Traccia della parte del corso di IAN sulla risoluzione numerica di PDE, a.a. 2013-2014

Dario A. Bini

21 maggio 2014

1 Introduzione

Questo documento contiene una traccia di parte degli argomenti trattati a lezione.

Le equazioni differenziali trattate sono del tipo

$$L[u] = f \tag{1}$$

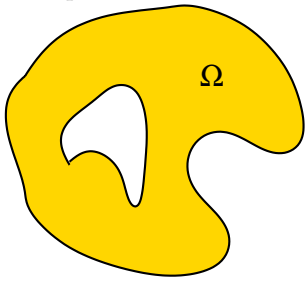
dove $u(x,y) : \overline{\Omega} \to \mathbb{R}$ è la funzione incognita definita sulla chiusura $\overline{\Omega}$ di un insieme $\Omega \subset \mathbb{R}^2$ aperto e connesso, $f(x,y) : \Omega \to \mathbb{R}$ è una funzione nota e sufficientemente regolare, $L[u]$ è un operatore differenziale lineare definito da

$$L[u] = A\frac{\partial^2 u}{\partial x^2} + B\frac{\partial^2 u}{\partial x \partial y} + C\frac{\partial^2 u}{\partial y^2} + D\frac{\partial u}{\partial x} + E\frac{\partial u}{\partial y} + Fu = f, \quad (x,y) \in \Omega$$

dove $A, B, C, D, E, F, f : \Omega \to \mathbb{R}$ sono funzioni assegnate sufficientemente regolari, u è la funzione incognita da Ω in $\mathbb{R}$. L'equazione (1) viene complementata con delle condizioni aggiuntive assegnate sulla frontiera $\partial\Omega$ di Ω (dette condizioni al contorno) che verranno specificate caso per caso e differiscono tra loro a seconda della specificità del problema. Si assume inoltre che la frontiera $\partial\Omega$ sia sufficientemente regolare.

Col termine " funzione sufficientemente regolare" si intende che la funzione con le sue derivate fino a un certo ordine siano continue sull'insieme $\bar{\Omega}$. Nella maggior parte dei casi significativi ci basta la continuità delle derivate fino al quarto ordine.

Esempio di dominio Ω



1.1 Classificazione

Premettiamo il seguente lemma che motiva la terminologia usata per la classificazione delle equazioni differenziali.

Lemma 1 *Dati $A, B, C \in \mathbb{R}$, l'insieme $\{(x, y) \in \mathbb{R}^2 : \quad Ax^2 + Bxy + Cy2 = 0\}$ è una ellisse, parabola o iperbole a seconda che $B^2 - 4AC$ sia minore, uguale o maggiore di zero.*

Procediamo allora alla seguente classificazione

Definizione 2 *Se le funzioni $A(x, y), B(x, y), C(x, y)$ sono tali che $B^2 - 4AC < 0 \ \forall \ (x, y) \in \Omega$, l'equazione (1) è detta ellittica, se $B^2 - 4AC = 0 \ \forall \ (x, y) \in \Omega$, l'equazione (1) è detta parabolica, se $B^2 - 4AC > 0 \ \forall \ (x, y) \in \Omega$, l'equazione (1) è detta iperbolica.*

In generale la natura dell'equazione può cambiare a seconda del punto (x, y) considerato del dominio.

1.2 Esempi

Alcuni esempi classici dalle applicazioni:

Problemi ellittici: Operatore di Laplace $L[u] = \Delta(u)$, dove

$$\Delta(u) = \frac{\partial^2}{\partial x^2} u(x, y) + \frac{\partial^2}{\partial y^2} u(x, y)$$

Equazione di Laplace $\Delta(u) = 0$, $(x, y) \in \Omega$.
Equazione di Poisson $\Delta(u) = f(x, y)$, $(x, y) \in \Omega$.
Condizioni al contorno di Dirichlet: $u(x, y) = g(x, y)$, $(x, y) \in \partial\Omega$, dove $\partial\Omega$ indica la frontiera di Ω.
Condizioni al contorno di Neumann, dove la frontiera $\partial\Omega$ di Ω è costituita da curve differenziabili a tratti:

$$\frac{\partial}{\partial \vec{n}} u(x, y) = g(x, y), \ \ (x, y) \in \partial\Omega \ \ \text{(derivata rispetto alla normale alla frontiera)}$$

Condizioni al contorno di Robin

$$\alpha \frac{\partial}{\partial \vec{n}} u(x, y) + \beta u(x, y) = g(x, y), \quad (x, y) \in \partial\Omega$$

Esempi di domini, $\Omega = (0, 1) \times (0, 1)$, oppure $\Omega = \{(x, y) : \ x^2 + y^2 < 1\}$.

Modello fisico: Ω: insieme racchiuso da una curva chiusa Γ nel piano; $u(x, y)$ è la quota del punto di coordinate (x, y) di una membrana elastica vincolata a passare per i punti $(x, y, g(x, y))$, $(x, y) \in \Gamma = \partial\Omega$ (filo chiuso nello spazio con proiezione Γ sul piano (x, y)), e soggetta ad una forza $f(x, y)$. Il modello è valido per superfici "non troppo sghembe" [2].

Problema della bolla di sapone o del *plateau* [2]: minimizzare l'area

$$A = \int\int \sqrt{1 + \left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2}\, dx\, dy$$

di una superficie $(x, y, u(x, y))$, $(x, y) \in \Omega$ vincolata su una curva chiusa sghemba nello spazio $(x, y, g(x, y))$, $(x, y) \in \Gamma$, dove Γ è la curva chiusa di $\mathbb{R}^2$ che racchiude il dominio $\Omega \subset \mathbb{R}^2$. L'equazione variazionale di Eulero-Lagrange è non lineare:

$$\frac{\partial^2 u}{\partial x^2}(1 + \left(\frac{\partial u}{\partial y}\right)^2) - 2\frac{\partial u}{\partial x}\frac{\partial u}{\partial y}\frac{\partial^2 u}{\partial x \partial y} + \frac{\partial^2 u}{\partial y^2}(1 + \left(\frac{\partial u}{\partial x}\right)^2) = 0$$

ed è approssimata da $\Delta u = 0$, se $u(x, y)$ "non è troppo sghemba", cioè se le derivate parziali prime di $u(x, y)$ sono "abbastanza piccole" così che l'errore causato dal rimuovere i quadrati delle derivate prime è trascurabile.

Modelli tridimensionali dell'equazione di Laplace descrivono la pressione $u(x, y, z)$ di un gas nel generico punto (x, y, z) di un dominio $\Omega \subset \mathbb{R}^3$ racchiuso da un involucro $\partial\Omega$

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

con condizioni al contorno miste [2].

Altri modelli fisici governati dall'equazione di Laplace riguardano la distribuzione della temperatura $u(x, y)$ su una porzione (lastra) di $\mathbb{R}^2$ racchiusa da una curva Γ, o nel caso tridimensionale la distribuzione della temperatura $u(x, y, z)$ su una porzione di spazio racchiuso da una superficie Γ [8].

Viene considerata una equazione di tipo ellittico anche l'equazione di *Sturm-Liouville* che è una equazione differenziale ordinaria con valori ai limiti:

$$\begin{aligned} &-[p(x)u'(x)]' + q(x)u(x) = f(x), \quad x \in (a, b) \\ &u(a) = u_a, \quad u(b) = u_b \end{aligned}$$

Problemi parabolici: Equazione del calore

$$\gamma \frac{\partial}{\partial t} u(x, t) - \frac{\partial^2}{\partial x^2} u(x, t) = 0, \quad 0 < x < \ell, \quad t \geq 0$$

dove $\gamma > 0$ e $u : [0, \ell] \times [0, +\infty) \to \mathbb{R}$, $\ell > 0$. Condizioni aggiuntive:

$$\begin{aligned} &u(x, 0) = g(x) \\ &u(0, t) = a(t), \quad u(\ell, t) = b(t) \end{aligned}$$

$a(t), b(t) : [0, +\infty) \to \mathbb{R}$, $g(x) : [0, \ell] \to \mathbb{R}$ funzioni assegnate.

Modello fisico: $u(x, t)$ è la temperatura al tempo t nel punto x di una sbarretta di metallo di estremi $0, \ell$ di cui è nota la temperatura iniziale $g(x)$, $0 \leq x \leq \ell$ (cioè al tempo $t = 0$), e la temperatura dei punti estremi $x = 0$, $x = \ell$ ad ogni istante t. La costante γ è tale che $\gamma = c\rho/k > 0$, dove k è la conducibilità termica, ρ la densità, c la capacità termica del materiale.

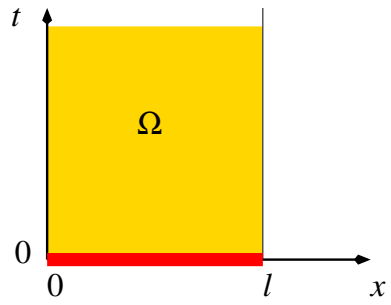


Figura 1: Dominio Ω nel caso dell'equazione del calore: $0 \le x \le \ell$, $0 \le t \le t_{\max}$.

Nel caso di un oggetto bidimensionale Ω (lastra di metallo) con contorno Γ l'equazione diventa

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - \gamma \frac{\partial u}{\partial t} = 0$$

Problemi iperbolici: Equazione delle onde

$$\frac{\partial^2}{\partial t^2} u(x,t) - \lambda \frac{\partial^2}{\partial x^2} u(x,t) = 0, \quad 0 < x < \ell, \quad t > 0$$

dove $\lambda > 0$. Condizioni aggiuntive:

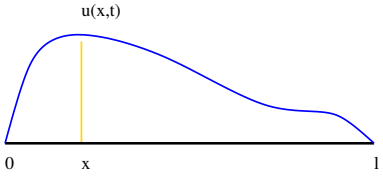
$$u(x,0) = g(x)$$

$$\frac{\partial}{\partial t} u(x,t)|_{t=0} = v(x)$$

$$u(0,t) = 0, \quad u(\ell,t) = 0$$

con $g(x)$, $v(t)$ funzioni assegnate.

Modello fisico: $u(x,t)$ è l' ordinata di un punto di ascissa x che sta su una corda elastica uniforme sottesa tra i punti $(0,0)$ e $(\ell,0)$. $g(x)$ è la posizione in cui viene abbandonata la corda all'istante iniziale e $v(x)$ è la velocita che il punto di ascissa x della corda ha al tempo iniziale. La costante λ è il rapporto tra la elasticità specifica e la densità di massa.



Problemi in dimensioni più alte (propagazioni di un'onda di una membrana elastica) portano all'equazione

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

Lo studio di onde di pressione conduce a equazioni del tipo

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

Osservazione 3 Soluzioni del tipo $u(x, y, z, t) = v(x, y, z) \cos \omega t$ in cui tutti i punti oscillano con medesima frequenza e fase esistono se

$$\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial z^2} = -\frac{\omega^2}{\lambda} v, \quad (x, y, z) \in \Omega$$
$$v(x, y, z) = 0, \quad (x, y, z) \in \partial\Omega.$$

Questo è un problema di autovalori per un operatore di tipo ellittico.

2 Metodi numerici di risoluzione

Raramente è possibile descrivere la soluzione di un problema differenziale, nel caso essa esista, in termini di funzioni elementari. È quindi necessario costruire dei metodi numerici per approssimare la soluzione con precisione arbitraria.

Possiamo classificare i metodi di risoluzione numerica nelle seguenti classi. Per semplicità diamo una descrizione sommaria in cui la funzione incognita $u(x)$ è intesa della sola variabile x e definita su $[0, 1]$, l'operatore differenziale è $L[u]$ e l'equazione differenziale è $L[u] = f$ con condizioni al contorno di Dirichlet omogenee. Svilupperemo più in dettaglio il primo di questi metodi per cui si fa riferimento alle pubblicazioni [5], [4], [7], [8] . Rimandiamo alle pubblicazioni [8], [4], [3], [6] per maggiori dettagli sugli altri approcci. Per le nozioni di algebra lineare numerica facciamo riferimento al libro [1].

1. *Differenze finite*: Fissato un intero $n > 0$ la funzione incognita $u(x)$ viene sostituita con un insieme finito di valori $u_i = u(x_i)$, $i = 0, 1, \ldots, n + 1$, che essa assume in un insieme assegnato di punti $x_0, \ldots, x_{n+1}$ equispaziati nel dominio con passo $h = 1/(n + 1)$. L'operatore differenziale L viene approssimato (discretizzato) da un operatore lineare L_n a dimensione finita (matrice) tale che $L_n((u_0, \ldots, u_{n+1}))_i = L[u]|_{x=x_i} + h^p \tau_i^{(n)}$, dove $\tau_i^{(n)}$ è limitato superiormente da una costante indipendente da n e p è un intero positivo. L'equazione $L[u] = f$ viene sostituita dal sistema lineare $L_n((v_0, \ldots, v_{n+1})) = (f(x_0), \ldots, f(x_n))$. La soluzione $(v_1, \ldots, v_n)$ sotto certe ipotesi approssima i valori $(u_1, \ldots, u_n)$.

   Questo è un approccio "alla Octave". Infatti, quando in Octave si vuole tracciare il grafico di una funzione assegnata $u(x)$ sull'intervallo $[a, b]$ viene costruita una griglia di punti equidistanziati $\texttt{x} = [\texttt{a} : \texttt{h} : \texttt{b}]$; distanziati di un passo `h` e poi viene tracciata la funzione calcolata su questi punti col comando $\texttt{plot}(\texttt{x}, \texttt{u}(\texttt{x}))$.

2. *Metodo di Galerkin*: Si ambienta il problema in uno spazio di Hilbert $\mathcal{H}$ di funzioni su cui è definito l'operatore differenziale e a cui la soluzione

appartiene. L'idea si basa sul fatto che se $L[u] = f$ allora per ogni $v \in \mathcal{H}$ vale $\langle L[u] - f, v\rangle = 0$.

Si sceglie un insieme di funzioni $\varphi_i \in \mathcal{H}$, $i = 1, 2, \ldots$, linearmente indipendenti e si approssima la soluzione con una combinazione lineare finita

$$v_n(x) = \sum_{i=1}^{n} \alpha_i \varphi_i(x).$$

Ad esempio, per un dominio $[a, b]$, e le condizioni al contorno sono di Dirichlet omogenee, le funzioni $\varphi_i(x)$, possono essere polinomi ortogonali oppure funzioni spline che verificano le condizioni al contorno omogenee. Fra tutte le funzioni $v_n(x) = \sum_{i=1}^{n} \alpha_i \varphi_i(x)$ si cerca quella per cui il residuo $f - L(v)$ è ortogonale a tutte le funzioni di $\mathcal{V}_n$ =span$(\varphi_1, \ldots, \varphi_n) \subset \mathcal{H}$. Quindi si impone la ortogonalità del residuo rispetto a un insieme di funzioni "test" assegnate in $\mathcal{V}_n$, ad esempio le stesse funzioni φ_i. Ciò conduce ad un sistema di equazioni lineari. Nel caso in cui le funzioni test siano le φ_i stesse il sistema diventa

$$\sum_{j=1}^{n} \langle \varphi_i, L(\varphi_j)\rangle \alpha_j = \langle f, \varphi_i\rangle, \quad i = 1, \ldots, n.$$

Se le funzioni φ_i sono autofunzioni dell'operatore lineare $L(\cdot)$ che soddisfano le condizioni al contorno omogenee allora la matrice del sistema è diagonale. Scegliendo funzioni spline a supporto compatto si ottiene un sistema con matrice a banda.

Come nel caso dell'approssimazione lineare di funzioni il problema principale è la determinazione di un insieme di funzioni φ_i, $i = 1, 2, \ldots$, per cui $\|v_n - u\| \to 0$ e tale che il sistema lineare finito sia di facile risoluzione.

3. *Metodo di Rayleigh-Ritz, elementi finiti*: Si fa l'ipotesi che l'operatore lineare $L(\cdot)$ sia autoaggiunto, cioè $\langle L(v), w\rangle = \langle v, L(w)\rangle$ per ogni $v, w \in \mathcal{H}$ e coercivo, cioè tale che $\langle v, L(v)\rangle \geq \gamma \langle v, v\rangle$ per ogni $v \in \mathcal{H}$, dove $\gamma > 0$ è una costante opportuna. In questo modo si riesce a riformulare il problema come problema di minimo di un funzionale. Vale cioè $u \in \mathcal{H}$ è soluzione se e solo se u minimizza il funzionale $F(v) = \frac{1}{2}\langle v, L(v)\rangle - \langle v, f\rangle$. Minimizzare questo funzionale equivale a minimizzare $\|v - u\|_L$ dove l'applicazione $\|v\|_L := \langle v, L(v)\rangle$ è una norma su $\mathcal{H}$, detta norma in energia, data la coercività di L.

   Come nel metodo di Galerkin, la soluzione si approssima come combinazione lineare $v(x) = \sum_i \alpha_i \varphi_i(x)$ di funzioni $\varphi_i(x)$. Il problema di minimo del funzionale trattato in termini discreti, cioè imponendo che il gradiente rispetto agli α_i sia nullo, o trattato con le tecniche gia viste nell'approssimazione di funzioni, imponendo cioè l'ortogonalità del resto allo spazio $\mathcal{V}_n$ generato dalle funzioni φ_i, conduce ad un analogo sistema lineare finito con matrice di elementi $\langle \varphi_i, L(\varphi_j)\rangle$.

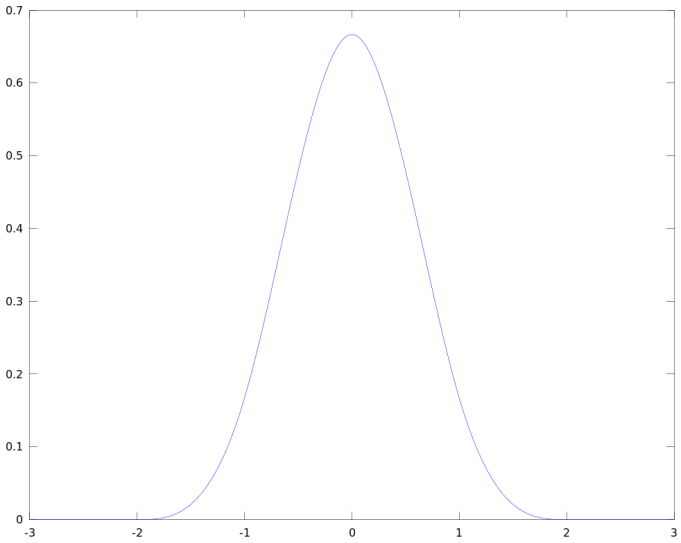


Figura 2: Spline cubica a supporto compatto

Nella tecnica degli elementi finiti il dominio viene discretizzato suddividendolo in elementi che possono essere segmenti di varia ampiezza, o, nel caso bidimensionale, in triangoli, e imponendo che ciascuna φ_i sia non nulla solo in pochi di questi elementi in modo che la matrice del sistema sia più sparsa possibile. Un esempio di scelta delle $\varphi_i(x)$ è costituito dalle funzioni spline ottenute discretizzando l'intervallo in sotto-intervalli $[x_i, x_{i+1}]$ di ampiezza h e traslando, centrandola nei vari punti x_i, la particolare funzione spline cubica a supporto compatto definita da

$$s(x) = \begin{cases} \frac{1}{6}(2+t)^3, & -2 < t \leq -1 \\ \frac{1}{6}(4 - 6t^2 - 3t^3) & -1 < t \leq 0 \\ \frac{1}{6}(4 - 6t^2 + 3t^3) & 0 < t \leq 1 \\ \frac{1}{6}(2-t)^3 & 1 < t < 2 \\ 0 & \text{altrove} \end{cases} \qquad \text{dove } t = x/h.$$

Un grafico di questa funzione è riportato in figura 2 In questo caso la matrice del sistema è a banda con 7 diagonali nonnulle.

Il grafico è stato ottenuto con la function di Octave riportata nel Listing 1

Listing 1: Script di Octave per tracciare il grafico della funzione B-Spline

```
function [s,x]=bspline(n)
% function [s,x]=bspline(n)
% calcola e traccia il grafico della B-spline cubica
% n: numero di punti in cui viene discretizzato ciascuno dei
    sottointervalli
% x: valori dei punti in cui la spline viene campionata
% s: valori della spline nei punti x
h=1/(n+1);
x0=[-3+h:h:-2];
x1=x0+1;
x2=x1+1;
x3=x2+1;
x4=x3+1;
x5=x4+1;
x=[x0,x1,x2,x3,x4,x5];
s0=x0*0;
s1=(1/6)*(2+x1).^3;
s2=(1/6)*(4-6*x2.^2-3*x2.^3);
s3=(1/6)*(4-6*x3.^2+3*x3.^3);
s4=(1/6)*(2-x4).^3;
s5=0*x5;
s=[s0,s1,s2,s3,s4,s5];
plot(x,s)
```

4. *Metodo di collocazione*: Anche in questo caso la soluzione viene approssimata con una funzione $v(x) = \sum_i \alpha_i \varphi_i(x)$ combinazione lineare di funzioni base $\varphi_i(x)$. Si impone che il residuo $f - L(v)$ si annulli in alcuni punti speciali del dominio. Cioè si interpola la funzione f con funzioni del tipo $\sum_{i=1}^{n} \alpha_i L(\varphi_i)$. Alla luce delle considerazioni fatte sulle costanti di Lebesgue e sulla relazione che intercorre tra l'errore di interpolazione l'errore di migliore approssimazione uniforme, è conveniente scegliere i nodi di collocazione in modo da contenere la crescita delle corrispondenti costanti di Lebesgue. Ad esempio i nodi di Chebyshev se il dominio è $[-1, 1]$ sono una scelta adatta.

3 Il metodo delle differenze finite

Descriviamo il metodo delle differenze finite partendo dal semplice *problema modello*

$$\begin{aligned} &u''(x) = f(x), \quad x \in (0,1) \\ &u(0) = a, \quad u(1) = b \end{aligned} \tag{2}$$

Si osserva che se $f(x) : [0,1] \to \mathbb{R}$ è continua allora la soluzione di (2) esiste ed è unica. Infatti, integrando l'espressione $u''(x) = f(x)$ si ha

$$u'(x) = \int_0^x f(t)dt + \gamma_1$$

Integrando nuovamente si ottiene

$$u(x) = \int_0^x \int_0^s f(t)dtds + \gamma_1 x + \gamma_2$$

Le costanti γ_1 e γ_2 vengono determinate imponendo le condizioni al contorno. Si osserva anche che se $f(x)$ è di classe $C^2[0,1]$ allora la soluzione $u(x)$ è di classe $C^4[0,1]$. La regolarità della soluzione ha un ruolo molto importante nel metodo delle differenze finite. le elaborazioni che svolgeremo nel seguito sono valide nell'ipotesi in cui la soluzione del problema differenziale esista e sia sufficientemente regolare.

Prima di trattare il metodo introduciamo alcune formule utili.

4 Discretizzazione degli operatori differenziali

4.1 Alcune formule

Sia $f(x) : [a,b] \to \mathbb{R}$ e si assuma che $f \in C^4[a,b]$, e $x, x+h, x-h \in [a,b]$. Allora vale

$$\begin{aligned} f(x+h) &= f(x) + hf'(x) + \tfrac{h^2}{2} f''(x) + \tfrac{h^3}{6} f^{(3)}(x) + \tfrac{h^4}{24} f^{(4)}(\xi) \\ f(x-h) &= f(x) - hf'(x) + \tfrac{h^2}{2} f''(x) - \tfrac{h^3}{6} f^{(3)}(x) + \tfrac{h^4}{24} f^{(4)}(\eta) \end{aligned} \tag{3}$$

dove $\xi \in (x, x+h)$, $\eta \in (x-h, x)$. Sommando le due formule (3) si ottiene

$$f(x+h) + f(x-h) = 2f(x) + h^2 f''(x) + \frac{h^4}{24}(f^{(4)}(\xi) + f^{(4)}(\eta))$$

da cui

$$f''(x) = \frac{1}{h^2}(f(x-h) - 2f(x) + f(x+h)) + \tau h^2, \quad |\tau| \leq \frac{1}{12} \max_{x\in[a,b]} |f^{(4)}(x)|. \tag{4}$$

Similmente sottraendo entrambi i membri di (3) si ottiene

$$f'(x) = \frac{1}{2h}(f(x+h) - f(x-h)) + \sigma h^2, \quad |\sigma| \leq \frac{1}{3} \max_{x\in[a,b]} |f^{(3)}(x)|. \tag{5}$$

La (5) è detta *differenza centrata* . Una approssimazione più scadente della derivata prima che coinvolge i valori $f(x)$ e $f(x+h)$ è data da

$$f'(x) = \frac{1}{h}(f(x+h) - f(x)) - \frac{1}{2}hf''(\xi), \quad \xi \in (x, x+h) \tag{6}$$

ed è detta *differenza in avanti*. Similmente l'espressione

$$f'(x) = \frac{1}{h}(f(x) - f(x-h)) - \frac{1}{2}hf''(\hat{\xi}), \quad \hat{\xi} \in (x-h, x)$$

ottenuta prendendo $-h$ al posto di h è detta *differenza all'indietro*. Approssimazioni più precise si possono ottenere se $f(x)$ ha maggior regolarità svolgendo sviluppi in serie di ordine più elevato.

4.2 Discretizzazione del problema modello

Si consideri l'equazione (2). Sia $n > 1$ un intero, $h = 1/(n+1)$, $x_i = ih$, $i = 0, 1, \ldots, n+1$. Dalla (4) si ottiene

$$\begin{aligned} &\frac{1}{h^2}(u(x_{i-1}) - 2u(x_i) + u(x_{i+1})) = f(x_i) - \tau_i h^2, \quad i = 1, \ldots, n, \\ &u(x_0) = a, \quad u(x_{n+1}) = b \end{aligned} \tag{7}$$

dove

$$\tau_i = \frac{1}{24}(u^{(4)}(\xi_i) + u^{(4)}(\eta_i)), \quad \xi_i \in (x_i, x_{i+1}), \quad \eta_i \in (x_{i-1}, x_i)$$

per cui

$$|\tau_i| \leq \frac{1}{12} \max_{x\in[a,b]} |u^{(4)}(x)|. \tag{8}$$

La formula fornisce una approssimazione dell'operatore differenziale $L[u] = u''$ con l'operatore alle *differenze finite* $L_n : \mathbb{R}^{n+2} \to \mathbb{R}^n$

$$L_n(\boldsymbol{u}^{(n)}) = \frac{1}{h^2}(u_{i-1} - 2u_i + u_{i+1})_{i=1:n} \tag{9}$$

dove abbiamo posto $u_i = u(x_i)$ e abbiamo indicato $\boldsymbol{u}^{(n)} = (u_i)_{i=0:n+1}$. In questo modo l'espressione (7) prende la forma

$$L_n(\boldsymbol{u}^{(n)}) = \boldsymbol{f}^{(n)} - h^2\boldsymbol{\tau}^{(n)}, \tag{10}$$

dove si è posto $\boldsymbol{f}^{(n)} = (f(x_i))_{i=1:n}$, $\boldsymbol{\tau}^{(n)} = (\tau_i^{(n)})_{i=1:n}$.

Per maggior rigore formale e per ricordare che le quantità introdotte dipendono da n, avremmo dovuto scrivere h_n al posto di h, $x_i^{(n)}$ al posto di x_i e $u_i^{(n)}$ al posto di u_i. Però, per non appesantire la notazione, eviteremo di riportare l'indice n fintanto che questo non crei ambiguità. Scriveremo quindi $\boldsymbol{u}$ al posto di $\boldsymbol{u}^{(n)}$. Comunque è utile osservare che il vettore $\boldsymbol{u}$ è di fatto da intendersi come una *successione di vettori* $\boldsymbol{u}^{(n)}$.

L'operatore $L_n(\boldsymbol{u})$ è individuato dalla matrice tridiagonale $n \times (n+2)$

$$L_n = -\frac{1}{h^2}\begin{bmatrix} -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \end{bmatrix}$$

per la quale useremo lo stesso simbolo L_n.

Il vettore di componenti $L[u]|_{x=x_i} - (L_n(\boldsymbol{u}))_i = h^2\tau_i$ è detto *errore locale di dicretizzazione*.

In generale, l'approssimazione alle differenze $L_n(\boldsymbol{u})$ che discretizza l'operatore differenziale $L[u]$ si dice *consistente di ordine* h^p se esiste una costante $\gamma > 0$, indipendente da n, tale che $\max_i |L[u]|_{x=x_i} - (L_n(\boldsymbol{u}))_i| \leq \gamma h^p$ per ogni $n > 1$. Per questa definizione l'approssimazione (9) dell'operatore $L[u] = u''$ è consistente di ordine h^2.

L'equazione differenziale (2) con condizioni al contorno di Dirichlet, ristretta ai valori $u_i = u(x_i)$ si può allora riscrivere come

$$-\frac{1}{h^2}\begin{bmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-1} \\ u_n \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{n-1} \\ f_n \end{bmatrix} - h^2\begin{bmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_{n-1} \\ \tau_n \end{bmatrix} - \frac{1}{h^2}\begin{bmatrix} a \\ 0 \\ \vdots \\ 0 \\ b \end{bmatrix}$$

essendo $u_0 = a$, $u_n = b$, dove $f_i = f(x_i)$.

Denotiamo con

$$A_n = -\frac{1}{h^2}\begin{bmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix} =: -\frac{1}{h^2}H_n$$

la matrice del sistema precedente. In questo modo possiamo scrivere il sistema come

$$A_n\boldsymbol{u}^{(n)} = \boldsymbol{f}^{(n)} - h^2\boldsymbol{\tau}^{(n)} - \frac{1}{h^2}(a\boldsymbol{e}_1^{(n)} + b\boldsymbol{e}_n^{(n)}) \tag{11}$$

dove abbiamo indicato con $\boldsymbol{e}_1^{(n)}$ e $\boldsymbol{e}_n^{(n)}$ il primo e l'ultimo vettore della base canonica di $\mathbb{R}^n$.

Rimuovendo la componente dell'errore locale di discretizzazione in (11) si ottiene il sistema

$$A_n \boldsymbol{v}^{(n)} = \boldsymbol{f}^{(n)} - \frac{1}{h^2}(a\boldsymbol{e}_1^{(n)} + b\boldsymbol{e}_n^{(n)}). \tag{12}$$

Si osserva che la matrice A_n del sistema (12) è dominante diagonale e irriducibile, quindi, per il terzo teorema di Gerschgorin è non singolare. Il sistema (12) ha allora una sola soluzione $\boldsymbol{v} = (v_i)$. Si osservi ancora che il termine noto è costituito da due componenti additive: la prima dipende dalla funzione $f(x)$, la seconda dipende dalle condizioni al contorno.

Per valutare l'accuratezza dell'approssimazione $\boldsymbol{v} = (v_i)$ ottenuta risolvendo il sistema (12) al posto del sistema (11) che ci fornirebbe la soluzione esatta valutata nei punti della discetizzazione, si introduce l'*errore globale* $\boldsymbol{\epsilon}^{(n)} = (\epsilon_i^{(n)})$, definito da $\boldsymbol{\epsilon}^{(n)} = \boldsymbol{v}^{(n)} - \boldsymbol{u}^{(n)}$.

Ci interessa dimostrare che per $n \to \infty$ la successione $\|\boldsymbol{v}^{(n)} - \boldsymbol{u}^{(n)}\|_\infty$ converge a zero. In questo modo, scegliendo valori di n arbitrariamente grandi, possiamo meglio approssimare i valori di $u(x)$ in un insieme di punti sempre più fitti e in modo sempre più accurato.

Sottraendo la (11) dalla (12) si ha

$$-\frac{1}{h^2}\begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix} \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{n-1} \\ \epsilon_n \end{bmatrix} = h^2 \begin{bmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_{n-1} \\ \tau_n \end{bmatrix}.$$

cioè

$$A_n \boldsymbol{\epsilon}^{(n)} = h^2 \boldsymbol{\tau}^{(n)}, \tag{13}$$

dove A_n è la matrice tridiagonale del sistema e dove i vettori $\boldsymbol{\epsilon}^{(n)}$, $\boldsymbol{\tau}^{(n)}$ hanno n componenti.

Quindi una espressione esplicita dell'errore globale è data da

$$\boldsymbol{\epsilon}^{(n)} = h^2 A_n^{-1} \boldsymbol{\tau}^{(n)}.$$

Se $\|\cdot\|$ è una norma vettoriale su $\mathbb{R}^n$ allora dalla relazione precedente si ottiene

$$\|\boldsymbol{\epsilon}^{(n)}\| \leq h^2 \|A_n^{-1}\| \cdot \|\boldsymbol{\tau}^{(n)}\| \tag{14}$$

dove $\|A_n^{-1}\|$ è la norma matriciale indotta.

In particolare, se $\|A_n\|$ è limitata superiormente da una costante σ indipendente da n allora $\|\boldsymbol{\epsilon}^{(n)}\| \leq \sigma h^2 \|\boldsymbol{\tau}^{(n)}\|$ e l'errore globale in norma converge a zero con lo stesso ordine di convergenza dell'errore locale.

Definizione 4 *Sia $\|\cdot\|$ una norma su $\mathbb{R}^n$. Lo schema alle differenze finite dato dal sistema lineare $L_n(\boldsymbol{v}) = \boldsymbol{f}$ si dice* stabile *in norma $\|\cdot\|$ se $\|A_n^{-1}\|$ è limitato superiormente da una costante indipendente da n. Si dice convergente in norma $\|\cdot\|$ con ordine h^p se $\lim \|\boldsymbol{\epsilon}^{(n)}\| = 0$ con ordine h^p.*

Se nella definizione di stabilità o di convergenza si omette il termine "in norma $\|\cdot\|$" allora si intende implicitamente che la stabilità o la convergenza è data in norma infinito.

Se un metodo consistente di ordine h^p è anche stabile in norma $\|\cdot\|$ allora esso è convergente con lo stesso ordine h^p in norma $\|\cdot\|$. Questa proprietà che può essere dimostrata sotto ipotesi generali, verrà da noi considerata caso per caso per i problemi modello studiati.

4.3 Analisi in norma 2

Per il problema modello è abbastanza facile fare una analisi della convergenza in norma 2. Nel seguito indichiamo con $\text{trid}_n(a,b,c)$ la matrice tridiagonale $n \times n$ con elementi uguali a b sulla diagonale principale, uguali a c sulla sopra-diagonale e uguali ad a sulla sotto-diagonale. Useremo la stessa notazione $\text{trid}_n(A,B,C)$ per denotare la matrice tridiagonale a blocchi con blocchi A, B, C di dimensione $m \times m$. Nel caso in cui gli elementi (o i blocchi) lungo le diagonali non fossero costanti useremo la notazione $\text{trid}_m(a_i, b_i, c_i)$ per denotare che la matrice tridiagonale ha sulla i-esima riga gli elementi a_i, b_i e c_i dove b_i è l'elemento diagonale, c_i è elemento sopra-diagonale mentre c_i è l'elemento sotto-diagonale. Similmente definiamo $\text{trid}(A_i, B_i, C_i)$. Indichiamo inoltre con $\text{diag}_n(d_1, \ldots, d_n)$ o più semplicemente con $\text{diag}(d_1, \ldots, d_n)$ la matrice diagonale con elementi diagonali $d_1, \ldots, d_n$.

Si ricordano i seguenti risultati che si possono dimostrare utilizzando un pò di trigonometria

Teorema 5 *Vale $S_n \text{trid}_n(1,0,1) S_n = \text{diag}(2c_1, \ldots, 2c_n)$, dove $c_i = \cos(i\pi/(n+1))$, e dove $S_n = (s_{i,j})$ è la matrice ortogonale simmetrica di elementi $s_{i,j} = \sqrt{\frac{2}{n+1}} \sin(ij\frac{\pi}{n+1})$.*

Inoltre vale $R_n^T(\text{trid}_n(1,0,1) + \boldsymbol{e}_n^{(n)} \boldsymbol{e}_n^{(n)T}) R_n = \text{diag}(2\hat{c}_1, \ldots, 2\hat{c}_n)$, dove $\hat{c}_i = \cos((2i-1)\frac{\pi}{2n+1})$, e dove $R_n = (r_{i,j})$ è una matrice ortogonale di elementi $r_{i,j} = \sqrt{\frac{4}{2n+1}} \sin(i(2j-1)\frac{\pi}{2n+1})$. Infine la matrice $\text{trid}_n(1,0,1) + \boldsymbol{e}_1^{(n)} \boldsymbol{e}_1^{(n)T}$ ha gli stessi autovalori di $\text{trid}_n(1,0,1) + \boldsymbol{e}_n^{(n)} \boldsymbol{e}_n^{(n)T}$ in quanto le due matrici sono simili.

Dim. (Cenno) Cerchiamo autovettori del tipo $\boldsymbol{v} = (v_i)$, $v_i = \sin(i\theta)$ per un opportuno θ. Posto $\boldsymbol{w} = \text{trid}_n(1,0,1)\boldsymbol{v}$, si ha $w_1 = \cos(2\theta) = 2\cos\theta \sin\theta = (2\cos\theta)v_1$, mentre $w_i = \cos((i-1)\theta) + \cos((i+1)\theta) = 2\cos\theta \sin(i\theta) = (2\cos\theta)v_i$, per $i = 2, \ldots, n-1$. Infine $w_n = \sin((n-1)\theta) = \sin(n\theta)\cos\theta + \cos(n\theta)\sin\theta$. Se $\theta = \pi i/(n+1)$ con i intero, risulta $w_n = (2\cos\theta)v_n$. Quindi $\lambda_i = 2\cos(\pi i/(n+1))$ è autovalore per $i = 1, \ldots, n$ corrispondente all'autovettore $v^{(i)}$ di componenti$v_j^{(i)} =$

$\sin(\pi ij/(n+1))$, $j = 1, \ldots, n$. Il fattore $\sqrt{2/(n+1)}$ rende gli autovettori di norma euclidea unitaria. Analogamente si procede per la matrice $\mathrm{trid}_n(1, 0, 1) + \boldsymbol{e}_n^{(n)} \boldsymbol{e}_n^{(n)T}$. In questo caso cambia solo la condizione su w_n che diventa $\sin(n\theta) = \sin(n\theta + 1)$. Poichè $\sin\alpha = \sin\beta$ se $\alpha + \beta$ è multiplo dispari di π, si ottiene $\theta = (2k-1)\pi/(2n+1)$. □

Segue che per la matrice tridiagonale $\mathrm{trid}_n(-1, 2, -1)$ vale la relazione

$$S\,\mathrm{trid}_n(-1, 2, -1)\,S = \mathrm{diag}(2 - 2c_1^{(n)}, \ldots, 2 - 2c_n^{(n)}), \quad c_i^{(n)} = \cos\frac{i\pi}{n+1}$$

Da questa proprietà segue che

$$\|A_n^{-1}\|_2 = h^2/\min_i(2 - 2c_i) = h^2/(2 - 2\cos\frac{\pi}{n+1}).$$

Tenendo presente che $1 - \cos x = x^2/2 + O(x^4)$ si ha

$$\|A_n^{-1}\|_2 = 1/\pi^2 + O(h^2).$$

Poiché arrestando lo sviluppo di $\cos x$ al terzo termine si ha $1 - \cos x = x^2/2 - \frac{1}{6}x^3 \sin\xi$ e poiché $\sin\xi \geq 0$ se $x \in [0, \pi]$ possiamo anche scrivere $1 - \cos x \leq x^2/2$. Per cui vale anche

$$\|A_n^{-1}\|_2 \leq 1/\pi^2.$$

Se definiamo

$$\|\boldsymbol{u}^{(n)}\| = \frac{1}{\sqrt{n+1}}(\sum_{i=1}^{n} u_i^2)^{1/2} = \frac{1}{\sqrt{n+1}}\|\boldsymbol{u}^{(n)}\|_2,$$

allora la norma matriciale indotta da questa norma coincide con la norma 2. Inoltre è facile verificare che

$$\lim_{n\to\infty} \|\boldsymbol{u}^{(n)}\| = (\int_0^1 u(x)^2 dx)^{1/2}$$

essendo $\frac{1}{n+1}\sum_{i=1}^{n} u_i^2$ l'approssimazione di $\int_0^1 u(x)^2 dx$ mediante la formula dei rettangoli.

Dalla (8) segue che

$$\|\boldsymbol{\tau}^{(n)}\| \leq \frac{\sqrt{n}}{\sqrt{n+1}}\frac{1}{12}\max|u^{(4)}(x)| < \frac{1}{12}\max|u^{(4)}(x)|.$$

Si ottiene allora da (14)

$$\|\boldsymbol{\epsilon}^{(n)}\| < \frac{1}{12\pi^2}h^2 \max_{x\in[0,1]}|u^{(4)}(x)|.$$

Cioè l'errore globale valutato nella norma $\|\cdot\|$ converge a zero come h^2. Questo non implica che $\|\boldsymbol{\epsilon}^{(n)}\|_\infty$ converge a zero come h^2. Poiché vale $\|\boldsymbol{u}^{(n)}\|_\infty \leq \|\boldsymbol{u}^{(n)}\|_2$, si ha che $\|\boldsymbol{u}^{(n)}\|_\infty \leq \sqrt{n+1}\|\boldsymbol{u}^{(n)}\|$, quindi

$$\|\boldsymbol{\epsilon}^{(n)}\|_\infty < \frac{1}{12\pi^2}h^{\frac{3}{2}} \max_{x\in[0,1]}|u^{(4)}(x)|.$$

Si ha quindi convergenza anche in norma ∞ ma la maggiorazione che otteniamo in questo modo non è stretta. Infatti si può dimostrare che vale $\|\boldsymbol{\epsilon}^{(n)}\|_\infty \leq \gamma h^2$, per una opportuna costante γ. Questo si dimostra nella prossima sezione.

4.4 Analisi in norma ∞

Per valutare la convergenza dello schema alle differenze in norma infinito è sufficiente valutare la norma infinito della matrice A_n^{-1}. Mostriamo due analisi diverse: un'analisi puramente matriciale e una basata sul *principio del massimo*. La prima è specifica del problema trattato, la seconda può essere esportata a casi più generali.

4.4.1 Analisi matriciale

Scriviamo $A_n = -\frac{1}{h^2}H_n$, dove $H_n = \text{trid}_n(-1, 2, -1)$ così che $A_n^{-1} = -h^2 H_n^{-1}$. Definiamo $\boldsymbol{e} = (1, 1, \ldots, 1)^T$, $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^T$, $\boldsymbol{e}_n = (0, \ldots, 0, 1)^T$, $\boldsymbol{p} = (1, 2, 3, \ldots, n)^T$ e $\boldsymbol{s} = (1, 4, 9, \ldots, n^2)^T$. Osserviamo che $H_n^{-1} \geq 0$ infatti $\frac{1}{2}H_n = I - \frac{1}{2}B_n$, $B_n = \text{trid}_n(1, 0, 1)$, quindi $(\frac{1}{2}H_n)^{-1} = \sum_{i=0}^{+\infty}(\frac{1}{2}B_n)^i \geq 0$, essendo $B_n \geq 0$ e $\rho(\frac{1}{2}B_n) < 1$. Dalla nonnegatività di H_n^{-1} e dalla definizione di $\|\cdot\|_\infty$ segue che $\|H_n^{-1}\|_\infty = \|H_n^{-1}\boldsymbol{e}\|_\infty$.

Inoltre, poiché $-(i-1) + 2i - (i+1) = 0$ per $i = 1, \ldots, n-1$, vale

$$H_n \boldsymbol{p} = (n+1)\boldsymbol{e}_n. \tag{15}$$

Analogamente, poiché $-(i-1)^2 + 2i^2 - (i+1)^2 = -2$ per $i = 1, \ldots, n-1$, mentre $-(n-1)^2 + 2n^2 = n^2 + 2n - 1$, vale

$$H_n \boldsymbol{s} = -2\boldsymbol{e} + (n+1)^2 \boldsymbol{e}_n,$$

che per la (15) dà $H_n\boldsymbol{s} = -2\boldsymbol{e} + (n+1)H_n\boldsymbol{p}$. Da cui, moltiplicando per H_n^{-1} si ricava

$$H_n^{-1}\boldsymbol{e} = -\frac{1}{2}\boldsymbol{s} + \frac{1}{2}(n+1)\boldsymbol{p}.$$

Dunque si ha

$$\|H_n^{-1}\|_\infty = \|H_n^{-1}\boldsymbol{e}\|_\infty = \frac{1}{2}\max_{i=1:n}|(n+1)i - i^2| \leq \frac{1}{8}(n+1)^2 \tag{16}$$

dove l'uguaglianza vale se n è dispari per $i = (n+1)/2$. Da cui si deduce che

$$\|A_n^{-1}\|_\infty \leq \frac{1}{8}.$$

Per la norma infinito dell'errore globale risulta allora dalla (14) che

$$\|\boldsymbol{\epsilon}^{(n)}\|_\infty \leq \frac{1}{8}h^2\|\boldsymbol{\tau}^{(n)}\|_\infty \leq \frac{1}{96}h^2 \max_{0\leq x\leq 1}|u^{(4)}(x)|.$$

4.4.2 Analisi mediante il principio del massimo

Si osserva che se $u''(x) \geq 0$ per $x \in [a,b]$, la funzione $u(x)$ è convessa in $[a,b]$ per cui il suo massimo viene necessariamente preso su un estremo. Similmente se $u''(x) \leq 0$ allora $u(x)$ è concava e il suo minimo viene preso su un estremo. Poiché $L_n(u)$ è la controparte discreta di $u''(x)$ ci si aspetta che se il vettore $\boldsymbol{u} = (u_i)_{i=0:n+1}$ è tale che $L_n(\boldsymbol{u}) \geq 0$ allora $\max(u_0, u_{n+1}) \geq u_i$, $i = 1, \ldots, n$. Analogamente, se $L_n(\boldsymbol{u}) \leq 0$ allora ci si aspetta che $\min(u_0, u_{n+1}) \leq u_i$, $i = 1, \ldots, n+1$. Vale infatti il seguente

Teorema 6 (Principio del massimo discreto) *Se* $\boldsymbol{u} = (u_i)_{i=0:n+1}$ *è tale che* $L_n(\boldsymbol{u}) \geq 0$ *allora* $\max(u_0, u_{n+1}) \geq u_i$, $i = 1, \ldots, n$. *Se* $L_n(\boldsymbol{u}) \leq 0$ *allora* $\min(u_0, u_{n+1}) \leq u_i$, $i = 1, \ldots, n$.

Dim. Sia $L_n(\boldsymbol{u}) \geq 0$ e si supponga per assurdo che u_k sia il massimo degli u_i e sia $u_k > u_0, u_{n+1}$. Allora dalla condizione $L_n(\boldsymbol{u}) \geq 0$ si ha

$$u_k \leq (u_{k-1} + u_{k+1})/2. \tag{17}$$

inoltre essendo $u_k \geq u_{k-1}, u_{k+1}$ risulta $u_k \geq (u_{k-1} + u_{k+1})/2$ per cui dalla (17) segue che $u_k = u_{k-1} = u_{k+1}$. Quindi anche u_{k+1} e u_{k-1} sono massimi. Ripetendo il ragionamento per u_{k-1} e per u_{k+1} si conclude che $u_i = u_0 = u_{n+1}$ per ogni i, che è assurdo. Analogamente si tratta il caso in cui $L_n(\boldsymbol{u}) \leq 0$. □

Basandoci sulla proprietà del massimo discreto soddisfatta da L_n dimostriamo ora la convergenza e la stabilità del metodo delle differenze finite per il problema modello. Successivamente diamo il risultato in forma più generale. Ricordiamo che l'errore globale $\boldsymbol{\epsilon}^{(n)} = (\epsilon_i^{(n)})_{i=1:n}$ è tale che $A_n \boldsymbol{\epsilon}^{(n)} = h^2 \boldsymbol{\tau}^{(n)}$. Inoltre, definendo $\widehat{\boldsymbol{\epsilon}}^{(n)} = (\widehat{\epsilon}_i^{(n)})_{i=0:n+1}$, dove $\widehat{\epsilon}_i^{(n)} = \epsilon_i^{(n)}$ per $i = 1, \ldots, n$, e $\widehat{\epsilon}_0^{(n)} = \epsilon_{n+1}^{(n)} = 0$, risulta $A_n \boldsymbol{\epsilon}^{(n)} = L_n(\widehat{\boldsymbol{\epsilon}}^{(n)})$.

Teorema 7 *Per l'errore globale* $\boldsymbol{\epsilon}^{(n)} = (\epsilon_i^{(n)})_{i=1:n}$ *vale*

$$\|\boldsymbol{\epsilon}^{(n)}\|_\infty \leq \frac{1}{2} h^2 \|\boldsymbol{\tau}^{(n)}\|_\infty$$

Dim. Si osserva che la funzione $w(x) = \frac{1}{2}x^2$ è tale che $L(w) = 1$. Questa proprietà è condivisa dalla controparte discreta $\boldsymbol{w}^{(n)} = (w_i^{(n)})_{i=0:n+1}$, $w_i^{(n)} = w(x_i) = \frac{1}{2}(hi)^2$. Infatti una semplice verifica mostra che $L_n(\boldsymbol{w}^{(n)}) = (1, \ldots, 1)^T = \boldsymbol{e}^{(n)}$. Si vede allora allora che, poiché per la (13) è $L_n(\widehat{\boldsymbol{\epsilon}}^{(n)}) = A_n(\boldsymbol{\epsilon}^{(n)}) = h^2 \boldsymbol{\tau}^{(n)}$, si ha

$$L_n(\pm\widehat{\boldsymbol{\epsilon}}^{(n)} + h^2 \|\boldsymbol{\tau}^{(n)}\|_\infty \boldsymbol{w}^{(n)}) = \pm h^2 \boldsymbol{\tau}^{(n)} + h^2 \|\boldsymbol{\tau}^{(n)}\|_\infty \boldsymbol{e}^{(n)} \geq 0.$$

Quindi, per il principio del massimo applicato ai vettori $\pm\widehat{\boldsymbol{\epsilon}}^{(n)} + h^2 \|\boldsymbol{\tau}^{(n)}\|_\infty \boldsymbol{w}^{(n)}$, risulta

$$\max_i(\pm\epsilon_i^{(n)} + h^2 \|\boldsymbol{\tau}^{(n)}\|_\infty w_i^{(n)}) \leq h^2 \|\boldsymbol{\tau}^{(n)}\|_\infty \max(w_0^{(n)}, w_{n+1}^{(n)}) = \frac{1}{2} h^2 \|\boldsymbol{\tau}^{(n)}\|_\infty$$

da cui

$$\pm\epsilon_i^{(n)} \leq \frac{1}{2}h^2\|\boldsymbol{\tau}^{(n)}\|_\infty - h^2\|\boldsymbol{\tau}^{(n)}\|_\infty w_i^{(n)} \leq \frac{1}{2}h^2\|\boldsymbol{\tau}^{(n)}\|_\infty, \quad i = 1, \ldots, n.$$

quindi $|\epsilon_i^{(n)}| \leq \frac{1}{2}\|\boldsymbol{\tau}^{(n)}\|_\infty h^2$. □

Si osservi che si è dimostrato che $\|\boldsymbol{\epsilon}^{(n)}\|_\infty \leq \frac{1}{2}\|h^2\boldsymbol{\tau}^{(n)}\|_\infty$ quando $\boldsymbol{\epsilon}^{(n)} = A_n^{-1}h^2\boldsymbol{\tau}^{(n)}$ qualunque sia $\boldsymbol{\tau}^{(n)} \neq 0$. Per cui vale $\|A_n^{-1}\boldsymbol{v}\|_\infty/\|\boldsymbol{v}\|_\infty \leq \frac{1}{2}$ qualunque sia $\boldsymbol{v} \neq 0$. Quindi, per definizione di norma indotta segue $\|A_n^{-1}\|_\infty \leq \frac{1}{2}$.

Il risultato di convergenza può essere espresso per una qualsiasi equazione differenziale del tipo

$$\begin{aligned} &L[u] = f, \quad x \in (0,1) \\ &u(0) = a,\ u(1) = b \end{aligned} \tag{18}$$

dove L è un operatore differenziale definito su uno spazio di funzioni sufficientemente regolari da $[0,1]$ in $\mathbb{R}$ che viene approssimato con un errore locale $O(h^k)$ da un operatore alle differenze finite dato dalla matrice L_n di dimensione $n \times (n+2)$. Cioè si assume che L_n sia tale che per ogni funzione $u(x)$ sufficientemente regolare, posto $z = L[u]$, valga

$$L_n\boldsymbol{u}^{(n)} = (z(x_1^{(n)}), \ldots z_n(x_n^{(n)}))^T + h^k\boldsymbol{\tau}^{(n)},$$

dove $h = 1/(n+1)$, $x_i^{(n)} = ih$, $\boldsymbol{u}^{(n)} = (u_i^{(n)})$, $u_i^{(n)} = u(x_i^{(n)})$ e $\|\boldsymbol{\tau}^{(n)}\|_\infty \leq M$ con M costante positiva.

Infatti, se ora denotiamo con $u(x)$ la soluzione di (18), $\boldsymbol{u}^{(n)} = (u(x_i^{(n)}))$ e $\boldsymbol{v}_i^{(n)} = (v_i^{(n)})$, la soluzione del sistema

$$L_n\boldsymbol{v}^{(n)} = \boldsymbol{f}^{(n)}, \quad \boldsymbol{f}^{(n)} = (f(x_i))_{i=1,n}, \quad v_0 = a,\ v_{n+1} = b$$

per cui

$$L_n(\widehat{\boldsymbol{\epsilon}}^{(n)}) = h^k\boldsymbol{\tau}^{(n)}, \quad \widehat{\boldsymbol{\epsilon}}^{(n)} = \boldsymbol{v}^{(n)} - \boldsymbol{u}^{(n)}.$$

Per l'errore globale $\widehat{\boldsymbol{\epsilon}}^{(n)}$ si ha il seguente risultato

Teorema 8 *Se valgono le seguenti proprietà*

i) $L_n\boldsymbol{y} \geq 0 \ \Rightarrow\ \max_i y_i = \max(y_0, y_{n+1})$*;*

ii) esiste $\boldsymbol{w}^{(n)} = (w_i^{(n)})_{i=0,n+1}$ *tale che* $w_i^{(n)} \geq 0$, $L_n\boldsymbol{w}^{(n)} \geq \gamma\boldsymbol{e}^{(n)}$, $w_0^{(n)}, w_{n+1}^{(n)} \leq \delta$*, con* γ *e* δ *costanti positive,*

allora $\|\widehat{\boldsymbol{\epsilon}}^{(n)}\|_\infty \leq h^k\frac{\delta}{\gamma}\|\boldsymbol{\tau}^{(n)}\|_\infty$.

Dim. Si considerino i vettori $\boldsymbol{y} = \pm\widehat{\boldsymbol{\epsilon}}^{(n)} + \frac{1}{\gamma}h^k\|\boldsymbol{\tau}^{(n)}\|_\infty\boldsymbol{w}^{(n)}$. Vale

$$L_n\boldsymbol{y} = \pm h^k\boldsymbol{\tau}^{(n)} + \frac{1}{\gamma}h^k\|\boldsymbol{\tau}^{(n)}\|_\infty L_n\boldsymbol{w}^{(n)} \geq \pm\boldsymbol{h}^k\tau^{(n)} + h^k\|\boldsymbol{\tau}^{(n)}\|_\infty\boldsymbol{e}^{(n)} \geq 0$$

Per cui, per il principio del massimo segue che $y_i \leq \max(y_0, y_{n+1})$. Poichè $\widehat{\epsilon}_0^{(n)} = \widehat{\epsilon}_{n+1}^{(n)} = 0$, vale $\max(y_0, y_{n+1}) \leq \frac{\delta}{\gamma} h^k \|\boldsymbol{\tau}^{(n)}\|_\infty$ da cui $y_i \leq \frac{\delta}{\gamma} h^k \|\boldsymbol{\tau}^{(n)}\|_\infty$, che per la definizione di y implica

$$\pm\boldsymbol{\epsilon}^{(n)} \leq \frac{\delta}{\gamma} h^k \|\boldsymbol{\tau}^{(n)}\|_\infty - \frac{1}{\gamma} h^k \|\boldsymbol{\tau}^{(n)}\|_\infty w_i^{(n)} \leq \frac{\delta}{\gamma} h^k \|\boldsymbol{\tau}^{(n)}\|_\infty.$$

□

L'esistenza di un vettore $\boldsymbol{w}^{(n)}$ con la proprietà ii) del teorema precedente si deduce generalmente dall'esistenza di una funzione $w(x)$ definita su $[0, 1]$ tale che $w(x) \geq 0$, $L[w] \geq \gamma > 0$, ponendo $w_i^{(n)} = w(x_i^{(n)})$.

4.4.3 Commenti sul principio del massimo

Osservazione 9 Il principio del massimo applicato a L_n è implicato da queste due proprietà:

- A_n è tale che $A_n^{-1} \leq 0$
- $L_n(\boldsymbol{e}^{(n)}) = 0$.

Infatti, $L_n(\boldsymbol{v}^{(n)}) \geq 0$ implica che

$$\begin{aligned}
0 \leq L_n(\boldsymbol{v}^{(n)}) =& A_n \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{n-1} \\ v_n \end{bmatrix} + (n+1)^2 \begin{bmatrix} v_0 \\ 0 \\ \vdots \\ 0 \\ v_{n+1} \end{bmatrix} \\
\leq& A_n \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{n-1} \\ v_n \end{bmatrix} + (n+1)^2 \max(v_0, v_{n+1}) \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \\
=& A_n \left(\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{n-1} \\ v_n \end{bmatrix} - \max(v_0, v_{n+1}) \boldsymbol{e}^{(n)} \right)
\end{aligned}$$

dove l'ultima uguaglianza segue dal fatto che $L_n(\boldsymbol{e}^{(n)}) = 0$. Cioè si ottiene che $A_n(\boldsymbol{v}^{(n)} - \max(v_0, v_{n+1}\boldsymbol{e}^{(n)})) \geq 0$. Poiché $-A_n^{-1} \geq 0$, moltiplicando entrambi i

membri della diseguaglianza precedente per $-A^{-1}$, segue che

$$\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{n-1} \\ v_n \end{bmatrix} \leq \max(v_0, v_{n+1})\boldsymbol{e}^{(n)}$$

che è il principio del massimo.

Osservazione 10 La condizione $L_n(\boldsymbol{e}^{(n)}) = 0$ nell'osservazione 9 può essere sostituita con la condizione più debole $L_n(\boldsymbol{e}^{(n)}) \leq 0$, affinché valga ancora il principio del massimo discreto.

Vale il seguente risultato che fornisce una condizione sufficiente facilmente verificabile affinché $A_n^{-1} \leq 0$

Teorema 11 *Se A_n è dominante diagonale e irriducibile e inoltre $a_{i,i} < 0$, $a_{i,j} \geq 0$ per $i \neq j$, allora A_n è invertibile e $A_n^{-1} \leq 0$.*

Dim. Vale $-A_n = D - B$, con $D = -\text{diag}(a_{1,1}, \ldots, a_{n,n})$, e $B \geq 0$. Inoltre $-A_n = D(I - D^{-1}B) = D(I - C)$ con $\rho(C) < 1$ per il terzo teorema di Gerschgorin. Per cui A_n è invertibile e $(-A_n)^{-1} = (I - C)^{-1}D^{-1} = (\sum_{i=0}^{+\infty} C^i)D^{-1} \geq 0$ essendo $C \geq 0$ e $\rho(C) < 1$. □

Osservazione 12 La classe di matrici scrivibili come $M = \alpha I - N$ con $N \geq 0$, cioè con elementi non negativi e $\alpha \geq \rho(N)$, dove $\rho(N)$ è il raggio spettrale di N, è nota come *classe delle M-matrici*. Se M è una M-matrice nonsingolare, poiché $\frac{1}{\alpha}M = I - \frac{1}{\alpha}N$, e $\rho(\frac{1}{\alpha}N) < 1$, la serie $\sum_{i=0}^{\infty}(\frac{1}{\alpha}N)^i$ è convergente e risulta $M^{-1} = \frac{1}{\alpha}\sum_{i=0}^{\infty}(\frac{1}{\alpha}N)^i \geq 0$. Segue dai teoremi di Gerschgorin che una matrice $M = \alpha I - N$, $N \geq 0$, irriducibile, dominante diagonale e con elementi diagonali positivi è una M-matrice non singolare per cui la sua inversa è non negativa.

4.5 Implementazione in Octave

È molto semplice scrivere un programma in Octave che risolve numericamente il problema modello $u''(x) = f(x)$ per $0 < x < 1$ con le condizioni $u(0) = u(1) = 0$, mediante il metodo delle differenze finite. Per poter trattare dimensioni elevate conviene adottare la modalità di rappresentazione delle matrici "sparse" fornita da Octave in cui vengono memorizzati solamente gli elementi diversi da zero di una matrice assieme ai loro indici. Per fare questo usiamo il comando `sparse`. Ad esempio

```
a=sparse(eye(n));
```

assegna alla variabile `a` la matrice identica di dimensione n rappresentata in modalità sparsa. Inoltre

Listing 2: Risoluzione del problema modello con condizioni omogenee di Dirichlet mediante differenze finite

```
function v=problema_modello(f)
n=length(f);
h=1/(n+1);
d=sparse(ones(n-1,1));
A=sparse(eye(n));
A=diag(d,1)+diag(d,-1)-2*A;
v=h^2*(A\f);
```

```
d=sparse(ones(n-1,1));
b=diag(d,1);
c=diag(d,-1);
```

costruiscono le matrici sparse `d` e `c` che hanno 1 rispettivamente nella sopra-diagonale e nella sotto-diagonale e zeri altrove.

In questo modo si può scrivere facilmente la function che risolve numericamente il problema modello con condizioni omogenee al contorno. Questa è riportata nel Listing 2

In figura 3 si riporta la soluzione ottenuta con $n = 1000$ e f funzione identicamente uguale a 1.

Una *function* commentata che tratta condizioni al bordo di Dirichlet non omogenee, in cui viene dato in uscita oltre che ai valori della funzione anche i punti della discretizzazione, è riportata nel Listing 3

La figura 4 riporta il grafico della soluzione ottenuta con $\alpha = 0$, $\beta = 0.2$ e con $f(x) = 1$ per $x < 1/2$, $f(x) = 0$ per $x \geq 1/2$.

4.6 Altre condizioni al contorno

Si consideri il problema

$$u''(x) = f(x), \quad x \in (0,1)$$
$$u'(0) = a, \quad u'(1) = b$$

Le condizioni al contorno di Neumann possono essere approssimate mediante la formula (5) nel modo seguente

$$\frac{1}{2h}(u_1 - u_{-1}) = a + \sigma_0 h^2, \quad \frac{1}{2h}(u_{n+2} - u_n) = b + \sigma_{n+1} h^2$$

dove $|\sigma_0|, |\sigma_{n+1}| \leq \frac{1}{3}\max|u'''(x)|$. Affiancando le precedenti equazioni alla condizione

$$\frac{1}{h^2}(u_{i-1} - 2u_i + u_{i+1}) = f_i + h^2\tau_i, \quad i = 0, \ldots, n+1,$$

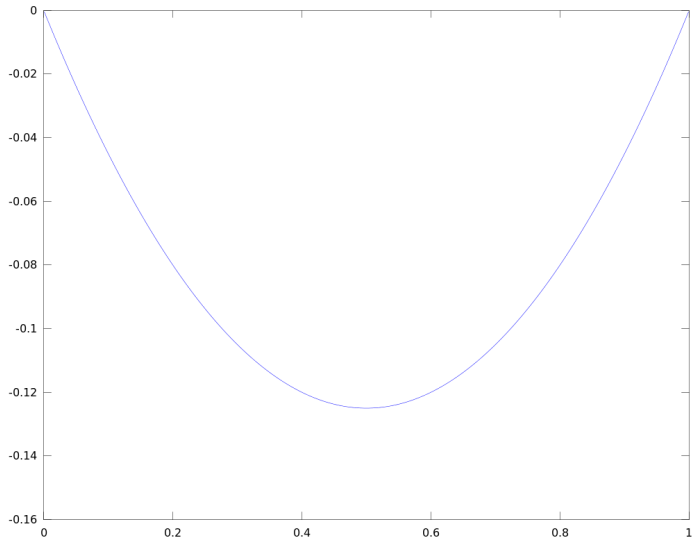


Figura 3: Soluzione del problema modello con $f(x) = 1$

Listing 3: Risoluzione del problema modello con condizioni di Dirichlet mediante differenze finite

```
function [v,x]=problema_modello1(f,alfa,beta)
% [v,x]=problema_modello(f,alfa,beta)
% risolve l''(x)=f(x), u(0)=alfa, u(1)=beta
% col metodo delle differenze finite sulla griglia x_i=ih, h=1/(n+1)
% f=(f_1,f_2,...,f_n), f_i=f(x_i),i=1,...,n
% v=(v_0,...,v_{n+1}) approssimazione della funzione u in x_0,...,x_{n+1}
% x=(x_0,...,x_{n+1})
n=length(f);
h=1/(n+1);
d=sparse(ones(n-1,1));
A=sparse(eye(n));
A=diag(d,1)+diag(d,-1)-2*A;
f(1)=f(1)-alfa/h^2;
f(n)=f(n)-beta/h^2;
w=h^2*(A\f);
v=[alfa;w;beta];
x=[0:n+1]/(n+1);
```

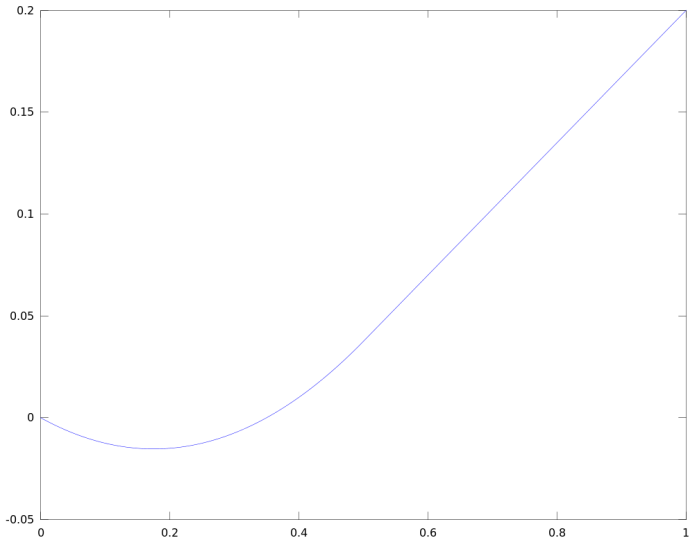


Figura 4: Soluzione del problema modello con $\alpha = 0$, $\beta = 0.2$, $f(x) = 1$ per $0 \leq x < 1/2$, $f(x) = 0$ per $1/2 \leq x \leq 1$.

si ottiene il sistema

$$-\frac{1}{h^2}\begin{bmatrix} h/2 & 0 & -h/2 & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & h/2 & 0 & -h/2 \end{bmatrix}\begin{bmatrix} u_{-1} \\ u_0 \\ \vdots \\ u_{n+1} \\ u_{n+2} \end{bmatrix} = \begin{bmatrix} a \\ f_0 \\ \vdots \\ f_{n+1} \\ b \end{bmatrix} + h^2 \begin{bmatrix} \sigma_0 \\ \tau_0 \\ \vdots \\ \tau_{n+1} \\ \sigma_{n+1} \end{bmatrix}$$

Si osservi che nel termine noto compare oltre all'errore locale di discretizzazione di L, anche l'errore locale dovuto alla discretizzazione delle condizioni al contorno di Neumann. Semplifichiamo il sistema precedente effettuando una combinazione lineare di righe. Più precisamente, moltiplicando la prima equazione per $2/h$, addizionando con la seconda e dividendo per 2, si ottiene

$$\frac{1}{h^2}(-u_0 + u_1) = \frac{1}{2}f_0 + a/h + \frac{1}{2}h^2\tau_0 + h\sigma_0$$

e analogamente moltiplicando l'ultima equazione per $2/h$, addizionando con la penultima e dividendo per 2, si ottiene

$$\frac{1}{h^2}(-u_{n+1} + u_n) = \frac{1}{2}f_{n+1} + b/h + \frac{1}{2}h^2\tau_{n+1} + h\sigma_{n+1}$$

La versione discreta del problema di Neumann diventa allora

$$-\frac{1}{h^2}\begin{bmatrix} 1 & -1 & & & \\ -1 & 2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{bmatrix}\begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_n \\ u_{n+1} \end{bmatrix} = \begin{bmatrix} \frac{1}{2}f_0 + a/h \\ f_1 \\ \vdots \\ f_n \\ \frac{1}{2}f_{n+1} + b/h \end{bmatrix} + \frac{1}{2}h^2\boldsymbol{\tau}^{(n)} + h\begin{bmatrix} \sigma_0 \\ 0 \\ \vdots \\ 0 \\ \sigma_{n+1} \end{bmatrix}$$

dove $\boldsymbol{\tau}^{(n)} = (\tau_0^{(n)}, \ldots, \tau_{n+1}^{(n)})^T$. Stavolta nel termine di errore compare, oltre all'errore locale di discretizzazione dell'operatore $L[u]$ anche l'errore dovuto alla discretizzazione della condizione di Neumann.

Si osservi che la matrice del sistema è singolare. Infatti il vettore $\boldsymbol{e}$ di tutte componenti uguali a 1 sta nel nucleo della matrice. La cosa non sorprende poiché il problema differenziale, se ha una soluzione, ne ha infinite; infatti le soluzioni differiscono per una costante additiva, così come accade per il problema discreto.

Supponiamo esista una soluzione e fissiamo la costante additiva imponendo che il valore di $u(x)$ sia assegnato in un estremo. Possiamo quindi ignorare una delle due condizioni di Neumann visto che che dipende linearmente dalle altre equazioni. Se imponiamo la condizione $u(0) = c$ e rimuoviamo la condizione di Neumann nell'estremo destro otteniamo il sistema

$$-\frac{1}{h^2}\begin{bmatrix} -1 & & & & \\ 2 & -1 & & & \\ -1 & 2 & \ddots & & \\ & \ddots & \ddots & -1 & \\ & & -1 & 2 & -1 \end{bmatrix}\begin{bmatrix} u_1 \\ \vdots \\ u_n \\ u_{n+1} \end{bmatrix} = \begin{bmatrix} \frac{1}{2}f_0 + \frac{a}{h} + \frac{1}{h^2}c \\ f_1 - \frac{1}{h^2}c \\ f_2 \\ \vdots \\ f_n \end{bmatrix} + \frac{h^2}{2}\begin{bmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_{n+1} \end{bmatrix} + h\begin{bmatrix} \sigma_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

che ha matrice non singolare. Denotando con T_n la matrice triangolare inferiore del sistema e indicando con $\boldsymbol{v}^{(n)}$ il vettore che risolve

$$T_n\boldsymbol{v}^{(n)} = \boldsymbol{b}^{(n)} \tag{19}$$

dove $\boldsymbol{b}^{(n)}$ è il vettore nella parte destra ottenuto rimuovendo la componente di errore, per l'errore globale $\boldsymbol{\epsilon}^{(n)} = \boldsymbol{u}^{(n)} - \boldsymbol{v}^{(n)}$ si ottiene

$$T_n\boldsymbol{\epsilon}^{(n)} = \frac{1}{2}h^2\begin{bmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_n \end{bmatrix} + h\begin{bmatrix} \sigma_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Si osservi che in questo modo stiamo risolvendo il problema *ai valori iniziali*

$$\begin{cases} u''(f) = f(x) & 0 < x < 1 \\ u(0) = c, \quad u'(0) = a \end{cases}$$

Si osservi ancora che lo schema alle differenze finite (19) è *esplicito* cioè la matrice del sistema è in forma triangolare per cui i valori v_i^n possono essere espressi in forma esplicita in funzione dei valori $v_j^{(n)}$ per $j < i$.

Per dimostrare la stabilità di questo schema basta dimostrare che $\|T_n^{-1}\|_\infty$ è limitata superiormente da una costante. Posto

$$Z = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ O & & 1 & 0 \end{bmatrix}$$

si osserva che $T_n = \frac{1}{h^2}(I - 2Z + Z^2) = \frac{1}{h^2}(I - Z)^2$. Per cui

$$T_n^{-1} = h^2(I - Z)^{-2} = h^2 \begin{bmatrix} 1 & & & \\ 1 & 1 & & \\ \vdots & \ddots & \ddots & \\ 1 & \dots & 1 & 1 \end{bmatrix}^2$$

cioè

$$T_n^{-1} = h^2 \begin{bmatrix} 1 & & & & \\ 2 & 1 & & & \\ 3 & 2 & 1 & & \\ \vdots & \ddots & \ddots & \ddots & \\ n+1 & n & \dots & 2 & 1 \end{bmatrix} \tag{20}$$

per cui

$$\|T_n^{-1}\|_\infty = h^2(1 + 2 + \cdots + (n+1)) = \frac{1}{(n+1)^2}\frac{(n+2)(n+1)}{2} \leq \frac{n+2}{2(n+1)} < 1,$$

inoltre $\boldsymbol{\epsilon}^{(n)} = \frac{1}{2}h^2 T_n^{-1}\boldsymbol{\tau}^{(n)} + hT_n^{-1}\sigma_0 \boldsymbol{e}_1^{(n)}$. Quindi, poiché

$$T_n^{-1}h \begin{bmatrix} \sigma_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = h^2\sigma_0 \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n+1} \end{bmatrix}, \quad x_i = ih,$$

l'errore globale rimane di ordine $O(h^2)$.

4.7 Condizioni al contorno di tipo misto

Nel caso in cui le condizioni al contorno siano del tipo misto

$$u'(0) = a, \quad u(1) = b$$

la matrice del sistema, di dimensione $(n+1) \times (n+1)$ prende la forma

$$-\frac{1}{h^2}\begin{bmatrix} 1 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix} =: -\frac{1}{h^2}K_n$$

mentre l'equazione differenziale discretizzata prende la forma

$$-\frac{1}{h^2}K_n\begin{bmatrix} u_0^{(n)} \\ u_1^{(n)} \\ \vdots \\ u_n^{(n)} \end{bmatrix} = \begin{bmatrix} \frac{1}{2}f_0 + \frac{1}{h}a \\ f_1 \\ \vdots \\ fn-1 \\ f_n - h^2 b \end{bmatrix} + h\sigma_0 \boldsymbol{e}_1 + h^2 \boldsymbol{\tau}^{(n)}.$$

Quindi la relazione tra errore globale ed errori locali diventa

$$-\frac{1}{h^2}K_n \boldsymbol{\epsilon}^{(n)} = h\sigma_0 \boldsymbol{e}_1 + h^2 \boldsymbol{\tau}^{(n)}$$

da cui

$$\boldsymbol{\epsilon}^{(n)} = (-\frac{1}{h^2}K_n)^{-1} h\sigma_0 \boldsymbol{e}_1 + (-\frac{1}{h^2}K_n)^{-1} h^2 \boldsymbol{\tau}^{(n)} \tag{21}$$

dove abbiamo indicato con $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^T$ il primo versore della base canonica di $\mathbb{R}^{n+1}$. Per i teoremi di Gerschgorin la matrice K_n è non singolare, inoltre per il teorema 5 l'autovalore più piccolo di K_n è limitato inferiormente da una costante indipendent da n per cui la norma $\|K_n^{-1}\|_2$ è limitata superiormente da una costante. Questo dimostra la stabilità in norma 2.

Per una analisi in norma infinito si usa l'espressione di K_n^{-1} in termini dell'inversa di $H = \text{trid}_{n+1}(-1, 2-1)$ data dalla formula di Sherman-Woodbury-Morrison (SWM). Infatti, poiché $K_n = H - \boldsymbol{e}_1 \boldsymbol{e}_1^T$, si ha

$$K_n^{-1} = H^{-1} + \frac{1}{1 - \boldsymbol{e}_1^T H^{-1} \boldsymbol{e}_1} H^{-1} \boldsymbol{e}_1 \boldsymbol{e}_1^T H^{-1}. \tag{22}$$

Ora esaminiamo separatamente le due quantità $(-\frac{1}{h^2}K_n)^{-1} h\sigma_0 \boldsymbol{e}_1$ e $(-\frac{1}{h^2}K_n)^{-1} h^2 \boldsymbol{\tau}^{(n)}$ che compongono l'errore globale $\boldsymbol{\epsilon}^{(n)}$ in (21).

Per la prima, usando la formula di SWM (22), si ha

$$\begin{aligned}(-\frac{1}{h^2}K_n)^{-1} h\sigma_0 \boldsymbol{e}_1 &= -h^3 \sigma_0 K_n^{-1} \boldsymbol{e}_1 = -h^3\sigma_0 (1 + \frac{\boldsymbol{e}_1^T H^{-1} \boldsymbol{e}_1}{1 - \boldsymbol{e}_1^T H^{-1} \boldsymbol{e}_1}) H^{-1} \boldsymbol{e}_1 \\ &= -h^3 \sigma_0 \frac{1}{1 - \boldsymbol{e}_1^T} H^{-1} \boldsymbol{e}_1.\end{aligned}$$

Essendo

$$H^{-1} \boldsymbol{e}_1 = \frac{1}{n+2}(n+1, n, \ldots, 2, 1)^T, \tag{23}$$

risulta

$$\boldsymbol{e}_1^T H^{-1} \boldsymbol{e}_1 = \frac{n+1}{n+2}, \tag{24}$$

da cui $(-\frac{1}{h^2}K_n)^{-1} h\sigma_0 H^{-1} \boldsymbol{e}_1 = -h^3\sigma_0 (n+2) H^{-1} \boldsymbol{e}_1 = -h^2 \sigma_0 \frac{n+2}{n+1} \boldsymbol{e}_1$, cioè

$$\|(-\frac{1}{h^2}K_n)^{-1} h\sigma_0 \boldsymbol{e}_1\|_\infty \leq |\sigma_0| h^2 \frac{n+2}{n+1}.$$

Per quanto riguarda la seconda quantità vale

$$\|(-\frac{1}{h^2}K_n)^{-1}h^2\boldsymbol{\tau}^{(n)}\|_\infty \leq \|(-\frac{1}{h^2}K_n)^{-1}\|_\infty h^2\|\boldsymbol{\tau}^{(n)}\|_\infty.$$

Inoltre una analisi in norma infinito ci fornisce

$$\|(-\frac{1}{h^2}K_n)^{-1}\|_\infty = h^2\|K_n^{-1}\|_\infty \leq h^2(\|H^{-1}\|_\infty + \frac{1}{2}(n+1)^2 \leq \frac{1}{8}+\frac{1}{2}=\frac{5}{8},$$

dove abbiamo usato (22), (23), (24) e (16). Per cui si ha stabilità del metodo, inoltre anche il secondo addendo è limitato superiormente da una costante per h^2. Questo dimostra la convergenza dello schema alle differenze con ordine h^2.

Una analisi di stabilità più semplice si ottiene osservando che per la matrice $K_n \in \mathbb{R}^{(n+1)\times(n+1)}$ vale la fattorizzazione LU

$$K_n = LL^T, \quad L = \begin{bmatrix} 1 & & & & \\ -1 & 1 & & & \\ 0 & -1 & 1 & & \\ \vdots & \ddots & \ddots & \ddots & \\ 0 & \dots & 0 & -1 & 1 \end{bmatrix}$$

Per cui $K_n^{-1} = L^{-T}L^{-1}$, L^{-1} è la matrice triangolare inferiore con tutti elementi uguali a 1, quindi $\|L^{-1}\|_\infty = \|L^{-T}\|_\infty = n+1$. Si deduce allora che

$$\|(\frac{1}{h^2}K_n)^{-1}\|_\infty \leq h^2\|L^{-1}\|_\infty\|L^{-T}\|_\infty = 1.$$

Inoltre, per quanto riguarda la parte di errore globale dato dalla condizione di Neuman, si ha

$$(\frac{1}{h^2}K_n)^{-1}h\sigma_0\boldsymbol{e}_1^{(n)} = h^3\sigma_0 L^{-T}\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix} = h^3\sigma_0\begin{bmatrix}1\\2\\\vdots\\n+1\end{bmatrix}$$

Da cui

$$\|(\frac{1}{h^2}K_n)^{-1})h\sigma_0\boldsymbol{e}_1^{(n)}\| = h^3|\sigma_0|(n+1) = h^2|\sigma_0|.$$

4.7.1 La formula di Sherman-Woodbury-Morrison

Se le matrici $n \times n$ A e B sono tali che

$$A = B + UV^T$$

dove U e V sono matrici $n \times m$, $m < n$, e se B è non singolare, allora A è non singolare se e solo se la matrice $m \times m$

$$S = I_m + V^T B^{-1} U$$

è non singolare e vale la formula di *Sherman-Woodbury-Morrison* per l'inversa di A

$$A^{-1} = B^{-1} - B^{-1}US^{-1}V^TB^{-1}.$$

Listing 4: Risoluzione del problema modello con condizioni al contorno miste mediante differenze finite

```
function [v,x]=miste(f,alfa,beta)
% [v,x]=miste(f,alfa,beta)
% risolve l''equazione u''(x)=f(x), u'(0)=alfa, u(1)=beta
% col metodo delle differenze finite sulla griglia x_i=ih, h=1/(n+1)
% f=(f_1,f_2,...,f_n), f_i=f(x_i),i=1,...,n
% v=(v_0,...,v_{n+1}) approssimazione della funzione u in x_0,...,x_{n+1}
% x=(x_0,...,x_{n+1})
n=length(f);
h=1/(n+1);
d=sparse(ones(n-1,1));
A=sparse(eye(n));
A=diag(d,1)+diag(d,-1)-2*A;
A(1,1)=-1;
f(1)=f(1)+alfa/h;
f(n)=f(n)-beta/h^2;
w=h^2*(A\f);
v=[w;beta];
x=[1:n+1]/(n+1);
```

4.8 Implementazione Octave

La function riportata nel Listing 4 fornisce una implementazione per il metodo delle differenze finite applicato al problema $u''(x) = f(x)$ con le condizioni miste $u'(0) = \alpha$, $u(1) = \beta$.

La figura 5 riporta il grafico della soluzione calcolata con $\alpha = 0$, $\beta = 0.2$ e con $f(x) = 1$ per $x < 1/2$, $f(x) = 0$ per $x \geq 1/2$.

Si noti la tangente orizzontale nell'estremo sinistro data dalla condizione $u'(0) = 0$.

5 Equazioni più generali

Trattiamo in questa sezione il caso dell'equazione $L[u] = f$ per $0 < x < 1$ con $u(0) = a$, $u(1) = b$, dove

$$L[u] = (c(x)u'(x))'$$

dove $c(x) : [0, 1] \to \mathbb{R}$ è una funzione sufficientemente regolare tale che $c(x) > 0$.

Analogamente al problema modello trattato nella sezione precedente, possiamo facilmente verificare che la soluzione dell'equazione differenziale esiste ed è sufficientemente regolare se lo è la funzione $f(x)$. Infatti, integrando l'espressione $L[u] = f(x)$ si ha

$$u'(x) = \frac{1}{c(x)} \int_0^x f(t)dt + \frac{\gamma_1}{c(x)}.$$

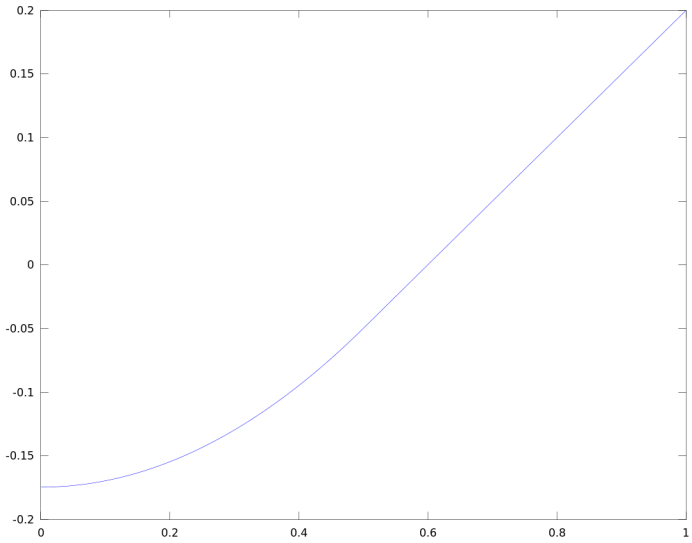


Figura 5: Soluzione del problema modello con condizioni miste e $\alpha = 0$, $\beta = 0.2$, $f(x) = 1$ per $0 \leq x < 1/2$, $f(x) = 0$ per $1/2 \leq x \leq 1$.

Integrando nuovamente si ottiene

$$u(x) = \int_0^x \frac{1}{c(s)} \int_0^s f(t)dtds + \int_0^x \frac{\gamma_1}{c(t)}dt + \gamma_2.$$

Le costanti γ_1 e γ_2 vengono determinate imponendo le condizioni al contorno. Si osserva anche che se $f(x)$ è di classe $C^2[0,1]$ e $c(x)$ è di classe $C^3[0,1]$ allora la soluzione $u(x)$ è di classe $C^4[0,1]$.

Dalla formula (5) di discretizzazione della derivata prima, con h sostituito da $h/2$, si ha

$$c(x)u'(x) = c(x)\frac{u(x+h/2) - u(x-h/2)}{h} + c(x)\frac{h^2}{48}(u'''(\xi) + u'''(\eta))$$

dove $\xi \in (x, x+h/2)$, $\eta \in (x-h/2, x)$.

Applicando nuovamente la (5) alla funzione $g(x) = c(x)u'(x)$ si ricava

$$(c(x)u'(x))' = \frac{g(x+h/2) - g(x-h/2)}{h} + \frac{h^2}{48}(g'''(\hat{\xi}) + g'''(\hat{\eta}))$$

con $\hat{\xi} \in (x, x+h/2)$, $\hat{\eta} \in (x-h/2, x)$. Componendo le due relazioni così ottenute

si ottiene

$$(c(x)u'(x))' = \frac{1}{h}\left(c(x+h/2)\frac{u(x+h)-u(x)}{h} - c(x-h/2)\frac{u(x)-u(x-h)}{h}\right) + h^2\tau(x) \tag{25}$$

dove $|\tau(x)|$ è limitato superiormente da una costante indipendente da n. Infatti, il valore di $h^2\tau$ è dato dalla somma di $h^2\hat{\tau} = \frac{h^2}{48}(g'''(\hat{\xi}) + g'''(\hat{\eta}))$ e di

$$h^2\tilde{\tau} = \frac{1}{h}(c(x+h/2)\frac{h^2}{48}(g'''(\xi_1) + g'''(\eta_1)) - c(x-h/2)\frac{h^2}{48}(g'''(\xi_2) + g'''(\eta_2)))$$

dove $\xi_1 \in (x+h/2, x)$, $\eta_1 \in (x, x+h/2)$, $\xi_2 \in (x-h/2, x)$, $\eta_2 \in (x, x+h/2)$. Non è difficile verificare, usando la regolarità delle funzioni coinvolte, che $|\tau_2|$ è limitato superiormente da una costante. Basta infatti applicare la relazione

$$\begin{aligned} c(x+h/2)g'''(\xi_1) - c(x-h/2)g'''(\xi_2) =& (c(x+h/2) - c(x-h/2))g'''(\xi_1) \\ &+ c(x-h/2)(g'''(\xi_1) - g'''(\xi_2)) \end{aligned}$$

e analogamente per la parte che coinvolge η_1 ed η_2, tenendo presente che $0 < \xi_1 - \xi_2 < h$ e $0 < \eta_1 - \eta_2 < h$.

Quindi, applicando la formula (25) nei punti $x_i^{(n)} = ih$ e denotando $c_i = c(x_i + h/2)$, si arriva alla discretizzazione

$$L_n = \frac{1}{h^2}\begin{bmatrix} c_0 & -(c_0+c_1) & c_1 & & & \\ & c_1 & -(c_1+c_2) & c_2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & c_{n-1} & -(c_{n-1}+c_n) & c_n \end{bmatrix}$$

$$A_n = \frac{1}{h^2}\begin{bmatrix} -(c_0+c_1) & c_1 & & & & \\ c_1 & -(c_1+c_2) & c_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & c_{n-2} & -(c_{n-2}+c_{n-1}) & c_{n-1} & \\ & & & c_{n-1} & -(c_{n-1}+c_n) \end{bmatrix}$$

Si osserva che, poiché $c(x) > 0$, la matrice $-A_n$ è una M-matrice non singolare, per cui $-A_n^{-1} \geq 0$. Inoltre vale $L_n(\boldsymbol{e}^{(n)}) = 0$, per cui alla luce delle osservazioni del paragrafo 4.4.3 per l'operatore discreto L_n vale il principio del massimo. Per dimostrare la convergenza e la stabilità dello schema alle differenze basta dimostrare che esiste una successione di vettori $\{\boldsymbol{w}^{(n)}\}_n$ tali che $\boldsymbol{w}^{(n)} \geq 0$, $w_0^{(n)}, w_{n+1}^{(n)} \leq \delta$ e $L_n(\boldsymbol{w}^{(n)}) \geq \gamma \boldsymbol{e}^{(n)}$ per una costante γ indipendente da n.

Per costruire $\boldsymbol{w}^{(n)}$ cerchiamo una funzione $w(x)$ tale che $L(w) = 1$ e poi definiamo $\boldsymbol{w}^{(n)}$ come il vettore di componenti $w(x_i^{(n)})$.

La condizione $L(w) = 1$ si riscrive come $(c(x)w'(x))' = 1$. Integrando si ottiene $c(x)w'(x) = x+\gamma_1$, da cui $w'(x) = (x+\gamma_1)/c(x)$. Integrando nuovamente si ottiene

$$w(x) = \int_0^x \frac{t+\gamma_1}{c(t)}dt + \gamma_2$$

Scegliamo $\gamma_1 = \gamma_2 = 0$ e si ottiene $w(x) = \int_0^x (t/c(t))dt$. Approssimando l'integrale con una sommatoria si ottiene il seguente candidato come $\boldsymbol{w}^{(n)}$:

$$w_i^{(n)} = h^2 \sum_{j=0}^{i} \frac{j}{c_{j-1}}, \quad i = 0, \ldots, n+1.$$

Poiché $(w_{i+1} - w_i)/h^2 = (i+1)/c_i$, ne segue allora

$$\begin{aligned}(L_n(\boldsymbol{w}^{(n)}))_i =& \frac{1}{h^2}(c_{i-1}w_{i-1} - (c_{i-1}+c_i)w_i + c_i w_{i+1}) \\ =& \frac{1}{h^2}(c_{i-1}(w_{i-1} - w_i) + c_i(w_{i+1} - w_i)) \\ =& (i+1)\frac{c_i}{c_i} - i\frac{c_{i-1}}{c_{i-1}} = 1.\end{aligned}$$

Si analizzi il caso dell'operatore di Sturm-Liouville

$$L[u] = -(p(x)u'(x))' + q(x)$$

dove $p(x), q(x) > 0$.

5.1 Approccio matriciale

Un modo diverso di studiare $\|A_n^{-1}\|_\infty$ si basa su un'analisi matriciale.

Si verifica facilmente che $-A_n = B_n + \frac{1}{h^2}c_n \boldsymbol{e}_n^{(n)} \boldsymbol{e}_n^{(n)T}$, dove

$$B_n = \frac{1}{h^2}\begin{bmatrix} c_0 + c_1 & -c_1 & & & & \\ -c_1 & c_1 + c_2 & -c_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -c_{n-2} & c_{n-2} + c_{n-1} & -c_{n-1} & \\ & & & -c_{n-1} & c_{n-1} \end{bmatrix}$$

Inoltre, una verifica diretta mostra che B_n si può fattorizzare come

$$B_n = \frac{1}{h^2}\begin{bmatrix} 1 & -1 & & \\ & 1 & \ddots & \\ & & \ddots & -1 \\ & & & 1 \end{bmatrix}\begin{bmatrix} c_0 & & & \\ & c_1 & & \\ & & \ddots & \\ & & & c_{n-1} \end{bmatrix}\begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}. \tag{26}$$

Si osserva infine che, poiché $c_i > 0$, $-A_n$ e B_n sono M-matrici non singolari essendo dominanti diagonali e irriducibili. Quindi le loro inverse hanno elementi positivi. Inoltre, dalla formula di Sherman-Woodbury-Morrison e dalle proprietà delle M-matrici segue che $0 \leq -A_n^{-1} \leq B_n^{-1}$. Quindi per la stabilità basta dimostrare che $\|B_n\|_\infty$ è limitata superiormente da una costante indipendente da n.

Poiché

$$\begin{bmatrix} 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \\ & & & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ & \ddots & \ddots & \vdots \\ & & 1 & 1 \\ & & & 1 \end{bmatrix}$$

si ha che $\|B_n\|_\infty \leq \frac{n^2}{(n+1)^2} \frac{1}{\min c_i} < 1/\min c_i$ che è limitato superiormente da una costante indipendente da n.

5.1.1 Analisi in norma 2

Si richiama la seguente proprietà delle matrici ad elementi non negativi: che se A e B sono matrici tali che $0 \leq a_{i,j} \leq b_{i,j}$ allora $\rho(A) \leq \rho(B)$. Inoltre si ricorda che per una matrice simmetrica A vale $\|A\|_2 = \rho(A)$.

Dunque dalla simmetria di A_n e B_n e dal fatto che $0 \leq -A_n^{-1} \leq B_n^{-1}$ segue $\| - A_n^{-1}\|_2 \leq \|B_n^{-1}\|_2$. Poichè

$$\begin{aligned} 0 \leq B_n^{-1} =& h^2 \mathrm{trid}_n(-1,1,0)^{-1} \mathrm{diag}(c_0, \dots c_{n-1})^{-1} \mathrm{trid}(0,1,-1)^{-1} \\ \leq& h^2 \frac{1}{\min c_i} \mathrm{trid}_n(-1,1,0)^{-1} \mathrm{trid}_n(0,1,-1)^{-1} \\ =& h^2 \frac{1}{\min c_i} (\mathrm{trid}_n(0,1,-1)\mathrm{trid}_n(-1,1,0))^{-1} \end{aligned}$$

vale $\|B_n^{-1}\|_2 \leq \frac{h^2}{\min c_i} \|K^{-1}\|_2$ dove

$$K = \mathrm{trid}_n(0,1,-1)\mathrm{trid}_n(-1,1,0) = \begin{bmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & 2 & -1 \\ & & -1 & 1 \end{bmatrix}$$

Per il teorema 5 gli autovalori della matrice K sono $2 - 2\cos(j\frac{2\pi}{n+1/2})$, $j = 1, \dots, n$ e corrispondono agli autovettori $\boldsymbol{v}_j = (\sin(ij\frac{2\pi}{n+1/2}))$. Il più piccolo autovalore è $2 - 2\cos(\frac{2\pi}{n+1/2}) = \frac{4\pi^2}{(n+1/2)^2} + O(h^4)$, per cui

$$\|K^{-1}\|_2 = \frac{(n+1/2)^2}{(n+1)^2} \frac{4}{\pi}^2 < \frac{4}{\pi^2} + O(h^2).$$

Listing 5: Risoluzione del problema $c(x)u'(x))' = f(x)$ con condizioni al contorno $u(0) = \alpha$, $u(1) = \beta$ mediante differenze finite

```
function [v,x]=generale(f,c,alfa,beta)
% [v,x]=generale(f,c,alfa,beta)
% risolve l'equazione (u'(x)c(x))'=f(x), u(0)=alfa, u(1)=beta
% col metodo delle differenze finite sulla griglia x_i=ih, h=1/(n+1)
% f=(f_1,f_2,...,f_n), f_i=f(x_i),i=1,...,n
% v=(v_0,...,v_{n+1}) approssimazione della funzione u in x_0,...,x_{n+1}
% c=(c_0,...,c_n) tale che c_i=c(x_i+h/2), i=0,...,n
% x=(x_0,...,x_{n+1})
n=length(f);
h=1/(n+1);
d0=sparse(c(1:n)+c(2:n+1));
d1=sparse(c(2:n));
A=-diag(d0)+diag(d1,1)+diag(d1,-1);
f(1)=f(1)-alfa*c(1)/h^2;
f(n)=f(n)-beta*c(n+1)/h^2;
w=h^2*(A\f);
v=[alfa;w;beta];
x=[0:n+1]/(n+1);
```

5.2 Implementazione in Octave

Nel Listing 5 è riportata la *function* Octave che risolve il problema $c(x)u'(x))' = f(x)$ con condizioni al contorno di Dirichlet $u(0) = \alpha$, $u(1) = \beta$.

La figura 6 riporta il grafico della soluzione del problema con $\alpha = \beta = 0$, $c(x) = x + 1/100$, $f(x) = 1$.

6 Il problema agli autovalori

Si consideri il "problema modello"

$$\begin{aligned} -u''(x) &= \lambda u(x), \quad 0 < x < 1 \\ u(0) &= 0, \ u(1) = 0 \end{aligned} \tag{27}$$

in cui occorre determinare gli scalari λ e le funzioni $u(x) : [0, 1] \to \mathbb{R}$, $u(x)$ non identicamente nulle, che verifichino (27). Gli scalari λ sono chiamati *autovalori* e le corrispondenti funzioni $u(x)$ *autofunzioni*. In questo caso speciale esiste una infinità numerabile di soluzioni date da $u_i(x) = \sin(\pi i x)$ e $\lambda_i = i^2\pi^2$, $i = 1, 2, \ldots$, per cui il problema della loro approssimazione di fatto non si pone. Usiamo però questo problema modello come paradigma per descrivere il suo trattamento numerico.

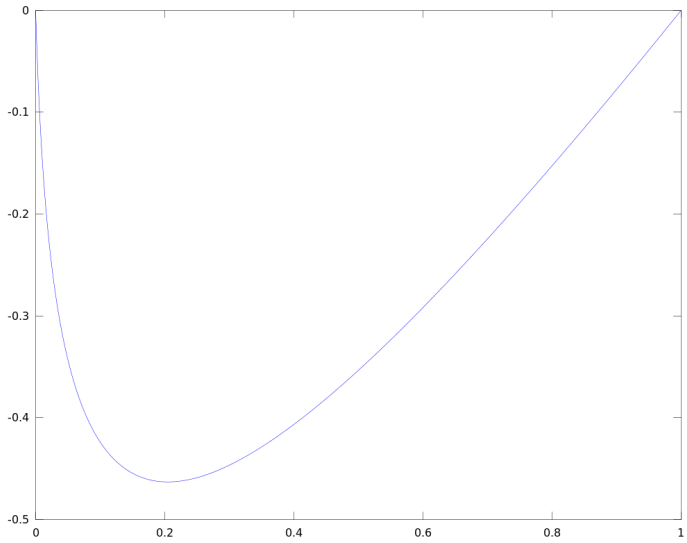


Figura 6: Soluzione del problema $c(x)u'(x))' = f(x)$, $u(0) = u(1) = 0$ con $f(x) = 1$ e $c(x) = x + 1/100$

Discretizzando l'intervallo $[0, 1]$ con i punti $x_i = ih$, $i = 0, \ldots, n+1$, $h = 1/(n+1)$, e utilizzando l'operatore discreto L_n si arriva alla seguente relazione

$$\frac{1}{h^2}\begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix} \boldsymbol{u}^{(n)} = \lambda \boldsymbol{u}^{(n)} + h^2 \boldsymbol{\tau}^{(n)} \tag{28}$$

dove $\tau^{(n)} = (\tau_i^{(n)})$ è tale che $|\tau_i^{(n)}| \leq \frac{1}{12} \max |u^{(4)}(x)|$, valida nel caso in cui $u(x)$ sia di classe $C^4[0, 1]$.

Si considera allora il problema

$$\frac{1}{h^2}\begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix} \boldsymbol{v}^{(n)} = \mu \boldsymbol{v}^{(n)} \tag{29}$$

e ci si chiede quanto le soluzioni del problema discreto (29) differiscano da quelle del problema continuo (27). Per questo è utile richiamare il teorema di Bauer-Fike. Ricordiamo che una norma assoluta $\|\cdot\|$ è tale che $\|\boldsymbol{x}\| = \|\boldsymbol{y}\|$

se $y_i = |x_i|$, $i = 1, \ldots, n$, inoltre ricordiamo che (vedi [1]) la norma matriciale indotta da una norma assoluta è tale che $\|D\| = \max |d_i|$ per ogni matrice diagonale $D = \mathrm{diag}(d_i)$. Osserviamo anche che le norme classiche 1,2 e ∞ sono norme assolute.

Teorema 13 (Bauer-Fike) *Siano A, B, F matrici $n \times n$ tali che $A = B + F$. Supponiamo che B sia diagonalizzabile, cioè $B = SDS^{-1}$ con D matrice diagonale. Sia inoltre $\|\cdot\|$ una norma assoluta. Allora, per ogni autovalore λ di A esiste un autovalore μ di B tale che*

$$|\lambda - \mu| \leq \|F\| \cdot \|S\| \cdot \|S^{-1}\|$$

Dim. Sia $A\boldsymbol{u} = \lambda \boldsymbol{u}$ per cui $(B - \lambda I)\boldsymbol{u} = -F\boldsymbol{u}$. Se λ è autovalore di B allora la tesi è vera. Se λ non è autovalore di B allora $B - \lambda I$ è non singolare e possiamo scrivere

$$\boldsymbol{u} = -(B - \lambda I)^{-1} F \boldsymbol{u}.$$

Questo implica che $\|(B - \lambda I)^{-1} F\| \geq 1$ da cui, poiché $(B - \lambda I)^{-1} = S(D - \lambda I)^{-1} S^{-1}$, si ha

$$1 \leq \|F\| \cdot \|S\| \cdot \|S^{-1}\| \cdot \|D - \lambda I\|.$$

Poiché $\|\cdot\|$ è una norma assoluta, si ha che $\|(D - \lambda I)^{-1}\| = 1/|\mu - \lambda|$ dove μ è un autovalore di B che ha distanza minima da λ. □

Si osservi che se S è ortogonale vale $\|S\|_2 = 1$. Ciò, unito al fatto che $\|\cdot\|_2$ è assoluta, implica il seguente

Corollario 14 *Nelle ipotesi del teorema precedente se B è simmetrica allora per ogni autovalore μ di B esiste un autovalore λ di A tale che*

$$|\lambda - \mu| \leq \|F\|_2$$

Per quanto riguarda lo studio della distanza tra l'autovettore $\boldsymbol{v}$ della matrice B e l'autovettore vettore $\boldsymbol{u}$ della matrice A premettiamo la definizione di decomposizione ai valori singolari (SVD) è la definizione di inversa generalizzata. Vale il seguente

Teorema 15 *Per ogni matrice $A \in \mathbb{R}^{m \times n}$ esistono matrici ortogonali $U \in \mathbb{R}^{m \times m}$ e $V \in \mathbb{R}^{n \times n}$ e una matrice diagonale $\Sigma = diag(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{m \times n}$ con $p = \min\{m, n\}$, e $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$, tali che $A = U\Sigma V^T$.*

La fattorizzazione $A = U\Sigma V^T$ è detta decomposizione ai valori singolari, le colonne di U e di V sono dette rispettivamente vettori singolari destri e sinistri. Gli scalari σ_i sono i valori singolari di A. Un risultato analogo vale per matrici ad elementi complessi dove U e V sono unitarie. Si osservi che le matrici $A^T A$ e AA^T hanno gli stessi autovalori non nulli che coincidono con i quadrati dei valori singolari non nulli. Inoltre le colonne di U e di V sono autovettori rispettivamente delle matrici AA^T e $A^T A$. Dalle colonne di V e U si ricavano facilmente delle basi ortogonali del nucleo di A e dello span di A.

Osservazione 16 *Se A è reale e simmetrica allora i valori singolari di A sono i valori assoluti degli autovalori di A. Se A è anche definita positiva i valori singolari di A coincidono con gli autovalori di A.*

Mediante la SVD si può definire il concetto di inversa generalizzata di una matrice $A \in \mathbb{R}^{m\times n}$ come $A^+ = V\Sigma^+U^T$, dove $\Sigma^+ = \text{diag}(\sigma_1^+, \ldots, \sigma_p^+) \in \mathbb{R}^{n\times m}$, e $\sigma_i^+ = 1/\sigma_i$ se $\sigma_i \neq 0$, $\sigma_i^+ = 0$ se $\sigma_i = 0$.

Vale il seguente risultato che permette di esprimere la soluzione di minima norma di un problema di minimi quadrati.

Teorema 17 *Dato il sistema $Ax = b$ con $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, il vettore $x^* = A^+b$ è il vettore di minima norma 2 che minimizza $\|Ax - b\|_2$.*

Teorema 18 *Nelle ipotesi del teorema di Bauer-Fike, sia $\boldsymbol{u} \in \mathbb{R}^n$ tale che, $(B + F)\boldsymbol{u} = \lambda\boldsymbol{u}$. Allora esistono un autovettore $\boldsymbol{v}$ e un autovalore μ di B tali che $B\boldsymbol{v} = \mu\boldsymbol{v}$ e*

$$\frac{\|\boldsymbol{v} - \boldsymbol{u}\|}{\|\boldsymbol{u}\|} \leq \|(B - \mu I)^+\| \cdot (\|F\| + |\lambda - \mu|) \leq \|(B - \mu I)^+\| \cdot \|F\|(1 + \|S\| \cdot \|S^{-1}\|)$$

dove $(B - \mu I)^+$ è l'inversa generalizzata di $B - \mu I$ e $S^{-1}BS = D$ è diagonale. In particolare, se B è simmetrica risulta

$$\frac{\|\boldsymbol{v} - \boldsymbol{u}\|_2}{\|\boldsymbol{u}\|_2} \leq \frac{2}{\min_{t\in\sigma(B), t\neq\mu} |t - \mu|}\|F\|_2.$$

dove $\sigma(B)$ è l'insieme degli autovalori di B.

Dim. Vale

$$(B + F - \lambda I)\boldsymbol{u} = 0$$
$$B\boldsymbol{v} - \mu\boldsymbol{v} = 0$$

Sottraendo entrambi i membri delle precedenti equazioni si ottiene

$$B(\boldsymbol{u} - \boldsymbol{v}) + F\boldsymbol{u} - (\lambda - \mu)\boldsymbol{u} - \mu(\boldsymbol{u} - \boldsymbol{v}) = 0.$$

Da cui

$$(B - \mu I)(\boldsymbol{u} - \boldsymbol{v}) = -F\boldsymbol{u} + (\lambda - \mu)\boldsymbol{u}$$

Quindi $F\boldsymbol{u}+(\lambda-\mu)\boldsymbol{u}$ sta nell'immagine di $B-\mu I$ e per il teorema 17 la soluzione $\boldsymbol{u} - \boldsymbol{v}$ di minima norma può essere scritta come

$$\boldsymbol{u} - \boldsymbol{v} = -(B - \mu I)^+(F - (\lambda - \mu)I)\boldsymbol{u}$$

inoltre vale

$$\|\boldsymbol{u} - \boldsymbol{v}\| \leq \|(B - \mu I)^+\|(\|F\| + |\lambda - \mu|)\|\boldsymbol{u}\|.$$

Da cui la prima parte della tesi. Se B è simmetrica, con autovalori β_i, si ha $\|S\|_2 \cdot \|S^{-1}\|_2 = 1$, per cui la tesi discende dall'osservazione 16; infatti i valori singolari di $(B - \mu I)^+$, per definizione di inversa generalizzata, sono $1/(\beta_i - \mu)$

se $\beta_i - \mu \neq 0$ e 0 altrimenti. $\square$

Siamo pronti per confrontare le soluzioni del problema continuo (27) con quelle del problema discreto (29). Infatti possiamo riscrivere (28) come

$$(B + F)\boldsymbol{u} = \lambda \boldsymbol{u}$$

dove si è indicato con $B = \frac{1}{h^2}\text{trid}(-1, 2, -1)$, e con $F = h^2\boldsymbol{\tau}\boldsymbol{u}^{(n)T}/(\boldsymbol{u}^{(n)T}\boldsymbol{u}^{(n)})$.

Applicando il teorema di Bauer-Fike con la norma 2, poiché

$$\|F\| = \frac{h^2}{\|\boldsymbol{u}\|^2}\|\boldsymbol{\tau}\| \cdot \|\boldsymbol{u}\| = h^2\|\boldsymbol{\tau}\|/\|\boldsymbol{u}\|,$$

si ha

$$|\lambda - \mu| \leq h^2 \frac{\|\boldsymbol{\tau}\|}{\|\boldsymbol{u}\|}. \tag{30}$$

Poiché $\tau_i = \frac{1}{24}(u^{(4)}(\xi_i) + u^{(4)}(\eta_i))$ con $\xi \in (x_i, x_{i+1})$ e $\eta_i \in (x_{i-1}, x_i)$, se $M = \max_{x\in[0,1]} |u^{(4)}(x)|$, allora $\frac{1}{\sqrt{n+1}}\|\boldsymbol{\tau}^{(n)}\|_2 \leq \frac{1}{12}M$, inoltre $\lim_n \frac{1}{\sqrt{n+1}}\|u(x)\|_2 = \left(\int_0^1 u(x)^2 dx\right)^{1/2}$ per cui il quoziente $\frac{\|\boldsymbol{\tau}\|_2}{\|\boldsymbol{u}\|_2}$ è limitato superiormente da una costante θ.

Inoltre, poiché gli autovalori di B sono $\beta_i = (n+1)^2(2 - 2\cos \pi i/(n+1)) = i^2\pi^2 + O(h^2)$, se $\mu = \beta_j$ allora $1/\min_{i\neq j} |\beta_i - \mu| = 1/((2j-1)\pi^2) + O(h^2)$ dove $1/((2j-1)\pi^2)$ non dipende da n. Per cui

$$\frac{\|\boldsymbol{v} - \boldsymbol{u}\|_2}{\|\boldsymbol{u}\|_2} \leq \gamma h^2 \frac{1}{(j - 1/2)\pi^2}\theta$$

per una costante $\gamma > 0$. Questo dimostra la convergenza della soluzione del problema discreto alla soluzione del problema continuo.

6.1 Implementazione in Octave

Nel Listing 6 è riportata la *function* Octave che risolve il problema agli autovalori $(c(x)u'(x))' = \lambda u(x)$ su $[0, 1]$ con condizioni al contorno di Dirichlet $u(0) = u(1) = 0$. In questo caso Octave non riesce a trarre vantaggio dalla sparsità della matrice per cui per valori moderatamente grandi di n si hanno tempi di calcolo elevati. Si suggerisce di provare con $n \leq 1000$.

La figura 7 riporta il grafico degli autovalori con $c(x) = x + 1/100$, mentre la figura 8 riporta le autofunzioni corrispondenti agli autovalori di modulo più piccolo.

7 Il caso multidimensionale

Si consideri il problema di Poisson con condizioni al contorno di Dirichlet

$$\begin{aligned} &\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y), \quad (x, y) \in \Omega = (0, 1) \times (0, 1) \\ &u(x, y) = g(x, y), \quad (x, y) \in \partial\Omega \end{aligned} \tag{31}$$

Listing 6: mediante differenze finite.]Soluzione del problema agli autovalori $(c(x)u'(x))' = \lambda u(x)$ sull'intervallo [0,1] mediante differenze finite.

```
function [v,lambda]=eig_generale(c)
% [v,lambda]=eig_generale(c)
% risolve il problema agli autovalori (u'(x)c(x))'=lambda u(x), u(0)=u(1)
    =0
% col metodo delle differenze finite sulla griglia x_i=ih, h=1/(n+1)
% v matrice le cui colonne approssimano le autofunzioni
% lambda vettore degli autovalori corrispondenti
% x=(x_0,...,x_{n+1})
n=length(c)-1;
h=1/(n+1);
d0=sparse(c(1:n)+c(2:n+1));
d1=sparse(c(2:n));
A=-diag(d0)+diag(d1,1)+diag(d1,-1);
A=A*(1/h^2);
[v,ei]=eig(A);
v=[zeros(1,n);v;zeros(1,n)];
lambda=diag(ei);
```

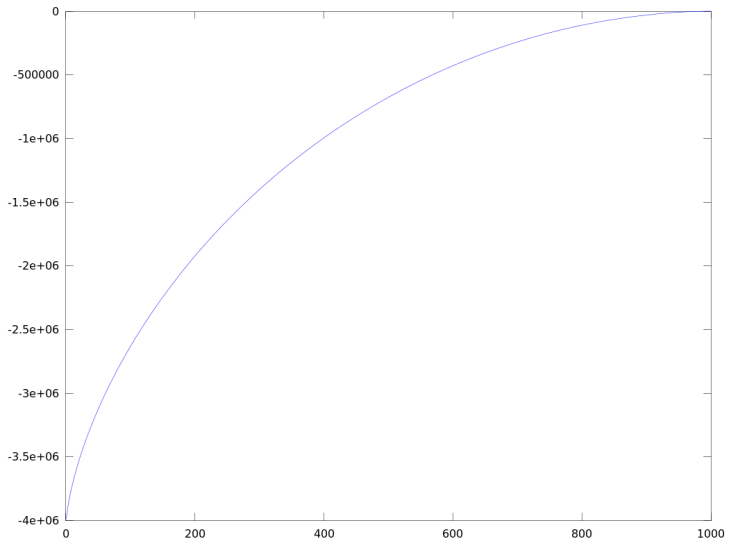


Figura 7: Autovalori del problema $c(x)u'(x))' = \lambda u(x)$, su [0,1] con $u(0) = u(1) = 0$ con e $c(x) = x + 1/100$

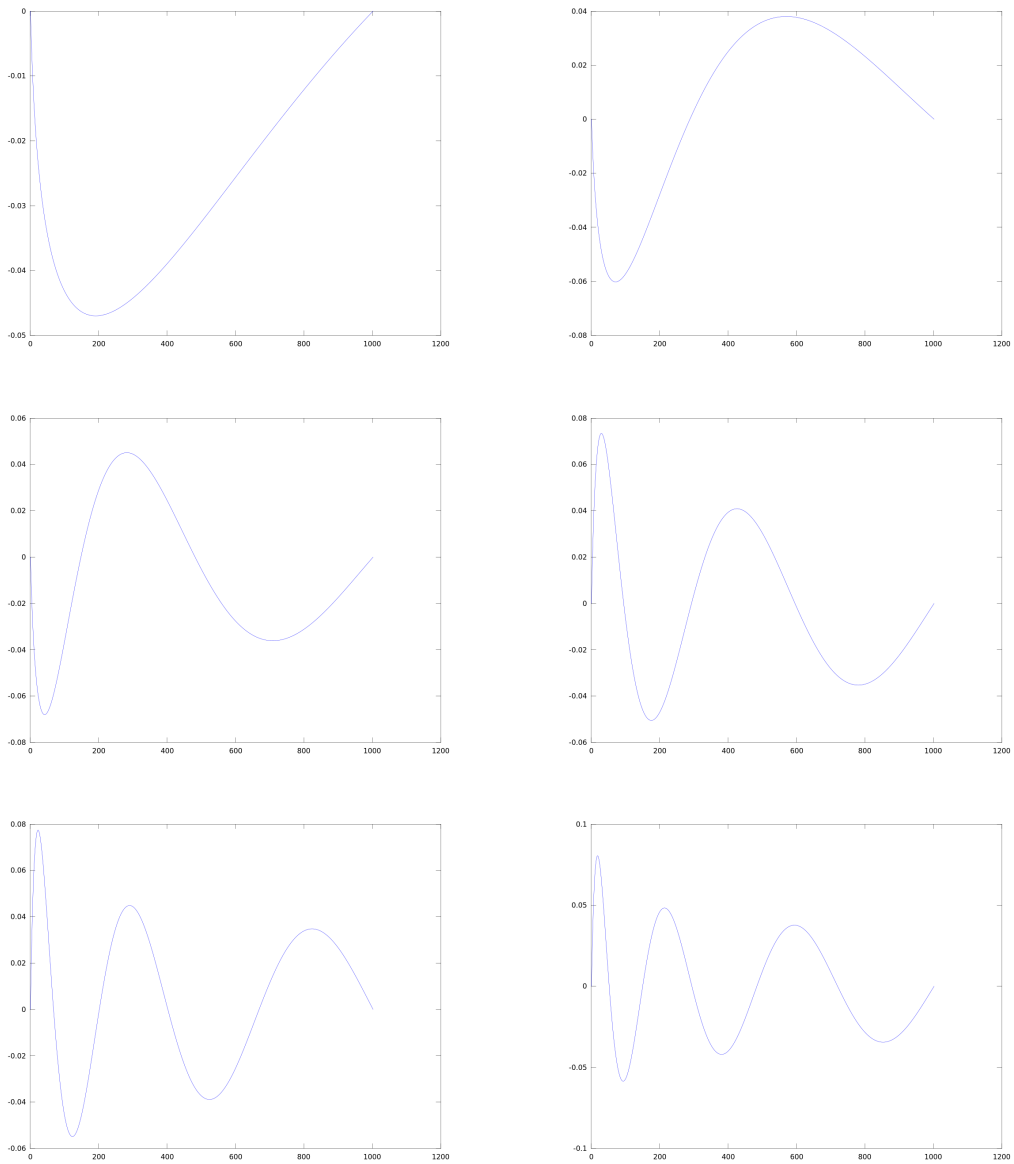


Figura 8: Prime 6 autofunzioni di $L(x) = (c(x)u'(x))'$ su $[0,1]$, con $c(x) = x + 1/100$, calcolate mediante differenze finite.

Si discretizzi il quadrato $[0,1]\times[0,1]$ con un reticolo di $(m+2)\times(n+2)$ punti (x_i,y_j), per $i=0,\ldots,m+1$, $j=0,\ldots,n+1$, dove $x_i=ih_x$, $y_j=jh_y$, con $h_x=1/(m+1)$, $h_y=1/(n+1)$.

Nell'ipotesi in cui la soluzione $u(x,y)$ è continua con le sue derivate parziali fino all'ordine 4, applicando la (4) alla funzione $u(x,y)$ vista separatamente come funzione della sola x e come funzione della sola y si ottiene

$$\begin{aligned}\frac{\partial^2 u}{\partial x^2}|_{(x_i,y_j)} =&\frac{1}{h_x}(u_{i-1,j}-2u_{i,j}+u_{i+1,j})+h_x^2\tau_{i,j}\\ \frac{\partial^2 u}{\partial y^2}|_{(x_i,y_j)} =&\frac{1}{h_y}(u_{i,j-1}-2u_{i,j}+u_{i,j+1})+h_y^2\nu_{i,j}\end{aligned} \tag{32}$$

dove $u_{i,j}=u(x_i,y_j)$ e $|\tau_{i,j}|\le\frac{1}{12}\max\left|\frac{\partial^2 u}{\partial x^2}\right|$, $|\nu_{i,j}|\le\frac{1}{12}\max\left|\frac{\partial^2 u}{\partial y^2}\right|$. Sommando le espressioni di (32) si arriva a

$$\begin{aligned}(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2})|_{(x_i,y_j)} =&\frac{1}{h_x^2}(u_{i-1,j}+u_{i+1,j})+\frac{1}{h_y^2}(u_{i,j-1}+u_{i,j+1})-2(\frac{1}{h_y^2}+\frac{1}{h_y^2})u_{i,j}\\ &+h_x^2\tau_{i,j}+h_y^2\nu_{i,j}.\end{aligned} \tag{33}$$

Per semplificare le notazioni, nel seguito supponiamo che $h_x=h_y=:h$ per cui la formula (33) diventa

$$(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2})|_{(x_i,y_j)}=\frac{1}{h^2}(u_{i-1,j}+u_{i+1,j}+u_{i,j-1}+u_{i,j+1}-4u_{i,j})+h^2\tau_{i,j}+h^2\nu_{i,j}. \tag{34}$$

La formula fornisce una approssimazione dell'operatore differenziale $\mathcal{L}[u]=\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}$ con l'operatore alle differenze finite

$$\mathcal{L}_{m,n}(U^{(m,n)})=\frac{1}{h^2}(u_{i-1,j}+u_{i+1,j}+u_{i,j-1}+u_{i,j+1}-4u_{i,j})_{i=1:m,j=1:n}$$

dove abbiamo indicato $U^{(m,n)}=(u_{i,j})_{i=0:m+1,j=0:n+1}$.

Un modo per descrivere sinteticamente questa formula, nota come formula dei cinque punti, è quello di usare la matrice 3×3 dei coefficienti di $u_{i+r,j+s}$ per $r,s=-1,0,1$ detta *stencil*

$$-\frac{1}{h^2}\begin{bmatrix}0&-1&0\\-1&4&-1\\0&-1&0\end{bmatrix}$$

Si osserva che la relazione che lega $\mathcal{L}_{m,n}$ con L_m e L_n è la seguente

$$\begin{aligned}\mathcal{L}_{m,n}(U^{(m,n)}) =& L_m \begin{bmatrix} u_{0,1} & \dots & u_{0,n} \\ u_{1,1} & \dots & u_{1,n} \\ \vdots & & \vdots \\ u_{m,1} & \dots & u_{m,n} \\ u_{m+1,1} & \dots & u_{m+1,n} \end{bmatrix} \\ &+ \begin{bmatrix} u_{1,0} & u_{1,1} & \dots & u_{1,n} & u_{1,n+1} \\ \vdots & \vdots & & \vdots & \vdots \\ u_{m,0} & u_{m,1} & \dots & u_{m,n} & u_{m,n+1} \end{bmatrix} L_n^T \end{aligned} \tag{35}$$

Infatti, applicare ad $u(x, y)$ la derivata seconda rispetto a x corrisponde nella versione discreta ad applicare l'operatore L_m a tutte le colonne di U escluse la prima e l'ultima che sono quelle di bordo; similmente applicare la derivata seconda rispetto a y corrisponde nel discreto ad applicare L_n a tutte le righe di U esclusa la prima e l'ultima che sono quelle di bordo.

L'equazione differenziale (31) con le condizioni al contorno di Dirichlet, ristretta ai valori $u_{i,j} = u(x_i, y_j)$, $i = 1, \dots, m$, $j = 1, \dots, n$ si può allora riscrivere come

$$\begin{aligned}-\frac{1}{h_x^2}\mathrm{trid}_m(-1,2,-1) &\begin{bmatrix} u_{1,1} & \dots & u_{1,n} \\ \vdots & & \vdots \\ u_{m,1} & \dots & u_{m,n} \end{bmatrix} \\ &- \frac{1}{h_y^2} \begin{bmatrix} u_{1,1} & \dots & u_{1,n} \\ \vdots & & \vdots \\ u_{m,1} & \dots & u_{m,n} \end{bmatrix} \mathrm{trid}_n(-1,2,-1) = B - h_x^2\tau^{(m,n)} - h_y^2\nu^{(m,n)}\end{aligned}$$

dove $B = F + \frac{1}{h_x^2}(\boldsymbol{g}_{:0}\boldsymbol{e}_1^{(n)T} + \boldsymbol{g}_{:n+1}\boldsymbol{e}_n^{(n)T}) + \frac{1}{h_y^2}(\boldsymbol{e}_1^{(n)}\boldsymbol{g}_{0:}^T + \boldsymbol{e}_m^{(n)}\boldsymbol{g}_{m+1:}^T)$, $F = (f_{i,j})_{i=1:m,j=1:n}$, $f_{i,j} = f(x_i, y_j)$ e abbiamo denotato $\boldsymbol{g}_{:0} = (g(x_i, 0))_{i=1:m}$, $\boldsymbol{g}_{:1} = (g(x_i, 1))_{i=1:m}$, $\boldsymbol{g}_{0:} = (g(0, y_i))_{i=1:n}$, $\boldsymbol{g}_{1:} = (g(1, y_i))_{i=1:n}$.

Rimuovendo la componente dell'errore locale si ottiene il sistema

$$-\frac{1}{h_x^2}H_m \begin{bmatrix} v_{1,1} & \dots & v_{1,n} \\ \vdots & & \vdots \\ v_{m,1} & \dots & v_{m,n} \end{bmatrix} - \frac{1}{h_y^2} \begin{bmatrix} v_{1,1} & \dots & v_{1,n} \\ \vdots & & \vdots \\ v_{m,1} & \dots & v_{m,n} \end{bmatrix} H_n = B \tag{36}$$

dove $H_m = \mathrm{trid}_m(-1, 2, -1)$, $H_n = \mathrm{trid}_n(-1, 2, -1)$.

È possibile scrivere tale sistema in forma standard $\mathcal{A}\boldsymbol{u} = \boldsymbol{b}$ se ordiniamo le componenti incognite $(u_{i,j})$ come un vettore $\boldsymbol{u}^{(m,n)}$ di mn componenti. Per questo utilizziamo l'operatore vec definito nel seguente modo

Definizione 19 *Data la matrice A di dimensione $m \times n$, il vettore $\boldsymbol{v} = \mathrm{vec}(A)$ è formato dagli elementi di A ordinati per colonne, cioè $v_{(j-1)m+i} = a_{i,j}$.*

Il vettore $\mathrm{vec}(A)$ è naturale vederlo come un vettore "a blocchi" dove il generico blocco in posizione j ha per elementi gli elementi della colonna j-esima

di A. Ad esempio, se

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$$

allora

$$\boldsymbol{a} = \begin{bmatrix} a_{11} \\ a_{21} \\ \hline a_{12} \\ a_{22} \\ \hline a_{13} \\ a_{23} \end{bmatrix}$$

Introduciamo il prodotto di Kronecker. Date la matrice A di dimensione $m \times n$ e la matrice B di dimensione $p \times q$, definiamo la matrice $A \otimes B$ come la matrice di dimensioni $mp \times nq$ che partizionata in blocchi $p \times q$ ha elementi $(a_{i,j}B)$. In particolare,

$$I_m \otimes B = \mathrm{diag}_m(B, B, \ldots, B)$$

mentre

$$A \otimes I_p = \begin{bmatrix} a_{11}I & \ldots & a_{1,n}I \\ \vdots & & \vdots \\ a_{m1}I & \ldots & a_{m,n}I \end{bmatrix}.$$

Formalmente se $H = A \otimes B$, allora $h_{(i-1)m+r,(j-1)n+s} = a_{i,j}b_{r,s}$, per $i = 1, \ldots, m$, $j = 1, \ldots, n$, $r = 1, \ldots, p$, $s = 1, \ldots, q$.

Si riportano le principali proprietà del prodotto di Kronecker la cui dimostrazione si lascia come esercizio:

$$\begin{aligned} \mathrm{vec}(AB) &= (I \otimes A)\mathrm{vec}(B) \\ \mathrm{vec}(AB) &= (B^T \otimes I)\mathrm{vec}(A) \\ (A \otimes C)(B \otimes D) &= (AB) \otimes (CD) \\ \det(A_m \otimes B_n) &= (\det A_m)^n (\det B_n)^m \\ (A_m \otimes B_n)^{-1} &= A_m^{-1} \otimes B_n^{-1} \\ (A_m \otimes B_n)^T &= A_m^T \otimes B_n^T \\ S_m^{-1} A_m S_m = \mathrm{diag}(\boldsymbol{a}),\ T_n^{-1} B_n T_n &= \mathrm{diag}(\boldsymbol{b}),\ \Rightarrow \\ (S_m \otimes T_n)^{-1}(A_m \otimes B_n)(S_m \otimes T_n) &= \mathrm{diag}(\boldsymbol{a}) \otimes \mathrm{diag}(\boldsymbol{b}) \end{aligned} \tag{37}$$

dove A, B, C, D sono matrici di dimensioni compatibili in modo che i prodotti AB e CD siano ben definiti, e dove A_m e B_n denotano matrici non singolari di dimensione rispettivamente $m \times m$ e $n \times n$.

Alla luce delle proprietà sopra riportate, l'equazione (36) si può scrivere nella forma

$$(I_n \otimes A_m + A_n \otimes I_m)\boldsymbol{v} = \boldsymbol{b} \tag{38}$$

dove $A_m = -\frac{1}{h_x^2}\mathrm{trid}_m(-1, 2, -1)$, $A_n = -\frac{1}{h_y^2}\mathrm{trid}_n(-1, 2, -1)$.

Denotiamo con $\mathcal{A}_{m,n} = I_n \otimes A_m + A_n \otimes I_m$. È facile vedere che la matrice $\mathcal{A}_{m,n}$ ha la seguente struttura

$$\mathcal{A}_{m,n} = -\frac{1}{h^2}\text{trid}_n(-I_m, H_m, -I_m), \quad H_m = \text{trid}_m(-1, 4, -1)$$

7.1 Analisi della stabilità e della convergenza

7.1.1 Analisi in norma 2

Utilizzando le proprietà del prodotto di Kronecker è abbastanza immediato dimostrare che la norma 2 della matrice $\mathcal{A}_{m,n}^{-1}$ è limitata superiormente da una costante. Per semplicità, ma comunque senza ledere la generalità, assumiamo $m = n$. Per le proprietà (37) gli autovalori della matrice $\mathcal{A}_{n,n} = I \otimes A_n + A_n \otimes I$ sono dati da $-\frac{1}{h^2}(2-2\cos\frac{\pi}{n+1}i+2-2\cos\frac{\pi}{n+1}j)$ e sono tutti negativi; il più piccolo autovalore in valore assoluto si ottiene con $i = j = 1$ e vale $\frac{1}{h^2}(4-4\cos\frac{\pi}{n+1})$. Poiché $\cos x = 1-x^2/2+O(x^4)$, si ha che il minimo dei valori assoluti degli autovalori di $\mathcal{A}_{n,n}$ è $\mu = 2\pi^2+O(h^2)$ per cui $\|\mathcal{A}_{n,n}^{-1}\|_2 = \rho(\mathcal{A}_{n,n}^{-1}) = 1/\mu = \frac{1}{2\pi^2}+O(h^2)$ che è limitato superiormente da una costante indipendente da n.

7.1.2 Analisi in norma infinito

In base ai commenti fatti nel paragrafo 4.4.3, per dimostrare la stabilità e la convergenza in norma infinito dello schema alle differenze per il problema di Poisson, basta far vedere che esiste una matrice $W = (w_{i,j})_{i=0:m+1,j=0:n+1} \geq 0$ tale che $\mathcal{L}_{m,n}(W)$ ha tutte componenti uguali a 1 e che vale il principio del massimo discreto per $\mathcal{L}_{m,n}$. Quest'ultima proprietà è verificata se $-\mathcal{A}_{m,n}^{-1} \geq 0$ e se $\mathcal{L}_{m,n}(E) = 0$, con E matrice di elementi uguali a 1.

Poiché $-\mathcal{A}_{m,n}$ è irriducibile e dominante diagonale con elementi diagonali positivi allora è una M-matrice non singolare (cf. l'osservazione 12) e quindi $-\mathcal{A}_{m,n}^{-1} \geq 0$. Inoltre, dalla relazione (35) applicata con $U = E$, segue che $\mathcal{L}_{m,n}(E) = 0$.

Per quanto riguarda l'esistenza di una matrice $W = (w_{i,j}) \geq 0$ tale che $\mathcal{L}_{m,n}(W)$ ha componenti unitarie, basta scegliere $W = \boldsymbol{w}^{(m)}\boldsymbol{e}^{(n+2)T}$, $\boldsymbol{w}^{(m)} = (\frac{1}{2}(h_m i)^2)_{i=0:m+1}$. Infatti, dalla (35) e dal fatto che $L_m(\boldsymbol{w}^{(m)}) = \boldsymbol{e}^{(m)}$ e $L_n(\boldsymbol{e}^{(n)}) = 0$ segue immediatamente la tesi.

Si osservi che la matrice W non è altro che la discretizzazione della funzione $w(x, y) = \frac{1}{2}x^2$ sul reticolo di punti (x_i, y_j).

7.2 Il caso di derivate miste

Si consideri l'operatore

$$a\frac{\partial^2 u}{\partial x^2} + 2b\frac{\partial^2 u}{\partial x \partial y} + c\frac{\partial^2 u}{\partial y^2}$$

dove $b^2 < ac$.

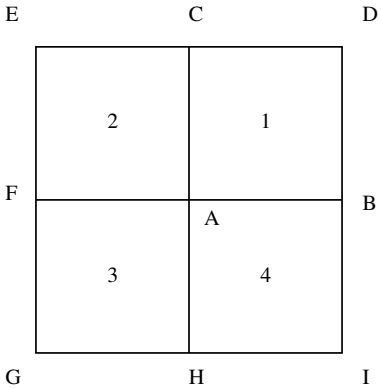
Sia $f(x)$ derivabile tre volte con continuità per cui

$$f'(x) = \frac{1}{h}(f(x+h) - f(x)) - \frac{h}{2}f''(x) - \frac{h^2}{6}f'''(\xi). \tag{39}$$

Si assuma per semplicità che $h_x = h_y =: h$. Dalla (39) applicata a $u(x,y)$ come funzione di x e come funzione di y con incremento h si ottiene

$$\begin{aligned}\frac{\partial^2 u}{\partial x \partial y} &= \frac{1}{h^2}\left(u(x+h,y+h) - u(x+h,y) - u(x,y+h) + u(x,y)\right) \\ &+ \frac{h}{2}\left(\frac{\partial^3 u}{\partial x \partial^2 y} + \frac{\partial^3 u}{\partial^2 x \partial y}\right) + O(h^2)\end{aligned} \tag{40}$$

Si consideri ora un generico punto del reticolo (x_i, y_j) e per semplicità si denoti con la lettera A. Si considerino i punti contigui che verranno denotati con le lettere dalla B alla I, e si numerino i quadrati che hanno un vertice in A da 1 a 4 come in figura:



Applicando la formula (40) nel punto A con incrementi $\pm h$ per la x e $\pm h$ per la y si ottengono le quattro diverse formule per approssimare la derivata mista $\frac{\partial^2 u}{\partial x \partial y}$

1. $\frac{1}{h^2}(u(\mathrm{D}) - u(\mathrm{C}) + u(\mathrm{A}) - u(\mathrm{B})) = \frac{\partial^2 u}{\partial x \partial y} + \frac{h}{2}(\frac{\partial^3 u}{\partial x \partial y^2} + \frac{\partial^3 u}{\partial x^2 \partial y}) + O(h^2)$
2. $\frac{1}{h^2}(u(\mathrm{C}) - u(\mathrm{E}) + u(\mathrm{F}) - u(\mathrm{A})) = \frac{\partial^2 u}{\partial x \partial y} + \frac{h}{2}(\frac{\partial^3 u}{\partial x \partial y^2} - \frac{\partial^3 u}{\partial x^2 \partial y}) + O(h^2)$
3. $\frac{1}{h^2}(u(\mathrm{A}) - u(\mathrm{F}) + u(\mathrm{G}) - u(\mathrm{H})) = \frac{\partial^2 u}{\partial x \partial y} + \frac{h}{2}(-\frac{\partial^3 u}{\partial x \partial y^2} - \frac{\partial^3 u}{\partial x^2 \partial y}) + O(h^2)$
4. $\frac{1}{h^2}(u(\mathrm{B}) - u(\mathrm{A}) + u(\mathrm{H}) - u(\mathrm{I})) = \frac{\partial^2 u}{\partial x \partial y} + \frac{h}{2}(-\frac{\partial^3 u}{\partial x \partial y^2} + \frac{\partial^3 u}{\partial x^2 \partial y}) + O(h^2)$

Una qualsiasi combinazione lineare delle quattro espressioni con coefficienti $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ tali che $\sum \alpha_i = 1$, fornisce una formula generale per approssimare

la derivata mista $\frac{\partial^2 u}{\partial x \partial y}$ con errore $O(h)$. Scegliendo però in modo opportuno i coefficienti è possibile annulare la componente $O(h)$ dell'errore e ottenere quindi un errore locale $O(h^2)$; questo si realizza scegliendo $\alpha_1 = \alpha_3$, $\alpha_2 = \alpha_4$. La formula generica con errore locale $O(h^2)$ può essere descritta con la matrice 3×3 dei coefficienti di $u_{i+r,j+s}$ per $r, s = -1, 0, 1$ (stencil)

$$\frac{a}{h^2}\begin{bmatrix} 0 & 0 & 0 \\ 1 & -2 & 1 \\ 0 & 0 & 0 \end{bmatrix} + 2\frac{b}{h^2}\begin{bmatrix} -\alpha_2 & \alpha_2 - \alpha_1 & \alpha_1 \\ \alpha_2 - \alpha_1 & 2(\alpha_2 - \alpha_1) & \alpha_2 - \alpha_1 \\ \alpha_1 & \alpha_2 - \alpha_1 & -\alpha_2 \end{bmatrix} + \frac{c}{h^2}\begin{bmatrix} 0 & 1 & 0 \\ 0 & -2 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

Se $b(x, y) > 0$ (il caso $b(x, y) < 0$ è trattato in modo analogo) scegliendo $\alpha_1 = \frac{1}{2}$, $\alpha_2 = 0$, si ottiene lo stencil

$$\frac{1}{h^2}\begin{bmatrix} 0 & c - b & b \\ a - b & -2(a + c - b) & a - b \\ b & c - b & 0 \end{bmatrix}$$

che genera una M-matrice se $0 < b < \min(a, c)$. Si osservi inoltre che la somma degli elementi dello stencil è nulla. Ciò permette di dimostrare la stabilità e la convergenza con ordine $O(h^2)$ di questo schema. Per questo è sufficiente determinare una matrice $W \geq 0$ tale che $\mathcal{L}_{m,n} \geq \gamma > 0$, dove γ è indipendente da m e da n. Per questo basta scegliere $w(x, y) = \frac{1}{2}x^2$, in modo che $\mathcal{L}(w) = a(x, y)$, $\gamma = \min a(x, y)$, e porre $W = (w_{i,j})$, $w_{i,j} = w(x_i, y_j)$.

Si osservi ancora che se $0 < b < \min(a, c)$ allora $b^2 < ac$ ma non viceversa.

7.3 Casi più generali

Si consideri il caso di

$$\mathcal{L}[u] = \frac{\partial}{\partial x}(a(x)\frac{\partial u}{\partial x}) + \frac{\partial}{\partial y}(b(y)\frac{\partial u}{\partial y}) \tag{41}$$

con $a(x) \geq \alpha > 0$, $b(y) \geq \beta > 0$.

Si verifichi che la matrice $\mathcal{A}_{m,n}$ ottenuta con condizioni al contorno di Dirichlet è

$$I_n \otimes \frac{1}{h^2}\text{trid}_m(a_{i-1}, -(a_{i-1} + a_i), a_i) + \frac{1}{h^2}\text{trid}_n(b_{i-1}, -(b_{i-1} + b_i), b_i) \otimes I_m$$

dove $a_i = a(x_i + h/2)$, $b_i = b(y_i + h/2)$.

L'analisi della stabilità in norma 2 si effettua valutando l'autovalore di minimo valore assoluto della matrice simmetrica $\mathcal{A}_{m,n}$. Usando le proprietà del prodotto di Kronecker, nell'ipotesi che $a(x, y) \geq \alpha > 0$ e $b(x, y) \geq \beta > 0$ si ottiene che l'autovalore di minimo valore assoluto di $\mathcal{A}_{m,n}$ è la somma degli autovalori di minimo valore assoluto di $\text{trid}_m(a_{i-1}, -(a_{i-1} + a_i), a_i)$ e di $\text{trid}_n(b_{i-1}, -(b_{i-1} + b_i), b_i)$ per cui la stabilità in norma 2 discende dall'analisi di stabilità in norma 2 nel caso monodimensionale.

Per l'analisi della stabilità in norma infinito si osserva che $-\mathcal{A}_{m,n}$ è una M-matrice e che $\mathcal{L}_{m,n}(E_{m,n}) = 0$ con $E_{m,n}$ matrice di elementi uguali a 1. Basta

quindi individuare una matrice $W_{m,n} \geq 0$ tale che $\mathcal{L}_{m,n}(W) \geq \gamma > 0$ con γ costante indipendente da m, n. Per questo basta scegliere $w(x,y) = \int_0^x \frac{t}{a(t)} dt$ e $W = \boldsymbol{w}^{(m)} \boldsymbol{e}^{(n)T}$ con $w_i^{(m)} = \sum_{j=0}^{i} \frac{(jh)^2}{a_{j-1}}$.

Si studi come può essere discretizzato l'operatore

$$\mathcal{L}[u] = \frac{\partial}{\partial x}(a(x,y)\frac{\partial u}{\partial x}) + \frac{\partial}{\partial y}(b(x,y)\frac{\partial u}{\partial y})$$

con $a(x,y) \geq \alpha > 0$, $b(x,y) \geq \beta > 0$.

Un altro caso interessante è il seguente

$$\mathcal{L}[u] = a(x,y)\frac{\partial^2 u}{\partial x^2} + b(x,y)\frac{\partial^2 u}{\partial y^2} \tag{42}$$

Si lascia come esercizio la sua discretizzazione e l'analisi di stabilità e convergenza.

7.4 Condizioni al contorno miste

Si consideri il problema di Poisson (31) con le seguenti condizioni al contorno miste

$$u(x,0) = g_1(x), \quad \frac{\partial u(x,1)}{\partial \vec{n}} = g_2(x)$$
$$u(0,y) = g_3(y), \quad \frac{\partial u(1,y)}{\partial \vec{n}} = g_4(y)$$

Si proceda come fatto nel caso monodimensionale e si dimostri che l'operatore discreto $\mathcal{A}_{m,n}$ ha la forma

$$\mathcal{A}_{m,n} = I_m \otimes A_n + A_m \otimes I_n$$

dove A_m e A_n sono le matrici che discretizzano $u''(x)$ con condizioni al contorno miste ottenute nel paragrafo 4.6.

7.5 Implementazione in Octave

Nel Listing 7 è riportata la *function* Octave che risolve il problema $\Delta u = f$ su un rettangolo con condizioni al contorno di Dirichlet. Il dominio è discretizzato con una griglia di $m \times n$ punti interni. La variabile `f` è una matrice di dimensione $(m+2) \times (n+2)$ che contiene sulla prima e ultima riga e sulla prima e ultima colonna i valori delle condizioni al contorno, mentre nella sua parte interna contiene i $m \times n$ valori di $f(x,y)$ nei punti della griglia. La variabile `v` è una matrice $(m+2) \times (n+2)$ che contiene la soluzione, inclusi anche i punti di bordo.

La figura 9 riporta il grafico della soluzione ottenuta scegliendo il termine noto $f = 0$ con valori al contorno dati da tre segmenti di retta. I valori di m e

Listing 7: Soluzione del problema di $\Delta u = f$ su un rettangolo mediante differenze finite.

```
function v = laplace(f)
% function v = laplace(f)
% risolve il problema di poisson sul rettangolo con termine noto f(x,y)
% f e' matrice (m+2)x(n+2), sulla prima e utima riga e colonna contiene
% le condizioni al bordo, nella parte rimanente contiene i valori di
% f(x,y) nei punti della griglia
% v e' una matrice (m+2)x(n+2) con i valori della soluzione, bordo
    incluso
[mm,nn]=size(f);
g=f(2:mm-1,2:nn-1);
m=mm-2; n=nn-2;
hm=1/(m+1);
hn=1/(n+1);
% completo il termine noto con le condizioni al contorno
g(1,:)=g(1,:)-f(1,2:nn-1)/hm^2;
g(m,:)=g(m,:)-f(mm,2:nn-1)/hm^2;
g(:,1)=g(:,1)-f(2:mm-1,1)/hn^2;
g(:,n)=g(:,n)-f(2:mm-1,nn)/hn^2;
% costruisco la matrice
dm=sparse(ones(m,1));
dn=sparse(ones(n,1));
dm1=sparse(ones(m-1,1));
dn1=sparse(ones(n-1,1));
Hn=(1/hn^2)*(-2*diag(dn)+diag(dn1,1)+diag(dn1,-1));
Hm=(1/hm^2)*(-2*diag(dm)+diag(dm1,1)+diag(dm1,-1));
A=kron(diag(dn),Hm)+kron(Hn,diag(dm));
% risolvo il sistema
v=A\vec(g);
v=reshape(v,m,n);
% completo alla dimensione originale
v=[f(1,:);f(2:m+1,1),v,f(2:m+1,n+2);f(m+2,:)];
```

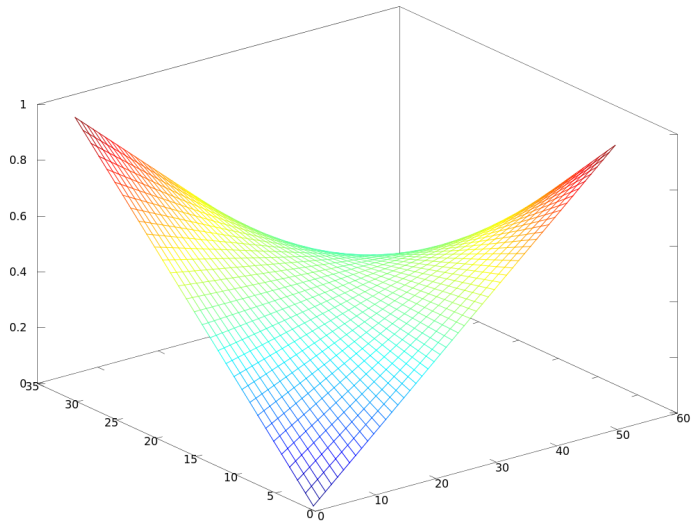


Figura 9: Configurazione di equilibrio ottenuta dall'equazione $\Delta u = 0$ con valori al contorno dati da tre segmenti di retta sghembi

n sono $m = 30$, $n = 50$. Il tempo di calcolo su un laptop con processore i3 è di 0.015 secondi.

Nella figura 10 si riporta il grafico della soluzione ottenuta con $f(x, y) = 1$ e condizioni omogenee al bordo. I valori scelti di m e n sono $m = n = 80$. Il tempo di calcolo è stato di 0.078 secondi.

8 Metodi variazionali (scritta da Federico Poloni)

8.1 Premessa

Sia B uno spazio di funzioni. Per evitare di generare confusione chiamando troppe cose con il nome "funzione", solitamente in analisi si preferisce chiamare *operatore* una mappa da B in sé (o comunque da uno spazio di funzioni a un altro spazio di funzioni), e *funzionale* una mappa da B a $\mathbb{R}$ (o $\mathbb{C}$). Notate anche che nella frase qui sopra, per evitare confusione, ho usato il termine *mappa*, che in fondo è un altro sinonimo di funzione.

Per evitare di avere troppe parentesi tonde, qui useremo le parentesi quadre per indicare l'applicazione di un operatore: per esempio, $L[f]$ indicherà l'applicazione dell'operatore L alla funzione f. La notazione $L[f](x)$, invece, indica che dobbiamo applicare l'operatore L alla funzione f, e valutare la funzione risultante nel punto x.

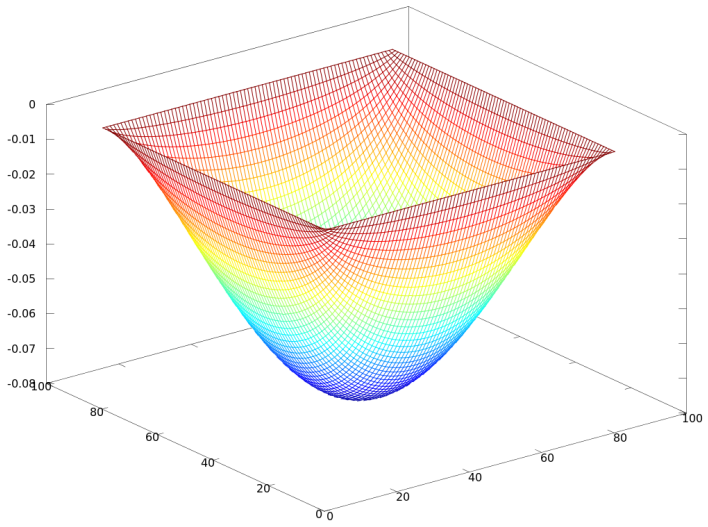


Figura 10: Configurazione di equilibrio ottenuta dall'equazione $\Delta u = f$ $f(x, y) = 1$ e con valori al contorno nulli.

8.2 Introduzione

I metodi alle differenze finite hanno due difetti:

1. Funzionano bene su griglie quadrate regolari, ma si adattano male a domini di forma più strana, per problemi in 2 o più dimensioni, che invece spesso capitano nelle applicazioni.

2. Richiedono un'alta regolarità della soluzione: per dimostrarne il funzionamento, abbiamo usato sviluppi di Taylor fino al IV ordine, richiedendo quindi che la funzione sia C^4 (e che $\left\|u^{(4)}\right\|_\infty$ non sia troppo grossa).

I metodi chiamati variazionali, o di Ritz-Galerkin, o degli elementi finiti, invece funzionano meglio sotto questi punti di vista.

8.3 Problema modello

Consideriamo il generico *problema di Sturm-Liouville*; cerchiamo $u : [a, b] \to \mathbb{R}$ di classe C^2 tale che

$$\begin{aligned} L[u] &= f; \\ u(a) &= \alpha \\ u(b) &= \beta, \end{aligned} \tag{43}$$

dove l'operatore L è definito come

$$L[v] := -\left[p(x)v'(x)\right]' + q(x)v(x)$$

per due funzioni $p(x) > 0, p \in C^1$, e $q(x) > 0$, $q \in C^0$ date. Nel seguito, quando possibile ometteremo i (x) dalle funzioni: per esempio l'equazione qui sopra sarà scritta in modo più compatto

$$L[v] = -[pv']' + qv.$$

Innanzitutto mi tolgo dai piedi le condizioni al contorno, cercando di rimpiazzarle con $u(a) = u(b) = 0$: per questo, scelgo una qualunque funzione $l(x) \in C^2$ che soddisfi $l(a) = \alpha$, $l(b) = \beta$ (per esempio c'è un polinomio di grado 1 che va bene), e noto che $\tilde{u} = u - l$ soddisfa

$$\begin{aligned} L[\tilde{u}] &= f - L[l] \\ \tilde{u}(a) &= 0 \\ \tilde{u}(b) &= 0 \end{aligned}$$

che è un problema dello stesso tipo ma con condizioni al bordo omogenee. Con qualche altro trucco (che non vediamo) è possibile trattare più o meno tutte le condizioni al contorno, incluse quelle di Neumann.

8.4 Forma debole

Consideriamo il prodotto scalare L^2 classico

$$\langle u, v \rangle = \int_a^b uv dx.$$

Nota che se u soddisfa (43), allora (integrando) soddisfa anche

$$\langle L[u], v \rangle = \langle f, v \rangle \quad \forall v \in C^1_{\text{a tratti}}, v(a) = v(b) = 0. \tag{44}$$

Nota che con quelle ipotesi su v possiamo scrivere

$$\langle L[u], v \rangle = -\int_a^b (pu')' v dx + \int_a^b quv dx = \int_a^b pu'v' dx + \int_a^b quv dx, \tag{45}$$

dove abbiamo integrato per parti il primo pezzo.

Con questa espressione per $\langle L[u], v \rangle$, nota che il problema (44) ha senso anche se u e v sono due funzioni soltanto $C^1_{\text{a tratti}}$ e nulle ai bordi. Anzi, possiamo chiedere solo che u' e v' esistano come derivate distribuzionali. Lo spazio delle funzioni

$$\{v : v \in L^2[a,b], v' \in L^2[a,b], v(a) = v(b) = 0\} =: H^1_0, \tag{46}$$

dove la derivata è fatta in senso distribuzionale, è detto uno *spazio di Sobolev*. Se non sapete cos'è una derivata distribuzionale, ignorate pure questa parte e prendete la (46) come definizione di H^1_0 solo con le funzioni C^1 a tratti, e tutto funzionerà lo stesso.

8.5 Proprietà di L

Su H_0^1, l'operatore L è simmetrico rispetto al prodotto scalare $\langle \cdot, \cdot \rangle$ (ovvio dalla (45)); possiamo anche provare che è definito positivo (in un certo senso).

Lemma 20 (versione 1D del lemma di Poincaré) *Esiste una costante C (che dipende solo dal dominio $[a, b]$) tale che per ogni $u \in H_0^1$ si ha*

$$\|u\|_N \leq C \|u'\|_{L^2}$$

per le due norme $N = \infty$ e $N = L^2$.

Dim. Usando Cauchy-Schwarz in modo furbo applicato col prodotto scalare ottenuto integrando da 0 a z, si ha $\langle u', 1 \rangle^2 \leq \langle 1, 1 \rangle \langle u', u' \rangle$, da cui $\forall z \in [a, b]$

$$u(z)^2 = (\int_a^z u' dx)^2 \leq (\int_a^z 1^2 dx)(\int_a^z u'^2 dx) \leq (z - a) \|u'\|_{L^2}^2 .$$

Visto che questa disuguaglianza vale per ogni $z \in [a, b]$, si ha

$$\|u\|_\infty^2 \leq (b - a) \|u'\|_{L^2}^2 .$$

Similmente, integrando entrambi i lati della disuguaglianza si ottiene

$$\|u\|_{L^2}^2 = \int_a^b u(z)^2 dz \leq \left(\int_a^b (z - a) dz \right) \|u'\|_{L^2}^2$$

□

Teorema 21 *Esistono costanti m, M che dipendono solo dal dominio tali che per ogni $u \in H_0^1$ si ha*

$$m \|u\|_N^2 \leq \langle L[u], u \rangle \leq M \|u'\|_N^2$$

Dim.

$$\begin{aligned} \langle L[u], u \rangle = \int_a^b p u'^2 dx + \int_a^b q u^2 dx &\geq (\min p) \int_a^b u'^2 dx \\ &\geq (\min p) \|u'\|_{L^2}^2 \geq (\min p) C^{-2} \|u\|_N^2 \end{aligned}$$

$$\begin{aligned} \langle L[u], u \rangle = \int_a^b p u'^2 dx + \int_a^b q u^2 dx &\leq (\max p) \|u'\|_{L^2}^2 dx + (\max q) \|u\|_{L^2}^2 \\ &\leq (\max p) \|u'\|_{L^2}^2 dx + (\max q) C^2 \|u'\|_{L^2}^2 \end{aligned}$$

□

8.6 Formulazione variazionale

Teorema 22 *Se u soddisfa (44), allora u è un punto di minimo (stretto) in H^1_0 del funzionale $F(v) := \langle L[v], v\rangle - 2\langle f, v\rangle$. In particolare, (44) ha al più una soluzione.*

Dim. Usando la simmetria di L,

$$F(v) + \langle L[u], u\rangle = \langle L[v-u], v-u\rangle \geq \|v-u\|_N \geq 0,$$

con l'ultimo $\geq$ sempre stretto a meno che $u = v$. □

8.7 Discretizzazione

Finora abbiamo dimostrato tante belle proprietà ma non abbiamo enunciato algoritmi per risolvere la (43). Lo facciamo ora: nella (44), rimpiazziamo H^1_0 con un suo sottospazio S con $\dim S = n < \infty$: cioè, cerchiamo $u_S \in S$ tale che

$$\langle L[u], v\rangle = \langle f, v\rangle \quad \forall v \in S. \tag{47}$$

Fissiamo una base di funzioni $(\phi_i(x))_{i=1}^n$ per S. Allora, u_S si scriverà in funzione della base come $u_S = \sum_{j=1}^n x_j\phi_j$. Scrivendo l'equazione (47) per $v = \phi_i$, $i = 1, 2, \ldots n$ (perché bastano?) otteniamo il sistema

$$\sum_{j=1}^{n} \langle \phi_i, L[\phi_j]\rangle\, x_j = \langle \phi_i, f\rangle, \quad i = 1, 2, \ldots, n. \tag{48}$$

È un sistema con matrice A_S simmetrica e positiva definita (perché $\langle f, L[f]\rangle \geq \|f\|^2_{L^2} > 0$ per ogni $0 \neq f \in S$). Quindi possiamo risolverlo e ricavare gli x_i, e quindi u_S.

Nota che il teorema 22 funziona (con la stessa dimostrazione) anche se rimpiazziamo H^1_0 con S, per cui u_S è il punto di minimo (stretto) del funzionale F su S.

8.8 Errore di discretizzazione

Teorema 23 (Lemma di Céa) *Esiste una costante (che dipende solo dal dominio) D per cui per ogni $v \in S$ si ha*

$$\|u - u_S\|_N \leq D\,\|u' - v'\|_N$$

per le norme $N = 2$, $N = \infty$.

Dim.

$$\begin{aligned} m\,\|u-u_S\|_N &\leq \langle L[u-u_S], u-u_S\rangle = \langle u, L[u]\rangle + F(u_s) \\ &\leq \langle u, L[u]\rangle + F(v) = \langle L[u-v], u-v\rangle \leq M\,\|u'-v'\|_N \end{aligned}$$

dove nel $\leq$ a cavallo dell'andata a capo abbiamo usato l'osservazione appena fatta che u_S è il minimo di F su S, e quindi $F(u_S) \leq F(v)$. □

Quindi l'errore globale sulla soluzione dipende solo dall'*errore di approssimazione* della derivata prima v nello spazio S: cioè, se scelgo un S dove la derivata prima di u può venire approssimata "bene" (cioè c'è una $v \in S$ tale che $v' - u'$ è piccolo), allora ottengo una buona stima della soluzione.

8.9 Esempio: Spline cubiche

Prendiamo come S lo spazio delle spline cubiche sui punti equispaziati $x_0, x_1, \ldots, x_N$. Ha dimensione $N + 1$ (infatti basta fissare i valori della u su $x_0, x_1, \ldots, x_N$ e possiamo costruire la sua spline). È possibile (non lo vediamo qui) costruire una sua base tale che per ogni i la funzione di base ϕ_i è diversa da zero solo nell'intervallo $[x_{i-2}, x_{i+2}]$. Ciò è bello perché implica che gli elementi della matrice del sistema $A_S = \langle \phi_i, L[\phi_j] \rangle$ sono nulli tutte le volte che $|j - i| > 2$ (è evidente dalla (45)). Quindi la matrice del sistema è pentadiagonale, e possiamo applicarci velocemente la maggior parte degli algoritmi di algebra lineare (eliminazione di Gauss in $O(n^2)$, un passo dei principali metodi iterativi in $O(n)$).

Per quanto riguarda l'errore globale, si può dimostrare (noi lo omettiamo) il seguente risultato:

Lemma 24 *Sia $u \in C^4$. La spline cubica che interpola u nei punti $x_0, x_1, \ldots, x_N$ soddisfa*

$$\|v' - u'\|_\infty \leq C \left\| u^{(4)} \right\|_\infty h^3$$

per una costante moderata C.

Questo + il lemma di Céa ci permettono di dire che l'errrore globale per questo metodo agli elementi finiti è al più $C \left\| u^{(4)} \right\|_\infty h^3$. Notare che per le differenze finite c'era un'espressione simile ma h^2 al posto di h^3, quindi gli elementi finiti si avvicinano meglio alla soluzione.

8.10 Esempio: funzioni lineari a tratti

Prendiamo come S lo spazio delle funzioni di approssimazione lineare nei punti $x_0, x_1, \ldots, x_N$ (cioè le funzioni lineari a tratti con i punti di non derivabilità nelle x_i). Una sua base è data dalle funzioni ϕ_i tali che $\phi_i(x_j) = \delta_{ij}$ ("hat functions"); con questa base la A_S risulta tridiagonale.

Notare che hanno la regolarità minima che serve per far funzionare gli elementi finiti (C^1 a tratti). Si può dimostrare (non difficile, ma lo omettiamo) che per la funzione v di approssimazione lineare per una funzione $u \in C^2$ soddisfa

$$\|v' - u'\|_\infty \leq C \left\| u^{(2)} \right\|_\infty h.$$

(notare che C^2 è la minima regolarità di u per cui ha senso porsi il problema (43)). Questo + lemma di Céa ci dicono che la soluzione converge con errore h.

8.11 Curiosità: cosa succede in più dimensioni

Tutti i risultati enunciati finora non richiedono particolari proprietà di $\mathbb{R}$, e infatti funzionano anche quando le funzioni sono definite su un aperto Ω di $\mathbb{R}^2$, per ogni operatore L per cui si riesce a dimostrare il teorema 21. Non serve che L sia simmetrico, ma se non lo è le dimostrazioni si complicano un po'. Nel caso simmetrico, quasi tutte le dimostrazioni che abbiamo visto funzionano pari pari (eccezione: lemma di Poincaré).

Sono possibili diverse scelte dello spazio S, a seconda del problema; la più comune è prendere una *triangolazione* del dominio, cioè una suddivisione di Ω in tanti piccoli triangoli, e considerare le funzioni che sono continue globalmente e lineari su ognuno dei triangoli. Con la base opportuna (le funzioni che valgono 1 in un vertice della triangolazione e 0 altrove), la matrice del sistema non ha più una struttura particolare come nel caso 1D, ma è comunque molto sparsa.

Le operazioni da fare nella pratica sono:

- Costruire la triangolazione del dominio. Non è difficile, ci sono algoritmi dedicati, ma talvolta non serve neppure applicarli perché molte applicazioni partono da programmi di progettazione al computer in cui il risultato è già naturalmente suddiviso in triangoli (o tetraedri nel caso 3D)—se qualcuno si interessa di computer-grafica 3D o di CAD ha già capito di cosa parliamo.

- Costruire la matrice e il termine noto del sistema (48). Qui ci sono molti integrali da fare. Nella pratica, questi integrali vengono calcolati utilizzando su ogni triangolo (o su ogni segmento $[x_i, x_{i+1}]$ di suddivisione dell'intervallo $[a, b]$ nel caso 1D) una formula di quadratura. Solitamente questa è la parte che porta via più tempo.

- Risolvere il sistema. Visto che di solito la matrice è molto sparsa, si usa un metodo iterativo (Jacobi, gradiente coniugato, GMRES), spesso accoppiato con precondizionatori sofisticati. Di solito si arresta il metodo iterativo con un residuo abbastanza alto: $\epsilon = 10^{-3}$, per esempio. Difatti, spesso, a causa degli errori già commessi "a monte" (nei dati sperimentali, nel modello, nella quadratura, errore globale nella soluzione dell'equazione differenziale), la precisione del risultato è già limitata a poche cifre significative.

9 Equazioni paraboliche

Si consideri l'equazione del calore

$$\gamma \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0, \quad \gamma > 0. \tag{49}$$

con condizioni

$$\begin{aligned} u(x,0) &= f(x), \quad 0 \le x \le \ell, \\ u(0,t) &= a(t), \quad u(\ell,t) = b(t), \quad t \ge 0. \end{aligned} \tag{50}$$

Il dominio $[0,\ell]\times[0,t_{\max}]$, si discretizza con $u_{i,j}=u(x_i,t_j)$, $x_i=i\Delta_x$, $\Delta_x=\ell/(n+1)$, $t_j=j\Delta_t$, $\Delta_t=t_{\max}/(m+1)$, $i=0,1,\ldots,n+1$, $j=0,1,2,\ldots,m+1$.

9.1 Una prima discretizzazione

Assumiamo che i dati $a(t)$, $b(t)$, $f(x)$ siano sufficientemente regolari in modo che la soluzione $u(x,t)$ con le sue derivate rispetto alla x e rispetto alla t fino al quart'ordine siano continue. Si approssima la derivata prima rispetto a t con la formula (6) e la derivata seconda rispetto a x mediante la formula (4). Si ottiene allora

$$\begin{aligned}\frac{\partial u}{\partial t} &= \frac{1}{\Delta_t}(u_{i,j+1}-u_{i,j})+\sigma_{i,j}\Delta_t\\ \frac{\partial^2 u}{\partial x^2} &= \frac{1}{\Delta_x^2}(u_{i-1,j}-2u_{i,j}+u_{i+1,j})+\tau_{i,j}\Delta_x^2\end{aligned}$$

dove $|\sigma_{i,j}|$ e $|\tau_{i,j}|$ sono limitati superiormente.

Sostituendo nella (49) e ponendo per semplicità $r=\frac{\Delta_t}{\gamma\Delta_x^2}$, si ottiene

$$u_{i,j+1}=r(u_{i+1,j}+u_{i-1,j})+(1-2r)u_{i,j}+\Delta_x^2\Delta_t\frac{\tau_{i,j}}{\gamma}-\sigma_{i,j}\Delta_t^2,\quad i=1,\ldots,n,\ \ j=0,1,\ldots,$$

dove $u_{0,j}=a(t_j)$, $u_{n+1,j}=b(t_j)$ e $u_{i,0}=f(x_i)$.

Si considera allora il sistema ottenuto rimuovendo l'errore locale di discretizzazione

$$\begin{aligned}v_{i,j+1} &= r(v_{i+1,j}+v_{i-1,j})+(1-2r)v_{i,j}\quad i=1,\ldots,n,\quad j=0,1,\ldots,m+1\\ v_{0,j} &= a(t_j),\ v_{n+1,j}=b(t_j),\ v_{i,0}=f(x_i).\end{aligned}$$

Ordinando le incognite "per colonne" cioè come

$$(v_{1,1},v_{2,1},\ldots,v_{n,1},\quad v_{1,2},v_{2,2},\ldots,v_{n,2},\quad\ldots\ ,\quad v_{1,m+1},v_{2,m+1},\ldots,v_{n,m+1})$$

si ottiene un sistema lineare la cui matrice è $\Delta_t\mathcal{A}_{m,n}$ dove

$$\mathcal{A}_{m,n}=\frac{1}{\Delta_t}\begin{bmatrix} I & & & \\ -T & I & & \\ & \ddots & \ddots & \\ & & -T & I\end{bmatrix},\quad T=\mathrm{trid}_n(r,-2r+1,r)$$

Per l'errore globale $\epsilon_{i,j}=u_{i,j}-v_{i,j}$ vale

$$\mathcal{A}_{m,n}\mathrm{vec}(\epsilon_{i,j})=\frac{\Delta_x^2}{\gamma}\mathrm{vec}(\tau_{i,j})-\Delta_t\mathrm{vec}(\sigma_{i,j})$$

Per dimostrare la stabilità e la convergenza del metodo occorre studiare la norma infinito della matrice $\mathcal{A}_{m,n}^{-1}$. Vale

$$\mathcal{A}_{m,n}^{-1}=\Delta_t\begin{bmatrix} I & & & \\ T & I & & \\ \vdots & \ddots & I & \\ T^m & \ldots & T & I\end{bmatrix}$$

La presenza di T^m nel blocco in basso a sinistra ci dice che condizione necessaria di stabilità è che $\rho(T) \leq 1$ altrimenti T^m divergerebbe esponenzialmente. Quindi condizione necessaria di stabilità è che gli autovalori di T siano in valore assoluto minori o uguali a 1. Cioè $|1-2r+2r\cos i\pi/(n+1)| \leq 1$. Ciò è verificato per ogni n se e solo se $r \leq 1/2$, cioè $\Delta_t/\Delta_x^2 \leq \gamma/2$. La stabilità e la convergenza sono subordinate ad una condizione sul rapporto Δ_t/Δ_x^2. Si parla per questo di stabilità condizionata.

Si osservi che se $r \leq 1/2$ la matrice T è non negativa e $\mathcal{A}_{m,n}$ è una M-matrice per cui $\mathcal{A}_{m,n}^{-1} \geq 0$. Vale allora $\|\mathcal{A}_{m,n}^{-1}\|_\infty = \|\mathcal{A}_{m,n}^{-1}\boldsymbol{e}^{(n(m+1))}\|_\infty$ dove $\boldsymbol{e}^{(n(m+1))}$ è il vettore con $n(m+1)$ componenti uguali a 1. Inoltre vale $T\boldsymbol{e}^{(n)} \leq \boldsymbol{e}^{(n)}$ per cui $T^j\boldsymbol{e}^{(n)} \leq \boldsymbol{e}^{(n)}$ quindi

$$\|\mathcal{A}_{m,n}^{-1}\|_\infty = \|\mathcal{A}_{m,n}^{-1}\boldsymbol{e}^{(n(m+1))}\|_\infty \leq \Delta_t \sum_{j=0}^{m} \|T^j\boldsymbol{e}^{(n)}\|_\infty \leq \Delta_t(m+1) = 1$$

e quindi per l'errore globale risulta

$$|\epsilon_{i,j}| \leq (\Delta_x^2 \max|\tau_{p,q}|/\gamma + \Delta_t \max|\sigma_{p,q}|)$$

9.2 Una discretizzazione più efficiente: il metodo di Crank-Nicolson

Si può ottenere un metodo dotato di convergenza incondizionata, cioè comunque Δ_t e Δ_x convergano a zero, con una piccola modifica del metodo mostrato nel paragrafo precedente.

Si osserva che per una funzione sufficientemente regolare $f(x)$ il quoziente $(f(x+h)-f(x))/h$ approssima la derivata prima di $f(x)$ in x con un errore $O(h)$ (confronta con (6)), però la stessa formula può essere vista come una applicazione della (5) con passo $h/2$ e quindi come una approssimazione di $f'(x+h/2)$ con errore $O(h^2)$. Ciò suggerisce di discretizzare la (49) uguagliando $\gamma(u_{i,j+1}-u_{i,j})/\Delta_t$ ad una approssimazione alle differenze finite della derivata seconda di $u(x,t)$ rispetto ad x però al tempo $t_j+\Delta_t/2$ anziché al tempo t_j come si era fatto nel paragrafo precedente. Per questo si usa come approssimazione della derivata seconda di $u(x,t)$ rispetto ad x al tempo $t_j+\Delta_t/2$ la media aritmetica delle due approssimazioni della derivata seconda ai tempi t_j e t_{j+1}, cioè $(u_{i-1,j}-2u_{i,j}+u_{i+1,j})/\Delta_x^2$ e $(u_{i-1,j+1}-2u_{i,j+1}+u_{i+1,j+1})/\Delta_x^2$. Infatti, in generale, per una funzione sufficientemente regolare $f(x)$ facendo uno sviluppo in serie di $f(x)$ con incremento h e $-h$ e prendendo la media aritmetica di entrambi i membri si ha che $(f(x-h)+f(x+h))/2 = f(x)+h^2(f''(\xi)+f''(\eta))/4$, dove $\xi \in (x+h)$, $\eta \in (x-h)$.

Si arriva quindi alla equazione alle differenze

$$\begin{aligned}\frac{1}{\Delta_t}(u_{i,j+1}-u_{i,j}) = \frac{1}{2\gamma\Delta_x^2}(&u_{i-1,j}-2u_{i,j}+u_{i+1,j}+u_{i-1,j+1}-2u_{i,j+1}+u_{i+1,j+1})\\ &+O(\Delta_x^2)+O(\Delta_t^2)\end{aligned}$$

che, ponendo $r = \Delta_t/(2\gamma\Delta_x^2)$, può essere riscritta come

$$\frac{1}{\Delta_t}\left[(1+2r)u_{i,j+1} - ru_{i-1,j+1} - ru_{i+1,j+1}\right] = \frac{1}{\Delta_t}\left[(1-2r)u_{i,j} + ru_{i-1,j} + ru_{i+1,j}\right] + O(\Delta_x^2) + O(\Delta_t^2)$$

Il sistema lineare ottenuto rimuovendo i termini con gli errori locali è dunque

$$\frac{1}{\Delta_t}\left((1+2r)v_{i,j+1} - rv_{i-1,j+1} - rv_{i+1,j+1}\right) = \frac{1}{\Delta_t}\left((1-2r)v_{i,j} + rv_{i-1,j} + rv_{i+1,j}\right) + O(\Delta_x^2) + O(\Delta_t^2)$$
$$v_{0,j} = a(t_j),\ v_{n+1,j} = b(t_j),\ v_{i,0} = f(x_i).$$

L'espressione per l'errore globale è quindi

$$\mathcal{A}_{m,n}\boldsymbol{\epsilon}^{(n(m+1))} = O(\Delta_x^2 + \Delta_t^2) \tag{51}$$

dove

$$\mathcal{A}_{m,n} = \frac{1}{\Delta_t}\begin{bmatrix} A & & & \\ -B & A & & \\ & \ddots & \ddots & \\ & & -B & A \end{bmatrix}$$

in cui $A = \mathrm{trid}_m(-r, 1+2r, -r)$, $B = \mathrm{trid}_m(r, 1-2r, r)$.

La convergenza del metodo in qualche norma si studia dando maggiorazioni alla norma di $\mathcal{A}_{m,n}^{-1}$. Si osserva che, posto $V = A^{-1}B$, risulta

$$\mathcal{A}_{m,n}^{-1} = \Delta_t \begin{bmatrix} I & & & \\ V & I & & \\ \vdots & \ddots & \ddots & \\ V^m & \dots & V & I \end{bmatrix} (I \otimes A^{-1}).$$

La presenza di del blocco V^m in basso a sinistra mostra che condizione necessaria di stabilità è $\rho(V) \leq 1$ altrimenti la matrice V^m divergerebbe esponenzialmente. Si verifica facilmente che $V = (I+rH)^{-1}(I-rH)$, dove $H = \mathrm{trid}_m(-1, 2, -1)$. Per cui gli autovalori di V sono

$$\frac{1-2r(1-c_i)}{1+2r(1-c_i)}, \quad c_i = \cos(\frac{\pi i}{n+1})$$

Il valore assoluto della espressione precedente è sempre minore di 1 qualunque sia il valore di r. Il massimo si ottiene per $c_i = \cos(\pi/(n+1))$ e vale $\rho(V) \doteq 1 - 2r(\pi/(n+1))^2$.

Questa proprietà può essere usata per dimostrare la convergenza incondizionata in norma 2. Infatti, partizionando $\boldsymbol{\epsilon}^{(n(m+1))}$ in blocchi $\boldsymbol{\epsilon}_i$ di lunghezza n, e denotando con $\boldsymbol{c} = (\boldsymbol{c}_i)$ il vettore a destra nel sistema (51), risulta

$$\boldsymbol{\epsilon}_k = \Delta_t \sum_{i=0}^{k-1} V^i A^{-1} \boldsymbol{c}_{k-i}.$$

Introducendo la norma $\|\boldsymbol{v}\| = \frac{1}{\sqrt{n}}\|\boldsymbol{v}\|_2$ per vettori di n componenti, segue che

$$\|\boldsymbol{\epsilon}_k\| \leq \Delta_t \sum_{i=0}^{k-1} \|V\|_2^i \|A^{-1}\|_2 \|\boldsymbol{c}_{k-i}\|$$

da cui

$$\|\boldsymbol{\epsilon}_k\| \leq \frac{\Delta_t}{1 - \|V\|_2} \|A^{-1}\|_2 \max_i \|\boldsymbol{c}_i\|.$$

Vale inoltre $\|A^{-1}\|_2 = 1/(1 + 2r(1 - \cos(\pi/(n+1)))) < 1$, e

$$\Delta_t/(1 - \|V\|_2) \doteq \Delta_t/(1 - (1 - 2r(\pi/(n+1))^2)) = \Delta_t/(2r\pi^2\Delta_x^2).$$

Poiché $r = \Delta_t/(\gamma\Delta_x^2)$, ne segue che $\Delta_t/(1 - \|V\|_2)$ è limitato superiormente da una costante indipendente da m e da n, da cui la convergenza del metodo di Crank-Nicolson in norma 2. Più precisamente esiste una costante $\theta > 0$ tale che $\|\boldsymbol{\epsilon}_k\| \leq \theta(\Delta_t^2 + \Delta_x^2)$.

Una dimostrazione elementare della stabilità in norma infinito si può dare sotto l'ipotesi $r \leq 1/2$. Infatti vale il seguente risultato.

Teorema 25 *Se $r \leq 1/2$ per le matrici $A = \mathrm{trid}_m(-r, 1+2r, -r)$ e $B = \mathrm{trid}_m(r, 1-2r, r)$, vale $A^{-1} \geq 0$, $B \geq 0$ inoltre*

$$\begin{aligned}
&A^{-1}\boldsymbol{e} \leq \boldsymbol{e} \\
&B\boldsymbol{e} \leq \boldsymbol{e} \\
&A^{-1}B\boldsymbol{e} \leq \boldsymbol{e} \\
&\sum_{j=0}^{m} (A^{-1}B)^j A^{-1}\boldsymbol{e} \leq (m+1)\boldsymbol{e}
\end{aligned}$$

da cui $\|\mathcal{A}_{m,n}^{-1}\|_\infty \leq 1$.

Dim. Posto $H = \mathrm{trid}(-1, 2, -1)$ vale $A = I + 2rH$, e $B = I - 2rH$. Per cui, se $r \leq 1/2$ è $B \geq 0$. Inoltre, essendo A una M-matrice dominante diagonale è $A^{-1} \geq 0$. Per quanto riguarda le rimanenti diseguaglianze vale $A\boldsymbol{e} = \boldsymbol{e} + r(\boldsymbol{e}_1 + \boldsymbol{e}_n) \geq \boldsymbol{e}$ da cui, poiché $A^{-1} \geq 0$ segue $A^{-1}\boldsymbol{e} \leq \boldsymbol{e}$. Inoltre $B\boldsymbol{e} = \boldsymbol{e} - r\boldsymbol{e}_1 - r\boldsymbol{e}_n \leq \boldsymbol{e}$. La terza diseguaglianza si ottiene componendo le due precedenti. La quarta diseguaglianza segue dalla terza. □

9.3 Il metodo θ

Si chiama metodo θ il metodo alle differenze finite che si ottiene combinando linearmente le approssimazioni della derivata seconda di u rispetto ad x al tempo t_j e t_{j+1} con peso θ e $1 - \theta$. Col parametro $\theta = 1/2$ si ottiene il metodo di Crank-Nicolson, per cui il metodo θ ne costituisce una generalizzazione.

9.4 Altre discretizzazioni

Si consideri lo schema alle differenze ottenuto mediante le formule

$$\frac{\partial^2 u(x,t)}{\partial x^2} = \frac{1}{\Delta_x^2}(u(x-\Delta_x,t) - 2u(x,t) + u(x+\Delta_x,t)) + \Delta_x^2\tau(\xi,t)$$
$$\frac{\partial u(x,t)}{\partial t} = \frac{1}{\Delta_t}(u(x,t) - u(x,t-\Delta_t)) + \Delta_t\sigma(x,\eta)$$

Si dimostri che lo schema è incondizionatamente stabile. Inoltre, per l'operatore alle differenze finite $L_\Delta(u_{i,j})$ ottenuto in questo modo vale il principio del massimo discreto, cioè se $L_\Delta(u_{i,j}) \leq 0$ nei punti della discretizzazione interni al dominio allora $u_{i,j}$ prende il massimo sul bordo.

Si verifichi che se la discretizzazione della derivata temporale è fatta mediante la formula

$$\frac{\partial u(x,t)}{\partial t} = \frac{1}{\Delta_t}(u(x,t+\Delta_t) - u(x,t)) + \Delta_t\tilde{\sigma}(x,\tilde{\eta})$$

allora per l'operatore alle differenze finite ottenuto in questo modo vale il principio del massimo discreto se $r = \Delta_t/(\gamma\Delta_x^2)$ è minore di $1/2$.

Si studi il metodo che si ottiene approssimando la derivata temporale con la formula (più precisa)

$$\frac{\partial u(x,t)}{\partial t} = \frac{1}{2\Delta_t}(u(x,t+\Delta_x) - u(x,t-\Delta_t)) + \Delta_t^2\hat{\sigma}(x,\hat{\eta}).$$

Si dimostri che la norma infinito dell'inversa della matrice che discretizza l'operatore diverge per $\Delta_t \to 0$ indipendentemente dal valore di $r = \Delta_t/(\gamma\Delta_x^2)$.

9.5 Caso tridimensionale

Nel modello in cui si studia la propagazione della temeperatura $u(x,y,t)$ dei punti (x,y) di una piastra Ω di cui si conosce ad ogni istante la temperatura g sul bordo $\partial\Omega$ e i valori iniziali della temperatura della piastra, l'equazione del calore prende la forma

$$\gamma\frac{\partial u(x,y,t)}{\partial t} - (\frac{\partial^2 u(x,y,t)}{\partial x^2} + \frac{\partial^2 u(x,y,t)}{\partial y^2}) = 0$$

con le condizioni al contorno

$$u(x,y,t) = g(x,y,t), \quad (x,y) \in \partial\Omega,$$
$$u(x,y,0) = f(x,y), \quad (x,y) \in \Omega.$$

Anche in questo caso si possono applicare gli analoghi dei metodi alle differenze finite esaminati nel paragrafo precedente.

10 Equazioni iperboliche

Consideriamo ora l'equazione delle onde

$$\frac{\partial^2 u}{\partial t^2} = \gamma \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1, \quad 0 < t < t_{\max}, \tag{52}$$

con le condizioni

$$\begin{aligned} &u(x,0) = f(x), \quad 0 < x < 1 \\ &\frac{\partial u(x,t)}{\partial t}|_{t=0} = g(x), \quad 0 < x < 1 \\ &u(0,t) = 0, \quad u(1,t) = 0. \end{aligned} \tag{53}$$

10.1 La soluzione di D'Alembert

L'equazione delle onde nella forma (52) ammette una forma esplicita della soluzione. Infatti, se $F(x)$ e $G(x)$ sono due funzioni derivabili due volte con continuità, allora $u(x,t) = F(x+ct) + G(x-ct)$ soddisfa l'equazione (52) con $c = \sqrt{\gamma}$. Basta quindi costruire F e G in modo che $u(x,t)$ soddisfi le condizioni (53). Si verifica che l'imposizione delle (53) porta a

$$u(x,t) = \frac{1}{2}[f(x+ct) + f(x-ct)] + \frac{1}{2c}\int_{x-ct}^{x+ct} g(\nu)d\nu$$

cioè, $F(x) = \frac{1}{2}(f(x) + \hat{g}(x))$, $G(x) = \frac{1}{2}(f(x) - \hat{g}(x))$ con $\hat{g}(x)' = \frac{1}{c}g(x)$.

Si osserva che il valore $u(x^*,t^*)$ della soluzione (x,t) nel punto (x^*,t^*) dipende unicamente dai valori $f(x^*+ct^*)$, $f(x^*-ct^*)$ e dai valori della funzione g nell'intervallo $[x^*-ct^*, x^*+ct^*]$. In altri termini, il valore di $u(x^*,t^*)$ non è influenzato dai valori delle condizioni iniziali al di fuori di $[x^*-ct^*, x^*+ct^*]$. Analogamente, i valori di $u(x^*,t^*)$ dipendono dai valori della "storia passata" di $u(x,t)$ racchiusa dalle due rette passanti per (x^*,t^*) e rispettivamente per i punti $(x^*-ct^*,0)$, $(x^*+ct^*,0)$ dette rette caratteristiche. Le rette hanno equazione $x-ct = x^*-ct^*$ e $x+ct = x^*+ct^*$. La figura 11 riporta tali rette e il dominio da esse racchiuso detto dominio di dipendenza.

10.2 Discretizzazione

Adottiamo l'equazione (52) come esempio per mostrare la risoluzione alle differenze finite di un problema iperbolico.

Discretizzando il dominio con i punti (x_i,t_j), $x_i = i\Delta_x$, $t_j = j\Delta_t$, $i = 0,\ldots,n+1$, $j = 0,1,\ldots,m+1$, $\delta_x = 1/(n+1)$, $\Delta_t = t_{\max}/(m+1)$ e approssimando le due derivate seconde con la formula (4) si ottiene

$$\frac{1}{\Delta_t^2}(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}) = \frac{1}{\Delta_t^2} r(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) + O(\Delta_t^2) + O(\Delta_x^2).$$

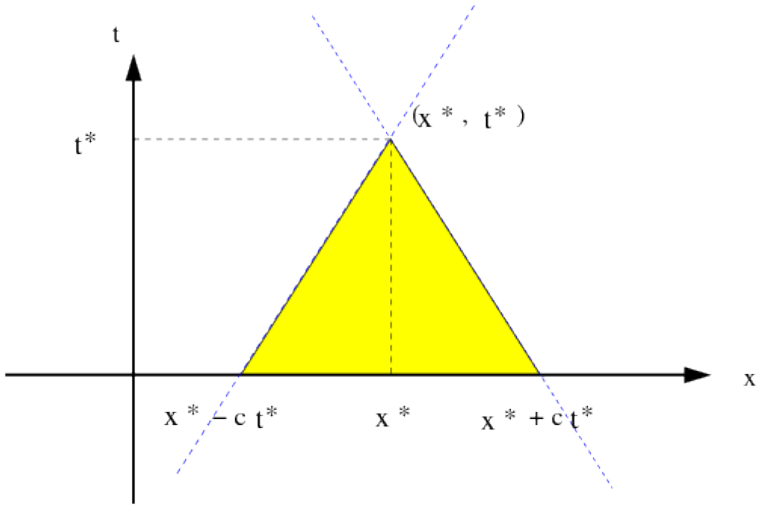


Figura 11: Dominio di dipendenza della soluzione dell'equazione delle onde

con $r = \gamma \left(\frac{\Delta_t}{\Delta_x}\right)^2$ e con le condizioni al bordo

$$\begin{array}{ll} u_{0,j} = u_{n+1,0} = 0 & j = 0, 1, \ldots, m+1 \\ u_{i,0} = f(x_i), & i = 1, \ldots, n \\ u_{i,1} = f(x_i) + \Delta_t g(x_i) + \frac{\Delta_t^2}{2}\gamma f''(x_i) + O(\Delta_t^3) & i = 1, \ldots, n \end{array} \tag{54}$$

La terza condizione si ottiene nel seguente modo mediante uno sviluppo in serie essendo $\partial u(x,0)/\partial x = g(x)$:

$$\begin{aligned} u_{i,1} = u(x_i, \Delta_t) &= f(x_i) + \Delta_t \frac{\partial u(x_i,0)}{\partial t} + \frac{\Delta_t^2}{2}\frac{\partial^2 u(x_i,0)}{\partial t^2} + O(\Delta_t^3) \\ &= f(x_i) + \Delta_t g(x_i) + \frac{\Delta_t^2}{2}\gamma \frac{\partial^2 u(x_i,0)}{\partial x^2} + O(\Delta_t^3) \\ &= f(x_i) + \Delta_t g(x_i) + \frac{\Delta_t^2}{2}\gamma f''(x_i) + O(\Delta_t^3) \end{aligned}$$

Se la derivata seconda di $f(x)$ non fosse disponibile basta approssimarla con una formula alle differenze con errore $O(\Delta_x^2)$.

Denotando con $\boldsymbol{u}^{(j)}$ il vettore di componenti $u_{i,j}$, per $i = 1, \ldots, n$, e con $\boldsymbol{f}$, $\boldsymbol{g}$, $\boldsymbol{f''}$ i vettori le cui componenti sono i valori delle corrispondenti funzioni $f(x), g(x), f''(x)$ in x_i per $i = 1, \ldots, n$, si ha

$$\begin{aligned} \frac{1}{\Delta_t^2}\boldsymbol{u}^{(1)} &= \frac{1}{\Delta_t^2}\boldsymbol{f} + \frac{1}{\Delta_t}\boldsymbol{g} + \frac{1}{2}\gamma \boldsymbol{f''} + O(\Delta_t^2) \\ \frac{1}{\Delta_t^2}\boldsymbol{u}^{(2)} &= \frac{1}{\Delta_t^2} V_n \frac{1}{\Delta_t^2}\boldsymbol{u}^{(1)} + \frac{1}{\Delta_t^2}\boldsymbol{f} + O(\Delta_t^2) + O(\Delta_x^2) \\ \frac{1}{\Delta_t^2}\boldsymbol{u}^{(j+1)} &= \frac{1}{\Delta_t^2} V_n \boldsymbol{u}^{(j)} - \frac{1}{\Delta_t^2}\boldsymbol{u}^{(j-1)} + + O(\Delta_t^2) + O(\Delta_x^2) \end{aligned}$$

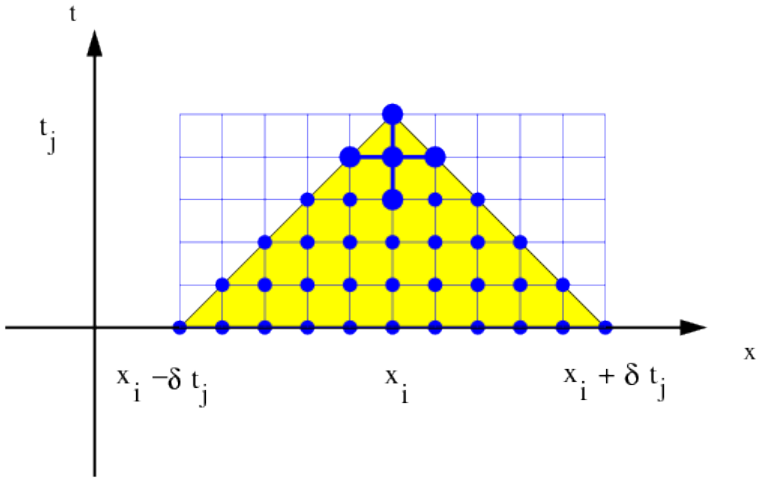


Figura 12: Dominio di dipendenza dell'operatore discreto per l'equazione delle onde, $\delta = \Delta_x/\Delta_t = \sqrt{\gamma/r}$

dove $V_n = \mathrm{trid}_n(r, 2-2r, r)$.

In forma matriciale si ottiene quindi il sistema

$$\frac{1}{\Delta_t^2}\begin{bmatrix} I_n & & & & \\ -V_n & I_n & & & \\ I_n & -V_n & I_n & & \\ & \ddots & \ddots & \ddots & \\ & & I_n & -V_n & I_n \end{bmatrix}\begin{bmatrix} \boldsymbol{u}^{(1)} \\ \vdots \\ \boldsymbol{u}^{(n)} \end{bmatrix} = \boldsymbol{b} + O(\Delta_t^2 + \Delta_x^2) \qquad (55)$$

dove il vettore $\boldsymbol{b}$ dipende dalle condizioni (54).

Lo schema alle differenze che si ottiene è uno schema esplicito nel senso che la soluzione discreta al tempo j si può esprimere in funzione delle soluzioni al tempo $j-1$ e $j-2$. Lo schema è rappresentato dallo stencil

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

e la soluzione nel punto (x_i, t_j) dipende dai valori delle soluzioni nei punti (x_{i-1}, t_{j-1}), (x_i, t_{j-1}), (x_{i+1}, t_{j-1}) e (x_i, t_{j-2}). In figura 12 viene riportata in forma grafica questa dipendenza assieme al dominio di dipendenza per l'operatore discreto dove si è posto $\delta = \Delta_x/\Delta_t = \sqrt{\gamma/r}$.

Si osservi che la soluzione discreta dipende dai valori delle funzioni $f(x)$ e $g(x)$, che danno le condizioni iniziali, nell'intervallo $[x_i - \delta t_j, x_i + \delta t_j]$, mentre la soluzione del problema continuo in (x_i, t_j) dipende dai valori nell'intervallo $[x_i - ct_j, x_i + ct_j]$. Se $\delta < c$ il secondo intervallo include strettamente il primo per cui esistono valori delle condizioni iniziali che alterano la soluzione della equazione

differenziale ma non alterano la soluzione dell'equazione alle differenze. Questo fatto ci fa capire che se $r > 1$ non ci può essere convergenza. Infatti dimostriamo ora che questa condizione è necessaria per la stabilità.

La stabilità e la convergenza dello schema dipendono dalla limitatezza della norma dell'inversa della matrice $\mathcal{A}_{m,n}$ del sistema (55). Supponiamo per semplicità m pari e si partizioni $\mathcal{A}_{m,n}$ in blocchi di dimensione $2n \times 2n$ in modo che risulti bidiagonale a blocchi. Vale allora

$$\mathcal{A}_n = \frac{1}{\Delta_t^2}\left(I_{m/2} \otimes (\begin{bmatrix} I_n & 0 \\ -V_n & I_n \end{bmatrix}\right) \begin{bmatrix} I_{2n} & & & \\ -H_{2n} & I_{2n} & & \\ & \ddots & \ddots & \\ & & -H_{2n} & I_{2n} \end{bmatrix}$$

dove

$$H_{2n} = \begin{bmatrix} I & 0 \\ -V_n & I \end{bmatrix}^{-1} \begin{bmatrix} I & -V_n \\ 0 & I \end{bmatrix} = \begin{bmatrix} I & -V_n \\ V_n & I - V_n^2 \end{bmatrix}$$

Poiché nell'inversa di $\mathcal{A}_{m,n}$ compare la matrice $H_{2n}^{n/2-1}$, una condizione necessaria di stabilità è $\rho(H_{2n}) \leq 1$. Si ha che $\rho(H_{2n}) \leq 1$ se e solo se gli autovalori λ di V_n sono tali che $-2 \leq \lambda \leq 2$. Usando il teorema 5 si deduce che tale condizione equivale a $r(1 - \cos(\pi i/(n+1))) \leq 2$, per $i = 1, \ldots, n$ e per ogni n, cioè $r \leq 1$.

Si può verificare che questa condizione di stabilità è anche sufficiente.

11 Note computazionali

I sistemi lineari ottenuti discretizzando equazioni differenziali alle derivate parziali col metodo delle differenze finite sono generalmente a banda (a blocchi) e sparsi. Nel caso in cui tali sistemi non siano triangolari (a blocchi) per cui un metodo di sostituzione può risolvere in modo efficiente il sistema, è preferibile usare un metodo iterativo. Nel caso delle equazioni di tipo ellittico in cui la matrice del sistema è una M-matrice simmetrica, e quindi definita positiva, il metodo iterativo più indicato è il metodo del gradiente coniugato (precondizionato). Tale metodo, così come molti altri, richiede ad ogni passo come operazione più costosa il calcolo del prodotto matrice vettore. Ad esempio, nel caso della matrice $\mathcal{A}_{m,n}$ che discretizza il laplaciano cambiato di segno su un rettangolo, il prodotto $\boldsymbol{y} = h^2\mathcal{A}_{m,n}\boldsymbol{x}$ si realizza con le seguenti semplici istruzioni nella sintassi tipo Octave.

```
for i=1:m
  for j=1:n
    y(i+1,j+1)=4*x(i+1,j+1)-x(i,j+1)-x(i+2,j+1)-x(i+1,j)-x(i+1,j+2);
  end
end
```

Si osservi che nel doppio ciclo `for` abbiamo dovuto aumentare di 1 i valori degli indici delle variabili `x` e `y`. Infatti, Octave richiede indici *positivi*. In

altri termini il valore di $u_{i,j}$ è memorizzato nella variabile $x(i+1, j+1)$ per $i = 0, \ldots, m+1$, $j = 0, \ldots, n+1$.

Si osservi ancora che nel doppio cilco `for` intervengono punti che stanno sul bordo del dominio. Questi possono essere rimossi assegnando il valore zero alle componenti di bordo di `x`. Nel caso in cui il metodo iterativo richieda il calcolo del residuo $\mathcal{A}_{m,n}x - b$ è conveniente assegnare ai valori di bordo della variabile `x` i valori al contorno del problema. In questo modo il calcolo svolto nel doppio cilco `for` non corrisponde alla sola moltiplicazione di $h\mathcal{A}_{m,n}$ per il vettore incognito ma include anche la parte del termine noto che contiene le condizioni al bordo.

Poiché Octave è un linguaggio interpretato, l'esecuzione di due cicli for annidati può richiedere un tempo di esecuzione elevato se i valori di m e n sono "moderatamente grandi". Un modo per ovviare a questo inconveniente è scrivere il doppio ciclo in forma "vettoriale" nel modo seguente:

```
y(2:m+1,2:n+1) = 4*x(2:m+1,2:n+1) - x(1:m,2:n+1) - x(3:m+2,2:n+1)
               - x(2:m+1,1:n) - x(2:m+1,3:n+2);
```

Nel caso in cui il dominio Ω non coincide col rettangolo ma è un suo sottoinsieme proprio, ad esempio un dominio di forma ad L o un quadrato con un foro al suo interno di forma quadrata, è possibile modificare leggermente il programma per il calcolo del prodotto matrice-vettore. Conviene introdurre una nuova variabile `dominio` tale che `dominio(i+1,j+1)` vale 1 se il punto (x_i, y_j) sta nel dominio e vale 0 altrimenti. Il doppio ciclo for si trasforma in modo semplice in

```
for i=1:m
 for j=1:n
  if dominio(i+1,j+1)
   y(i+1,j+1)=4*x(i+1,j+1)-x(i,j+1)-x(i+2,j+1)-x(i+1,j)-x(i+1,j+2);
  end
 end
end
```

Anche in questo caso è possibile dare una versione in forma "vettorizzata" del doppio ciclo for procedendo come nel caso del rettangolo, facendo seguire il calcolo dall'operazione `y=dominio .* y;` che serve a riportare a zero i valori di y al di fuori del dominio di interesse. Si ricorda che l'operatore `.*` esegue il prodotto elemento a elemento delle due matrici.

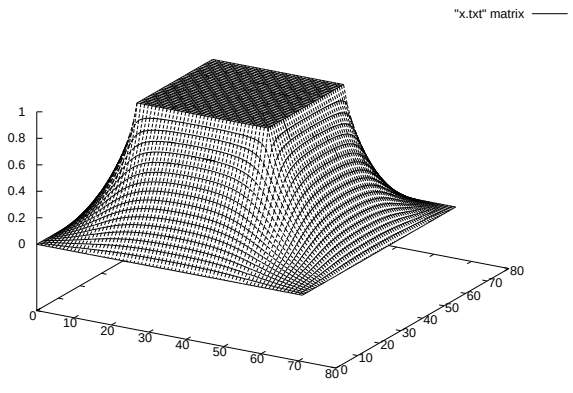
Un metodo iterativo per risolvere il sistema $\mathcal{A}x = b$, che si basa sul calcolo del residuo, è il metodo di Richardson definito da

$$x^{(k+1)} = x^{(k)} - \alpha(\mathcal{A}x^{(k)} - b)$$

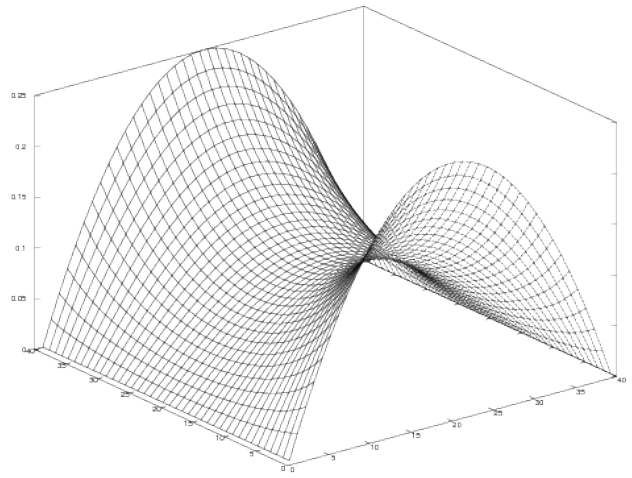
a partire da un vettore iniziale $x^{(0)}$, generalmente $x^{(0)} = 0$. La convergenza del metodo si ha se il raggio spettrale della matrice di iterazione $I - \alpha\mathcal{A}$ è minore di 1. Se la matrice $\mathcal{A}$ è definita positiva e i suoi autovalori sono compresi in

$(0, \lambda_{max})$, allora la scelta $\alpha \leq 1/\lambda_{max}$ garantisce la convergenza. Non avendo informazioni sullo spettro di $\mathcal{A}$, si può sempre scegliere $\alpha = 1/\|\mathcal{A}\|$ per una qualsiasi norma matriciale indotta $\|\cdot\|$.

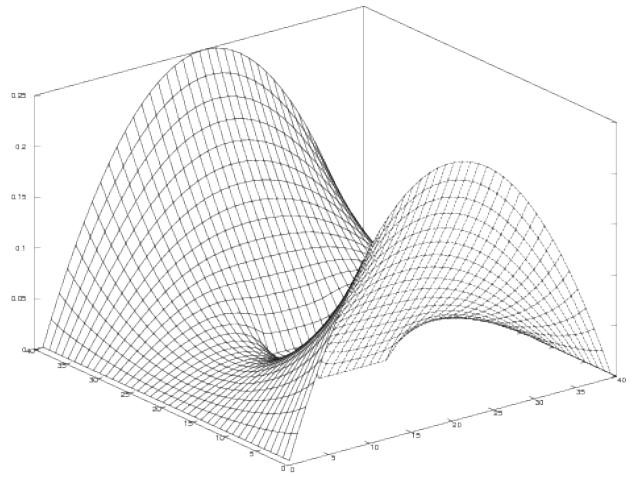
La figura che segue mostra la configurazione di una bolla di sapone calcolata col metodo del gradiente coniugato applicato al sistema lineare ottenuto discretizzando l'equazione di Laplace sul dominio $[0,3] \times [0,3] \backslash [1,2] \times [1,2]$ con le condizioni $u(x,y) = 0$ sul bordo del quadrato maggiore e $u(x,y) = 1$ sul bordo del quadrato minore.



La figura successiva mostra una analoga configurazione in cui il dominio è il quadrato $[0,1] \times [0,1]$ e la funzione è nulla su due lati opposti del quadrato mentre vale $x(1-x)$ sugli altri due lati opposti.



La figura che segue differisce dalla precedente poichè al dominio è stato tolto un quadrato centrale al bordo del quale la funzione prende il valore nullo.



Per una panoramica dei metodi iterativi e per una descrizione algoritmica che permette facili implementazioni, si suggerisce il documento [9].

Si riportano le *function* nella sintassi di Octave che svolgono il calcolo della configurazione di equilibrio risolvendo l'equazione di Laplace $\Delta u = 0$. Si osservi che, poiché l'equazione è omogenea, il parametro h non ha un ruolo effettivo nel sistema di equazioni. Infatti esso non compare nell'iterazione.

La *function* 9 calcola la posizione di equlibrio nel caso di un dominio arbitario contenuto in un dominio rettangolare. Il dominio in cui viene trattato il

Listing 8: Soluzione del problema di $\Delta u = f$ su un rettangolo mediante differenze finite.

```
function u = membrana(v)
  % u = membrana(v) risolve l'equazione di Laplace sul rettangolo
  % [0 (m+1)h,0,(n+1)h], h=1/(min(m,n)+1)
  % cioe' trova la posizione di equilibrio di una membrana elastica col
  % bordo vincolato a stare su una curva chiusa nello spazio a proiezione
  % rettangolare
  % INPUT: matrice V mxn con prima e ultima riga e colonna contenenti le
  % quote dei punti sul bordo della membrana. I valori dei restanti
  % elementi sono i valori della approssimazione iniziale da cui parte il
  % metodo iterativo di risoluzione del sistema lineare che discretizza
  % l'equazione di Laplace.
  % OUTPUT: la matrice mxn U che contiene i valori della funzione
  % soluzione che da' la configurazione di equilibrio

  % L'equazione di Laplace, Laplaciano(u)=0 viene discretizzata col
  % metodo delle differenze finite che produce il sistema lineare Au=b.
  % Il metodo usato per risolvere il sistema lineare e' l'iterazione di
  % Richardson: u_{k+1}=u_k-alfa*(Au_k-b) dove alfa=1/||A||
  % Il residuo Au-b viene calcolato in "modo vettoriale"
  % Il termine noto appare implicitamente attraverso le condizioni al
  % contorno date dalle righe e colonne estreme della matrice di input v

maxit=2000; epsi=1.e-6;
m=size(v)(1); n=size(v)(2);
n=n-2; m=m-2;
u=v; r=u;
for it=1:maxit
  r(2:m+1,2:n+1)=4*u(2:m+1,2:n+1)-u(1:m,2:n+1)-u(3:m+2,2:n+1);
  r(2:m+1,2:n+1)= r(2:m+1,2:n+1)-u(2:m+1,1:n)-u(2:m+1,3:n+2);
  err=max(max(abs(r(2:m+1,2:n+1))));
  disp([it err]);
  u(2:m+1,2:n+1)=u(2:m+1,2:n+1)-r(2:m+1,2:n+1)/8;
  if err<epsi
    break
  end
end
mesh(u); % disegna la posizione di equilibrio della membrana
print "bolla.jpg" -djpg % salva l'immagine in un file nel formato jpg
```

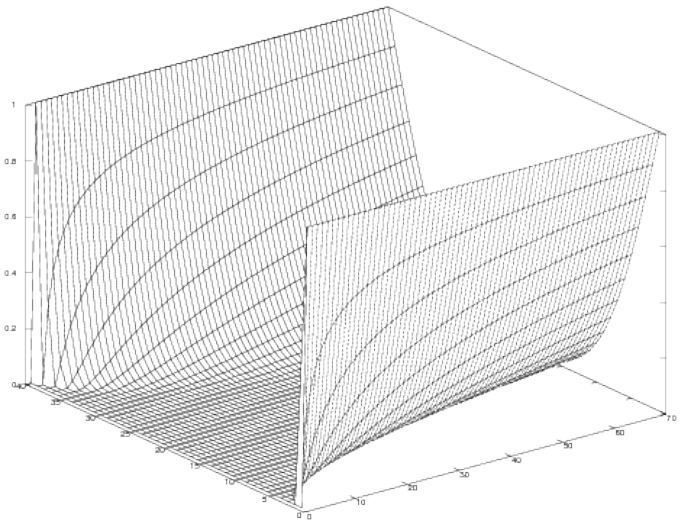


Figura 13: Soluzione dell'equazione del calore: caso stabile

problema è individuato dalla matrice `dominio` i cui elemnti valgono 1 nei punti interni al dominio, valgono 0 altrove. L'iterazione di Richardson si ripete in modo analogo al caso del dominio rettangolare con la differenza che l'iterazione non viene aggiornata nei punti fuori dal dominio e che i punti esterni al dominio non contribuiscono al prodotto matrice vettore se non quelli di bordo. Ciò si realizza moltipicando

La *function* riportata nel Listing 10 calcola la posizione di equilibrio di una membrana elastica vincolata a un bordo chiuso nello spazio che viene sollecitata mediante una forza ortogonale al dominio. In questo caso il parametro h interviene nel calcolo dei valori della forza per creare il termine noto del sistema lineare.

La *function* riportata nel listing 11 risolve l'equazione del calore col metodo del paragrafo 9.1 che è stabile se $r < 1/2$.

Calcolando la soluzione con $T = 10$, $\gamma = 800$, $u(0,t) = u(1,t) = 1$, $u(x,0) = 0$, e scegliendo $n = 38$, $m = 70$ si ha stabilità essendo $r = 0.27554$ e si ottiene la figura 13. Scegliendo invece $m = 35$ si ha $r = 0.55919$ per cui il metodo non è stabile. La soluzione calcolata in questo caso è mostrata in figura 14.

Riferimenti bibliografici

[1] D. Bini, M. Capovani, O. Menchi, *Metodi Numerici per l'Algebra Lineare*. Zanichelli, Bologna, 1987.

Listing 9: Soluzione del problema di $\Delta u = f$ su un rettangolo mediante differenze finite.

```
function u = membrana1(v, dominio)
  % u = membrana1(v,dominio) risolve l'equazione di Laplace su un dominio
  % contenuto in un rettangolo
  % INPUT: matrice V mxn con elemnti di bordo contenenti le quote dei
  % valori di bordo della membrana
  % matrice DOMINIO: vale 1 nei punti interni al dominio e vale 0
  % nei punti esterni o di bordo.
  % OUTPUT: la matrice mxn U che contiene i valori della funzione
  % soluzione che da' la configurazione di equilibrio
  % Il residuo Au-b viene calcolato in "modo vettoriale" anche sui punti
  % esterni al dominio e viene "ripulito" moltiplicandolo elemento a
  % elemento con la matrice DOMINIO
  % Il termine noto appare implicitamente attraverso le condizioni al
  % contorno date dalle righe e colonne estreme della matrice v

maxit=2000; epsi=1.e-6;
m=size(v)(1); n=size(v)(2);
n=n-2; m=m-2;
u=v;
r=u;
for it=1:maxit
  r(2:m+1,2:n+1)=4*u(2:m+1,2:n+1)-u(1:m,2:n+1)-u(3:m+2,2:n+1);
  r(2:m+1,2:n+1)= r(2:m+1,2:n+1)-u(2:m+1,1:n)-u(2:m+1,3:n+2);
  r=r.*dominio;
  err=max(max(abs(r(2:m+1,2:n+1))));
  disp([it err]);
  u(2:m+1,2:n+1)=u(2:m+1,2:n+1)-r(2:m+1,2:n+1)/8;
  if err<epsi
    break
  end
end
mesh(u); % disegna la posizione di equilibrio della membrana
print "bolla.jpg" -djpg % salva l'immagine in un file nel formato jpg
```

Listing 10: Soluzione del problema di $\Delta u = f$ su un rettangolo mediante differenze finite.

```
function u = membrana2(v, dominio,forza)
  % u = membrana2(v, dominio, forza) risolve l'equazione di Poisson su un
  % dominio contenuto nel rettangolo [0 (m+1)h,0,(n+1)h],
  % h=1/(min(m,n)+1) cioe' trova la posizione di equilibrio di una
  % membrana elastica col bordo vincolato a stare su curve nello spazio
  % definite sul bordo del dominio e sottoposta ad una forza ortogonale
  % al dominio
  % INPUT: matrice V mxn con prima e ultima riga e colonna contenenti le
  % quote dei punti sul bordo della membrana. I valori dei restanti
  % elementi sono i valori della approssimazione iniziale da cui parte
  % il metodo iterativo di risoluzione del sistema lineare che
  % discretizza l'equazione di Laplace.
  % matrice DOMINIO che vale 1 se nei punti interni al dominio e vale 0
  % nei punti esterni o di bordo.
  % matrice FORZA che in posizione (i,j) ha il valore della forza che
  % viene applicata alla membrana nel punto (i,j) diretta lungo l'asse z
  % OUTPUT: la matrice mxn U che contiene i valori della configurazione
  % di equilibrio
  % L'equazione di Poisson, Laplaciano(u)=forza, viene discretizzata col
  % metodo delle differenze finite che produce il sistema lineare Au=b.
  % Il metodo usato per risolvere il sistema lineare e' l'iterazione di
  % Richardson: u_{k+1}=u_k-alfa*(Au_k-b) dove alfa=1/||A||
  % Il residuo Au-b viene calcolato in "modo vettoriale" anche sui punti
  % esterni al dominio e viene "ripulito" moltiplicandolo elemento a
  % elemento con la matrice DOMINIO

maxit=2000; epsi=1.e-6;
m=size(v)(1); n=size(v)(2);
n=n-2; m=m-2; h=1/(min(m,n)+1);
u=v; r=u;
for it=1:maxit
  r(2:m+1,2:n+1)=4*u(2:m+1,2:n+1)-u(1:m,2:n+1)-u(3:m+2,2:n+1);
  r(2:m+1,2:n+1)= r(2:m+1,2:n+1)-u(2:m+1,1:n)-u(2:m+1,3:n+2);
  r=r + h^2*forza;
  r=r.*dominio;
  err=max(max(abs(r(2:m+1,2:n+1))));
  disp([it err]);
  u(2:m+1,2:n+1)=u(2:m+1,2:n+1)-r(2:m+1,2:n+1)/8;
  if err<epsi
    break
  end
end
mesh(u); % disegna la posizione di equilibrio della membrana
print "bolla.jpg" -djpg % salva l'immagine in un file nel formato jpg
```

Listing 11: Soluzione del problema di $\Delta u = f$ su un rettangolo mediante differenze finite.

```
function u = calore(gamma,T,a,b,v)
% FUNCTION u=calore(gamma,T,a,b,v) risolve l'equazione del calore
% gamma u_t -u_xx =0, u(0,t)=a(t), u(1,t)=b(t), u(x,0)=v
% per 0<=t<=T con il metodo delle differenze finite
% La derivata prima e' approssimata con una differenza in avanti
% il metodo e' stabile solo se r=Dt/(gamma*Dx^2)<1/2
% INPUT: gamma costante positiva
% T valore massimo del tempo
% a vettore di m componenti con i valori della soluzione nell'estremo
% sinistro
% b vettore di m componenti con i valori della soluzione nell'estremo
% destro
% v vettore di n componenti con i valori iniziali della temperatura della
     sbarretta
m=length(a); Dt=T/(m-1);
n=length(v)-2; Dx=1/(n+1);
u=zeros(n+2,m);
u(1,:)=a; u(n+2,:)=b; u(:,1)=v;
r=Dt/(gamma*Dx^2);
if r>0.5
  disp("attenzione il metodo non e' stabile")
  r
end
for j=1:m-1
  for i=2:n+1
    u(i,j+1)=r*(u(i+1,j)+u(i-1,j))+(1-2*r)*u(i,j);
  end
end
```

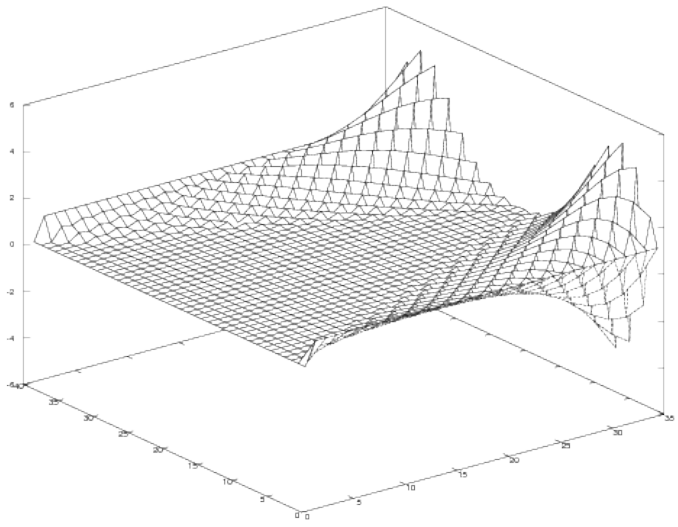


Figura 14: Soluzione dell'equazione del calore: caso instabile

[2] Garrett Birkhoff and Robert E. Lynch, *Numerical Solution of Elliptic Problems*. SIAM, Philadelphia 1984.

[3] J.F. Botha and G.F Pinder, *Fundamental Concepts in the Numerical Solution of Differential Equations*, Jhon Wiley & Sons, 1983.

[4] Eugene Isaacson and Herbert Bishop Keller, *Analysis of Numerical Methods*. Jhon Wiley & Sons, Inc., New York, 1966.

[5] A. R. Mitchell and D.F. Griffiths, *The Finite Difference Method in Partial Differential Equations*. Jhon Wiley & Sons, New York 1980.

[6] Alfio Quarteroni and Alberto Valli, *Numerical Approximation of Partial Differential Equations*. Springer Series in Computational Mathematics, Berlin 1997.

[7] Josef Stoer, Roland Bulirsch, *Introduzione all'Analisi Numerica (2)*. Zanichelli, Bologna 1975.

[8] Gerald Wheatley, Applied Numerical Analysis, Pearson, Addison Wesley, New York 2004.

[9] Richard Barret et Al., Templates For the Solution of Linear Systems: Building Blocks For Iterative Methods. `http://www.netlib.org/linalg/html_templates/Templates.html`