DIFFERENZE FINITE

Il metodo delle differenze finite consiste nell'approssimare il valore della derivata di una funzione u in un punto $\widetilde{x}$ (per il quale sarebbe necessario conoscere tutti i valori della funzione (quindi infiniti) in un intorno di $\widetilde{x}$), con un'espressione che ne tenga in conto solo un numero finito (spesso molto piccolo). Si passa cioè dall'operazione di limite a quella di rapporto incrementale. Ciò permette, ad esempio, di trasformare un'equazione alle derivate parziale in un problema algebrico. In particolare se il problema di partenza è lineare, si ottiene un sistema lineare del tipo $A\boldsymbol{x} = \boldsymbol{b}$, con A matrice sparsa, la cui sparsità dipende dal numero di valori usati nell'approssimazione delle derivate.

1 Derivazione delle formule

1.1 Sviluppo in serie di Taylor

Il più classico approccio per determinare le approssimazioni alle differenze finite ed analizzarne l'errore, consiste nello sviluppare la funzione $u(x)$ in serie di Taylor in un intorno del punto $\widetilde{x}$, di troncare opporunamente tale sviluppo ed eventualmente di combinare linearmente varie espressioni relative a diversi sviluppi ottenuti considerando diversi intorni. Vediamo nel seguito alcune applicazioni di tale approccio.

Dato $h > 0$, si considerino i due seguenti sviluppi:

$$u(\widetilde{x}+h) = u(\widetilde{x}) + hu'(\widetilde{x}) + \frac{1}{2}h^2u''(\widetilde{x}) + \frac{1}{6}h^3u'''(\widetilde{x}) + O(h^4), \tag{1}$$

$$u(\widetilde{x}-h) = u(\widetilde{x}) - hu'(\widetilde{x}) + \frac{1}{2}h^2u''(\widetilde{x}) - \frac{1}{6}h^3u'''(\widetilde{x}) + O(h^4). \tag{2}$$

Si ricorda che una quantità è un $O(h^p)$ se va a zero quando $h \to 0$ rapidamente almeno tanto quanto ci va h^p. Dalle due precedenti espressioni si ricavano

facilmente le due seguenti approssimazioni del primo ordine per la derivata prima:

$$D_+u(\widetilde{x}) = \frac{u(\widetilde{x}+h) - u(\widetilde{x})}{h} = u'(\widetilde{x}) + O(h) \tag{3}$$

e

$$D_-u(\widetilde{x}) = \frac{u(\widetilde{x}) - u(\widetilde{x}-h)}{h} = u'(\widetilde{x}) + O(h). \tag{4}$$

Infatti dalla (1) dividendo per h si ottiene $(u(\widetilde{x}+h) - u(\widetilde{x}))/h = u'(\widetilde{x}) + 1/2hu''(\widetilde{x}) + 1/6h^2u'''(\widetilde{x}) + O(h^3)$. Essendo $\widetilde{x}$ fissato, i valori $u''(\widetilde{x})$ e $u'''(\widetilde{x})$ sono dei numeri costanti indipendenti da h; si ottiene quindi la (3). Analogo ragionamento vale per determinare la (4). Similmente, sottraendo la (2) alla (1) e dividendo per $2h$, si ottiene

$$D_cu(\widetilde{x}) = \frac{u(\widetilde{x}+h) - u(\widetilde{x}-h)}{2h} = u'(\widetilde{x}) + O(h^2), \tag{5}$$

cioè un'approssimazione della derivata prima del secondo ordine.

Per ottenere delle espressioni approssimanti le derivate di ordine superiore, si possono combinare gli operatori differenziali discreti relativi agli ordini inferiori. Ad esempio, per ottenere un'approssimazione della derivata seconda, si può scrivere:

$$D^2(\widetilde{x}) = D_+(D_-u(\widetilde{x})) = \frac{1}{h}[D_-u(\widetilde{x}+h) - D_-u(\widetilde{x})] = \tag{6}$$

$$= \frac{1}{h}[(\frac{u(\widetilde{x}+h) - u(\widetilde{x})}{h}) - (\frac{u(\widetilde{x}) - u(\widetilde{x}-h)}{h})] = \frac{1}{h^2}[u(\widetilde{x}+h) - 2u(\widetilde{x}) + u(\widetilde{x}-h)].$$

Per determinarne l'errore, si sostituiscono nell'espressione trovata gli sviluppi in serie di Taylor opportunamente troncati. Ad esempio, nella (6), sostituendo la (1) e la (2), si ottiene:

$$D^2(\widetilde{x}) = u''(\widetilde{x}) + O(h^2) \tag{7}$$

Un metodo alternativo per determinare un'espressione che approssimi una certa derivata è il *metodo dei coefficienti incogniti*. Esso consiste nel considerare una combinazione lineare dei valori di cui si dispone e di imporre delle condizioni sui coefficienti in modo da assicurare la migliore accuratezza. Ad esempio, si vuole trovare un'approssimazione della derivata prima in $\widetilde{x}$ conoscendo i valori $u(\widetilde{x})$, $u(\widetilde{x}-h)$ e $u(\widetilde{x}-2h)$. Si consideri la seguente combinazione:

$$D_1u(\widetilde{x}) = au(\widetilde{x}) + bu(\widetilde{x}-h) + cu(\widetilde{x}-2h).$$

Considerando lo sviluppo $u(\widetilde{x}-2h) = u(\widetilde{x}) - 2hu'(\widetilde{x}) + 1/2(2h)^2u''(\widetilde{x}) - 1/6(2h)^3u'''(\widetilde{x}) + O(h^4)$ e la (2), si ottiene

$$D_1u(\widetilde{x}) = (a+b+c)u(\widetilde{x})-(b+2c)hu'(\widetilde{x})+\frac{1}{2}(b+4c)h^2u''(\widetilde{x})-\frac{1}{6}(b+8c)h^3u'''(\widetilde{x})+O(h^4),$$

da cui, per ottenere la massima accuratezza, si possono imporre le seguenti condizioni:

$$\begin{cases} a+b+c=0 \\ b+2c=-1/h \\ b+4c=0 \end{cases}$$

Risolvendo il precedente sistema si ottiene la seguente espressione

$$D_1u(\widetilde{x}) = \frac{1}{2h}[3u(\widetilde{x}) - 4u(\widetilde{x}-h) + u(\widetilde{x}-2h)]$$

Sostituendo in quest'ultima gli sviluppi in serie di Taylor per $u(\widetilde{x}-h)$ e $u(\widetilde{x}-2h)$, si ottiene il seguente errore

$$D_1u(\widetilde{x}) = u'(\widetilde{x}) + O(h^2).$$

1.2 Interpolazione

Un modo per ottenere schemi alle differenze finite più generali consiste nel derivare (esattamente) un opportuno interpolatore della funzione $u(x)$ a partire da un numero finito di informazioni. In base al tipo di interpolazione utilizzata si ottengono diversi schemi. In particolare nel seguito faremo riferimento a 3 tipi di interpolazione.

1.2.1 Interpolazione composita lagrangiana - Differenze finite classiche

Si suddivida il dominio di intresse $[a,b]$ in $N+1$ nodi $a = x_0, x_1, \ldots, x_N = b$ e si raggruppino questi ultimi in gruppi di $k+1$ (con k fissato) in modo che i nodi estremi di due intervalli adiacenti I_l e I_{l+1} coincidano. Si definisce *interpolatore composito lagrangiano (ICL) di grado k* la funzione $\Pi_h^k u(x)$ che soddisfa le seguenti proprietà:

$$\begin{cases} \Pi_h^k u(x_j) = u(x_j) & \forall j = 0, \ldots, N \\ \Pi_h^k u|_{I_l} \in \mathbb{P}^k & \forall I_l \end{cases}$$

Tale interpolatore è unico e soddisfa la seguente stima:

$$\|u - \Pi_h^k u\|_\infty \leq Ch^{k+1}\|u^{(k+1)}\|_\infty$$

dove si è indicato con $\|g\|_\infty = \max_{x\in[a,b]} |g(x)|$ e a patto di assumere che $\|u^{(k+1)}\|_\infty$ sia una quantità finita. Questo ci assicura la convergenza uniforme dell' ICL alla funzione u con velocità pari (almeno) a $k+1$. Ogni volta che si deriverà l'ICL per ottenere uno schema alle differenze finite si perderà un grado di esattezza (ad esempio, un' approssimazione per la derivata prima avrà ordine di esattezza k).

Ponendo $k = 1$ si consideri l'interpolazione composita *lineare* ottenuta raccordando linearmente i valori della funzione u in corrispondenza dei nodi. Si consideri il generico elemento I_j costituito dai nodi x_j e x_{j+1} e si ponga $u(x_i) = u_i$ e $p(x) = \Pi_h^1 u|_{I_j}$. Si faccia inoltre una traslazione dell'asse y in modo che $x_j = 0$ e $x_{j+1} = h$, ove $h = x_{j+1} - x_j$. La retta passante per i due punti $(0, u_j)$ e (h, u_{j+1}) ha espressione

$$p(x) = \frac{u_{j+1} - u_j}{h}x + u_j$$

e quindi si ottiene (per la funzione u si mantiene l'originario asse y):

$$u'(x_j) \simeq D_j = p'(0) = \frac{u_{j+1} - u_j}{h} \tag{8}$$

$$u'(x_{j+1}) \simeq D_{j+1} = p'(h) = \frac{u_{j+1} - u_j}{h} \tag{9}$$

dove con D_i si è indicata l'approssimazione della derivata prima nel nodo x_i. Le due espressioni precedenti costituiscono un'approssimazione del primo ordine (infatti l'ICL ha ordine di esattezza pari a 2) della derivata prima in un nodo costruita utilizzando il valore della funzione nel nodo in questione ed in quello immediatamente alla sua destra (Eulero in avanti) o alla sua sinistra (Eulero indietro) rispettivamente.

Si ponga ora $k = 2$ e si consideri il generico elemento I_l formato dai nodi x_{j-1}, x_j e x_{j+1}. Sia $p(x) = \Pi_h^2 u|_{I_l}$ l'ICL di secondo ordine relativo all'intervallo I_l e si trasli l'asse y in modo che $x_{j-1} = -h, x_j = 0$ e $x_{j+1} = h$. Si ottiene la seguente espressione per la parabola interpolatrice:

$$p(x) = \frac{u_{j+1} - 2u_j + u_{j-1}}{2h^2}x^2 + \frac{u_{j+1} - u_{j-1}}{2h}x + u_j.$$

Derivando, si ottengono quindi le seguenti 3 espressioni:

$$u'(x_j) \simeq D_j = p'(0) = \frac{u_{j+1} - u_{j-1}}{2h} \tag{10}$$

$$u'(x_{j-1}) \simeq D_{j-1} = p'(-h) = \frac{-u_{j+1} + 4u_j - 3u_{j-1}}{2h}$$

$$u'(x_{j+1}) \simeq D_{j+1} = p'(h) = \frac{3u_{j+1} - 4u_j + u_{j-1}}{2h} \tag{11}$$

Tutte e tre queste approssimazioni della derivata prima hanno grado di esattezza pari a 2 (infatti l'ICL ha ordine di esattezza pari a 3). Si noti come si siano ritrovate le approssimazioni derivate nella precedente sezione.

Per ottenere un'approssimazione della derivata seconda si deriva due volte l'ICL. Ad esempio nel caso precedente si ottiene:

$$f''(x_j) \simeq p''(0) = \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} \tag{12}$$

A differenza di quanto ci si possa aspettare, questa approssimazione della derivata seconda ha grado di esattezza pari a 2. Più in generale infatti, derivando due volte l'ICL si perde solo un grado di esattezza se k è pari.

1.2.2 Interpolazione hermitiana - Differenze finite compatte

Per costruire un polinomio interpolatore di grado N, si potrebbe pensare di utilizzare $N + 1$ informazioni concernenti non solo i valori della u, ma anche delle sue derivate, qualora disponibili. In questo modo si possono costruire interpolatori compositi di un certo grado utilizzando un minor numero di nodi rispetto agli ICL (*interpolatore hermitiano*, IH). Questo tipo di interpolazione non è sempre ben posta. Se però i nodi dove sono assegnate le derivate sono un sottoinsieme di quelli ove si assegnano i valori della funzione, allora l'interpolatore di Hermite esiste.

Tuttavia, se l' interpolazione hermitiana viene utilizzata per ottenere un'approssimazione alle differenze finite della derivata di ordine r, i valori $u^{(r)}(x_j)$ non sono disponibili e devono essere considerati incogniti. Ad esempio, volendo derivare un'approssimazione della derivata prima, si potrebbe costruire un interpolatore hermitiano di grado 4 su 3 nodi, utilizzando i valori $u_{j-1}, u_j, u_{j+1}, u'_{j-1}$ e u'_{j+1} (avendo posto $u'_i = u'(x_i)$). Tuttavia, queste ultime due informazioni non sono disponibili. Di conseguenza si sostituiscono i valori u'_{j-1} e u'_{j+1} con

le approssimazioni (incognite) D_{j-1} e D_{j+1} rispettivamente. Traslando l'asse y come in precedenza, si ottiene:

$$\begin{cases} p(x) = a + bx + cx^2 + dx^3 + ex^4 \\ p'(x) = b + 2cx + 3dx^2 + 4ex^3 \end{cases}$$

Imponendo che siano soddisfate le 5 condizioni, si ottiene:

$$\begin{cases} p(-h) = a - bh + ch^2 - dh^3 + eh^4 = u_{j-1} \\ p(0) = a = u_j \\ p(h) = a + bh + ch^2 + dh^3 + eh^4 = u_{j+1} \\ p'(-h) = a - 2bh + 3ch^2 - 4dh^3 = D_{j-1} \\ p'(h) = a + 2bh + 3ch^2 + 4dh^3 = D_{j+1} \end{cases}$$

Risolvendo questo sistema, si ottiene

$$b = -\frac{1}{4}(D_{j+1} + D_{j-1}) + \frac{3}{4}(u_{j+1} - u_{j-1})$$

da cui

$$u'(x_j) \simeq D_j = p'(0) = b = -\frac{1}{4}(D_{j+1} + D_{j-1}) + \frac{3}{4}(u_{j+1} - u_{j-1})$$

Scrivendo le precedenti espressioni per ogni nodo x_i, si ottiene un sistema lineare $C\boldsymbol{d} = T\boldsymbol{u}$ nelle incognite D_i con C matrice tridiagonale e $\boldsymbol{u}$ il vettore di componenti $u(x_i)$. Risolvendo tale sistema, si ottiene un'approssimazione della derivata prima con grado di esattezza pari a 4. Tuttavia, applicando tali schemi alle equazioni differenziali, il sistema per determinare le formule compatte non va risolto (si veda Sezione 2.1).

Derivando quindi un interpolatore hermitiano, si possono ottenere schemi alle differenze finite di ordine N utilizando un numero di nodi $M < N + 1$ (da cui il nome di differenze finite *compatte*). Il prezzo da pagare consiste nel dover risolvere un sistema lineare per ottenere le approssimazioni cercate a differenza degli schemi classici per cui le espressioni delle approssimazioni sono esplicite.

1.2.3 Interpolazione polinomiale su nodi di Gauss - Differenze finite pseudo-spettrali

Si consideri l'intepolazione composta lagrangiana. Ponendo $k = N$, si ottiene l' *interpolatore polinomiale (IP) di grado N* $\Pi_N u(x)$. L'IP può avere problemi

di instabilità per N sufficientemente grande. Esistono infatti delle funzioni $u(x)$ per cui $\|\Pi_N u - u\|_\infty$ non va a zero se $N \to \infty$. Famoso è il controesempio di Runge, che mostra come l'IP di grado sufficientemente grande della funzione $1/(1+x^2)$ sul dominio $[-1,1]$ sia instabile, nel senso che presenta grandi oscillazioni in prossimità degli estremi. Un rimedio per ovviare a tali fenomeni consiste nel considerare una ubicazione dei nodi non uniforme. In generale, interpolando sui *nodi di Gauss* (radici di opportuni polinomi), si dimostra che l'IP è stabile. In particolare, molto utilizzati sono i nodi di *Gauss-Chebichev* (radici dei polinomi di Chebichev):

$$x_j^c = -cos\Big(\frac{(2j+1)\pi}{2(N+1)}\Big)$$

La precedente espressione si riferisce all'intervallo $[-1,1]$, ma può essere estesa facilmente a un intervallo qualsiasi $[a,b]$. Come si può notare, i nodi di Gauss-Chebichev sono più fitti vicino agli estremi, laddove sorgono i problemi di instabilità. Un'altra possibilità molto utilizzata nei cosiddetti *elementi spettrali* è costituita dai nodi di *Gauss-Legendre*.

In generale gli estremi dell'intervallo non sono nodi di Gauss. Se si volessero includere anche gli estremi per costruire l'IP, si parla di nodi di *Gauss-Lobatto* (in particolare di nodi di Gauss-Lobatto-Chebichev e Gauss-Lobatto-Legendre). L'utilizzo dei nodi di Gauss-Lobatto non garantisce solo la stabilità dell'interpolatore, ma migliora anche l'accuratezza: infatti mentre l'IP su nodi equispaziati ha grado di esattezza pari a N, nel senso che interpola esattamente polinomi di grado N, l'IP costruito sui nodi di Gauss-Lobatto ha grado di esattezza pari a $2N-1$.

Derivando l'IP costruito sui nodi di Gauss (o di Gauss-Lobatto) si ottengono approssimazioni delle derivate su tali nodi che danno luogo agli schemi alle differenze finite *pseudo-spettrali*.

2 Applicazione alle equazioni differenziali

Si vogliono nel seguito fornire degli esempi di applicazione delle espressioni precedentemente ricavate per discretizzare le equazioni differenziali. Data la funzione $f(x)$, si consideri la seguente equazione

$$\begin{cases} -u''(x) = f(x) \quad 0 < x < 1 \\ u(0) = 0 \quad u(1) = 0. \end{cases} \tag{13}$$

L'equazione è semplicissima e in realtà si presta ad essere risolta analiticamente. Basta infatti integrare due volte sul dominio:

$$u'(x) = \int_0^x -f(s)ds + C_1$$

$$u(x) = \int_0^x \Big(\int_0^\xi -f(s)ds \Big) d\xi + C_1 x + C_0$$

Imponendo che siano soddisfatte le condizioni al bordo, si determinano le costanti:

$$\begin{cases} C_0 = 0 \\ C_1 = \int_0^1 \Big(\int_0^\xi f(s)ds \Big) d\xi \end{cases}$$

Ponendo $F(\xi) = \int_0^\xi f(s)ds$, si ottiene integrando per parti:

$$\int_0^1 F(\xi)d\xi = [\xi F(\xi)]_0^1 - \int_0^1 \xi f(\xi)d\xi = F(1) - \int_0^1 \xi f(\xi)d\xi = \int_0^1 (1-\xi)f(\xi)d\xi$$

e

$$\int_0^x F(\xi)d\xi = [\xi F(\xi)]_0^x - \int_0^x \xi f(\xi)d\xi = F(x) - \int_0^x \xi f(\xi)d\xi = \int_0^x (x-\xi)f(\xi)d\xi.$$

Sosituendo, si ottiene:

$$u(x) = \Big(\int_0^1 F(\xi)d\xi \Big) x - \int_0^x F(\xi)d\xi = x \int_0^1 (1-\xi)f(\xi)d\xi - \int_0^x (x-\xi)f(\xi)d\xi$$

e, spezzando il primo integrale,

$$u(x) = \int_0^x \xi(1-x)f(\xi)d\xi + \int_x^1 x(1-\xi)f(\xi)d\xi.$$

Introducendo la *funzione di Green*

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

si ottiene la soluzione esplicita:

$$u(x) = \int_0^1 G(\xi, x) f(\xi) d\xi.$$

Passando ad una discretizzazione che utilizzi le formule alle differenze finite, si consideri una discretizzazione dell'intervallo $[0,1]$ di passo uniforme h e nodi x_j, con $j = 0, \ldots, m+1$, e si denoti con T_h tale dominio computazionale. Si indichi inoltre con U_j l'approssimazione incognita di $u(x_j)$. Si consideri inoltre come approssimazione della derivata seconda l'espressione data dalla (6). Si può allora sostituire l'equazione (13) con un sistema di equazioni algebriche ottenute scrivendo per ogni nodo x_j:

$$-\frac{1}{h^2}(U_{j+1} - 2U_j + U_{j-1}) = f(x_j). \tag{14}$$

In altre parole si *colloca* l'equazione in un punto e la si sosituisce con una sua approssimazione. Ciò fa sì che il metodo delle differenze finite sia, a differenza degli elementi finiti, un metodo di *collocamento*.

Volendo trattare diverse condizioni al bordo, si distinguono di seguito tre casi che portano a tre diversi problemi discreti.

2.1 Condizioni di Dirichlet

Si considerino le seguenti condizioni al bordo:

$$\begin{cases} u(0) = \alpha \\ u(1) = \beta \end{cases}$$

Si noti che in questo caso si conoscono i valori $U_0 = \alpha$ e $U_{m+1} = \beta$. Avendo quindi m incognite, si scrive l'equazione (14) nei nodi $x_1, \ldots, x_m$ ottenendo così m equazioni. Si ottiene un sistema lineare nella forma $A\boldsymbol{u}_h = \boldsymbol{F}$, con $\boldsymbol{u}_h = (U_1, \ldots, U_m)$ il vettore delle incognite,

$$A = \frac{1}{h^2}\begin{bmatrix} 2 & -1 & 0 & \ldots & & \\ -1 & 2 & -1 & 0 & \ldots & \\ 0 & -1 & 2 & -1 & 0 & \ldots \\ \vdots & & \ddots & \ddots & \ddots & \\ & \ldots & 0 & -1 & 2 & -1 \\ & & \ldots & 0 & -1 & 2 \end{bmatrix}$$

e

$$\boldsymbol{F} = \begin{bmatrix} f(x_1) + \alpha/h^2 \\ f(x_2) \\ f(x_3) \\ \vdots \\ f(x_{m-1}) \\ f(x_m) + \beta/h^2 \end{bmatrix}.$$

Infatti scrivendo l'equazione (14) nel nodo x_1 si ottiene:

$$-\frac{1}{h^2}(U_2 - 2U_1 + U_0) = f(x_1).$$

Il termine U_0, essendo noto, può essere portato a destra. Ciò spiega il fatto che la prima riga della matrice A e la prima componente del vettore $\boldsymbol{F}$ abbiamo una forma diversa rispetto alle altre. Analogo ragionamento vale scrivendo l'equazione (14) nel nodo x_m.

Si noti come se si fosse utilizzato uno schema compatto per discretizzare la derivata seconda, ottenuto risolvendo il sistema $C\boldsymbol{d}_2 = T\boldsymbol{u}_h$, con $\boldsymbol{d}_2$ il vettore delle approssimazioni incognite della derivata seconda, si sarebbe ottenuto $C^{-1}T\boldsymbol{u}_h = \boldsymbol{F}$, da cui il sistema lineare $T\boldsymbol{u}_h = C\boldsymbol{F}$. In questo caso la matrice T ha uno *stencil* a parità di accuratezza più piccolo rispetto a quello della matrice A. Al contario, la matrice S derivante dall'applicazione delle formule pseudo-spettrali è piena e quindi il costo computazionale richiesto è molto più grande. In compenso l'ordine di accuratezza di tali formule (e quindi della soluzione discreta, vedi Sezione 3) è molto alto, in particolare se $f \in C^\infty$ si ha una convergenza esponenziale.

Completato il percorso di discretizzazione, che ha portato ad un sistema lineare con matrice tridiagonale, si passi alla analisi del problema discreto, che, sostanzialmente, consiste nel chiedersi:

a) Il problema discreto, ovvero il sistema lineare precedentemente ricavato, è ben posto?

b) $\boldsymbol{u}_h \to u$ per $h \to 0$, ovvero la soluzione discreta *converge* a quella esatta?

Si noti come la soluzione discreta sia un vettore e che, se il suo limite esiste, esso sia una funzione poichè $T_h \to [0, 1]$ per $h \to 0$. In generale per rispondere al primo quesito è sufficiente mostrare che la matrice A è *non singolare*. In

questo caso specifico, essendo $\mathbf{0} \neq \boldsymbol{z} \in \mathbb{R}^m$, si ottiene:

$$\boldsymbol{z}^T A \boldsymbol{z} = [z_1\, z_2 \dots z_m] \frac{1}{h^2} \begin{bmatrix} 2 & -1 & 0 & \dots & & \\ -1 & 2 & -1 & 0 & \dots & \\ 0 & -1 & 2 & -1 & 0 & \dots \\ \vdots & & \ddots & \ddots & \ddots & \\ & \dots & 0 & -1 & 2 & -1 \\ & & \dots & 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{bmatrix} =$$

$$= \frac{1}{h^2} [z_1\, z_2 \dots z_m] \begin{bmatrix} 2z_1 - z_2 \\ -z_1 + 2z_2 - z_3 \\ \vdots \\ -z_{j-1} + 2z_j - z_{j+1} \\ \vdots \\ -z_{m-1} + 2z_m \end{bmatrix} =$$

$$= \frac{1}{h^2} \Big(z_1^2 + z_m^2 + \sum_{i=1}^{m-1} (z_i - z_{i+1})^2 \Big) > 0 \tag{15}$$

Ciò significa che la matrice A è definita positiva, che i suoi autovalori sono tutti positivi e che quindi è invertibile. Si fa notare come la proprietà fondamentale sia la *definizione in segno*, non la positività. Infatti sono le matrici indefinite o semidefinite in segno a risultare problematiche.

Per quanto riguarda il quesito b) sulla convergenza si rimanda alla prossima sezione.

2.2 Condizioni miste

Si considerino le seguenti condizioni al bordo:

$$\begin{cases} u'(0) = \sigma \\ u(1) = \beta \end{cases}$$

si assegni cioè nel nodo x_0 una condizione di Neumann. Si noti che in questo caso si ha un'incognita in più rispetto al caso precedente, in quanto il valore di U_0 non è assegnato. Si ha quindi bisogno di un'equazione in più che tenga in conto della condizione di Neumann. Si potrebbe ad esempio considerare la seguente approssimazione del tipo (3) per la condizione in x_0:

$$\frac{U_1 - U_0}{h} = \sigma$$

che, posta come prima equazione, porta al seguente sistema lineare nelle incognite $(U_0, U_1, \ldots, U_m)$:

$$\frac{1}{h^2}\begin{bmatrix} 1 & -1 & 0 & \ldots & & \\ -1 & 2 & -1 & 0 & \ldots & \\ 0 & -1 & 2 & -1 & 0 & \ldots \\ \vdots & & \ddots & \ddots & \ddots & \\ & \ldots & 0 & -1 & 2 & -1 \\ & & \ldots & 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} U_0 \\ U_1 \\ U_2 \\ \vdots \\ U_{m-1} \\ U_m \end{bmatrix} = \begin{bmatrix} -\sigma/h \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{m-1}) \\ f(x_m) + \beta/h^2 \end{bmatrix}.$$

Anche in questo caso si può verificare che la matrice A è definita positiva.

Un'altra scelta (che come si vedrà in seguito porta ad una migliore approssimazione in termini di errore) è quella di considerare per la condizione di Neumann un'approssimazione centrata del tipo (5):

$$\frac{1}{2h}(U_1 - U_{-1}) = \sigma, \tag{16}$$

che introduce un'ulteriore incognita associata ad un nuovo nodo (nodo fantasma). Per eliminare tale incognita si noti come nel nodo x_0 sia possibile collocare l'equazione:

$$-\frac{1}{h^2}(U_{-1} - 2U_0 + U_1) = f(x_0).$$

Eliminando U_{-1} dalla (16) e sostituendola nella precedente, si ottiene la seguente equazione da inserire nel sistema lineare:

$$\frac{1}{h^2}(U_0 - U_1) = -\sigma/h + \frac{1}{2}f(x_0).$$

2.3 Condizioni di Neumann

Si consideri infine il caso in cui si assegni su entrambi i nodi esterni una condizione di Neumann:

$$\begin{cases} u'(0) = \sigma_0 \\ u'(1) = \sigma_1 \end{cases}$$

Utilizzando per entrambe le condizioni la seconda delle strategie analizzate nel caso precedente, si perviene al seguente sistema lineare:

$$\frac{1}{h^2}\begin{bmatrix} 1 & -1 & 0 & \dots & & \\ -1 & 2 & -1 & 0 & \dots & \\ 0 & -1 & 2 & -1 & 0 & \dots \\ \vdots & & \ddots & \ddots & \ddots & \\ & \dots & 0 & -1 & 2 & -1 \\ & & \dots & 0 & -1 & 1 \end{bmatrix}\begin{bmatrix} U_0 \\ U_1 \\ U_2 \\ \vdots \\ U_m \\ U_{m+1} \end{bmatrix} = \begin{bmatrix} -\sigma_0/h + 1/2 f(x_0) \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_m) \\ \sigma_1/h + 1/2 f(x_{m+1}) \end{bmatrix}.$$

Tuttavia, in questo caso la matrice A è singolare per cui il sistema può o ammettere infinite soluzioni oppure non ammetterne, in base al fatto che il termine noto appartenga o meno all'immagine di A. Infatti in questo caso si ottiene:

$$\boldsymbol{z}^T A \boldsymbol{z} = \frac{1}{h^2}\sum_{i=0}^{m}(z_i - z_{i+1})^2 \geq 0.$$

La matrice A è semidefinita positiva, infatti per $\boldsymbol{z} = \beta \boldsymbol{e}_i$, con $\beta \neq 0$ ed $\boldsymbol{e}_i$ il vettore di componenti δ_{ij} (*delta di Kroenecher*), si ottiene $\boldsymbol{z}^T A \boldsymbol{z} = 0$ con $\boldsymbol{z} \neq \boldsymbol{0}$. Ciò era da aspettarsi, in quanto la funzione incognita al bordo compare solo sotto segno di derivata. La causa della non risolvibilità del precedente sistema lineare non è quindi da ricercarsi nel sistema numerico scelto, bensì nel problema continuo che è mal posto. Ciò può essere anche spiegato dando un'interpretazione fisica dell'equazione (13). Essa potrebbe rappresentare l'equazione stazionaria che descrive la distribuzione di temperatura in una sbarra soggetta ad una sorgente di calore $f(x)$. Le condizioni di Dirichlet corrispondono ad assegnare il valore della temperatura in un estremo, mentre quelle di Neumann corrispondono ad assegnarne il flusso di calore entrante o uscente. Assegnando due condizioni di Neumann omogenee (cioè imponendo che la sbarra sia termicamente isolata) ci si aspetta che con termine sorgente nullo qualsiasi soluzione del tipo $u = c$, con c costante, risolva il problema posto. In tal caso infatti il termine $\boldsymbol{F}$ (cioè il vettore nullo) appartiene all'immagine di A. Se $f \neq 0$, ci si aspetta che il problema non ammetta soluzione dal momento che si sta aggiungendo (o togliendo) calore ad un sistema isolato, la cui temperatura crescerà quindi (o diminuirà) indefinitivamente.

3 Consistenza, stabilità e convergenza

In questa sezione si indagherà la convergenza della soluzione approssimata alla soluzione esatta. Non si farà dunque più riferimento al problema con entrambe le condizioni di Neumann.

Partendo dalle considerazioni sugli ordini di accuratezza delle formule di approssimazione delle derivate, si potrebbe pensare di asserire che poichè l'errore commesso approssimando la derivata seconda con la (6) è un $O(h^2)$, anche l'errore di approssimazione fra la vera soluzione $u(x)$ del problema (13) e quella approssimata sia del secondo ordine. Tuttavia tale affermazione va giustificata in quanto non ovvia. Si introduca a questo proposito la quantità

$$\tau_j = -\frac{1}{h^2}(u(x_{j+1}) - 2u(x_j) + u(x_{j-1})) - f(x_j)$$

che rappresenta l'errore commesso nel nodo x_j dalla vera soluzione se calata nel problema discreto ed è denominato *errore di troncamento locale*. Introducendo gli sviluppi in serie di Taylor nella (6) e supponendo che la soluzione $u(x)$ sia sufficientemente regolare, si ottiene, utilizzando la (7):

$$\|\tau_j\| = \| - (u''(x_j) + O(h^2)) - f(x_j)\| = \frac{h^2}{12}\|u^{(iv)}(x_j)\| = \frac{h^2}{12}\|f''(x_i)\| = O(h^2).$$

Il dato f deve essere quindi almeno C^2 per poter fare le considerazioni che seguono. Denotando con $\boldsymbol{u}$ il vettore dei valori $u(x_j)$ e con $\boldsymbol{\tau}$ quello dei valori τ_j (errore di troncamento globale), si può scrivere $\boldsymbol{\tau} = A\boldsymbol{u} - \boldsymbol{F}$ da cui si ottiene $A\boldsymbol{E} = -\boldsymbol{\tau}$, avendo denotato con $\boldsymbol{E} = \boldsymbol{u}_h - \boldsymbol{u}$ il vettore dell' errore globale. Si noti come per ricavare quest'ultimo sistema si sia utilizzata la linearità del problema di partenza. Siamo interessati a far vedere che $\|\boldsymbol{E}\| \to 0$ per una qualche norma se $h \to 0$. Quest'ultima è nota come proprietà di *convergenza* del metodo. Se il problema numerico è ben posto, si può scrivere $\boldsymbol{E} = -A^{-1}\boldsymbol{\tau}$. Applicando le norme si ottiene:

$$\|\boldsymbol{E}\| = \|A^{-1}\boldsymbol{\tau}\| \leq \|A^{-1}\|\,\|\boldsymbol{\tau}\|. \tag{17}$$

Per assicurare la convergenza dobbiamo dunque verificare che:

- $\|\boldsymbol{\tau}\| \to 0$ per una qualche norma. Questa proprietà è nota come *consistenza* del metodo numerico.
- $\|A^{-1}\| \leq C$, con C indipendente da h. Questa proprietà è nota come *stabilità* del metodo numerico e consiste nel verificare che la matrice A^{-1} non esploda facendo tendere h a 0

3.1 Consistenza

Dato un vettore $\boldsymbol{W} \in \mathbb{R}^n$, si considerano le seguenti norme:

$$\begin{cases} \|\boldsymbol{W}\|_\infty = max_j|W_j| \\ \|\boldsymbol{W}\|_2 = \sqrt{\sum_{i=1}^n |W_j|^2} \end{cases}$$

Nel caso di condizioni di Dirichlet si ottiene $\tau_j = O(h^2)\,\forall j$ e quindi $\|\boldsymbol{\tau}\|_\infty = O(h^2)$ e $\|\boldsymbol{\tau}\|_2 = O(h^2)$. Nel caso invece di condizioni al bordo miste, seguendo la prima delle due strategie proposte, si ottiene

$$\tau_0 = \frac{1}{h}(u(x_1) - u(x_0)) - \sigma = (u'(x_0) + O(h)) - \sigma = O(h),$$

quindi $\|\boldsymbol{\tau}\|_\infty = O(h)$ e $\|\boldsymbol{\tau}\|_2 = O(h^p)$, con $1 \leq p < 2$. Seguendo invece la seconda strategia si ottiene $\tau_0 = O(h^2)$, da cui $\|\boldsymbol{\tau}\|_\infty = O(h^2)$ e $\|\boldsymbol{\tau}\|_2 = O(h^2)$. La consistenza è dunque verificata in tutti i casi considerati.

3.2 Stabilità

Un primo modo per verificare la stabilità nella norma 2 è specifico al problema considerato con condizioni di Dirichlet. Poichè la matrice A^{-1} è definita positiva, la sua norma 2 è data dall'autovalore massimo in modulo, che coincide con l'inverso dell'autovalore minimo di A. Si può dimostrare che gli autovalori di A sono dati da

$$\lambda_k = \frac{2}{h^2}(\cos(k\pi h) - 1)$$

Per $h \to 0$, si può scrivere:

$$\cos(k\pi h) \simeq 1 - \frac{k^2\pi^2}{2} + O(h^4)$$

da cui l'autovalore di modulo minimo ($k = 1$) è:

$$|\lambda_1| \simeq \pi^2 + O(h^2).$$

Per $h \to 0$, si può quindi scrivere la seguente stima:

$$\|A^{-1}\|_2 = \frac{1}{|\lambda_1|} \simeq \frac{1}{\pi^2} \tag{18}$$

Un secondo modo, più generale, per ottenere una stima di stabilita' di un problema con condizioni di Dirichlet (per semplicità poste omogenee) consiste nel dare una interpretazione *operatoriale* al problema discreto, in maniera analoga a quanto si farà per gli *elementi finiti*. Indichiamo con $V_{h,0}$ lo spazio delle funzioni di griglia (ossia definite solo sui nodi) nulle agli estremi. Si consideri quindi l'operatore L_h tale che:

$$L_h v_h(x_j) = -\frac{v_h(x_{j-1}) - 2v_h(x_j) + v_h(x_{j+1})}{h^2}$$

per ogni $v_h \in V_{h,0}$, in modo che sia

$$L_h u_h(x_j) = f(x_j)$$

dove con u_h si è indicata la funzione di $V_{h,0}$ associata al vettore $\boldsymbol{u}_h$. Ponendo $v_h(x_j) = v_{jh}$, introduciamo la norma $\|v_h\|_*^2 = \Big(h/2v_{0h}^2 + h/2v_{mh}^2 + \sum_{i=1}^{m-1} hv_{ih}^2\Big)$. Ricordando la (15), si ottiene:

$$(L_h u_h, u_h) = \boldsymbol{u}_h^T A \boldsymbol{u}_h = \boldsymbol{u}_h^t \boldsymbol{F} \geq \sum_{i=1}^{m-1} \Big(\frac{U_{i+1} - U_i}{h}\Big)^2.$$

Quest'ultima quantità definisce una norma (che indicheremo con $|||\cdot|||$) per le funzioni di $V_{h,0}$ (si fa notare che non sarebbe una norma per funzioni che non si annullano al bordo). Si ottiene quindi:

$$|||u_h|||^2 \leq (L_h u_h, u_h) \leq \|f\|_* \|u_h\|_*$$

Si può inoltre dimostrare che $\|v_h\|_* \leq 1/\sqrt{2}|||v_h|||$, da cui si ottiene la stima di stabilità voluta:

$$\|u_h\|_* \leq \frac{1}{2}\|f\|_* \tag{19}$$

3.3 Convergenza

Come anticipato, la convergenza di un metodo numerico applicato a un problema lineare è garantita dalla consistenza e dalla stabilità. Si può anche dimostrare che tali condizioni sono necessarie per ottenere la convergenza. In particolare, utilizzando la stima (18) e ricordando la (17), si ottiene:

$$\|\boldsymbol{E}\|_2 \leq \|A^{-1}\|_2 \, \|\boldsymbol{\tau}\|_2 \leq \frac{1}{\pi^2}\|\boldsymbol{\tau}\|_2 = O(h^2)$$

da cui si ottiene la convergenza della soluzione discreta a quella esatta con ordine di convergenza uguale a 2 per il problema di Dirichlet. Poichè si può dimostrare che anche la norma della matrice inversa associata al problema con condizioni miste è limitata da una costante indipendente da h, si ottiene che la velocità di convergenza coincide con l'errore di troncamento globale e quindi dipende dalla scelta fatta per discretizzare la condizione di Neumann.

Utilizzando invece la (19), dall'equazione dell'errore scritta in forma operatoriale $L_h e_h = \tau$, si ottiene:

$$\|\boldsymbol{E}\|_* = \|e_h\|_* \leq \frac{1}{2}\|\tau\|_* = O(h^2)$$

con le stesse considerazioni di prima circa il problema misto.

Si passi a considerare la norma infinito. Innanzitutto si osserva che:

$$h\|v_h\|_\infty \leq \sum_{i=1}^{m-1} h v_{hi}^2 = \|v_h\|_*^2$$

da cui si ottiene:

$$\|e_h\|_\infty \leq \frac{1}{\sqrt{h}}\|e_h\|_* \leq \frac{1}{\sqrt{h}} C h^2 \leq C h^{3/2}.$$

C'e quindi una perdita di accuratezza rispetto all'errore di troncamento (sub-ottimalità). Seguiamo dunque una strada alternativa per cercare di dimostrare qualcosa di meglio. E' immediato osservare che, almeno da un punto di vista formale, le funzioni di Green introdotte precedentemente sono soluzioni dei problemi $-u'' = \delta_i$, con δ_i la delta di Dirac centrata in x_i. La controparte discreta scalata di h di questo problema diventa:

$$A\boldsymbol{g}_{ih} = \boldsymbol{b}_i$$

con $\boldsymbol{b}_i$ di componenti $b_{i,j} = \delta_{ij}$ (delta di Kroenecher). Si ottiene esplicitamente $\boldsymbol{g}_{ih} = h\sum_{k=1}^{m-1} G(x_k, x_i)$. Inoltre si può mostrare che $\boldsymbol{g}_{ih}$ sono le colonne della matrice A^{-1} e quindi si ottiene:

$$\|A^{-1}\|_\infty \leq \|\boldsymbol{g}_{ih}\|_\infty \leq h(m-1) \leq 1$$

da cui si recupera l'ottimalità della convergenza anche per la norma infinito.

4 Altre questioni

La bontà della soluzione discreta ottenuta dipende principalmente da due fattori:

- Il numero di informazioni k usate per costruire la formula di approssimazione della derivata.
- Il passo di discretizzazione h.

E' chiaro che l'ordine di accuratezza dell'errore di troncamento e quindi in linea di principio l'ordine di convergenza, dipendono dal numero k. In particlare più alto è k maggiore (più precisamente *non minore*) sarà l'ordine di accuratezza. Il prezzo da pagare e' che dobbiamo risolvere un sistema più *costoso* in quanto la matrice da invertire è meno sparsa. D'altra parte, riducendo il passo h non si migliora l'ordine di accuratezza, ma si riduce l'errore. Ciò è vero fino ad un certo punto in quanto da una parte prima o poi ci si scontrerà con gli errori di macchina e dall'altra al diminuire di h spesso il numero di condizionamento della matrice A aumenta e ciò può portare ad una soluzione imprecisa (oltre che a maggiori oneri computazionali). Un altro fattore che determina la scelta di uno schema è la natura fisica del problema. Ad esempio in un'equazione di diffusione trasporto, se il termine di trasporto b è positivo (negativo), cioè se il trasporto avviene da sinistra (destra) verso destra (sinistra), si preferirà uno schema decentrato all'indietro (in avanti) per discretizzare la derivata prima.