $x = a$ e $x = b$:

$$\varphi(t) = f_1(t) + \frac{x-a}{b-a}\left[f_2(t) - f_1(t)\right] = f_2(t) + \frac{b-x}{b-a}\left[f_1(t) - f_2(t)\right].$$

Sostituendo nella (8.81) la u con $v + \varphi$, si ottiene un'equazione ad essa analoga nella quale la funzione incognita v soddisfa condizioni di omogeneità in $x = a$ e $x = b$. Più precisamente si ottiene il seguente problema:

$$\begin{cases} \dfrac{\partial v}{\partial t} = \dfrac{\partial}{\partial x}\left(k(x,t)\dfrac{\partial v}{\partial x}\right) - a_0(x,t)v + h(x,t), \quad a \le x \le b,\ 0 \le t \le T, \\ v(a,t) = v(b,t) = 0, \quad 0 \le t \le T, \\ v(x,0) = \hat{g}(x), \quad a \le x \le b, \end{cases} \tag{8.82}$$

dove $h(x,t)$ e $\hat{g}(x)$ sono funzioni note ottenute dalla (8.81) per semplice sostituzione della u con la $v + \varphi$. Indicata con w una funzione test dello spazio di Sobolev $H_0^1[a,b]$ ($w \in H_0^1[a,b]$), moltiplicando primo e secondo membro dell'equazione differenziale (8.82) per w e integrando in $[a,b]$ otteniamo

l'equazione

$$\int_a^b \frac{\partial v}{\partial t} w(x)\,dx = \left[k(x,t)\frac{\partial v}{\partial x} w\right]_a^b - \int_a^b k(x,t)\frac{\partial v}{\partial x} w'(x)\,dx \\ - \int_a^b a_0(x,t) v w\,dx + \int_a^b h(x,t) w\,dx.$$

Di conseguenza, tenendo conto del fatto che $w(a) = w(b) = 0$, si ottiene la forma variazionale

$$\int_a^b \frac{\partial v}{\partial t} w(x)\,dx = -\int_a^b k(x,t)\frac{\partial v}{\partial x} w'(x)\,dx - \int_a^b a_0(x,t) v w\,dx + \int_a^b h(x,t) w\,dx. \tag{8.83}$$

Per proiettare la (8.83) in uno spazio finito dimensionale, suddiviso l'intervallo $[a,b]$ mediante i punti nodali $x_i = a + ih$, $i = 0, 1, \ldots, n+1$, $h = \frac{b-a}{n+1}$, costruiamo la base n-dimensionale di splines lineari $\{H_i(x)\}_{i=1}^n$, contenuta in $H_0^1[a,b]$.

Approssimiamo quindi la funzione incognita v con la funzione

$$v_h(x,t) = \sum_{j=1}^n v_{h,j}(t) H_j(x), \tag{8.84}$$

dove, per la cardinalità delle $\{H_j(x)\}_{j=1}^n$, $v_{h,j}(x) = v_h(x_j, t)$, $j = 1, 2, \ldots, n$. Sostituendo la v_h nella (8.83) e ponendo $w(x) = H_i(x)$, per $i = 1, 2, \ldots, n$ otteniamo il sistema

$$\sum_{j=1}^n v'_{h,j}(t) \int_a^b H_j(x) H_i(x)\,dx + \sum_{j=1}^n v_{h,j}(t) \int_a^b k(x,t) H'_j(x) H'_i(x)\,dx \\ = -\sum_{j=1}^n v_{h,j}(t) \int_a^b a_0(x,t) H_j(x) H_i(x)\,dx + \int_a^b h(x,t) H_i(x)\,dx, \tag{8.85}$$

nel quale $H'_i(x)$ e $H'_j(x)$ indicano derivate in senso debole delle rispettive funzioni. Ponendo per $i, j = 1, 2, \ldots, n$

$$(\boldsymbol{M}_h)_{ij} = \int_a^b H_j(x) H_i(x)\,dx,$$
$$\boldsymbol{K}_{ij}(t) = \int_a^b \left[k(x,t) H'_j(x) H'_i(x) - a_0(x,t) H_j(x) H_i(x)\right] dx,$$
$$\boldsymbol{b}_{h,i}(t) = \int_a^b h(x,t) H_i(x)\,dx,$$

il sistema (8.85) assume la seguente forma matriciale:

$$\boldsymbol{M}_h \boldsymbol{v}_h' + \boldsymbol{K}_h \boldsymbol{v}_h = \boldsymbol{b}_h, \tag{8.86}$$

dove le matrici $\boldsymbol{M}_h$ e $\boldsymbol{K}_h$ sono chiaramente simmetriche e a banda, dato che $H_i H_j$, come pure $H_i' H_j'$, sono nulle per $|i-j| \geq 2$. La matrice $\boldsymbol{M}_h$ è anche non singolare in quanto è la matrice di Gram relativa alle splines lineari $\{H_i\}_{i=1}^n$ [Teorema 5.2].

Il sistema (8.86) è un sistema di ODEs lineari del primo ordine. Di conseguenza per l'unicità della soluzione occorre fissare la condizione iniziale per ciascuna delle n funzioni $(v_h)_i(t)$, $i = 1, 2, \ldots, n$. Questo è immediato, dato che $v(x, 0) = \hat{g}(x)$, $a \leq x \leq b$. Infatti, essendo $v_h(x, t)$ l'approssimante cercata di $v(x, t)$, è naturale richiedere che

$$(\boldsymbol{v}_h)_i = v_h(x_i, 0) = \hat{g}(x_i), \qquad i = 1, 2, \ldots, n. \tag{8.87}$$

Sulla risoluzione numerica del sistema (8.86) esistono numerose tecniche. Per una panoramica dei metodi più diffusi si rinvia ai libri di Analisi Numerica [17, 28], mentre per una visione più completa e specialistica si rinvia ai libri [10, 11, 31].

Esercizio 8.12 Illustrare il procedimento per la risoluzione del seguente problema parabolico:

$$\begin{cases} \dfrac{\partial u}{\partial t} = \dfrac{\partial}{\partial x}\left([2+\cos xt] \dfrac{\partial u}{\partial x} \right) + (1+xt)\dfrac{\partial u}{\partial x} - (x \sin t)u, \ x \in [-4, 4], \ t \in [0, 5], \\ u(-4, t) = f_1(t), \quad u(4, t) = f_2(t), \\ u(x, 0) = g(x). \end{cases}$$

Iniziamo con il ricondurre il problema alla forma canonica (omogeneità in $u(-4, t)$ e $u(4, t)$). A tale scopo poniamo

$$u(x, t) = v(x, t) + \varphi(x, t), \qquad \varphi(x, t) = f_1(t) + \frac{x+4}{8}\left[f_2(t) - f_1(t) \right],$$

da cui segue che $v(-4, t) = v(4, t) = 0$. Sostituendo nell'equazione la u abbiamo

$$\begin{aligned} \frac{\partial v}{\partial t} + f'(t) + \frac{x+4}{8}\left[f_2'(t) - f_1'(t) \right] = \frac{\partial}{\partial x}\left[(2+\cos xt)\left(\frac{\partial v}{\partial x} + \frac{1}{8}(f_2(t) - f_1(t)) \right) \right] \\ + (1+xt)\left(\frac{\partial v}{\partial x} + \frac{1}{8}(f_2(t) - f_1(t)) \right] - (x \sin t)(v + \varphi). \end{aligned}$$

Da cui, riordinando e indicando con $h(x,t)$ una funzione nota segue la forma canonica

$$\begin{cases} \dfrac{\partial v}{\partial t} = \dfrac{\partial}{\partial x}\left([2+\cos xt]\dfrac{\partial v}{\partial x}\right) + (1+xt)\dfrac{\partial v}{\partial x} - (x\,\sin t)v + h(x,t), \\ v(-4,t) = v(4,t) = 0, \\ v(x,0) = g(x) - \varphi(x,0). \end{cases}$$

Indicata quindi con $w \in H_0^1[-4,4]$ una funzione test, moltiplicando primo e secondo membro per w e integrando in $(-4,4)$ si ottiene

$$\begin{aligned} \int_{-4}^{4} \frac{\partial v}{\partial t} w(x)\,dx = \int_{-4}^{4} \frac{\partial}{\partial x}\left([2+\cos xt]\frac{\partial v}{\partial x}\right) w(x)\,dx + \int_{-4}^{4} (1+xt)\frac{\partial v}{\partial x} w(x)\,dx \\ - \int_{-4}^{4} (x\,\sin t) v(x,t) w(x)\,dx + \int_{-4}^{4} h(x,t) w(x)\,dx. \end{aligned}$$

Integrando per parti il secondo e terzo integrale e tenendo conto della invertibilità della derivazione e integrazione nel primo, essendo $w(-4,t) = w(4,t) = 0$, otteniamo

$$\begin{aligned} \frac{d}{dt}\int_{-4}^{4} v(x,t) w(x)\,dx = &- \int_{-4}^{4} (2+\cos xt)\frac{\partial v}{\partial x} w'(x)\,dx \\ &- \int_{-4}^{4} (1+xt) v(x,t) w'(x)\,dx \\ &- \int_{-4}^{4} (x\,\sin t) v(x,t) w(x)\,dx + \int_{-4}^{4} h(x,t) w(x)\,dx. \end{aligned}$$

La funzione v è la soluzione debole del problema nella forma canonica se e solo se è soluzione del precedente problema, qualunque sia $w \in H_0^1[-4,4]$, con le condizioni $v(-4,t) = v(4,t) = 0$ e $v(x,0) = g(x) - w(x,0)$. Si tratta ora di approssimare la soluzione nel senso degli elementi finiti.

A tal fine, costruita una base $\{H_i(x)\}_{i=1}^n$ di splines lineari, si approssima $v(x,t)$ con

$$v_n(x,t) = \sum_{i=1}^{n} a_i(t) H_i(x),$$

dove $a_i(t) = v_n(x_i,t)$. Di conseguenza,

$$\frac{\partial v_n}{\partial t} = \sum_{j=1}^{n} a_j'(t) H_j(x), \frac{\partial v_n}{\partial x} = \sum_{j=1}^{n} a_j(t) H_j'(x).$$

Ponendo $w = H_i$, $i = 1, 2, \ldots, n$, si ottiene il sistema

$$\sum_{j=1}^{n} \left(\int_{-4}^{4} H_i(x) H_j(x)\, dx \right) a_j'(t) + \sum_{j=1}^{n} a_j(t) \int_{-4}^{4} \big[(2 + \cos xt) H_i'(x) H_j'(x) + (1 + xt) H_i(x) H_j'(x) + (x \sin t) H_i(x) H_j(x) \big]\, dx = \int_{-4}^{4} h(x,t) H_i(x)\, dx,$$

rappresentabile nella forma

$$\sum_{j=1}^{n} \boldsymbol{M}_{ij} a_j'(t) + \sum_{j=1}^{n} \boldsymbol{K}_{ij}(t) a_j(t) = \boldsymbol{b}_i(t), \qquad i = 1, 2, \ldots, n,$$

che, nella notazione matriciale diventa

$$\boldsymbol{M}\boldsymbol{a}'(t) + \boldsymbol{K}(t)\boldsymbol{a}(t) = \boldsymbol{b}(t).$$

Siamo così a un sistema differenziale di primo ordine di n equazioni nelle n incognite $\{a_i(t)\}_{i=1}^{n}$. Per la sua risoluzione occorre fissare le condizioni iniziali, ottenibili imponendo che sia

$$v_n(x_i, 0) = \sum_{j=1}^{n} a_j(0) H_j(x_i) = a_i(0) = g(x_i) - w(x_i), \qquad i = 1, 2, \ldots, n.$$

Ci si riconduce, in tal modo, alla risoluzione del problema di Cauchy

$$\begin{cases} \boldsymbol{M}\, \boldsymbol{a}'(t) + \boldsymbol{K}(t)\boldsymbol{a}(t) = \boldsymbol{b}(t), \\ a_i(0) = g(x_i) - w(x_i), \quad i = 1, 2, \ldots, n. \end{cases}$$

Per le tecniche per la sua risoluzione si veda il Capitolo 2.

Caso parabolico con due variabili spaziali. Consideriamo ora il seguente BVP:

$$\begin{cases} \dfrac{\partial u}{\partial t} - \nabla \cdot (k\, \nabla u) + a_0(x, y; t) u = f(x, y; t), \quad (x, y) \in \Omega,\ 0 \le t \le T, \\ u(x, y; t) = g(x, y; t), \quad \text{per } (x, y) \in \partial\Omega \text{ e } 0 \le t \le T, \\ u(x, y; 0) = f_1(x, y), \quad \text{per } (x, y) \in \Omega. \end{cases} \tag{8.88}$$

Come nel caso precedente, si trasforma inizialmente il problema dato in uno analogo con condizioni di omogeneità sulla frontiera.

A tale scopo associamo al dominio Ω una mesh Ω_h, esattamente come nel caso ellittico. Indicato con $z_j = (x_j, y_j)$, $j = 1, 2, \ldots, N_v$, l'insieme dei suoi punti nodali, costruiamo una funzione ausiliaria $F(x, y; t)$ che soddisfa le seguenti condizioni di interpolazione:

$$F(x, y; t) = \begin{cases} g(z_j; t), & \text{per } z_j \in \partial\Omega_h, \\ 0, & \text{per ogni punto nodale interno.} \end{cases}$$

Per la sua definizione analitica, ad ogni punto nodale z_j su $\partial\Omega$, $j = 1, 2, \ldots, \overline{n}_r$, associamo una box-spline che assume il valore 1 in z_j e zero in tutti gli altri punti nodali della mesh Ω_h. Tale funzione può pertanto essere così definita:

$$F(x, y; t) = \sum_{j=1}^{\overline{n}_r} g(z_j; t)\varphi_j(x, y). \tag{8.89}$$

Di conseguenza, ponendo $u = v + F$, è evidente che $v(x, y; t) = 0$ per $(x, y) \in \partial\Omega_h$ e $t \in [0, T]$. Effettuando tale sostituzione nella (8.88), otteniamo un BVP del tipo

$$\begin{cases} \dfrac{\partial v}{\partial t} - \nabla \cdot (k\, \nabla v) + a_0(x, y; t)v = h(x, y; t), \quad \text{su } \Omega, \\ v(x, y; t) = 0, \quad \text{su } \partial\Omega \times [0, T], \\ v(x, y; 0) = \hat{f}_1(x, y), \text{per } (x, y) \in \Omega, \end{cases} \tag{8.90}$$

dove v è la funzione incognita, mentre h e $\hat{f}_1$ sono funzioni note derivanti dalle funzioni note dalla (8.88), dalla F e sue derivate parziali F_x e F_y (considerate in senso debole). Indicata quindi con w una funzione test dello spazio di Sobolev $H_0^1(\Omega)$, moltiplichiamo primo e secondo membro dell'equazione in (8.90) per w e integriamo in Ω. Applicando quindi la prima identità di Green, otteniamo l'equazione

$$\begin{aligned} \int_\Omega \frac{\partial v}{\partial t} w\, dxdy - \int_{\partial\Omega} k\, \frac{\partial v}{\partial n} w\, d\sigma + \int_\Omega k(x, y; t)\nabla v \cdot \nabla w\, dxdy \\ + \int_\Omega a_0(x, y; t)vw\, dxdy = \int_\Omega h(x, y; t)w\, dxdy, \end{aligned} \tag{8.91}$$

dalla quale, ricordando che $w \in H_0^1(\Omega)$,

$$\begin{aligned} \int_\Omega \frac{\partial v}{\partial t} w\, dxdy + \int_\Omega k(x, y; t)\nabla v \cdot \nabla w\, dxdy \\ + \int_\Omega a_0(x, y; t)vw\, dxdy = \int_\Omega h(x, y; t)w\, dxdy, \end{aligned} \tag{8.92}$$

per ogni $w \in H_0^1(\Omega)$.

Per proiettare la (8.92) in uno spazio finito dimensionale, che consenta di approssimare la soluzione mediante gli elementi finiti, ad ogni punto nodale interno z_i, $i = 1, 2, \ldots, n_v$, della mesh Ω_h associamo la box-spline $\varphi_i(x, y)$ che assume il valore 1 in z_i e zero in tutti gli altri punti nodali (interni e di frontiera) di Ω_h. La soluzione v del problema (8.92) viene quindi approssimata con la funzione

$$v_h(x, y; t) = \sum_{j=1}^{n_v} v_{h,j}(t)\varphi_j(x, y) \tag{8.93}$$

nella quale, per la cardinalità delle box-splines utilizzate, $v_{h,j}(t) = v_h(z_j; t)$, $j = 1, 2, \ldots, n_v$. Sostituendo quindi la v_h nella (8.92) e ponendo $w = \varphi_i(x, y)$, $i = 1, 2, \ldots, n_v$, per $i = 1, 2, \ldots, n_v$ otteniamo il sistema di ODEs

$$\begin{aligned}\sum_{j=1}^{n_v} v'_{h,j}(t) \int_{\Omega_h} \varphi_j(x, y)\varphi_i(x, y)\, dxdy + \sum_{j=1}^{n_v} \int_{\Omega_h} k(x, y; t)\nabla\varphi_i \cdot \nabla\varphi_j\, dxdy \\ + \sum_{j=1}^{n_v} v_{h,j}(t) \int_{\Omega_h} a_0(x, y; t)\varphi_j(x, y)\varphi_i(x, y)\, dxdy = \int_{\Omega_h} h(x, y; t)\varphi_i(x, y)\, dxdy.\end{aligned} \tag{8.94}$$

Ponendo infine

$$\begin{aligned}(\boldsymbol{M}_h)_{ij} &= \int_{\Omega_h} \varphi_i(x, y)\varphi_j(x, y)\, dxdy, \\ (\boldsymbol{K}_h)_{ij} &= \int_{\Omega_h} [k(x, y; t)\nabla\varphi_i(x, y) \cdot \nabla\varphi_j(x, y) + a_0(x, y; t)\varphi_i(x, y)\varphi_j(x, y)]\, dxdy, \\ \boldsymbol{b}_{h,i}(t) &= \int_{\Omega_h} h(x, y; t)\varphi_i(x, y)\, dxdy,\end{aligned}$$

il sistema (8.94) può essere scritto nella forma matriciale

$$\boldsymbol{M}_h\boldsymbol{v}'_h(t) + \boldsymbol{K}_h(t)\boldsymbol{v}_h(t) = \boldsymbol{b}_h(t), \tag{8.95}$$

nel quale le matrici $\boldsymbol{M}_h$ e $\boldsymbol{K}_h(t)$ sono a banda, in conseguenza dei supporti minimali della φ_i e φ_j. La matrice $\boldsymbol{M}_h$ è non singolare essendo la matrice di Gram delle box-splines [Teorema 5.2]. Per l'unicità delle soluzione è chiaramente necessario la condizione iniziale per ogni componente $v_{h,j}(t)$ del vettore $\boldsymbol{v}_h(t)$. Tali condizioni, tenuto conto della condizione $v(x, y; 0) = \hat{f}_1(x, y)$, per $(x, y) \in \Omega$, sono le seguenti:

$$v_{h,j}(0) = \hat{f}_1(z_j) = \hat{f}_1(x_j, y_j), \qquad j = 1, 2, \ldots, n_v.$$

Caso iperbolico in una variabile spaziale. In questo caso il modello è:

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} = \dfrac{\partial}{\partial x}\left(k(x,t)\dfrac{\partial u}{\partial x}\right) - a_0(x,t)u + f(x,t), \\ \quad a \le x \le b,\ 0 \le t \le T, \\ u(a,t) = f_1(t), \quad u(b,t) = f_2(t), \quad 0 \le t \le T, \\ u(x,0) = g_1(x), \quad u_t(x,0) = g_2(x), \quad a \le x \le b. \end{cases} \tag{8.96}$$

Esso differisce, in particolare, dall'analogo parabolico per la presenza della velocità iniziale $u_t(x,0) = g_2(x)$. Anche in questo caso, ponendo $u(x,t) = v(x,t) + \varphi(x,t)$, dove φ è l'interpolante lineare di $f_1(t)$ e $f_2(t)$ in $[a,b]$, il problema iniziale viene trasformato nel seguente (con condizioni di omogeneità della v in $x = a$ e $x = b$):

$$\begin{cases} \dfrac{\partial^2 v}{\partial t^2} = \dfrac{\partial}{\partial x}\left(k(x,t)\dfrac{\partial v}{\partial x}\right) - a_0(x,t)v + h(x,t), \quad a \le x \le b,\ 0 \le t \le T, \\ v(a,t) = v(b,t) = 0, \quad 0 \le t \le T, \\ v(x,0) = \hat{g}_1(x), \quad v_t(x,0) = \hat{g}_2(x), \quad a \le x \le b, \end{cases} \tag{8.97}$$

dove $h(x,t)$, $\hat{g}_1(x)$ e $\hat{g}_2(x)$ sono funzioni note derivanti dalla semplice sostituzione nel modello (8.96) della u con $v+\varphi$. Indicata quindi con $w \in H_0^1[a,b]$ una funzione test, moltiplicando primo e secondo membro della (8.97) per w, integrando in $[a,b]$ e tenendo conto della condizione $w(a) = w(b) = 0$, otteniamo la forma variazionale

$$\int_a^b \frac{\partial^2 v}{\partial t^2} w(x)\,dx + \int_a^b k(x,t)\frac{\partial v}{\partial x} w'(x)\,dx - \int_a^b a_0(x,t)vw\,dx = \int_a^b h(x,t)w(x)\,dx, \tag{8.98}$$

per ogni $w \in H_0^1[a,b]$.

A questo punto, per risolvere la (8.98) con il metodo degli elementi finiti, approssimiamo la v con la

$$v_h(x,t) = \sum_{j=1}^{n} v_{h,j}(t) H_j(x), \tag{8.99}$$

dove $v_{h,j}(x,t) = v_h(x_j,t)$ e H_j è la j-esima spline relativa alla decomposizione di $[a,b]$. Sostituendo nella (8.98) v_h alla v e H_i, $i = 1,2,\ldots,n$, alla w, per

$i = 1, 2, \ldots, n$ otteniamo il sistema di ODEs del secondo ordine

$$\sum_{j=1}^{n} v''_{h,j}(t) \int_a^b H_i(x) H_j(x)\, dx + \sum_{j=1}^{n} v_{h,j}(t) \int_a^b k(x,t) H'_i(x) H'_j(x)\, dx$$
$$- \sum_{j=1}^{n} v_{h,j}(t) \int_a^b a_0(x,t) H_i(x) H_j(x)\, dx = \int_a^b h(x,t) H_i(x)\, dx. \qquad (8.100)$$

Utilizzando quindi la stessa notazione usata nel problema parabolico, otteniamo il seguente sistema di n ODE del secondo ordine:

$$\boldsymbol{M}_h \boldsymbol{v}''_h(t) + \boldsymbol{K}_h \boldsymbol{v}_h(t) = \boldsymbol{b}_h(t), \qquad (8.101)$$

per la cui unicità occorre fissare i valori di $v_{h,i}(0)$ e $(\frac{\partial v_h}{\partial t})_i(0)$. Questi si ottengono facilmente dalle condizioni iniziali fissate per il modello. Più precisamente risulta

$$v_{h,i}(0) = \hat{g}_1(x_i), \quad \text{e} \quad \left(\frac{\partial v_h}{\partial t}\right)_i (0) = \hat{g}_2(x_i), \qquad i = 1, 2, \ldots, n.$$

Le precedenti considerazioni evidenziano che non esistono sostanziali differenze nell'applicazione del metodo agli elementi finiti alla risoluzione dei problemi parabolico e iperbolico con una variabile spaziale. La differenza più importante riguarda l'ordine della ODE che si deve risolvere nei due casi.

Considerazioni del tutto analoghe valgono nella risoluzione degli stessi tipi di problemi con due variabili spaziali.

Per i riferimenti alle tecniche di risoluzione del sistema (8.101) si rinvia al Capitolo 2.

Appendice A

RISULTATI ESSENZIALI DI ANALISI FUNZIONALE

Spazi normati. Sia X uno spazio lineare complesso (o reale). Una funzione $\|\cdot\| : X \to \mathbb{R}$ che soddisfa le seguenti proprietà:

(1)	$\|\varphi\| \geq 0$	(positività)
(2)	$\|\varphi\| = 0$ se e solo se $\varphi = 0$	(definitezza)
(3)	$\|\alpha\varphi\| = \|\alpha\| \, \|\varphi\|$	(omogeneità)
(4)	$\|\varphi + \psi\| \leq \|\varphi\| + \|\psi\|$	(disuguaglianza triangolare)

per qualunque $\varphi, \psi \in X$ e per ogni $\alpha \in \mathbb{C}$ (oppure $\mathbb{R}$) è definita *norma* su X. Dalle (3)-(4) segue immediatamente che $|\|\varphi\| - \|\psi\|| \leq \|\varphi - \psi\|$.

In uno spazio normato, per distanza di φ da ψ si intende la $\|\varphi - \psi\|$.

Convergenza. Una successione $\{\varphi_n\}_{n=1}^{\infty}$ di elementi di uno spazio normato X è detto *convergente* in X se esiste un elemento $\varphi \in X$ tale che

$$\lim_{n\to\infty} \|\varphi_n - \varphi\| = 0,$$

ossia se, per ogni $\varepsilon > 0$, esiste un intero $n(\varepsilon)$ tale che $\|\varphi_n - \varphi\| < \varepsilon$ per ogni $n > n(\varepsilon)$.

Successione di Cauchy. Una successione $\{\varphi_n\}_{n=1}^{\infty}$ di elementi di uno spazio normato X è detta di Cauchy se, per ogni $\varepsilon > 0$, esiste un intero $n(\varepsilon)$ tale che $\|\varphi_n - \varphi_m\| < \varepsilon$ per tutti gli $n, m > n(\varepsilon)$, ossia se $\lim_{n,m\to\infty} \|\varphi_n - \varphi_m\| = 0$.

Completezza. Un sottospazio lineare U di uno spazio normato X è detto *completo* se ogni successione di Cauchy di elementi di U converge ad un elemento di U.

Spazio di Banach. Uno spazio normato X è uno *spazio di Banach* se esso è completo.

Continuità. Una funzione $A : U \subset X \to Y$ che trasforma gli elementi di un sottoinsieme U di uno spazio normato X in elementi di uno spazio normato Y è detta *continua* in $\varphi \in U$ se $\lim_{n\to\infty} A\varphi_n = A\varphi$ per ogni successione $\{\varphi_n\}_{n=1}^{\infty} \subset U$ con $\lim_{n\to\infty} \varphi_n = \varphi$.

La funzione A è detta continua se essa è continua in ogni $\varphi \in U$. La precedente definizione può essere espressa anche nel modo seguente: una funzione $A : U \subset X \to Y$ è continua in φ se per ogni $\varepsilon > 0$ esiste $\delta(\varepsilon, \varphi)$ tale che $\|A\varphi - A\psi\|_Y < \varepsilon$ per tutti i $\psi \in U$ con $\|\varphi - \psi\|_X < \delta$.

Continuità Uniforme. La funzione A è detta *uniformemente continua* se δ dipende unicamente da ε, ossia se per ogni $\varepsilon > 0$ esiste un $\delta(\varepsilon)$ tale che $\|A\varphi - A\psi\|_Y < \varepsilon$ per tutte le φ e ψ con $\|\varphi - \psi\|_X < \delta$.

Esempi di spazi di Banach.

1. Indicato con Ω un insieme chiuso e limitato di $\mathbb{R}^n$ ($\Omega \subset \mathbb{R}^n$), sia $C(\Omega)$ l'insieme delle funzioni continue in Ω. Allora, indicato con $\mathbb{R}^+$ l'insieme dei numeri reali nonnegativi, la funzione $\|\cdot\|_\infty : C(\Omega) \to \mathbb{R}^+$, con
$$\|f\|_\infty = \max_{\boldsymbol{x}\in\Omega} |f(\boldsymbol{x})|,$$
introduce una norma completa in $C(\Omega)$, per cui lo spazio $C(\Omega)$, dotato di tale norma, è uno spazio normato completo e quindi uno spazio di Banach.

2. La suddetta definizione si può generalizzare ai sottoinsiemi Ω di $\mathbb{R}^n$ che non sono necessariamente chiusi e limitati. In tal caso, indicato con $\hat{C}(\Omega)$ l'insieme delle funzioni continue e **limitate** in Ω, la funzione $\|\cdot\|_\infty : \hat{C}(\Omega) \to \mathbb{R}^+$, con
$$\|f\|_\infty = \sup_{\boldsymbol{x}\in\Omega} |f(\boldsymbol{x})|,$$
introduce una norma completa in $\hat{C}(\Omega)$. Di conseguenza, lo spazio $\hat{C}(\Omega)$, dotato di tale norma, è uno spazio di Banach.

3. Indicato con Ω un sottoinsieme misurabile in $\mathbb{R}^n$, sia $L^2(\Omega)$ lo spazio delle funzioni al quadrato sommabili (nel senso di Lebesgue) in Ω. Supponiamo ora che due funzioni φ e ψ in $L^2(\Omega)$, che assumano valori diversi soltanto su un sottoinsieme di Ω di misura nulla, vengano identificate, dato che $\int_\Omega |\varphi - \psi|\, d\boldsymbol{x} = 0$. Sotto tale ipotesi, la funzione $\|\cdot\|_2 : L^2(\Omega) \to \mathbb{R}^+$, con
$$\|f\|_2 = \left(\int_\Omega |f(\boldsymbol{x})|^2 \, d\boldsymbol{x} \right)^{\frac{1}{2}},$$

definisce una norma completa in $L^2(\Omega)$, che pertanto costituisce uno spazio di Banach.

Sfera aperta e chiusa. Per un elemento φ di uno spazio normato X e un numero positivo r, l'insieme $B(\varphi, r) = \{\psi \in X : \|\varphi - \psi\| < r\}$ definisce la *sfera aperta* di raggio r e centro φ. L'insieme $B[\varphi, r] = \{\psi \in X : \|\varphi - \psi\| \leq r\}$ definisce la *sfera chiusa* di raggio r e centro φ.

Insieme aperto. Un sottoinsieme di uno spazio normato X è definito *aperto* se per ogni $\varphi \in U$ esiste un $r > 0$ tale che $B(\varphi; r) \subset U$.

Parte interna. La *parte interna* $\dot{U}$ di un sottoinsieme U di uno spazio normato X è il sottoinsieme aperto più grande contenuto in U. Esso consiste in tutti i punti $\varphi \in U$ per cui esiste un numero $r = r(\varphi)$ tale che $B(\varphi, r) \subset U$.

Insieme chiuso. Un sottoinsieme U di uno spazio normato X è definito *chiuso* se esso contiene tutti i limiti di tutte le successioni con termini in U e limiti in X.

Chiusura. La *chiusura* $\overline{U}$ di un sottoinsieme U di uno spazio normato X (in X) è l'insieme di tutti i limiti delle successioni con termini in U e limiti in X. Essa è pertanto il sottoinsieme chiuso più piccolo di X contenente U.

Frontiera. La *frontiera* ∂U di un sottoinsieme U di uno spazio normato X è l'insieme di tutti gli elementi di X che sono limiti sia di una successione con termini in U, sia di una successione con termini in $X \setminus U$. Infatti

$$\partial U = \overline{U} \cap \overline{(X \setminus U)} = \partial(X \setminus U).$$

Densità e separabilità. Un insieme U è definito *denso* in V se $V \subset \overline{U}$, cioè se ogni elemento di V è il limite di una successione convergente di elementi di U. Uno spazio normato X è detto *separabile* se contiene un sottoinsieme numerabile denso di X.

Limitatezza. Un sottoinsieme U di uno spazio normato X è detto *limitato* se esiste una costante positiva C tale che $\|\varphi\| \leq C$ per tutti i $\varphi \in U$. In altre parole, un sottoinsieme U di uno spazio normato X è limitato se esso è sottoinsieme di una sfera con raggio $r \in \mathbb{R}^+$.

Prodotto scalare (prodotto interno). Sia X uno spazio lineare complesso (o reale). Allora una funzione $(\cdot, \cdot) : X \times X \to \mathbb{C}$ (oppure $\mathbb{R}$) soddisfacente le

seguenti proprietà:

(1)	$(\varphi, \varphi) \geq 0$	(positività)
(2)	$(\varphi, \varphi) = 0$ se e solo se $\varphi = 0$	(definitezza)
(3)	$(\varphi, \psi) = \overline{(\psi, \varphi)}$	(simmetria)
(4)	$(\alpha\varphi + \beta\psi, \chi) = \alpha(\varphi, \chi) + \beta(\psi, \chi)$	(linearità)

per tutte le $\varphi, \psi, \chi \in X$ e $\alpha, \beta \in \mathbb{C}$ (oppure $\mathbb{R}$) è definita *prodotto scalare* o *prodotto interno* (il soprasegno indica il complesso coniugato). Dalle (3)-(4) segue immediatamente la sesquilinearità, o bilinearità

$$(\varphi, \alpha\psi + \beta\chi) = \overline{\alpha}(\varphi, \psi) + \overline{\beta}(\varphi, \chi).$$

Norma indotta. Ogni prodotto scalare induce una norma, così definita

$$\|\varphi\| = \sqrt{(\varphi, \varphi)}$$

per ogni $\varphi \in X$. Per ogni coppia di elementi φ e ψ di X, vale inoltre la cosiddetta *disuguaglianza di Schwartz*

$$|(\varphi, \psi)| \leq \|\varphi\| \, \|\psi\|.$$

Spazio pre-Hilbert. Per *spazio pre-Hilbert* si intende uno spazio lineare dotato di prodotto interno.

Spazio di Hilbert. Si definisce *spazio di Hilbert* uno spazio pre-Hilbert completo rispetto alla norma indotta dal suo prodotto scalare.

Esempi di spazi di Hilbert.

1. Lo spazio $L^2(\Omega)$ è uno spazio di Hilbert rispetto al prodotto interno

$$(\varphi, \psi) = \int_\Omega \varphi(\boldsymbol{x})\overline{\psi(\boldsymbol{x})}\, d\boldsymbol{x}.$$

2. Lo spazio l^2 delle successioni complesse $\{x_n\}_{n=1}^\infty$ è uno spazio di Hilbert rispetto al suo prodotto interno

$$(\{x_n\}_{n=1}^\infty, \{y_n\}_{n=1}^\infty) = \sum_{n=1}^\infty x_n \, \overline{y}_n.$$

Bibliografia

[1] C. de Boor, *A Practical Guide to Splines*, Revised edition, Applied Mathematical Sciences **27**, Springer, 2001.

[2] S.C. Brennes and L.R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer, New York, 1994.

[3] J.C. Butcher, *The Numerical Analysis of Ordinary Differential Equations*: *Runge-Kutta and General Linear Methods*, Wiley, Chichester, 1987.

[4] K. Chandrasekharan, *Classical Fourier Transform*, Universitext, Springer Verlag, Berlin, 1989.

[5] Ph. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978. Reprinted as Classics in Applied Mathematics **40**, SIAM, Philadelphia, 2002.

[6] Pietro Contu, C. van der Mee, and Sebastiano Seatzu, *Fast and Effective Finite Difference Method for* $2D$ *Photonic Crystals*, Communications in Applied and Industrial Mathematics (CAIM) **2**(2), 2011, 18 pp. doi: 10.1685/journal.caim.374.

[7] Pietro Contu, C. van der Mee, and Sebastiano Seatzu, *A finite element frequency domain method for* $2D$ *photonic crystals*, Journal of Computational and Applied Mathematics **236**, 3956–3966 (2012). doi: 10.1016/j.cam.2012.02.041

[8] M.S. Gockenbach, *Partial Differential Equations*. *Analytical and Numerical Methods*, SIAM, Philadelphia, 2002.

[9] M.S. Gockenbach, *Understanding and Implementing the Finite Element Method*, SIAM, Philadelphia 2006; 2° ed., 2011.

[10] G.H. Golub and G. Meurant, *Matrices, Moments, and Quadrature with Applications*, Princeton Series in Applied Mathematics, Princeton University Press, Princeton, 2010.

[11] G.H. Golub and Ch.F. Van Loan, *Matrix Computations*, John Hopkins Studies in the Mathematical Sciences, John Hopkins University Press, Baltimore, 1983.

[12] D. Greenspan and V. Casulli, *Numerical Analysis for Applied Mathematics, Science and Engineering*, Addison-Wesley, Redwood City (CA), 1988.

[13] M.R. Hestenes and E.L. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards Section B **49**, 409–432 (1952).

[14] A. Iserles, *A First Course in the Numerical Analysis of Differential Equations*, Cambridge Texts in Applied Mathematics **15**, Cambridge University Press, Cambridge, 1996.

[15] J.D. Joannopoulos, R.D. Meade, and J.N. Winn, *Photonic Crystals, Molding the flow of light*, Princeton University Press, Princeton, 2006.

[16] C.T. Kelley, *Iterative Methods for Linear and Nonlinear Equations*, SIAM, Philadelphia, 1995.

[17] R. Kress, *Numerical Analysis*, Graduate Texts in Mathematics **181**, Springer, New York, 1998.

[18] A. Mitchell and D. Griffiths, *The Finite Difference Method in Partial Differential Equations*, Wiley-Blackwell, New York, 1980.

[19] P.V. O'Neil, *Advanced Engineering Mathematics*, Fourth Edition, Brooke-Cole Publishing Company, Pacific CA93950, USA, 2002.

[20] J.M. Ortega, *Numerical Analysis. A Second Course*, Classics in Applied Mathematics **3**, SIAM, Philadelphia, 1990.

[21] L.M. Ortega and W.C. Rheinboldt, *Iterative Solutions of Nonlinear Equations in Several Variables*, Advanced Press, New York, (1970).

[22] A. Quarteroni, *Modellistica Numerica per Problemi Differenziali*, Springer, Milano, 2007.

[23] P.A. Raviart e J.M. Thomas, *Introduzione all'Analisi Numerica delle Equazioni alle Derivate Parziali*, Masson, Milano, 1989.

[24] G. Rodriguez e S. Seatzu, *Introduzione alla Matematica Applicata e Computazionale*, Pitagora Editore, Bologna, 2010.

[25] H.L. Royden, *Real Analysis*, 3° ed., Macmillan, New York, 1988.

[26] S. Seatzu e P. Contu, *Equazioni alle Derivate Parziali*, Pitagora Editore, Bologna, 2012.

[27] M.R. Spiegel, *Analisi di Fourier*, Collana Schaum's, McGraw-Hill, Milano, 1994.

[28] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Texts in Applied Mathematics **12**, Springer, Berlin, 1992.

[29] G. Strang, *Linear Algebra and Matrix Theory*, Wiley, New York, 1970.

[30] G. Strang and G.J. Fix, *An Analysis of the Finite Element Method*, Wellesley-Cambridge Press, Wellesley (MA), 1988.

[31] L.N. Trefethen and D. Bau III, *Numerical Linear Algebra*, SIAM, Philadelphia, 1997.

[32] J. Weidmann, *Spectral Theory of Ordinary Differential Operators*, Lecture Notes in Mathematics **1258**, Springer, Berlin, 1987.