

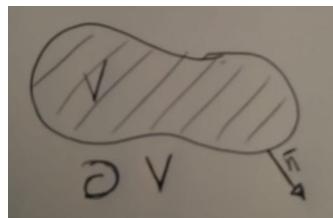
# Lezioni Analisi Numerica II modulo

Sebastiano Boscarino

Dipartimento di Matematica e Informatica  
Università di Catania, ITALY

March 12, 2020

## Equazione del calore



Bilancio dell'energia

$$\frac{d}{dt} \int_V \mathcal{E} d^3x = - \int_{\partial V} \vec{q} \cdot \vec{n} dS,$$

Teorema della divergenza di Gauss

$$\int_{\partial V} \vec{q} \cdot \vec{n} dS = \int_V \nabla \cdot \vec{q} d^3x$$

l'operatore di divergenza  $\nabla \cdot$  (nabla) (in coordinate cartesiane è definito da:

$$\nabla = \mathbf{i} \frac{\partial}{\partial x} + \mathbf{j} \frac{\partial}{\partial y} + \mathbf{k} \frac{\partial}{\partial z}$$

$$\int_V \left( \rho \frac{\partial e}{\partial t} + \nabla \cdot \vec{q} \right) d^3x = 0, \quad \mathcal{E} = \rho e. \quad (1)$$

Relazione vera per ogni  $V$ :

$$\rho \frac{\partial e}{\partial t} + \nabla \cdot \vec{q} = 0$$

incognite 4:  $e$ ,  $\vec{q}$ , ed una equazione, non siamo in grado di risolvere questo problema.

Legge di Fourier  $\vec{q} = -k\nabla T$ ,  $k$  conducibilità termica (dipende dal materiale). La quantità  $\nabla$  operatore di gradiente definito come,  $\phi(x, y, z)$  quantità scalare:

$$\nabla \phi = \left( \frac{\partial \phi}{\partial x}, \frac{\partial \phi}{\partial y}, \frac{\partial \phi}{\partial z} \right)^T.$$

All'equazione (1) si può considerare anche la presenza di un termine di sorgente  $\int_V s(\vec{x}, t)d^3x$  l'equazione diventa:

$$\rho \frac{\partial e}{\partial t} + \nabla \cdot \vec{q} = s(\vec{x}, t). \quad (2)$$

Vale la seguente relazione per la densità di energia interna  $e(T)$ :

$$\frac{de}{dT} = c_v$$

dove  $c_v$  = calore specifico.

Sostituiamo in (2) otteniamo:

$$\rho c_v \frac{\partial T}{\partial t} - \nabla \cdot (k \nabla T) = s,$$

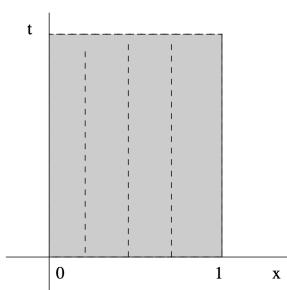
Se  $\rho c_v$  = costante, otteniamo:

$$\frac{\partial T}{\partial t} = \nabla \cdot (K \nabla T) + \Phi,$$

con  $K = \frac{k}{\rho c_v}$ , (coefficiente di diffusione)  $\Phi = \frac{s}{\rho c_v}$ .

Problema ben posto:  $K > 0$  coefficiente positivo; Problema IBVP, ovvero a valori iniziali e condizioni al bordo.

Questa equazione l'andiamo a studiare in un dominio limitato, sia nello spazio  $[a, b]$ , (es.  $[0, 1]$ ) che nel tempo  $[0, t]$ .



IVBP in 1 dimensione:

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x} \left( K(x) \frac{\partial u}{\partial x} \right) + \Phi,$$

con condizioni iniziale  $u(x, 0) = u_0(x)$  e condizioni al bordo:

- ▶ condizioni di Dirichelet:

$$u(a, t) = u_a(t), \quad u(b, t) = u_b(t),$$

- ▶ Condizioni di Neumann:

$$u(a, t) = u_a(t), \quad \frac{\partial u}{\partial x}(b, t) = u'_b(t),$$

- ▶ condizioni di Robin:

$$\alpha u(b, t) + \beta \frac{\partial u}{\partial x}(b, t) = r_b,$$

Equazione stazionaria (condizioni al contorno non dipendono dal tempo, e la soluzione non dipende dal tempo!)

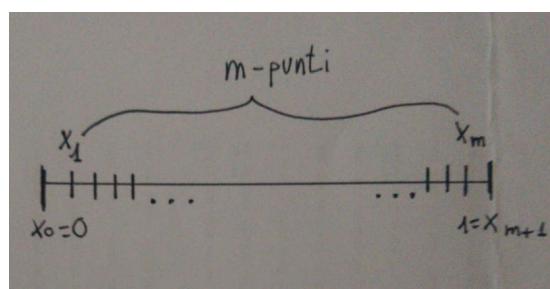
$$u_{xx} = f(x), \quad u(a) = u_a, \quad u(b) = u_b.$$

Soluzioni: se  $f \in C^k \rightarrow$  soluzione di classe  $u \in C^{k+2}$ .

Il modo più semplice per discretizzare questa equazione è attraverso le differenze finite.

Consideriamo l'intervallo  $[0, 1]$ , con  $x_0 = 0$ ,  $x_{m+1} = 1$ ,  $h = 1/(m+1)$ , con  $x_j = jh$ ,  $j = 0, \dots, m+1$ , ( $m$  punti interni,  $m+1$  intervalli,  $m+2$  punti totali).

Sia  $u'_j \approx u(x_j)$ .



Consideriamo l'espansione di Taylor delle funzioni:

$$u(x_{j+1}) = u(x_j + h),$$

$$u(x_{j-1}) = u(x_j - h)$$

$$u(x+h) = u(x) + hu' + \frac{h^2}{2}u'' + \frac{h^3}{6}u''' + \frac{h^4}{24}u^{(IV)} + \dots$$

$$u(x-h) = u(x) - hu' + \frac{h^2}{2}u'' - \frac{h^3}{6}u''' + \frac{h^4}{24}u^{(IV)} + \dots$$

Sommando termine a termine e dividendo per  $h^2$

$$\frac{u(x+h) - 2u(x) + u(x-h)}{h^2} = u''(x) + \frac{h^2}{12}u^{(IV)} + \dots$$

Scritta l'equazione nei punti  $x_j$ :

$$u''(x_j) = \frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{h^2} - \frac{h^2}{12}u^{(IV)}(\xi_j) \quad \xi_j \in [x_{j-1}, x_{j+1}],$$

quindi abbiamo:

$$\frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{h^2} = f(x_j) + \frac{h^2}{12}u^{(IV)}(\xi_j) \quad \xi_j \in [x_{j-1}, x_{j+1}],$$

Segue per l'approssimazione dell'equazione:

$$\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} = f(x_j), \quad \forall j = 1, \dots, m.$$

Quindi segue:

$$\begin{cases} \frac{-2u_1 + u_2}{h^2} = f_1 - \frac{u_a}{h^2}, & j = 1 \\ \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} = f_j - \frac{u_a}{h^2}, & j = 2, \dots, m-1 \\ \frac{u_{m-1} - 2u_m}{h^2} = f_m - \frac{u_b}{h^2}, & j = m \end{cases}$$

Introducendo i vettori  $U = (u_1, \dots, u_m)$  e  $U^e = (u(x_1), \dots, u(x_m))$  In forma vettoriale il nostro sistema diventa

$$AU^e = F + \tau_h.$$

Sistema per eq. numerica  $AU = F$  con:

$$A = \begin{pmatrix} -2 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \cdots & \cdots & 1 & -2 \end{pmatrix}$$

e

$$F = \begin{pmatrix} f(x_1) - \frac{u_a}{h^2} \\ f(x_2) \\ \vdots \\ f(x_m) - \frac{u_b}{h^2} \end{pmatrix}, \quad \tau_h = \frac{h^2}{12} \begin{pmatrix} u^{(IV)}(\xi_1) \\ u^{(IV)}(\xi_2) \\ \vdots \\ u^{(IV)}(\xi_m) \end{pmatrix}$$

Il primo problema che si pone dal punto di vista matematico è la convergenza:  $U \rightarrow U^e$ , per  $h \rightarrow 0$ .

Consideriamo la differenza abbiamo:

$$A(U^e - U) = \tau_h, \rightarrow U^e - U = A^{-1}\tau_h$$

passando alle norme:

Diciamo la norma-2, quindi abbiamo:

$$\|U^e - U\|_2 \leq \|A^{-1}\|_2 \|\tau_h\|_2$$

con

$$\|U^e - U\|_2^2 = \frac{1}{m} \sum_{j=1}^m |U_j^e - U_j|^2.$$

Se vogliamo che  $\|U^e - U\|_2 \rightarrow 0$ , allora  $\tau_h \rightarrow 0$ , per  $h \rightarrow 0$ , (ovvero proviamo la consistenza del metodo!)

Sia  $L_h$  un operatore discreto, applicando tale operatore alla soluzione esatta abbiamo una certa quantità detta errore di discretizzazione  $d_h(x)$ , se questa quantità in qualche norma, è maggiorata da una funzione che tende a zero, quando  $h \rightarrow 0$ , allora il metodo è detto consistente.

Allora abbiamo per la soluzione esatta, considerando la norma-2 e  $f \in C^2[0, 1]$ , esiste ed è finita la derivata  $u^{IV}$ :

$$L_h U^e = AU^e - F = \tau_h,$$

$$\|\tau_h\|_2^2 \leq \frac{h^2}{12} \|u^{IV}\|_2^2 = \frac{h^2}{12} \|f_{xx}\|_2^2, \quad \|f_{xx}\|_2^2 = \frac{1}{m} \sum_{j=1}^m |f_{xx}(\xi_j)|^2$$

abbiamo da  $h \rightarrow 0$

$$\|\tau_h\|_2 \rightarrow 0.$$

Il metodo è consistente.

Da

$$\|U^e - U\|_2 \leq \|A^{-1}\|_2 \|\tau_h\|_2$$

per provare la convergenza allora bisogna provare che  $\|A^{-1}\|_2 \leq M$ ,  $M$  costante indipendente da  $h$ , ovvero che il metodo sia stabile.

Da questo segue che se il metodo é consistente e stabile allora é convergente.

Ricordiamo che la morma-2 di una matrice  $A \in \mathbb{C}^{n \times n}$

$$\|A\|_2^2 = \max(\lambda(A^* A)),$$

$A^*$  trasposta coniugata (la trasposta nel caso reale).  $A$  simmetrica e reale,  $A = A^T$ ,

$$\|A\|_2^2 = \max(\lambda(A)) = \rho(A),$$

$\rho(A)$  raggio spettrale, allora:

$$\|A\|_2^2 = \max(\lambda(A^{-1})) = \rho(A^{-1}) = \max\left(\frac{1}{|\lambda(A)|}\right) = \frac{1}{\min|\lambda(A)|}.$$

Si cerca quindi l'autovalore di modulo minimo:  $Av = \lambda v$ , ovvero

$$\frac{r_{j+1} - 2r_j + r_{j-1}}{h^2} = \lambda r_j, \quad j = 2, \dots, m-1$$

# Lezioni Analisi Numerica II modulo

Sebastiano Boscarino

Dipartimento di Matematica e Informatica  
Università di Catania, ITALY

March 21, 2020

Per  $j = 1$

$$\frac{r_1 - 2r_1 + r_0}{h^2} = \lambda v_1,$$

Per  $j = m$

$$\frac{r_{m+1} - 2r_m + r_{m-1}}{h^2} = \lambda v_m,$$

Gli elementi  $r_0$  ed  $r_{m+1}$  non sono nella matrice quindi  $r_0 = r_{m+1} = 0$ , e quindi posso valutare:

$$\frac{r_{j+1} - 2r_j + r_{j-1}}{h^2} = \lambda r_j, \quad j = 1, \dots, m,$$

Questa equazione rappresenta un'equazione alle differenze del secondo ordine a coefficienti costanti con c.i.  $r_0 = 0$ ,  $r_{m+1} = 0$ , (scelgo:  $a^j = r_j$ ,  $a$  è della forma  $e^{i\xi}$ , sostituendo all'equazione di secondo grado in  $a$ ). e la soluzione è una combinazione lineare di:  $e^{ij\xi}$ ,  $e^{-ij\xi}$  (2 funzioni linearmente indipendenti). Allora la soluzione che cerchiamo come combinazioni lineari di funzioni trigonometriche con argomento  $i\xi$ :

$$r_j = A \cos(j\xi) + B \sin(j\xi), \quad j = 1, \dots, m$$

Quindi abbiamo un'equazione discreta e dalle condizioni iniziali  $r_0 = 0$ ,  $r_{m+1} = 0$  segue

$$r_0 = 0, \rightarrow A = 0, \quad r_{m+1} = 0 \rightarrow B \sin((m+1)\xi) = 0.$$

Naturalmente  $B \neq 0$ , imponiamo  $\sin((m+1)\xi) = 0 \rightarrow (m+1)\xi = k\pi$ , quindi escludiamo  $(m+1)\xi = k\pi$ ,  $k \neq 0$ . Da cui abbiamo:  $\xi_k = \frac{k\pi}{m+1}$ . La componente  $j$ -esima del  $k$ -esimo autovettore:

$$r_j^{(k)} = B \sin(j\xi_k), \quad k = 1, \dots, m$$

$B$  arbitraria (gli autovettori sono definiti a meno di una cost. multiplativa).

$$\begin{aligned} r_{j+1} &= \sin((j+1)\xi_k) = \sin(j\xi_k) \cos(\xi_k) + \cos(j\xi_k) \sin(\xi_k), \\ r_{j-1} &= \sin((j-1)\xi_k) = \sin(j\xi_k) \cos(\xi_k) - \cos(j\xi_k) \sin(\xi_k), \end{aligned}$$

$$(r_{j+1} + r_{j-1}) = 2 \sin(j\xi_k) \cos(\xi_k)$$

$$\frac{2 \sin(j\xi_k) \cos(\xi_k) - 2 \sin(j\xi_k)}{h^2} = \lambda_k \sin(j\xi_k), \quad k = 1, \dots, m,$$

segue

$$\lambda_k = \frac{2(\cos(\xi_k) - 1)}{h^2} \quad k = 1, \dots, m,$$

$\cos(\alpha) \leq 1$ ,  $\forall \alpha$ , il piú piccolo autovalore in modulo è esattamente  $\lambda_1$ ,  $k = 1$ , ovvero quello il cui coseno è più vicino a 1:  $\xi_1 = \frac{\pi}{m+1}$ .

$$\lambda_1 = \frac{-2 \left(1 - \cos\left(\frac{\pi}{m+1}\right)\right)}{h^2} = \frac{-4 \sin^2\left(\frac{h\pi}{2}\right)}{h^2} = -4 \left(\frac{\pi h}{2}\right)^2 \cdot \frac{1}{h^2} + \mathcal{O}(h^2)$$

dove  $1 - \cos(\alpha) = 2 \sin(\alpha/2)$ ,  $h = 1/(m+1)$  e  $\sin\left(\frac{\pi h}{2}\right) \approx \frac{\pi h}{2}$ . Segue:

$$\min |\lambda_k(A_h)| = |\lambda_1| \approx \pi^2, \Rightarrow \rho(A_h^{-1}) \approx \frac{1}{\pi^2}.$$

Abbiamo quindi:

$$\|E\|_2 = \|U^e - U\|_2 \leq \|A_h^{-1}\|_2 \|\tau_h\|_2 \rightarrow 0, \quad h \rightarrow 0.$$

Caso norma infinito, in generale vale  $\|A\|_\infty \leq \sqrt{m} \|A\|_2 \approx \frac{\sqrt{m}}{\pi^2}$ , con  $A$  di ordine  $m$ ,

$$\|E\|_\infty = \frac{1}{\sqrt{h}} \|U^e - U\|_2$$

tende l'errore come  $\mathcal{O}(h^{\frac{3}{2}})$ .

Cosa succede invece se al bordo abbiamo condizioni di Neumann? Esempio, destra Dirichlet e a sinistra Neumann

$$\begin{aligned} u_{xx} &= f, \\ u'(0) &= \sigma, \\ u(1) &= \beta. \end{aligned}$$

Esistono diverse tecniche per discretizzare questa equazione.

## Tecnica I

Calcolo la derivata con il rapporto incrementale (accuratezza del primo ordine  $\mathcal{O}(h)$ ):

$$\left\{ \begin{array}{l} \frac{u_1 - u_0}{h} = \sigma, \\ \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} = f_j, \quad j = 1, \dots, m-1 \\ \frac{u_{m-1} - 2u_m}{h^2} = f_m - \frac{\beta}{h^2}, \quad j = m \end{array} \right.$$

Il sistema diventa  $A_h U = F$ :

$$A_h = \frac{1}{h^2} \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \cdots & \cdots & 1 & -2 \end{pmatrix}, \quad F = \begin{pmatrix} \sigma \\ \frac{f(x_1)}{h} \\ \vdots \\ f(x_{m-1}) \\ f_m - \frac{\beta}{h^2} \end{pmatrix}.$$

Matrice simmetrica. Gli autovalori del Laplaciano con condizioni di Dirichlet sono tutti negativi (strettamente), quindi se la matrice è simmetrica, segue matrice definita negativa.

$A_h$  matrice a predominanza diagonale stretta, simmetrica e definita negativa (ultima linea della matrice vale il maggiore stretto)!, ed è una matrice irriducibile; (riducibile: se, esiste una matrice di cambiamento di base  $B$  tale che:  $B^{-1}AB$  matrice triangolare a blocchi).

## Tecnica II

Approssimiamo la derivata con una differenza centrale  $\frac{u_1 - u_{-1}}{2h} = \sigma$ , approssimazione  $\mathcal{O}(h^2)$ . In questo caso abbiamo aggiunto una incognita in più  $x_{-1}$ .

$$\left\{ \begin{array}{l} \frac{u_1 - u_{-1}}{2h} = \sigma, \\ \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} = f_j, \quad j = 0, \dots, m-1 \\ \frac{u_{m-1} - 2u_m}{h^2} = f_m - \frac{\beta}{h^2}, \quad j = m \end{array} \right.$$

$m+2$  equazioni ed  $m+2$  incognite. Ma l'incognita in più si può evitare se consideriamo:

$$\frac{u_{-1} - 2u_0 + u_1}{h^2} = f_0,$$

Sfruttiamo la C.N.  $u_{-1} = u_1 - 2h\sigma$  otteniamo:



$$\left\{ \begin{array}{l} \frac{u_1 - u_0}{h^2} = \frac{f_0}{2} + \frac{\sigma}{h} \\ \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} = f_j, \quad j = 1, \dots, m-1 \\ \frac{u_{m-1} - 2u_m}{h^2} = f_m - \frac{\beta}{h^2}, \quad j = m \end{array} \right.$$

## Tecnica III

Utilizziamo tre punti per valutare la derivata prima.

$$u_1 = u_0 + hu' + \frac{1}{2}h^2u'' + \frac{1}{6}h^3u''' + \dots$$

$$u_2 = u_0 + 2hu' + 2h^2u'' + \frac{8}{6}h^3u''' + \dots$$

moltiplichiamo la prima per 4 e facciamo la differenza, otteniamo:

$$\frac{2u_1 - \frac{1}{2}u_2 - \frac{3}{2}u_0}{h} = u' - \frac{2}{6}h^2u'''$$

Quindi:



$$\frac{2u_1 - \frac{1}{2}u_2 - \frac{3}{2}u_0}{h} = \sigma$$

$$A_h = \begin{pmatrix} -\frac{3}{2} & 2 & -\frac{1}{2} & \cdots & 0 \\ 1 & -2 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \cdots & \cdots & 1 & -2 \end{pmatrix},$$

La matrice non é piú tridiagonale.

C. di N. da entrambi i lati  $u'(0) = \sigma_0$ ,  $u'(1) = \sigma_1$ , (Tecnica II) abbiamo:

$$A_h = \frac{1}{h^2} \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \cdots & \cdots & 1 & -1 \end{pmatrix}, \quad F = \begin{pmatrix} \frac{1}{2}f(x_0) + \frac{\sigma_0}{h} \\ f_1 \\ \vdots \\ f_m \\ \frac{1}{2}f_{m+1} - \frac{\sigma_1}{h} \end{pmatrix}.$$

Notiamo che la somma degli elementi sulle righe é zero; se moltiplichiamo a sinistra o a destra, per un vettore di 1, fa zero, ovvero

$$A_h \mathbf{1} = 0$$

ovvero, autovettore costante  $\mathbf{1} = (1, \dots, 1)^T$  corrispondente ad autovalore nullo. Il determinante di  $A_h$  é nullo (Matrice singolare). Abbiamo:



$$\mathbf{1}^T A_h U = \mathbf{1}^T F, \Rightarrow 0 = \mathbf{1}^T F$$

Vale la condizione di compatibilitá: se  $\sum_j F_j = 0 \Rightarrow$  esiste soluzione di  $A_h U = F$ .

Se  $\bar{U}$  é soluzione  $\Rightarrow$  anche  $\bar{U} + \alpha \mathbf{1}$  é soluzione  $\forall \alpha$ .

Dal punto di vista continuo quando abbiamo un'equazione di Poisson con condizioni di N, se integriamo  $u''(x) = f$  in  $x$  abbiamo

$$\int_0^1 f(x) dx = u'(1) - u'(0), \Rightarrow \int_0^1 f(x) dx = \sigma_1 - \sigma_0,$$

detta *condizione di compatibilitá* (esistono  $\infty^1$  soluzioni, altrimenti non esistono soluzioni).

Notiamo che  $\sum_{j=1}^m F_j = 0, \Rightarrow \frac{1}{2}f_0 + \frac{1}{2}f_{m+1} + \sum_{j=1}^m f_j = \sigma_1 - \sigma_0 h$ , ovvero

$$h \left( \frac{1}{2}f_0 + \frac{1}{2}f_{m+1} + \sum_{j=1}^m f_j h \right) = \sigma_1 - \sigma_0$$

formula dei trapezi composita applicata a  $\int_0^1 f(x) dx$ . Versione del pr. di compatibilitá nel discreto analogo al continuo.



Generalizzazione. Sia

$$\begin{cases} a(x)u'' + b(x)u' + c(x)u = f(x) \\ u(0) = \alpha, \quad u(1) = \beta. \end{cases}$$

Dividiamo l'intervallo in  $m + 1$  punti,  $m$  intervalli. Utilizziamo discretizzazione centrale:

$$\begin{cases} a_j \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + b_j \frac{u_{j+1} - u_{j-1}}{2h} + c_j u_j = f_j, \\ a(x_j) \frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{h^2} + \dots \\ + b(x_j) \frac{u(x_{j+1}) - u(x_{j-1})}{2h} + c(x_j)u(x_j) = f(x_j) + \tau_j \end{cases}$$

con  $\tau_j = \mathcal{O}(h^2)$ .

Con  $f \in C^2[0, 1]$  soluzione  $u \in C^1[0, 1]$  e  $a(x)$   $b(x)$   $c(x) \in C^1[0, 1]$ .

Allora sia

$$A_h U = F_h.$$

La stabilità dipende dalle proprietà di questa matrice:

$$A_h = \begin{pmatrix} -\frac{2a_1}{h^2} + c_1 & \frac{a_1}{h^2} + \frac{b_1}{2h} & 0 & \dots & 0 \\ \frac{a_2}{h^2} - \frac{b_2}{2h} & -\frac{2a_2}{h^2} + c_2 & \frac{a_2}{h^2} + \frac{b_2}{h^2} & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots \\ 0 & \dots & \dots & \dots & \dots \end{pmatrix},$$

$$F = \begin{pmatrix} f_1 - \frac{a_1\alpha}{h^2} + \frac{b_1\alpha}{2h} \\ f_2 \\ \vdots \\ f_{m-1} \\ f_m - \frac{a_m\beta}{h^2} - \frac{b_m\beta}{2h} \end{pmatrix}.$$

La stabilitá si puó desumere da:

- ▶ Se  $a(x) > 0$ ,  $b = 0$ ,  $c(x) \leq 0$ : Matrice  $A_h$  é irriducibilmente diagonalmente dominante e tutti gli autovalori  $\lambda(A_h) < 0$  (matrice definita negativa).
- ▶ Se  $a(x) < 0$ ,  $b(x) = 0$ ,  $c(x) \geq 0$ :  $A_h$  definita positiva:  $A_h \geq A_h^0 \rightarrow (A_h)^{-1} \leq (A_h^0)^{-1}$ .
- ▶ Se  $b(x) \neq 0$ , La matrice  $A_h$  non é piú a predominanza diagonale (dipende da cosa fa  $c(x)$ ), se  $h \rightarrow 0$ ,  $a$  conta molto di piú di  $b$  e  $c$ . Se  $b(x)$  diventa molto grande, rispetto a  $a$  ci possono essere problemi di stabilitá.

Consideriamo adesso l'equazione:

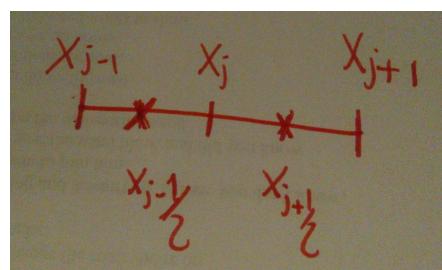
$$(k(u)u')' = f,$$

$$k(x)u'' + k'(x)u' = f$$

con  $a = k$ ,  $b = k'$ ,  $c = 0$  e  $k(x)$  condicibilitá termica positiva.

Discretizziamo la seconda equazione:  $A_h U = F_h$  con

$$(k(x)u')'|_{x_j} \approx \frac{(ku')|_{j+1/2} - (ku')|_{j-1/2}}{h} = \frac{k_{j+1/2} \left( \frac{u_{j+1} - u_j}{h} \right) - k_{j-1/2} \left( \frac{u_j - u_{j-1}}{h} \right)}{h}$$



di solito:  $k_{j+1/2} = \frac{k_j + k_{j+1}}{2}$ .

$$A_h U = F_h,$$

La matrice é data da:

$$A_h = \begin{pmatrix} -\left(k_{\frac{1}{2}} + k_{\frac{3}{2}}\right) & k_{\frac{3}{2}} & 0 & \cdots & 0 \\ k_{\frac{3}{2}} & -\left(k_{\frac{3}{2}} + k_{\frac{5}{2}}\right) & k_{\frac{5}{2}} & \cdots & 0 \\ 0 & k_{\frac{5}{2}} & -\left(k_{\frac{5}{2}} + k_{\frac{7}{2}}\right) & k_{\frac{7}{2}} & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots \end{pmatrix},$$

matrice simmetrica e la somma algebrica degli elementi sulle righe é uguale a zero. Se poi  $k$  é sempre maggiore di zero,  $A_h$  é irriducibilmente e diagonalmente dominante.  $A_h$  definita negativa.

L'operatore differenziale  $(k(x)u')'$  é autoaggiunto.

(Se  $A$  é operatore aggiunto  $(Au, v) = (u, A^T v)$ , Se  $A$  é autoaggiunto  $\Rightarrow A = A^T$  se  $(Au, v) = (u, Av)$ )

Sia:

$$\int_0^1 (k(x)u')' v dx,$$

$$\text{da } (k(x)u'v)' = (k(x)u')' v + k(x)u'v'.$$

$$\begin{aligned} \int_0^1 (k(x)u'v)' dx - \int_0^1 k(x)u'v' dx &= [k(x)u'v]_0^1 - \int_0^1 k(x)u'v' dx = \\ &= [k(x)u'v]_0^1 - [k(x)v'u]_0^1 + \int_0^1 (k(x)v')' u dx \end{aligned}$$

Se  $u$  e  $v$  al bordo valgono zero, l'operatore é autoaggiunto.

A livello discreto si traduce in una matrice simmetrica.

A livello continuo si ha anche

$$\int_0^1 (k(x)u')' dx = k(1)u'(1) - k(0)u'(0) = \int_0^1 f(x)dx$$

ottengo delle relazioni tra i flussi, cioè la somma dei flussi è uguale all'integrale di sorgente (proprietà conservativa).

A livello discreto

$$(k(x)u')'|_{x_j} \approx \frac{k_{j+1/2} \left( \frac{u_{j+1} - u_j}{h} \right) - k_{j-1/2} \left( \frac{u_j - u_{j-1}}{h} \right)}{h} = f_j$$

sommendo su  $j$  e dividendo per  $h$ :

$$k_{3/2} \frac{(u_2 - u_1)}{h} - k_{1/2} \frac{(u_2 - u_1)}{h} + k_{5/2} \frac{(u_3 - u_2)}{h} - k_{3/2} \frac{(u_2 - u_1)}{h} + \dots$$

$$\dots + k_{m+1/2} \frac{(u_{m+1} - u_m)}{h} = h \sum_j f_j$$

$$k_{m+1/2} \frac{(u_{m+1} - u_m)}{h} - k_{1/2} \frac{(u_2 - u_1)}{h} = h \sum_j f_j$$

Se discretizzo alcune proprietà qualitative vengono mantenute. Se  $k > 0$  con problema di Dirichlet ed  $f = 0$ , vale un principio del massimo per l'equazione:

$$(k(x)u')' = 0, \quad u(0) = \alpha, \quad u(1) = \beta$$

ovvero, il massimo della  $u$  si trova agli estremi. Anche questa proprietà viene garantita.

Alcune discretizzazioni permettono di preservare delle proprietà qualitative della soluzione analitica.

Se discretizziamo  $(k(x)u')' = f$  è conservativa, altrimenti discretizzare  $k(x)u'' + k(x)'u' = f$  è non-conservativa.

Consideriamo l'equazione di convezione-diffusione

$$u_t + bu_x = \mu u_{xx} + \phi$$

$\mu$  coefficiente di diffusione,  $b$ , (grandezza che ha la dimensione di una velocità) e sia  $[0, L]$  l'intervallo. Consideriamo soluzione stazionaria e dividiamo tutto per  $b$ , abbiamo

$$-\frac{1}{Pe} u_{xx} + u_x = \bar{\phi}$$

con  $\bar{\phi} = \phi/b$ , e

$$Pe = \frac{bL}{\mu}$$

$Pe$  é detto numero di Peclet, cioé (*termine convettivo/termine diffusivo*) ed  $L$  lunghezza.

Se  $Pe$  grande, la convezione domina, altrimenti domina la parte diffusiva. Matlab code  $[0, 1]$ ,  $k = \frac{1}{Pe}$ , e  $\bar{\phi} = 0$ , con condizioni al bordo di D.  $u(0) = \alpha$ ,  $u(1) = \beta$ . Se  $k = 0$ , la soluzione esatta é una retta di coefficiente angolare 1.

Se applico il metodo alle difference finite ottengo:

$$-k \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + \frac{u_{j+1} - u_{j-1}}{2h} = \bar{\phi}_j$$

c. contorno  $u_0 = \alpha$ ,  $u_{m+1} = \beta$ .

Questa é ancora un'equazione alle difference del II ordine e cerchiamo soluzioni dell'equazione omogenea associata:

$$-k \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + \frac{u_{j+1} - u_{j-1}}{2h} = 0,$$

che sono della forma  $u_j = z^j$  (dove  $z$  é della forma  $e^{\alpha}$ ):

$$-k \frac{z^{j+1} - 2z^j + z^{j-1}}{h^2} + \frac{z^{j+1} - z^{j-1}}{2h} = -\frac{k}{h} \frac{z^{j+1} - 2z^j + z^{j-1}}{h} + \frac{z^{j+1} - z^{j-1}}{2} = 0.$$

Raccogliendo i termini:

$$\left(-q + \frac{1}{2}\right)z^{j+1} + 2qz^j - \left(-q + \frac{1}{2}\right)z^{j-1} = 0,$$

con  $q = \frac{k}{q}$ , quindi:

$$\left(-q + \frac{1}{2}\right)z^2 + 2qz - \left(-q + \frac{1}{2}\right) = 0,$$

le cui soluzioni sono  $z_1 = 0$ ,  $z_2 = \frac{q\left(2 + \frac{1}{q}\right)}{q\left(2 - \frac{1}{q}\right)} = \frac{(p+2)}{(p-2)}$ ,

con  $p = \frac{1}{q} = \frac{h}{k} = hPe$ .

Soluzione generale dell'equazione differenziale:  $u_j = A + B\left(\frac{p+2}{p-2}\right)^j$ .

Se  $p \rightarrow 2$  allora  $\left(\frac{p+2}{p-2}\right) \rightarrow \infty$ .

Inoltre se  $h$  è sufficientemente piccolo il termine  $\left(\frac{p+2}{p-2}\right)$  è piccolo (relativamente prox a 1), invece se fisso  $h$  e diminuisco  $k$ , il rapporto puó avvicinarsi a 2.

Quindi se  $p \geq 2$ , la soluzione potrebbe oscillare ed é quello che succede nel codice Matlab.

Si forma un boundary layer nell'estremo di sinistra e la soluzione diventa oscillante ( $k \gg h$  e  $p > 2$ ).

Per evitare questo problema utilizziamo una discretizzazione di tipo *upwind*

$$\frac{\partial u}{\partial x}|_{x_i} \approx \begin{cases} \frac{u_j - u_{j-1}}{h} & b > 0, \\ \frac{u_{j+1} - u_j}{h} & b < 0, \end{cases} \quad (1)$$

$$-k \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + \frac{u_j - u_{j-1}}{h} = \bar{\phi}_j$$

Osserviamo che in questo caso la matrice é:

$$A_h = \begin{pmatrix} \frac{2k}{h^2} + \frac{1}{h} & -\frac{k}{h^2} & 0 & \cdots & 0 \\ -\frac{k}{h^2} - \frac{1}{h} & \frac{2k}{h^2} + \frac{1}{h} & -\frac{k}{h^2} & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots \end{pmatrix},$$

la discretizzazione upwind la matrice é irriducibile diagonalmente dominante, simmetrica e autovalori tutti strettamente maggiori di zero, quindi definita positiva.

Quindi il metodo upwind é molto piú stabile delle differenze centrali ma meno accurato!

## Metodi di ordine elevato

Consideriamo approssimazioni di ordine alto per le derivate per avere metodi di alto ordine.

Ricaviamo, per esempio, una formula centrale di ordine 4, utilizzando l'approssimazione di secondo ordine per la derivata seconda.

$$I : \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} = u''(x_j) + \frac{1}{12}u'''(x_j)h^2 + \mathcal{O}(h^4),$$

raddoppiamo il passo

$$II : \frac{u_{j+2} - 2u_j + u_{j-2}}{(2h)^2} = u''(x_j) + \frac{1}{12}u'''(x_j)(2h)^2 + \mathcal{O}(h^4),$$

allora

$$\frac{4I - II}{3} = 3u''(x_j) + \mathcal{O}(h^4)$$

ovvero otteniamo un'approssimazione di ordine 4 per la derivata seconda:

$$u''(x_j) \approx \frac{u_{j+2} + 16u_{j+1} - 30u_j + 16u_{j-1} - u_{j-2}}{12h^2}$$

il metodo appena usato é un caso particolare del metodo di *Estrapolazione*. Qui si sono considerati 5 punti per lo stencil  $(x_{j-2}, x_{j-1}, x_j, x_{j+1}, x_{j+2})$ .

# Lezioni Analisi Numerica II modulo

Sebastiano Boscarino

Dipartimento di Matematica e Informatica  
Università di Catania, ITALY

March 24, 2020

## Equazione ellittica in due dimensioni

Eq. Ellittica a coefficienti costanti in 2D ha la forma:

$$a_1 u_{xx} + a_2 u_{xy} + a_3 u_{yy} + a_4 u_y + a_5 u_x + a_6 u = f,$$

$(x, y) \in \Omega \subset \mathbb{R}^2$ , e  $u(x, y) : \Omega \rightarrow \mathbb{R}$ .

Supponiamo che siano soddisfatte tutte le condizioni di regolarità.

Per determinare il carattere dell'equazione i coefficienti che contano sono quelli di ordine massimo.

Consideriamo la forma quadratica associata:  $a_1\xi^2 + a_2\xi\eta + a_3\eta^2$  (luogo geometrico), discriminante dell'equazione:  $a_2^2 - 4a_1a_3$

- ▶  $\geq 0$ : (segni opposti  $a_1$  e  $a_3$ ) CASO IPERBOLICO;
- ▶  $= 0$ : CASO PARABOLICO;
- ▶  $\leq 0$  (stessi segni  $a_1$  e  $a_3$ ) CASO ELLITTICO;

Caso  $\leq 0$  ELLITTICO, l'equazione più semplice è l'equazione di Laplace  $u_{xx} = 0$ , se c'è il termine not fallora eq. di Poisson.

Vogliamo risolvere in  $\Omega$  il seguente problema di Poisson in 2D:

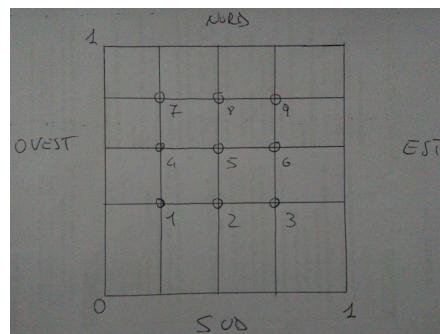
$$u_{xx} + u_{yy} = f,$$

con  $Bu = g$ , su  $\partial\Omega$  cond. al bordo e  $B$  un operatore del bordo. Come nel caso 1D possiamo assegnare c.c. di D. N. o Robin.

Su  $\Gamma = \partial\Omega$  si possono avere allora:

- ▶ C.d.D.  $u(x, y) = g_D(x, y)$ ,  $(x, y) \in \Gamma_D$ ;
- ▶ C.d.N.  $\frac{\partial u}{\partial n}(x, y) = g_N(x, y)$ ,  $(x, y) \in \Gamma_N$ ;
- ▶ C.d.R.  $\alpha_D u(x, y) + \alpha_N \frac{\partial u}{\partial n}(x, y) = g_R(x, y)$ ,  $(x, y) \in \Gamma_R$ ;

Sceglieremo al momento il prob. di Poisson, dominio semplicemente connesso (dominio un rettangolo  $[0, 1] \times [0, 1]$ ) e cond. di D.



Ho dato cond. di D. conosco i valori di  $f$  al bordo e le incognite sono nei punti interni  $x_i = ih$ ,  $y_j = jh$ ,  $i, j = 0, \dots, m+1$ , punti interni  $i, j = 1, \dots, m$ ,  $i, j = 0, m+1$  punti sul bordo.

Discretizzando la derivata seconda in due dimensioni è:

$$\frac{\partial^2 u}{\partial x^2}|_{i,j} = \frac{u(x_{i+1}, y_j) - 2u(x_i, y_j) + u(x_{i-1}, y_j)}{h^2} - \frac{h^2}{12} \frac{\partial^4 u}{\partial x^4}(\xi_i, y_j)$$

$$\frac{\partial^2 u}{\partial x^2}|_{i,j} \approx \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2}$$

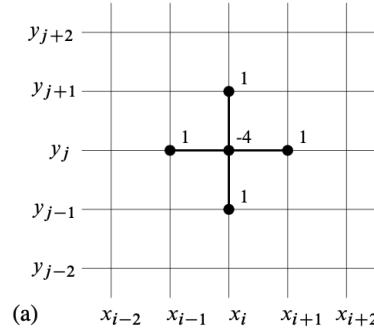
$h = x_{i+1} - x_i = y_{j+1} - y_j$ , quindi per l'equazione di P. ho la discretizzazione:

$$\frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2} + \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{h^2} = f_{i,j}, \quad i, j = 1, \dots, m$$

Riscriviamo il tutto come

$$\frac{1}{h^2}(u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1} - 4u_{ij}) = f_{i,j}, \quad i,j = 1, \dots, m$$

e questo schema alle differenze finite può essere rappresentato dallo stencil a 5 punti



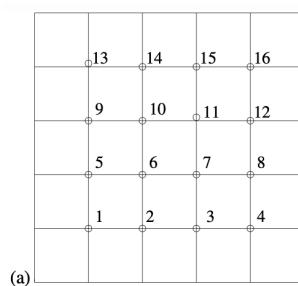
Quindi incognite  $u_{i,j}$  a ogni  $m_1 \cdot m_2 = m^2$  punti griglia, per  $i,j = 1, \dots, m$  con  $h = 1/(m+1)$ . avremo quindi un sistema lineare di  $m^2$  incognite.

Mettiamo tutto sotto forma matriciale e ordiniamo le incognite e equazioni come:

$$A_h U = f_h,$$

con  $A_h \in \mathbb{R}^{m^2 \times m^2}$  matrice sparsa (molti dei suoi elementi sono zero). A differenza del caso 1D, qui dobbiamo scegliere come ordinare le incognite, per esempio: Ordiniamo per colonna (come in figura). Lungo le colonne (dal basso):  $u_{11}, u_{21}, u_{31}, \dots, u_{m1}$ , seconda colonna:  $u_{12}, u_{22}, u_{32}, \dots, u_{m2}$  e così via. Il vettore allora sarà:

$$U = \begin{pmatrix} u^{[1]} \\ u^{[2]} \\ \vdots \\ u^{[m]} \end{pmatrix}, \quad u^{[j]} = \begin{pmatrix} u_{1j} \\ u_{2j} \\ \vdots \\ u_{mj} \end{pmatrix}, \quad j = 1, \dots, m$$



Questa scelta porta ad una matrice della forma:

$$A_h = \frac{1}{h^2} \begin{pmatrix} G & I & & & \\ I & G & I & & \\ & I & G & I & \\ & & \ddots & \ddots & \ddots \\ & & & I & G \end{pmatrix}, \quad (1)$$

che è una  $m \times m$  matrice tridiagonale a blocchi in cui ogni blocco  $G$  e  $I$  è esso stesso  $m \times m$  matrice:

$$G = \begin{pmatrix} -4 & 1 & & & \\ 1 & -4 & 1 & & \\ & 1 & -4 & 1 & \\ & & \ddots & \ddots & \ddots \\ & & & 1 & -4 \end{pmatrix}, \quad (2)$$

e  $I$  è la matrice identità. Chiaramente questa matrice ha una buona struttura, e gli elementi 1 nella matrice identità sono separati dalle diagonali da  $m - 1$  zeri, poiché questi coefficienti corrispondono ai punti griglia che giacciono sopra e sotto al punto centrale dello stencil e quindi sono nella precedente o successiva riga di incognite.

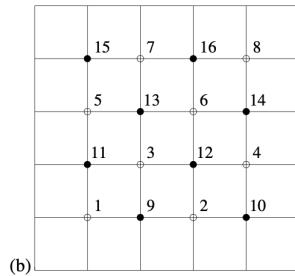
LA matrice (1) può essere definita in MATLAB dai comandi:

1. `I = eye(m);`
2. `e = ones(m,1);`
3. `G = spdiags([e, -4 * e, e], [-1, 0, 1], m, m);`
4. `S = spdiags([e, e], [-1, 1], m, m);`
5. `A = spdiags(kron(I,T) + kron(S,I))/h^2;`

`kron(X,Y)` is the Kronecker tensor product of  $X$  and  $Y$

$$\begin{bmatrix} X(1,1) * Y & X(1,2) * Y & X(1,3) * Y \\ X(2,1) * Y & X(2,2) * Y & X(2,3) * Y \end{bmatrix}$$

Un'altra alternativa che ha un vantaggio nel contesto di metodi iterativi é usare la tecnica *red-black ordering* (vedi figura nero-bianco):



Qui i quattro vicini del punto centrale (bianco-nero) sono punti (nero-bianchi), e viceversa con i colori. Questo conduce a un'equazione di matrice della forma:

$$\begin{pmatrix} D & H \\ H^T & D \end{pmatrix}, \begin{pmatrix} u_{\text{bianco-nero}} \\ u_{\text{nero-bianco}} \end{pmatrix} = \begin{pmatrix} f_{\text{bianco-nero}} \\ -f_{\text{nero-bianco}} \end{pmatrix}$$

dove  $D = -\frac{4}{h^2}I$  é matrice diagonale di dimensione  $m^2/2$  e  $H$  é una matrice a bande della stessa dimensione con quattro diagonali non zero.

Nota: Quando metodi diretti come metodo di eliminazione di Gauss, sono usato per risolvere sistemi lineari, uno tipicamente vuole ordinare le equazioni e le incognite per cercare di ridurre il problema del fill-in durante il processo di eliminazione. Questo é fatto automaticamente dal *back-slash* operatore di Matlab.

## ANALISI DI CONVERGENZA

Usiamo esattamente lo stesso approccio usato in 1D. L'errore locale di troncamento  $\tau_{ij}$  in  $(i,j)$  punto griglia definito come

$$\tau_{ij} = \frac{1}{h^2} (u(x_{i-1}, y_j) + u(x_{i+1}, y_j) + u(x_i, y_{j-1}) + u(x_i, y_{j+1}) - 4u(x_i, y_j)) = f(x_i, y_j),$$

dove in questo caso, *splittando* nelle direzioni  $x$  e  $y$

$$\tau_{ij} = \frac{1}{12} h^2 (\partial_x^4 u(\xi_i, y_j) + \partial_y^4 u(x_i, \eta_j)) + \mathcal{O}(h^4).$$

Abbiamo, facendo la differenza tra soluzione esatta e approssimata:

$$A_h E_h = -\tau_h,$$

con  $E_{ij} = u_{ij} - u(x_i, y_j)$ . Segue sempre:  $\|E_h\| \leq \|A_h^{-1}\| \|\tau_h\|$ . Quindi per studiare la convergenza del metodo dovremo maggiorare  $\|A_h^{-1}\|$  con un termine indipendente da  $h$ , cioé il metodo sia stabile. Nella norma-2 dobbiamo esplicitamente valutare il raggio spettrale della matrice, come fatto in 1D, gli autovalori e gli autovettori di  $A_h$  (ovvero:  $A_h U = \lambda U$ ) saranno qui indicizzati da due parametri:  $p$  e  $q$ . Il  $(p, q)$  autovettore

$$u_{ij}^{p,q} = \sin(p\pi ih) + \sin(q\pi jh), \quad p, q = 1, \dots, m, \quad i, j = 1, \dots, m$$

ha  $m^2$  elementi e il corrispondente autovalore ha la forma:

$$\lambda_{p,q} = \frac{2}{h^2}((\cos(p\pi h) - 1) + (\cos(q\pi h) - 1))$$

$h = 1/(m+1)$ ,  $\Omega = [0, 1]^2$ . Questi autovalori sono strettamente negativi ( $A$  é definita negativa). L'autovalore piú vicino all'origine é  $\lambda_{1,1}$  (l'autovalore piú lontano dall'origine é quello con  $p = q = m$ ). Quindi:

$$\lambda_{11} = \frac{4}{h^2}(\cos(\pi h) - 1)$$

$$\text{e da } (1 - \cos(x)) = 2(\sin(\frac{x}{2}))^2 \approx \frac{1}{2}x^2,$$

$$\lambda_{11} = \frac{4}{h^2} \left( -\frac{1}{2}\pi^2 h^2 + \mathcal{O}(h^4) \right) = -2\pi^2 + \mathcal{O}(h^2),$$

segue:

$$\rho(A_h^{-1}) = \frac{1}{\min_{1 \leq p,q \leq m} |\lambda_{pq}(A_h)|} = \frac{1}{\lambda_{11}} \approx \frac{1}{2\pi^2}$$

allora  $\|E_h\|_2 \leq \frac{1}{\pi^2} \|\tau_h\|_2$ , quindi il metodo é stabile e inoltre per  $h \rightarrow 0$  é consistente e quindi convergente.

Per quanto riguarda l'autovalore massimo:  $\lambda_{mm}$ , abbiamo:

$$\lambda_{11} = \frac{4}{h^2}(\cos(m\pi h) - 1)$$

$$\begin{aligned} \cos\left(\frac{m\pi}{m+1}\right) - 1 &= \cos\left(\frac{\pi}{m+1} + \frac{m\pi}{m+1} - \frac{\pi}{m+1}\right) - 1 = \dots \\ \cos(\pi - \pi h) - 1 &= -\cos(\pi h) - 1 = (1 - \cos(\pi h)) - 2, \end{aligned}$$

dove  $\cos(\pi - \alpha) = -\cos(\alpha)$ .

$$\lambda_{mm} = \frac{2}{h^2} \left( (-4) + \frac{1}{2}\pi^2 h^2 + \mathcal{O}(h^4) \right) = -\frac{8}{h^2} + \mathcal{O}(h^2)$$

e quindi otteniamo per il numero di condizionamento:

$$\mu(A) \approx \frac{4}{\pi^2 h^2} = \mathcal{O}\left(\frac{1}{h^2}\right).$$

Per  $h \rightarrow 0$ , il n.di cond. diventa grande, e quindi la matrice diventa malcondizionata, e questo é responsabile del rallentamento di metodi iterativi quando risolviamo il sistema lineare.

## Accuratezza di alto ordine

In 1D abbiamo:

$$u_{xx}|_{x_i} \approx \frac{1}{h^2} (-u_{i-2} + 16u_{i-1} - 30u_i + 16u_{i+1} - u_{i+2})$$

Questo tipo di discretizzazione coinvolge 5 punti nello stencil, e questo vuol dire che per le condizioni al contorno ho bisogno di due punti vicini  $x_{-1}, x_{-2}$  (punti fantasma), quindi ho due problemi:

- ▶ matrice meno sparsa;
- ▶ cond. al bordo più difficile da discretizzare;

Per superare queste difficoltà facciamo la seguente considerazione.

Per l'errore abbiamo:

$$\Delta_h u(x_i) = u_{xx}(x_i) + \frac{h^2}{12} \partial_x^4 u(x_i) + \mathcal{O}(h^4)$$

Se  $f$  regolare allora  $u_{xx} = f \Rightarrow u_{xxxx} = f_{xx}$  ovvero:

$$\Delta_h u(x_i) = f(x_i) + \frac{h^2}{12} f_{xx}(x_i) + \mathcal{O}(h^4)$$

Per cui non cambio l'operatore ma cambio l'equazione, questa mi dice che l'operatore Laplaciano  $\Delta_h$  ha un'accuratezza del quarto ordine.

Ragioniamo adesso nel caso 2D. Consideriamo l'equazione in 2D:

$$\Delta u = f, \quad \Delta = \partial_{xx} + \partial_{yy},$$

con

$$\Delta_h u = \Delta u + \frac{h^2}{12} (\partial_x^4 u + \partial_y^4 u) + \mathcal{O}(h^4).$$

Adesso andiamo a scegliere nuove variabili:

$$\xi = \frac{x+y}{\sqrt{2}}, \quad \eta = \frac{y-x}{\sqrt{2}},$$

segue:

$$\xi - \eta = \frac{2x}{\sqrt{2}} \Rightarrow y = \frac{\xi - \eta}{\sqrt{2}}, \quad \eta + \xi = \frac{2y}{\sqrt{2}} \Rightarrow x = \frac{\eta + \xi}{\sqrt{2}},$$

quindi segue:

$$\frac{\partial u}{\partial \xi} = \underbrace{\frac{\partial u}{\partial x} \frac{\partial x}{\partial \xi}}_{\frac{1}{\sqrt{2}}} + \underbrace{\frac{\partial u}{\partial y} \frac{\partial y}{\partial \xi}}_{\frac{1}{\sqrt{2}}} = \frac{u_x + u_y}{\sqrt{2}}$$

e:

$$\frac{\partial^2 u}{\partial \xi^2} = \frac{u_{xx} + u_{yy} + 2u_{xy}}{2}, \quad \frac{\partial^2 u}{\partial \eta^2} = \frac{u_{xx} + u_{yy} - 2u_{xy}}{2}$$

e

$$\frac{\partial^2 u}{\partial \xi^2} + \frac{\partial^2 u}{\partial \eta^2} = \frac{1}{2} (u_{xx} + u_{yy})$$

Quindi avrei due laplaciani discreti:  $\Delta_h$ , e  $\tilde{\Delta}_h$  (derivata rispetto ad assi inclinati di 45gradi): con

$$\tilde{\Delta}_h = \frac{1}{2h^2} (u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1} - 4u_{ij})$$

e queste sono due discretizzazioni consistenti del laplaciano.

Qualunque combinazione convessa:  $(\omega\Delta_h + (1 - \omega)\tilde{\Delta}_h) u$  si ha:

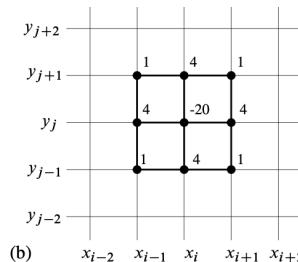
$$\Delta u = (\omega\Delta_h + (1 - \omega)\tilde{\Delta}_h) u + \frac{h^2}{12} (\omega(\partial_x^4 u + \partial_y^4 u) + (1 - \omega)(\partial_\xi^4 u + \partial_\eta^4 u)) + \mathcal{O}(h^4)$$

Opportune scelte di  $\omega$ , per esempio  $\omega = \frac{2}{3}$  abbiamo un Laplaciano a 9 punti abbiamo:

$$\nabla_9^2 u_{ij} = \frac{1}{6} (4\Delta_h + 2\tilde{\Delta}_h),$$

esplicitamente:

$$\begin{aligned} \nabla_9^2 u_{ij} = \frac{1}{6h^2} & (4u_{i-1,j} + 4u_{i+1,j} + 4u_{i,j-1} + 4u_{i,j+1} \\ & + u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1} - 20u_{ij}) \end{aligned}$$



Se noi applichiamo  $\nabla_9^2$  alla soluzione vera  $u(x_i, y_j)$  e espandiamo in serie di Taylor noi otteniamo :

$$\nabla_9^2 u(x_i, y_j) = \Delta u + \frac{1}{12} h^2 (u_{xxxx} + u_{yyyy} + 2u_{xxyy}) + \mathcal{O}(h^4)$$

A prima vista questa discretizzazione  $\nabla_9^2$  non sembra tanto meglio di quella a 5-punti perché l'errore è dell'ordine sempre di  $\mathcal{O}(h^2)$ .

Adesso ritornando all'equazione in 2D e cercando di fare un ragionamento analogo a 1D per aumentare l'ordine partiamo da:

$$\Delta u = f, \quad \Delta = \partial_{xx} + \partial_{yy},$$

e ricordiamo che:

$$\Delta_h u = \Delta u + \frac{h^2}{12} (\partial_x^4 u + \partial_y^4 u) + \mathcal{O}(h^4).$$

e non sappiamo come legare queste le derivate ( $\partial_x^4 u + \partial_y^4 u$ ) con le derivate della  $f$ .



Quindi per fare un ragionamento analogo al precedente in 2D valutiamo:

$$\Delta\Delta u = (u_{xx} + u_{yy})_{xx} + (u_{xx} + u_{yy})_{yy} = (u_{xxxx} + u_{yyyy}) + 2u_{xxyy}$$

ovvero assomigliante alle derivate dell'errore della discretizzazione  $\nabla_9^2$ .

Quindi se noi risolviamo  $\Delta u = f$  allora segue applicando di nuovo  $\Delta$  come sopra:

$$u_{xxxx} + u_{yyyy} + 2u_{xxyy} = \Delta f,$$

quindi noi possiamo valutare il termine dominante nell'errore di troncamento ( $u_{xxxx} + u_{yyyy} + 2u_{xxyy}$ ) con  $\Delta f$ , senza conoscere la funzione  $u$  del problema ovvero abbiamo:

$$\nabla_9^2 u(x_i, y_i) = f(x_i, y_i) + \frac{h^2}{12} \Delta f + \mathcal{O}(h^4).$$

in particolare se  $f = 0$  questo temine  $f(x_i, y_i) + \frac{h^2}{12} \Delta f$  svanisce e ottengo un accuratezza esatta del 4 ordine.

# Lezioni Analisi Numerica II modulo

Sebastiano Boscarino

Dipartimento di Matematica e Informatica  
Università di Catania, ITALY

March 31, 2020

## EQUAZIONE DEL CALORE

$$\rho c_v \frac{\partial u}{\partial t} = \frac{\partial}{\partial x} \left( k \frac{\partial u}{\partial x} \right)$$

con  $\mu = 1/\rho c_v$  e  $k$  coefficienti che potrebbero dipendere da  $x$ .

Caso semplificato  $\mu > 0$  costante e  $k = 1$ , il dominio dove consideriamo è:  $\Omega_T = [a, b] \times [0, T]$ .

Problema a valori iniziali (IVP) con condizioni al contorno  $\Rightarrow$  Problema ben posto (esiste ed è unica la soluzione e dipende con continuità dai dati).

$u(x, 0) = u_0(x)$ ,  $x \in [a, b]$  e condizioni al contorno: C.D., C. N, C.R.

In generale, potrei pensare di risolvere l'equazione del calore su tutto  $\mathbb{R}$ .

Se  $u_0 \in C^0(\mathbb{R})$ ,  $u_0$  limitata, la soluzione del problema

$$\frac{\partial u}{\partial t} = \mu \frac{\partial^2 u}{\partial x^2}, \quad u(x, 0) = u_0, \quad \mu > 0 \quad (1)$$

la soluzione del problema  $u \in C^\infty(\mathbb{R})$ , cioè l'equazione del calore ha un'effetto regolarizzante sulla soluzione. Anche se scelgo la  $u_0$  non regolare, a tratti, o con un numero finito di discontinuità, dopo un po' di tempo la soluzione si regolarizza. Con costante  $\mu$  negativa, problema mal posto.

Tecnica analitica: 1. Sep. Variabili, 2. Funzione di Green, 3. Trasf. di Fourier.

Sia dato il problema (1) con  $u_0$  funzione periodica di periodo  $L$ , ( $u(x + L, t) = u(x, t)$ ) e sviluppiamo in serie di Fourier (dato che il periodo è sviluppabile in serie di Fourier)

$$u_0(x) = \sum_{-\infty}^{\infty} c_j e^{i2\pi j \frac{x}{L}}$$

Se  $u_0$  è reale i coefficienti  $c_j = \bar{c}_j$ , avrò funzioni trigonometriche. L'equazione (1) ammette soluzione *singolo moto di Fourier* che è del tipo:

$$u(x, t) = c(t) e^{ikx},$$

dove  $k_j = 2\pi j/L$ . Inserisvo questa soluzione nell'equazione del calore:

$$c'(t) e^{ikx} = -\mu c k^2 e^{ikx}, \Rightarrow c'(t) = -\mu c k^2.$$

$c(t) = c(0) e^{-\mu k^2 t}$ , segue che la soluzione dell'equazione de calore (soluzione dipendente da  $k$ ),

$$u_k(x, t) = c_k e^{-\mu k^2 t} e^{ikx}$$

una sua qualunque c. lineare lo è ancora.



Se siamo nel caso periodico:

$$e^{ikx} = e^{ik(x+L)}, \quad e^{ikL} = 1, \quad kL = 2\pi j,$$

$j$  multiplo intero cioè  $k_j = 2\pi j/L$ .

Quindi la soluzione generale è:

$$u(x, t) = \sum_{-\infty}^{\infty} c_j e^{-\mu k_j^2 t} e^{ik_j x}$$

$t = 0$  riottengo  $u_0(x)$ . Esistono delle tecniche per calcolare i coefficienti  $c_j$ , per esempio la tecnica FFT.

Discretizziamo il mio problema con il *Metodo delle linee*: discretizzo prima spazialmente (approssimo la derivata seconda)

$$\frac{\partial^2 u}{\partial x^2} \approx \frac{u(x_{i+1}, t) - 2u(x_i, t) + u(x_{i-1}, t)}{h^2} + \frac{1}{12} h^2 \partial_x^4 u(\xi_i, t)$$

con questa discretizzazione passo da un sistema di PDEs a uno di ODEs.

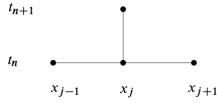
$$\frac{du_i}{dt} = \mu \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2}, \quad u_0(t) = g_a(t), \quad u_{m+1}(t) = g_b(t), \quad u_i(0) = u_0(x_i),$$

$$i = 1, \dots, m$$



Applichiamo per la discretizzazione in tempo Eulero esplicito.

$$u_j^n \approx u(x_j, t^n)$$



con  $t^n = n\Delta t$ . Discretizzando in spazio e tempo (*fully discrete*).

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = \mu \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2}$$

Definiamo il seguente operatore differenziale  $L$  tale che:

$$Lu = 0, \quad L = \frac{\partial}{\partial t} - \mu \frac{\partial^2}{\partial x^2}$$

mentre l'operatore discreto è:

$$L_{\Delta t, \Delta x} u(x, t) = \frac{u(x, t + \Delta t) - u(x, t)}{\Delta t} - \mu \frac{u(x + \Delta x, t) - 2u(x, t) + u(x - \Delta x, t)}{\Delta x^2},$$

ed utilizziamo questo operatore quando vogliamo studiare: consistenza, stabilità e convergenza di un metodo.

## Consistenza

Il nostro metodo è consistente se l'errore locale di discretizzazione  $\tau_{\Delta t, \Delta x}$  tende a zero per  $\Delta t, \Delta x \rightarrow 0$ , with

$$L_{\Delta t, \Delta x} u(x, t) = Lu(x, t) + \tau_{\Delta t, \Delta x},$$

Assumiamo  $u(x, t)$  funzione regolare, e considerando lo sviluppo di Taylor delle seguenti quantità  $u(x, t + \Delta t)$  rispetto a  $t$  e  $u(x + \Delta x, t), u(x - \Delta x, t)$  rispetto a  $x$  abbiamo

$$u(x, t + \Delta t) = u(x, t) + u_t \Delta t + \frac{\Delta t^2}{2} u_{tt} + \frac{\Delta t^3}{6} u_{ttt} + \dots,$$

e

$$u(x + \Delta x, t) = u(x, t) + u_x \Delta x + \frac{\Delta x^2}{2} 2u_{tt} + \frac{\Delta x^3}{6} u_{xxx} + \frac{\Delta x^4}{24} u_{xxxx} \dots,$$

otteniamo sostituendo tutte le quantità:

$$\tau_{\Delta t, \Delta x} = \left( u_t + \frac{\Delta t}{2} u_{tt} + \frac{\Delta t^2}{6} u_{ttt} + \dots \right) - \mu \left( u_{xx} + \frac{\Delta x^2}{12} u_{xxxx} + \dots \right)$$

Poiché  $u_t = \mu u_{xx}$  e  $u_{tt} = \mu u_{txx} = \mu^2 u_{xxxx}$  abbiamo:

$$\tau_{\Delta t, \Delta x} = \frac{\mu}{2} \left( \Delta t - \frac{1}{6} \Delta x^2 \right) u_{xxxx} + \mathcal{O}(\Delta t^2, \Delta x^4)$$

cioé, il metodo ha un ordine di consistenza  $\mathcal{O}(\Delta t, \Delta x^2)$ , ovvero del secondo ordine nello spazio e primo ordine di accuratezza nel tempo.

Se scelgo un passo  $\Delta t$  e  $\Delta x$  tale che:

$$\mu \Delta t = \frac{\Delta x^2}{6},$$

allora

$$\tau_{\Delta t, \Delta x} = \mathcal{O}(\Delta t^2, \Delta x^4).$$

Dal Teorema di Lax sappiamo che consistenza + stabilità implica convergenza.

## Stabilità

Quello che otteniamo è un sistema di ODEs della forma:

$$\frac{du}{dt} = Au, \quad u \in \mathbb{R}^{n \times n}, \quad A \in \mathbb{R}^{m \times m}.$$

Studiamo gli autovalori di  $A$ , i.e.,  $\lambda(A)$ . Supponiamo che  $\lambda \in \mathbb{C}$ ,  $\operatorname{Re}(\lambda) < 0$ . Caso scalare

$$\frac{du}{dt} = \lambda u,$$

E. E.  $z = \lambda \Delta t \in D := \{z \in \mathbb{C} : |z + 1| \leq 1\}$ .

La matrice  $A$ :

$$A = \begin{pmatrix} -2 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \cdots & \cdots & 1 & -2 \end{pmatrix}$$

Matrice con condizione di Dirichlet, gli autovalori abbiamo visto sono:

$$\lambda_p = -2 \frac{\mu}{\Delta x^2} (1 - \cos(p\pi\Delta x)) = -4 \frac{\mu}{\Delta x^2} (\sin^2(p\pi\Delta x)), \quad p = 1, 2, \dots, m.$$

con  $1 - \cos(\alpha) = 2 \sin^2(\alpha/2)$  e il più distante autovalore dall'origine è:  $\lambda_m \approx -4/\Delta x^2$ , quindi richiediamo che  $-4/\Delta x^2 \in D$ .

Se noi assumiamo il metodo di Eulero esplicito vogliamo che da:  $\lambda_p \Delta t = -4 \frac{\mu}{\Delta x^2} (\sin^2(p\pi\Delta x))$ ,  $p = 1, 2, \dots, m$ . richiediamo che:  $|1 + \lambda_p \Delta t| \leq 1$ , ovvero  $-2 \leq -4\Delta t/\Delta x^2 < 0$ , (Tutti gli autovalori stanno sull'asse negativa delle  $x$  in quanto hanno tutti parte immaginaria nulla!), quindi: abbiamo:

$$\frac{4\mu\Delta t}{\Delta x^2} \leq 2, \Rightarrow \mu\Delta t \leq \frac{1}{2}\Delta x^2,$$

ovvero condizione CFL di stabilità del metodo Eulero esplicito.

Un'altra maniera per ottenere questa condizione è passare dai modi di Fourier, la stabilità non dipende dalle condizioni al contorno ma dipende da come sono relazionati spazio e tempo.

Utilizziamo condizioni al contorno periodiche, così da utilizzare espansione di Fourier.

# Lezioni Analisi Numerica II modulo

Sebastiano Boscarino

Dipartimento di Matematica e Informatica  
Università di Catania, ITALY

April 10, 2020

## Analisi di Von Neumann

L'analisi di Von Neumann per la stabilità del metodo è basato sull'analisi di Fourier e quindi limitato a PDEs lineari a coefficienti costanti. Inoltre, l'analisi di Von Neumann è usata per studiare la stabilità di problemi con condizioni al contorno periodiche (periodic boundary conditions).

Dato il metodo di Eulero esplicito per l'equazione del calore:

$$u_j^{n+1} = u_j^n + \mu \frac{\Delta t}{\Delta x^2} (u_{j+1}^n - 2u_j^n + u_{j-1}^n)$$

l'idea è di cercare una soluzione della forma:

$$u_j^n = \rho^n e^{ij\xi}$$

con  $\xi$  parametro di Fourier e  $\rho$  fattore di amplificazione.

La ragione di base è che queste funzioni  $e^{ij\xi}$ , con numero d'onda  $\xi = \text{costante}$ , sono autofunzioni dell'operatore differenziale  $\partial_x$ :

$$\partial_x e^{ij\xi} = i\xi e^{ij\xi},$$

e quindi per ogni operatore differenziale a coefficiente costante.

Qui  $i = \sqrt{-1}$  e  $j$  è il punto griglia.

Sostituendo quindi abbiamo:

$$\rho^{n+1} e^{ij\xi} = \rho^n e^{ij\xi} + \mu \frac{\Delta t}{\Delta x^2} \left( \rho^n e^{i(j+1)\xi} - 2\rho^n e^{ij\xi} + \rho^n e^{i(j-1)\xi} \right),$$

ovvero:

$$\rho^{n+1} e^{ij\xi} = \rho^n e^{ij\xi} + \mu \frac{\Delta t}{\Delta x^2} (e^{i\xi} - 2 + e^{-i\xi}) \rho^n e^{ij\xi},$$

da  $\cos(\xi) = \frac{e^{i\xi} + e^{-i\xi}}{2}$  abbiamo:

$$\rho = 1 + \mu \frac{2\Delta t}{\Delta x^2} (\cos(\xi) - 1),$$

in generale  $\rho(\xi)$  potrebbe essere anche un numero complesso.

Il modulo di  $\rho$ ,  $|\rho|$ , mi dice se il moto cresce o decresce, adesso poiché tutti i moti di Fourier decrescono, tranne quello costante, allora si vuole che anche tutti i moti di F. della sol. num. decrescano, tranne quello costante!

Cioé imponiamo che:

$$|\rho| \leq 1, \Rightarrow -1 \leq \rho \leq 1,$$

Allora:

$$-1 \leq 1 + \mu \frac{2\Delta t}{\Delta x^2} (\cos(\xi) - 1) \leq 1$$

abbiamo

$$\mu \frac{2\Delta t}{\Delta x^2} (1 - \cos(\xi)) \leq 2$$

ovvero

$$\mu \frac{\Delta t}{\Delta x^2} \leq \frac{1}{1 - \cos(\xi)}$$

Poiché  $-1 \leq \cos(\xi) \leq 1$ , per ogni  $\xi$ , allora abbiamo:

$$\frac{\mu \Delta t}{\Delta x^2} \leq \frac{1}{2}$$

Caso peggiore quando  $\cos(\xi) = 1$ .

Notiamo che: se la PDEs non é lineare a coefficienti costanti allora l'analisi di Von Neumann non si puó utilizzare, e la tecnica é basta sull'analisi della matrice del sistema, ovvero sia:

$$U^{n+1} = BU^n,$$

ponendo  $a = \frac{\mu\Delta t}{\Delta x^2}$ ,

$$A = \begin{pmatrix} 1 - 2a & a & 0 & \cdots & 0 \\ a & 1 - 2a & a & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & a \\ 0 & \cdots & \cdots & a & 1 - 2a \end{pmatrix}$$

Il metodo é stabile quando tutti gli autovalori di  $B$  sono  $\leq 1$ . Qui cond. nec. e suff. affinché il metodo sia stabile se  $\lambda(B) \leq 1$ .

Usando il metodo di Eulero abbiamo ottenuto un metodo del primo ordine in tempo e secondo ordine in spazio, ma globalmente primo ordine. Chiaramente se applichiamo metodi RK del secondo ordine quello che ci aspettiamo sono schemi del secondo ordine in tempo  $\Delta t^2$  e in spazio  $\Delta x^2$ , globalmente secondo ordine.

Ma chiaramente, per la stabilitá, avremo sempre delle restrizioni nel passo temporale molto piú piccolo di quello spaziale, tipo caso problema stiff (quando aumentiamo il numero di punti griglia). quindi abbiamo bisogno di metodi impliciti che di solito sono incondizionatamente stabili.

Il piú semplice metodo del primo ordine implicito é il metodo di Eulero implicito. Otteniamo

$$u_j^{n+1} = u_j^n + \mu \frac{\Delta t}{\Delta x^2} \left( u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1} \right)$$

Il metodo di Eulero implicito é incondizionatamente stabile. Questo metodo sarà accurato al primo ordine globale ( $\Delta t, \Delta x^2$ ).

Per il metodo al secondo ordine abbiamo il metodo di Crank-Nicolson (in ODE metodo dei trapezi). Nel campo delle ODEs il metodo dei trapezi é un metodo implicito A-stabile (incondizionatamente stabile).

$$u_j^{n+1} = u_j^n + \mu \frac{\Delta t}{\Delta x^2} \left( \frac{1}{2} \delta^2 u_j^n + \frac{1}{2} \delta^2 u_j^{n+1} \right),$$

con  $\delta^2 u_j^n = u_{j+1}^n - 2u_j^n + u_{j-1}^n$ . Proviamo che é un metodo del secondo ordine globale e incondizionatamente stabile.

Accuratezza. Indichiamo con:  $k = \Delta t$ ,  $h = \Delta x$ . Scriviamo l'operatore discreto di CN:

$$L_{hk}^{cn} u_j^n = \frac{u_j^{n+1} - u_j^n}{k} - \frac{\mu}{\Delta x^2} \left( \frac{1}{2} \delta^2 u_j^n + \frac{1}{2} \delta^2 u_j^{n+1} \right),$$

Applichiamo tale operatore al caso continuo:

$$L_{hk}^{cn} u(x, t) = \frac{u(x, t+k) - u(x, t)}{\Delta t} - \frac{\mu}{h^2} \left( \frac{1}{2} \delta^2 u(x, t) + \frac{1}{2} \delta^2 u(x, t+k) \right).$$

Sviluppiamo in serie di Taylor:

$$\frac{\delta^2 u(x, t)}{h^2} = \frac{\partial^2 u}{\partial x^2} + \frac{h^2}{12} \partial_x^4 u(\xi, t)$$

allora

$$\frac{\delta^2 u(x, t+k)}{h^2} = \frac{\partial^2 u(x, t+k)}{\partial x^2} + \frac{h^2}{12} \partial_x^4 u(\tilde{\xi}, t+k)$$

sostituendo:

$$L_{hk}^{cn} u(x, t) = u_t + \frac{k}{2} u_{tt} + \frac{k^2}{6} u_{ttt} - \frac{\mu}{2} (u_{xx} + u_{xx} + k u_{txx}) + \dots,$$

abbiamo:

$$L_{hk}^{cn} u(x, t) = u_t - \mu u_{xx} + \frac{k}{2} u_{tt} - \frac{\mu k}{2} k u_{txx} + \mathcal{O}(k^2, h^2),$$

allora se la  $u$  soddisfa l'equazione del calore segue:  $u_t - \mu u_{xx} = 0$  e

$$\frac{k}{2} \frac{\partial}{\partial t} (u_t - \mu u_{xx}) = 0,$$

per cui abbiamo  $L_{hk}^{cn} u(x, t) = \mathcal{O}(k^2, h^2)$ . Metodo del secondo ordine nello spazio e nel tempo.

## $\theta$ -Metodo

$$u_j^{n+1} = u_j^n + \frac{\mu \Delta t}{\Delta x^2} \left( (1-\theta) \delta^2 u_j^n + \theta \delta^2 u_j^{n+1} \right), \quad 0 \leq \theta \leq 1,$$

- ▶  $\theta = 0$  Eulero esplicito;
- ▶  $\theta = 1$  Eulero implicito;
- ▶  $\theta = 1/2$  CN;

Analisi di stabilità di Von Neumann, cerchiamo cioè:  $u_j^n = \rho^n e^{ij\xi}$ . Sostituendo e semplificando per  $\rho^n e^{ij\xi}$  abbiamo:

$$\rho = 1 + c(2(1-\theta)(\cos(\xi) - 1) + \theta 2\rho(\cos(\xi) - 1))$$

$$\text{con } c = \frac{\mu \Delta t}{\Delta x^2}.$$

$$\rho(1 + 2c\theta\rho(1 - \cos(\xi))) = 1 - 2c(1 - \theta)(1 - \cos(\xi))$$

posto  $x = 2c(1 - \cos(\xi)) \geq 0$ ,  $c \geq 0$ ,  $1 - \cos(\xi) \geq 0$ , si ha

$$\rho = \frac{1 - (1 - \theta)x}{1 + \theta x},$$

Se  $\theta \geq \frac{1}{2}$  ( $\theta \rightarrow 1$ ),  $\Rightarrow |\rho| < 1$ , incondizionatamente stabile.

Se  $\theta \leq \frac{1}{2}$ , proviamo  $\rho \geq -1$

$$\frac{1 - (1 - \theta)x}{1 + \theta x} \geq -1, \Rightarrow 2 \geq (1 - 2\theta)x$$

con  $(1 - 2\theta) > 0$ , quindi  $x \leq \frac{2}{1 - 2\theta}$  dove  $x = 2c(1 - \cos(\xi))$  (caso peggiore quando questo termine è grande),  $2c$ .

$$2c \leq \frac{2}{2(1 - 2\theta)}, \Rightarrow c \leq \frac{1}{2(1 - 2\theta)}$$

segue

$$\mu \frac{\Delta t}{\Delta x^2} \leq \frac{1}{2(1 - 2\theta)}$$

Per cui  $\theta < 1/2$ , restrizione temporale metodo del primo ordine come Eulero esplicito.

Caso generale

$$u_t = \mu(x) \frac{\partial}{\partial x} \left( K(x) \frac{\partial u}{\partial x} \right),$$

Osserviamo che: per calcolare un' approssimazione della derivata seconda applicando due volte la derivata prima:

$$\frac{\partial u}{\partial x} \Big|_{x_i} \approx \frac{u_{j+1} - u_{j-1}}{2h},$$

allora:

$$\approx \mu \frac{\partial}{\partial x} \left( \frac{u_{j+1} - u_{j-1}}{2h} \right) \approx \frac{\mu \left( \frac{u_{j+2} - u_j}{2h} - \frac{u_j - u_{j-2}}{2h} \right)}{2h} = \mu \frac{u_{j+2} - 2u_j + u_{j-2}}{4h^2},$$

Stencil con i punti  $x_{j-2}, x_j, x_{j+2}$ , approssimazione non buona!

Se faccio invece una composizione della derivata prima (approssimata al primo ordine) nei valori:  $u_{j+1}, u_j$  e  $u_j$  e  $u_{j-1}$ , abbiamo:

$$\mu \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2}. \quad (1)$$

Invece calcoliamo le differenze centrali nei punti  $x_{j+1}$  e  $x_{j-1}$ , prendiamo la derivata nei  $j - 1/2$  e  $j + 1/2$ , (centrati in  $x_j$ )

$$\frac{\partial u}{\partial x}|_{x_i} \approx \frac{u_{j+1/2} - u_{j-1/2}}{h},$$

componendo abbiamo esattamente: (1).

Tornando al caso generale abbiamo allora:

$$\frac{du_j}{dt} = \mu(x_i) \frac{1}{h} (f_{j+1/2} - f_{j-1/2})$$

con

$$f_{j+1/2} = K(x_{j+1/2}) \left( \frac{u_{j+1} - u_j}{h} \right),$$

allora:

$$\frac{du_j}{dt} = \mu(x_j) \frac{1}{h^2} (K(x_{j+1/2})(u_{j+1} - u_j) - K(x_{j-1/2})(u_j - u_{j-1})),$$

$$\frac{du_j}{dt} = \frac{\mu(x_j)}{h^2} (K(x_{j+1/2})u_{j+1} - u_j) - (K(x_{j+1/2}) + K(x_{j-1/2})) + K(x_{j-1/2})u_{j-1})$$

Matrice con C.d.D. matrice simmetrica def. negativa:

$$A = \begin{pmatrix} -\left(K_{\frac{1}{2}} + K_{\frac{3}{2}}\right) & K_{\frac{3}{2}} & 0 & \cdots & 0 \\ K_{\frac{3}{2}} & -\left(K_{\frac{3}{2}} + K_{\frac{5}{2}}\right) & K_{\frac{5}{2}} & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \vdots & \vdots \end{pmatrix}$$

Per dimostrare che la matrice é definita negativa:

$$u \in \mathbb{R}^m : u^T A u < 0, \forall u.$$

Nel caso continuo abbiamo:

$$u_t = \mu \frac{\partial^2 u}{\partial x^2}, \Rightarrow uu_t = \mu \frac{\partial^2 u}{\partial x^2} u \Rightarrow \int_a^b u \frac{\partial^2 u}{\partial x^2} dx = \mu \int_a^b u \frac{\partial^2 u}{\partial x^2} u dx$$

ovvero

$$\frac{\partial}{\partial t} \int_a^b \frac{1}{2} u^2 dx = -\mu \int_a^b (u_x)^2 dx,$$

con  $(u_x u)_x = u_{xx} u + u_x^2$  segue  $\mu \int_a^b (u_x u)_x dx = \mu \int_a^b (u_x u)_x dx - \mu \int_a^b u_x^2 dx$

ovvero, da un problema del tipo:

$$u_t = \mu u_{xx}, \quad u(x, 0) = u_0, \quad u(a) = u(b) = 0,$$

segue

$$\frac{\partial}{\partial t} \int_a^b \frac{1}{2} u^2 dx = -\mu \int_a^b (u_x)^2 dx,$$

ovvero l'integrale al secondo membro è negativo, ovvero un altro modo per dire che l'operatore differenziale è definito negativo.

Caso metodo di Eulero Implicito abbiamo:

$$U^{n+1} = (I - \Delta t A_h)^{-1} U^n$$

con

$$A = \begin{pmatrix} -(c_1 + c_2) & c_2 & 0 & \cdots & 0 \\ c_2 & -(c_2 + c_3) & c_3 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \vdots & \vdots \end{pmatrix}$$

Pero il teorema di Gershgorin allora i centri di  $(I - \Delta t A_h)$  sono:  $1 + \frac{\Delta t}{h^2} (c_j + c_{j+1})$  e raggio:  $(c_j + c_{j+1}) \frac{\Delta t}{h^2}$ . Gli autovalori sono compresi nell'unione di questi cerchi, ma dato che abbiamo a che fare con autovalori reali allora stanno nell'intervallo:  $[1, 1 + 2 \frac{(c_j + c_{j+1}) \Delta t}{h^2}]$ . Se considero il  $R = \max_j (c_j + c_{j+1}) \frac{\Delta t}{h^2}$  allora

$$1 \leq \lambda \leq 1 + 2R,$$

ma a noi ci servono gli inversi  $\lambda((I - \Delta t A_h)^{-1}) \leq 1$ .

Per il secondo ordine globale si utilizza il metodo di CN.

Quindi per la stabilità si ha:

$$U^{n+1} = U^n + \frac{\Delta t}{2} (A_h U^n + A_h U^{n+1})$$

ovvero:

$$\left( I - \frac{\Delta t}{2} A_h \right)^{-1} U^{n+1} = \left( I + \frac{\Delta t}{2} A_h \right) U^n$$

Cerco un oggetto del tipo:  $U^n = \rho^n U$ , con  $U$  autovettore di  $A_h$ :  $A_h U = \lambda U$   
sostituendo

$$\left( I - \frac{\Delta t}{2} A_h \right) \rho^{n+1} U = \left( I + \frac{\Delta t}{2} A_h \right) \rho^n U$$

$$\left( I - \frac{\Delta t}{2} A_h \right) \rho U = \left( I + \frac{\Delta t}{2} A_h \right) U$$

da  $A_h U = \lambda U$  segue

$$\rho \left( 1 - \frac{\Delta t}{2} \lambda \right) U = \left( 1 + \frac{\Delta t}{2} \lambda \right) U$$

$$\rho = \frac{1 + \frac{\Delta t}{2} \lambda}{1 - \frac{\Delta t}{2} \lambda}$$

Dato che i  $\lambda$  di  $A_h$  sono negativi (matrice def. negativa),  $|\rho| \leq 1$ .

Il teorema di Lax per la convergenza (consistenza + stabilità) non funziona per problemi non lineari come:

$$u_t = \mu(x) \partial_x (K(u) \partial_x u),$$

con  $K(u) \geq 0$ . Problema della stabilità semplificato con  $K$  fuori dalla derivata e applico E.E.:

$$U^{n+1} = U^n + \Delta t K(U) \nabla^2 U^n$$

qui la matrice  $\mathcal{B}$  dipende dal tempo  $U(t)$ , allora  $U^{n+1} = \mathcal{B} U^n$  con  $\mathcal{B} = I - \Delta t K(U^n) A_h$ ,  $K(U^n)$  matrice diagonale:

$$A = \begin{pmatrix} 1 - 2\Delta t \frac{K(u_1^n)}{h^2} & \Delta t \frac{K(u_1^n)}{h^2} & 0 & \cdots & 0 \\ \Delta t \frac{K(u_2^n)}{h^2} & 1 - 2\Delta t \frac{K(u_2^n)}{h^2} & \Delta t \frac{K(u_1^n)}{h^2} & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \vdots & \vdots \end{pmatrix}$$

Quindi le condizioni di stabilità richiedono in questo caso è

$$\max_j \frac{\Delta t}{h^2} K(u_j^n) \leq \frac{1}{2}$$

ovvero  $\Delta t$  dipende da  $n$ , cioè nel passo da  $n$  al passo  $n+1$  devo garantire questa condizione di stabilità *locale*. Ovvero che non ci siano amplificazioni spurious.

Caso metodo implicito.

Il piú semplice E.I.

$$U^{n+1} = U^n + \Delta t K(U^{n+1}) A_h U^{n+1},$$

incognita qui compare anche nella funzione  $K$ , cioè porta ad un sistema di equazioni non lineari, e si può risolvere con il metodo di Newton,

$$F(U) = U - U^n - \Delta t K(U^{n+1}) A_h U, \quad J(U^{(k)}) \Delta U = F(U^{(k)}), \text{ sistema lineare}$$

con jacobiano  $J_{ij}(U) = \frac{\partial F_i}{\partial U_j}$  invertibile, quindi:

$$U^{(k+1)} = U^{(k)} + \Delta U, \quad \frac{\|\Delta U\|}{\|U\|} < TOL.$$

In generale se voglio discretizzare nello spazio e nel tempo problemi come:

$$u_t = \mu(x) \partial_x (K(u) \partial_x u),$$

posso capire quali dei termini è responsabile della mancanza della stabilità, in questo caso il termine responsabile (come caso E.E.) è il termine  $\partial_x^2 u$  in quanto nella discretizzazione contiene  $1/h^2$  e se  $h \rightarrow 0$  (oppure  $N \rightarrow \infty$ ) il problema diventa stiff.

Ma solo impliciti abbiamo eccessivo spreco di calcolo computazionale, (full implicit).

Quindi esistono dei metodi, che trattano i termini "difficili" (nel senso che dominano il prob. della stabilità), ovvero nel nostro caso allora possiamo pensare di trattare  $K(U)$  esplicitamente, e  $\partial_x^2 u$  implicitamente.

Caso piú semplice E.I.+E.E:

$$U^{n+1} = U^n + \Delta t K(U^n) A_h U^{n+1}$$

si può provare che questo metodi è consistente e incondizionatamente stabile. Metodi semi-impliciti.

# Lezioni Analisi Numerica II modulo

Sebastiano Boscarino

Dipartimento di Matematica e Informatica  
Università di Catania, ITALY

March 21, 2021



## Equazione del calore in 2D

In 2D l'equazione del calore é data da (caso semplificato):

$$u_t = \mu(u_{xx} + u_{yy})$$

in  $\Omega \times [0, T]$ , con  $\Omega = [a, b] \times [c, d]$  dominio rettangolare, con condizione iniziale  $u(x, y, 0) = \eta(x, y)$  e condizioni al contorno  $Bu = g$  (tipo D. oppure N.)

Sia  $u_{ij}^n \approx u(x_i, y_j, t^n)$ , con  $x_i = a + \Delta i$ ,  $y_j = c + \Delta j$  e  $t^n = t^0 + n\Delta t$ , Possiamo discretizzare lo spazio usando il Laplaciano discreto, per esempio (visto lezioni precedenti):

$$\delta_h^2 u_{ij} = \mu \left( \frac{1}{\Delta x^2} \delta_x^2 u_{ij} + \frac{1}{\Delta y^2} \delta_y^2 u_{ij} \right)$$

con  $\delta_x^2 u_{ij} = u_{i-1,j} + u_{i+1,j} - 2u_{i,j}$ , e  $\delta_y^2 u_{ij} = u_{i,j-1} + u_{i,j+1} - 2u_{i,j}$ .

1) Caso Eulero esplicito ho sempre un metodo globalmente al primo ordine. Per quanto riguarda la stabilitá anche nel caso 2D ci aspettiamo una restrizione nel passo temporale. Analisi di VonNeumann:

$$u_{jk}^n = \rho^n e^{ij\xi} e^{ik\eta},$$

e sostituendo nell'equazione di Eulero:  $u_{jk}^{n+1} = u_{jk}^n + \mu \Delta t \left( \frac{1}{\Delta x^2} \delta_x^2 u_{jk}^n + \frac{1}{\Delta y^2} \delta_y^2 u_{jk}^n \right)$   
abbiamo:



$$\rho = 1 + \mu \Delta t \left( \frac{2}{\Delta x^2} (\cos(\xi) - 1) + \frac{2}{\Delta y^2} (\cos(\eta) - 1) \right),$$

Vogliamo che  $|\rho| \leq 1$ , otteniamo come caso 1D,

$$2\mu \Delta t \left( \frac{1}{\Delta x^2} + \frac{1}{\Delta y^2} \right) \leq 1,$$

Se  $\Delta x = \Delta y$  allora  $4\mu \frac{\Delta t}{\Delta x^2} \leq 1$ .

2) Caso C.N. Consistenza del secondo ordine, analoga al caso 1D.

$$u_{jk}^{n+1} = u_{jk}^n + \frac{\mu \Delta t}{2} \left( \left( \frac{1}{\Delta x^2} \delta_x^2 u_{jk}^n + \frac{1}{\Delta y^2} \delta_y^2 u_{jk}^n \right) + \left( \frac{1}{\Delta x^2} \delta_x^2 u_{jk}^{n+1} + \frac{1}{\Delta y^2} \delta_y^2 u_{jk}^{n+1} \right) \right).$$

Per la stabilità

$$\rho = 1 - \underbrace{\frac{\mu \Delta t}{2} \left( \frac{2}{\Delta x^2} (1 - \cos(\xi)) + \frac{2}{\Delta y^2} (1 - \cos(\eta)) \right)}_Q (1 + \rho),$$

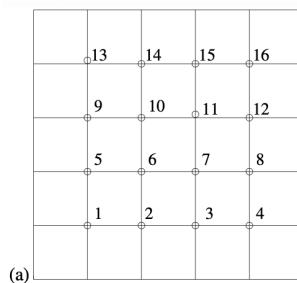
$$\rho = \frac{1 - \mu \Delta t Q}{1 + \mu \Delta t Q}.$$

Posto,  $h = \Delta x = \Delta y$ , sia:

$$L_h u_{ij} = \frac{1}{h^2} (\delta_x^2 u_{ij} + \delta_y^2 u_{ij}),$$

Supponiamo che abbiamo cond. di Dirichlet, quindi se abbiamo termini di bordo, dato che li conosciamo, li mettiamo come termine noti.

Se  $U$  rappresenta la soluzione nei punti interni:



$$U^{n+1} = U^n = \mu \frac{\Delta t}{2} (L_h U^n + L_h U^{n+1})$$

con  $\Omega_h$  griglia di  $\mathbb{R}^m \times \mathbb{R}^m$ , e  $\Gamma_h$  solo punti di frontiera quindi:

$L_h : \Omega_h \cup \Gamma_h \rightarrow \Omega_h$  ovvero da  $\mathbb{R}^{(m+2) \times (m+2)}$  a  $\mathbb{R}^{m \times m}$ . Siano  $U_\Gamma$  punti sulla frontiera, che conosco.

$$U^{n+1} = U^n + \frac{\mu \Delta t}{2} (\tilde{L}_h U^n + \tilde{L}_h U^{n+1}) + \frac{\mu \Delta t}{2} (B U_\Gamma^n + B U_\Gamma^{n+1})$$

quindi:

$$\left( I - \frac{\mu \Delta t}{2} \tilde{L}_h \right) U^{n+1} = \left( I + \frac{\mu \Delta t}{2} \tilde{L}_h \right) U^n + \frac{\mu \Delta t}{2} (B U_\Gamma^n + B U_\Gamma^{n+1})$$

la matrice  $\left( I - \frac{\mu \Delta t}{2} \tilde{L}_h \right)$  é simmetrica a predominanza diagonale stretta (vale anche per C. di N.) con

$$(I - \mu \frac{\Delta t}{2} \tilde{L}_h) = \frac{1}{h^2} \begin{pmatrix} D & -\mu \frac{\Delta t}{2} I & & \\ -\mu \frac{\Delta t}{2} I & D & -\mu \frac{\Delta t}{2} I & \\ & -\mu \frac{\Delta t}{2} I & D & -\mu \frac{\Delta t}{2} I \\ & & \ddots & \ddots & \ddots \\ & & & -\mu \frac{\Delta t}{2} I & D \end{pmatrix},$$

che é una  $m \times m$  matrice tridiagonale a blocchi in cui ogni blocco  $D$  e  $-\mu \frac{\Delta t}{2} I$  é esso stesso  $m \times m$  matrice:

$$D = \begin{pmatrix} 1 + 2\mu \frac{\Delta t}{h^2} & -\mu \frac{\Delta t}{2h^2} & & \\ -\mu \frac{\Delta t}{2h^2} & 1 + 2\mu \frac{\Delta t}{h^2} & -\mu \frac{\Delta t}{2h^2} & \\ & -\mu \frac{\Delta t}{2h^2} & 1 + 2\mu \frac{\Delta t}{h^2} & -\mu \frac{\Delta t}{2h^2} \\ & & \ddots & \ddots & \ddots \\ & & & \ddots & \ddots & \ddots \end{pmatrix},$$

Matrice con 5 diagonali, in 2D diventa piú complicato con questo tipo di matrici, per velocizzare il calcolo si puó usare un metodo numerico chiamato *Alternative Direction Implicit ADI* che vedremo in seguito.

*Metodo ADI* Vogliamo risolvere l'equazione:

$$u_t = \mu(u_{xx} + u_{yy}), \quad u(x, y, 0) = u_0(x, y).$$

Supponiamo che  $\Delta x = \Delta y = h$  e scriviamo l'equazione dal passo  $n$  al passo  $n + 1$ . Allora:

$$IMP - EXP : u_{jk}^{n+1/2} = u_{jk}^n + \mu \frac{1}{2\Delta x^2} \left( \delta_x^2 u_{jk}^{n+1/2} + \frac{1}{\Delta y^2} \delta_y^2 u_{jk}^n \right), \quad (1)$$

passo  $x$  implicito, passo  $y$  esplicito, (1) metodo C.N. direzione  $x$ , (2) m.C.N.direzione  $y$ :

$$EXP - IMP : u_{jk}^{n+1} = u_{jk}^{n+1/2} + \mu \frac{1}{2\Delta x^2} \left( \delta_x^2 u_{jk}^{n+1/2} + \frac{1}{\Delta y^2} \delta_y^2 u_{jk}^{n+1} \right), \quad (2)$$

passo  $x$  esplicito, passo  $y$  implicito.

Qual é il vantaggio del metodo ADI? Se ad esempio ho un dominio rettangolare: avrò  $m$  sistemi triangolari lungo la  $x$ , ed  $m$  sistemi triangolari lungo la  $y$ .

Nel caso (1) il termine esplicito, matrice costante costa  $m$  (se coefficienti variabili  $2m$ ), termine implicito costa  $5m$  (5 diagonali diverse da zero), quindi con  $7m$  operazioni risolvo la (1), cioè un costo lineare, analogamente per la (2). Quindi per avere la soluzione al tempo  $n+1$  da  $n$  costo totale:  $14m^2$ .

Questo approccio da sistemi tridiagonali disaccoppiati per risolvere ogni passo  $n$ .

Stabilità. Analisis di Von Neumann:  $u_{jk}^n = \rho^n e^{ij\xi} e^{ij\eta}$ , e  $u_{jk}^{n+1/2} = \rho^{n+1/2} e^{ij\xi} e^{ij\eta}$ . sostituendo nella (1):  $\rho^{1/2} = 1 + a(\rho^{1/2} 2(\cos(\xi) - 1) + 2(\cos(\eta) - 1))$  con  $a = \frac{\mu \Delta t}{2h^2}$ . Sostituendo anche nella (2) invece:

$$\rho = \rho^{1/2} + a(\rho^{1/2} 2(\cos(\xi) - 1) + 2(\cos(\eta) - 1)\rho).$$

Riscrivendo l'ultima espressione sostituendo  $\rho^{1/2}$ :

$$\rho = \frac{(1 - 2a(1 - \cos(\xi)))(1 - 2a(1 - \cos(\eta)))}{(1 + 2a(1 - \cos(\eta)))(1 + 2a(1 - \cos(\xi)))}$$

questa espressione è del tipo:

$$\frac{1-x}{1+x} \frac{1-y}{1+y}, \quad x, y \in [0, +\infty].$$

e segue  $|\rho| \leq 1$ , metodo incondizionatamente stabile.



Proviamo che il Metodo ADI è del secondo ordine. Potremmo utilizzare lo sviluppo in serie di Taylor della soluzione esatta e sostituire nell'operatore differenziale come fatto in precedenza. MA per la simmetria dei due passi, comunque, l'errore locale introdotto nel secondo passo cancella quasi esattamente l'errore introdotto nel primo passo, così che il metodo combinato è accurato al secondo ordine globalmente.

Un metodo alternativo è di fare la differenza tra il metodo di C.N. e ADI è del terzo ordine! (ovvero cancella l'errore al secondo ordine...)

Somma e differenza tra (1) e (2):

$$Somma : \quad u_{jk}^{n+1} = u_{jk}^n + \mu \frac{\Delta t}{h^2} \left( \delta_x^2 u_{jk}^{n+1/2} + \frac{1}{2} \delta_y^2 (u_{jk}^n + u_{jk}^{n+1}) \right)$$

$$Differenza : \quad u_{jk}^{n+1/2} = \frac{u_{jk}^{n+1} + u_{jk}^n}{2} - \mu \frac{\Delta t}{4h^2} \left( \delta_y^2 (u_{jk}^n + u_{jk}^{n+1}) \right)$$

sostituendo  $u_{jk}^{n+1/2}$  e facendo la differenza con il metodo di C.N. otteniamo:

$$y_{jk}^{n+1,CN} - y_{jk}^{n+1,ADI} = \mu \frac{\Delta t}{h^2} \delta_x^2 \left( \mu \frac{\Delta t}{4h^2} \delta_y^2 (u_{jk}^{n+1} - u_{jk}^n) \right)$$



da

$$\frac{\delta_x^2}{h^2} \approx \frac{\partial^2}{\partial x^2}, \quad \frac{\delta_y^2}{h^2} \approx \frac{\partial^2}{\partial y^2}$$

e

$$\frac{\partial u}{\partial t} \approx \frac{u_{jk}^{n+1} - u_{jk}^n}{\Delta t},$$

segue

$$y_{jk}^{n+1,CN} - y_{jk}^{n+1,ADI} \approx \mu^2 \frac{\Delta t^3}{4} \frac{\partial}{\partial t} \frac{\partial^2}{\partial x^2} \frac{\partial^2}{\partial y^2} u + \dots,$$

Questo dimostra che il metodo ADI é in accordo con il metodo di C.N. per termini  $\Delta t^3$  dopo un passo, cioé il metodo ADI é del secondo ordine nello spazio e nel tempo.

## Splitting method

Sia data l'equazione

$$\partial_t u = Au + Bu, \quad u(0) = u_0,$$

con  $A$  e  $B$  operatore lineare (es. matrici quadrati), soluzione esatta:  
 $u(\Delta t) = e^{(A+B)\Delta t} u_0$ .

Metodo dello *splitting*, calcolo:

$$u_t = Au, \quad u(0) = u_0,$$

e ottengo  $\tilde{u}(\Delta t) = e^{A\Delta t} u_0$ , poi risolvo:

$$u_t = Bu, \quad u(0) = \tilde{u}(\Delta t),$$

e ottengo la soluzione  $\bar{u}(\Delta t)$ , dove ho utilizzato come condizione iniziale la soluzione ottenuta al passo precedente.

Facciamo il confronto tra le due soluzioni  $u(\Delta t)$  e  $\bar{u}(\Delta t)$ .

$$\bar{u}(\Delta t) = e^{Bt} \tilde{u}(\Delta t) = e^{B\Delta t} e^{A\Delta t} u_0,$$

Se  $A$  e  $B$  sono quantitá scalari valgono le proprietá delle potenze. Se  $A$  e  $B$  matrici non é vero che le due soluzioni sono guali, perché in generale le matrici  $A$  e  $B$  commutano.

Per calcolare l'esponenziale di matrici sia  $A \in \mathbb{R}^{n \times n}$ , é definita mediante il seguente sviluppo:

$$e^A = I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \cdots + \frac{A^n}{n!} + \cdots = \sum_{k=0}^{+\infty} \frac{A^k}{k!}$$

dove  $A^0 = I$ , matrice identitá. Questa definizione dal punto di vista analitico va bene dato che la funzione esponenziale converge sempre ( $\|e^A\|$  é sempre convergente). Ma chiaramente é antipatico da calcolare, in quanto serie!

Allora prendiamo  $A\lambda = \lambda u$ , se  $u$  autovettore della matrice  $A$ , allora:

$$e^A u = \sum_{n=1}^{\infty} \frac{A^n}{n!} u = \sum_{n=1}^{\infty} \frac{\lambda^n}{n!} u = \sum_{n=1}^{\infty} \frac{e^\lambda}{u},$$

e se  $A$  é diagonalizzabile ( $n$  autovettori linearmente indipendenti)  $A = V\Lambda V^{-1}$ , con  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ , con la  $j$ -esima colonna di  $V$  é il corrispondente autovalore di  $\lambda_j$ , allora:

$$e^A = V e^\Lambda V^{-1},$$

in quanto se  $x \in \mathbb{R}^n$ ,  $x = \sum_j \alpha_j u_j$  allora  $e^A x = \sum_j \alpha_j e^{\lambda_j} u_j$ , e vale per ogni  $x$ , con  $e^\Lambda = \text{diag}(e^{\lambda_1}, \dots, e^{\lambda_m})$

Se  $A$  non é diagonalizzabile, allora é sempre possibile avere una decomposizione di Jordan  $A = VJV^{-1}$ , dove  $J$  é una matrice a blocchi con  $J_k = \text{diag}(\lambda_k) \in \mathbb{C}^{m_k \times m_k}$ ,  $m_1 + m_2 + \cdots + m_k = m$ .

Se le due matrici  $A$  e  $B$  commutano allora:

$$e^{A+B} = e^A e^B,$$

oppure in generale  $e^{A+B} \neq e^A e^B$ . Per cui in questo ultimo caso si commette un errore, detto *errore di splitting*.

Adesso se volessimo risolvere l'equazione

$$u_t = u_{xx} + u_{yy} + f(x, y)$$

e volendo integrare per un tempo molto lungo, la soluzione dipenderá solo da  $x$  e  $y$  si avrà una soluzione stazionaria che soddisfa l'equazione

$$-\Delta u = f.$$

Quindi utilizzando un metodo di splitting abbiamo:

$$u_t = u_{xx} + f/2, \quad u_t = u_{yy} + f/2$$

Utilizzando un metodo di splitting avremo i due sistemi sopra evidenziati. Se studiamo l'errore locale dopo un passo  $\Delta t$  di un metodo splitting abbiamo:

$$err = u(\Delta t) - \bar{u}(\Delta t),$$

sviluppo in serie di Taylor per

$$\begin{aligned} e^{(A+B)\Delta t} &= I + \Delta t(A + B) + \frac{\Delta t^2}{2}(A + B)^2 + \dots \\ e^{B\Delta t} e^{A\Delta t} &= \left( I + \Delta tA + \frac{\Delta t^2}{2}A^2 + \dots \right) \left( I + \Delta tB + \frac{\Delta t^2}{2}B^2 + \dots \right) = \\ &= I + \Delta t(A + B) + \Delta t^2 \left( \frac{1}{2}B^2 + BA + \frac{1}{2}A^2 \right) + \mathcal{O}(\Delta t^3) \\ e^{(A+B)\Delta t} &= I + \Delta t(A + B) + \frac{\Delta t^2}{2} \left( \frac{1}{2}A^2 + B^2 + AB + BA \right) + \mathcal{O}(\Delta t^3) \\ err &= u(\Delta t) - \bar{u}(\Delta t) = \frac{\Delta t^2}{2}(AB - BA) + \mathcal{O}(\Delta t^3) \end{aligned}$$

errore locale del secondo ordine, quindi metodo del primo ordine! Se il primo termine lo facciamo per un tempo  $\frac{\Delta t}{2}$ :

$$e^{A\frac{\Delta t}{2}} e^{B\Delta t} e^{A\frac{\Delta t}{2}}$$

sottraendo otteniamo:  $u(\Delta t) - \bar{u}(\Delta t) = e^{(A+B)\Delta t} - e^{A\frac{\Delta t}{2}} e^{B\Delta t} e^{A\frac{\Delta t}{2}} = \mathcal{O}(\Delta t^3)$   
ovvero secondo ordine globale (Strang Splitting).



Dal metodo di Strang splitting per il sistema

$$u_t = (A + B)u,$$

allora da

$$u_1 = e^{-A\frac{\Delta t}{2}} e^{-B\Delta t} e^{-A\frac{\Delta t}{2}} u_0$$

e

$$u_{n+1} = e^{-A\frac{\Delta t}{2}} e^{-B\Delta t} e^{-A\frac{\Delta t}{2}} u_n$$

otteniamo:

$$u_{n+1} = e^{-A\frac{\Delta t}{2}} e^{-B\Delta t} e^{-A\frac{\Delta t}{2}} e^{-A\frac{\Delta t}{2}} e^{-B\Delta t} e^{-A\frac{\Delta t}{2}} u_{n-1}$$

cioé con  $N$  soluzioni di  $u_t = Bu$  e  $N + 1$  soluzioni di  $u_t Au$ , con

$$e^{-A\frac{\Delta t}{2}} e^{-B\Delta t} \dots e^{-A\frac{\Delta t}{2}} e^{-B\Delta t} e^{-A\frac{\Delta t}{2}},$$

cioé sfalsando questi esponenziali abbiamo gratis un metodo di ordine 2.



In generale abbiamo visto che i vari metodi hanno una struttura:

$$-A_j u_{j+1} + B_j u_j - C_j u_{j-1} = D_j, \quad j = 1, \dots, m$$

ricordiamo che condizioni di Dirichlet per  $j = 1$  e  $j = m$  compaiono elementi da mettere come termine noto.

Le quantità  $A_j, B_j, C_j > 0$ , e  $B_j > A_j + C_j$  (matrice a predominanza diagonale stretta), questa relazione vale per sistemi ellitici, mentre si ha il maggiore uguale per problemi parabolici.

La tecnica per risolvere sistemi tridiagonali: Algoritmo di Thomas (riportare un sistema tridiagonale a bidiagonale);

Si pone:  $u_j = E_j u_{j+1} + F_j$ , riscriviamo  $u_{j-1} = E_{j-1} u_j + F_{j-1}$ , sostituendo nell'equazione di partenzaabbiamo:

$$-A_j u_{j+1} + (B_j - C_j E_{j-1}) u_j = D_j + C_j F_{j-1}, \quad j = 1, \dots, m$$

e ricavando  $u_j$  otteniamo:

$$u_j = \frac{A_j}{B_j - C_j E_{j-1}} u_{j+1} + \frac{D_j + C_j F_{j-1}}{B_j - C_j E_{j-1}}$$

quindi abbiamo trovato due espressioni per  $u_j$  quindi ...

Da queste due quantità di ricorrenza ottengo:

$$E_j = \frac{A_j}{B_j - C_j E_{j-1}}$$

$$F_j = \frac{D_j + C_j F_{j-1}}{B_j - C_j E_{j-1}}$$

Da  $u_0 = 0$ ,  $u_{m+1} = 0$  (C.D.) segue:  $E_0 = 0$ ,  $F_0 = 0$ . Da questa si calcolano per ricorrenza tutte le altre:  $E_j$   $j = 1, \dots, m$  e poi ricavare  $F_j$ . Infine dalla relazione:  $u_j = E_j u_{j+1} + F_j$  partendo dalle  $E_j$  e  $F_j$  troviamo le  $u_j$  (ricordiamo che conosciamo il valore  $u_{m+1} = 0$  C.D.) allora:

for  $j = m : -1 : 1$

$$u_j = E_j u_{j+1} + F_j$$

end

Costo? Per calcolare  $E_j$  2flop,  $F_j$  2flop, e la  $u_j$  1flop, totale  $5 * m$ .

Se la matrice è a predominanza stretta, ovvero  $B_j > A_j + C_j$  si dimostra che l'algoritmo è stabile, ovvero le quantità  $E_j$  e  $F_j$

VARIANTE: Abbiamo una variante dell'algoritmo di Thomas.

Supponiamo che abbiamo una matrice che non è esattamente tridiagonale ma differisce dall'esserlo per una matrice di rango 1. Esattamente, supponiamo di risolvere un sistema lineare  $Ax = b$ , con  $A = T + u \cdot v^T$ , con  $u \cdot v^T$  matrice di rango 1.

Formula di Sherman-Morrison (l'idea è di non sapere risolvere facilmente un sistema del tipo  $Ax = b$ ).

ESEMPIO: Supponiamo di avere un Laplaciano con condizioni periodiche, e sia

$$T = \begin{pmatrix} \ddots & \ddots & & & \bullet \\ \ddots & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots \\ \bullet & & & \ddots & \ddots \end{pmatrix},$$

con  $u = (u_1, 0, 0, \dots, u_m)^T$ ,  $v = (v_1, 0, 0, \dots, v_m)^T$  e sia

$$A = T + u \cdot v^T = \begin{pmatrix} u_1 v_1 + t_{11} & \ddots & & u_1 v_m \\ \ddots & \ddots & \ddots & \\ & \ddots & \ddots & \\ u_m v_1 & \ddots & \ddots & u_m v_m + t_{mm} \end{pmatrix},$$

Imponiamo:

$$u_1 v_m = a_{1m}, \quad u_m v_1 = a_{m1}$$

2 eqs. e 4 incognite  $\infty^2$  soluzioni. Tra queste infinite soluzioni impongo anche:  $a_{11} = t_{11} + u_1 v_1$ ,  $a_{11} = t_{mm} + u_m v_m$ , allora abbiamo 4 eqs. in 6 incognite ( $t_{11}, t_{mm}, u_1, v_1, u_m, v_m$ ), ovvero  $\infty^2$  gradi di libertà.

Allora posso scegliere  $u$  e  $v$  tale che  $T$  sia a predominanza diagonale (così da utilizzare l'algoritmo di Thomas)

In generale un sistema lineare, si può scrivere come una matrice "facilmente invertibile" e una matrice di rango 1.

Allora risolviamo il sistema  $Ax = (T + uv^T)x = b$ , e cerco soluzioni del tipo:  $x = y + k$ , con  $k$  nuova incognita.

sostituendo abbiamo:

$$(T + uv^T)(y + k) = b, \quad Ty + Tk + uv^T(y + k) = b$$

ovvero

$$Tk = -u \underbrace{v^T(y + k)}_{\alpha} \alpha, \quad Tk = -\alpha u,$$

introduco adesso un nuovo vettore  $z$  e risolvo:  $Tz = u$  allora da queste due equazioni ricavo:  $k = -\alpha z$ , sostituendo

$$Tk = -u \underbrace{v^T(y + k)}_{\alpha}, \quad -\alpha Tz = -uv^T(y - \alpha z)$$

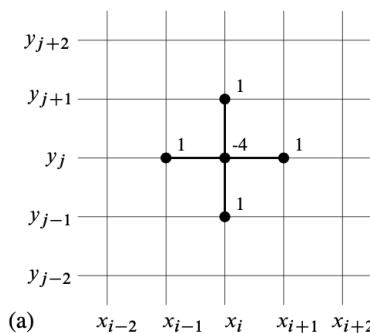
Da  $Tz = u$  ottengo:  $-\alpha Tz = -uv^T(y - \alpha z)$ , da cui, dividendo per  $-u$  ottengo:

$$\alpha = v^T(y - \alpha z), \quad \alpha = v^T y - \alpha v^T z, \Rightarrow \alpha = \frac{v^T y}{1 + v^T z}.$$

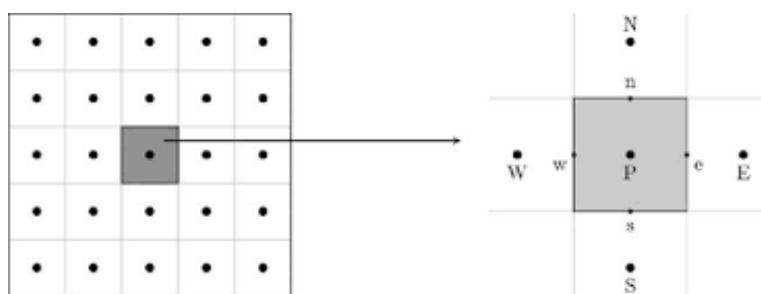
Per cui la formula di Sherman-Morrison posso scriverle:

$$\alpha = \frac{v^T y}{1 + v^T z}, \quad Ty = b, \quad Tz = u, \quad x = y - \alpha z.$$

Abbiamo sempre considerato la discretizzazione del dominio nei punti griglia, questo in modo particolare conviene se abbiamo condizioni di D. (VERTEX-CENTER DISCRETIZATION) abbiamo  $m \times m$  incognite.



Alternativa: CELL-CENTERED DISCRETIZATION, incognite al centro della cella,  $m \times m$ , celle,  $m \times m$  incognite.



Quindi avrei caso 2D eq. ellittica:

$$-u_W - u_N - u_S - u_E + 4u_P = f_P h^2,$$

ma  $u_W$  non ce l'ho quindi se vogliamo imporre Dirichlet basta considerare una interpolazione del punto W e P e prendo il punto di bordo e impongo la condizione di Dir. qui facciamo caso

$$\frac{u_W + u_P}{2} = g_D$$

e trovo  $u_W$ , se voglio Neumann,  $\frac{\partial u}{\partial n} = g_N$  si ottiene come:

$$\frac{u_W - u_P}{h} = g_N$$

quindi ho introdotto un punto "fantasma"  $u_W$  lo determino dalle condizioni al contorno e lo sostituisco sopra così da avere sempre un sistema di  $m$  equazioni e di  $m$  incognite, ovvero C.D. ho

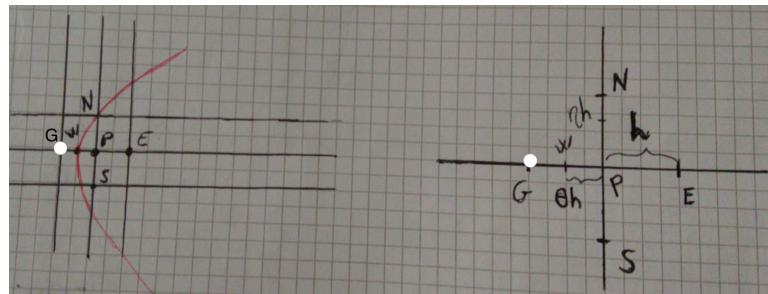
$$-u_N - u_S - u_E + 5u_P = f_P h^2 + 2g_D$$

Newmann  $-u_N - u_S - u_E + 3u_P = f_P h^2 + hg_N$ . Notiamo che se faccio la somma dei punti sulla stessa riga con C.N. la somma algebrica fa zero, ovvero matrice singolare, mentre C.D. matrice irriducibile diagonalmente dominante.



In queste discretizzazioni abbiamo sempre messo punti fantasma e tolto questo se la geometria è regolare, ma se le geometrie non sono regolari (domini arbitrari) allora non possiamo togliere i punti fantasma e avrei:  $N_i + N_g$  punti interni e punti fantasma come incognite e stesso numero di equazioni.

### GHOST POINT METHODS.



Sia  $G$  il punto "ghost" (fantasma), e utilizziamo una estrapolazione lineare  $u_W = \theta u_G + (1 - \theta)u_P$ , quando  $\theta \rightarrow 0$ ,  $W = P$ .



Adesso consideriamo la discretizzazione della derivata seconda in  $u_P$

$$u''_P \approx \frac{u_G - 2u_P + u_E}{h^2}$$

sostituendo

$$u_G = \frac{u_W - (1 - \theta)u_P}{\theta}$$

ottengo

$$u''_P \approx \frac{u_W - (1 + \theta)u_P + \theta u_E}{\theta h^2}$$

si può provare che il metodo fornisce un'accuratezza del secondo ordine anche se l'approssimazione non è consistente, ovvero come se stessi valutando un problema diverso ma con un errore del termine noto del secondo ordine. Ovvero risolviamo un problema che differisce da quello originale di un ordine di accuraterzza di ordine due in  $h$ , quindi alla fine il metodo fornisce un metodo del secondo ordine comunque.

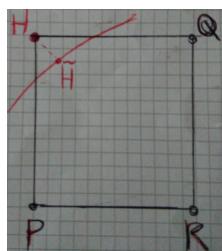
Approssimazione utilizzando tre punti, ho

$$u_G = \frac{2u_W + (2\theta^2 - 2)u_P + (1 - \theta^2)}{\theta^2 + \theta}$$

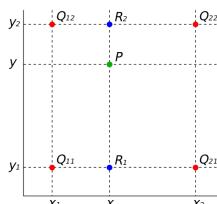
Quindi avrò incognite fantasma e incognite interne.



Alternativamente in questo caso se volessi assegnare invece di 2 punti per determinare il valore lungo  $H - P$ , come prima, possiamo lavorare con un solo valore, ovvero assegno solo il valore di  $H$  alla sua proiezione ortogonale  $\tilde{H}$  nella frontiera  $\Gamma$  e impongo che l'interpolazione di  $u$  in  $\tilde{H}$  soddisfi le condizioni di Dirichlet.



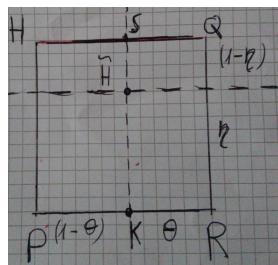
Dati i punti  $H, Q, R, P$  calcolo una approssimazione bilinare (interpolazione bilinare, interpolazione lineare nelle due direzioni)



e impongo che:  $\mathcal{L}(\tilde{H}; u_H, u_Q, u_P, u_R) = g_D(\tilde{H})$ .



Allora voglio interpolare nel punto  $\tilde{H}$ , il modo piú semplice per farlo é come prodotto cartesiano di interpolazioni in una dimensione.



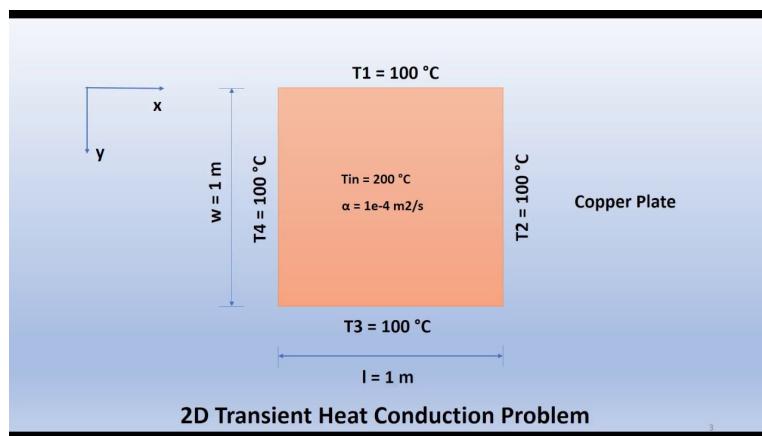
Lungo  $x$ :  $u_K = u_P\theta + (1 - \theta)u_R$ ,  $u_S = u_H\theta + (1 - \theta)u_Q$ , alla fine trovo il valore in  $\tilde{H}$  come:

$u_K(1 - \eta) + u_S\eta = u_P\theta(1 - \eta) + u_R(1 - \theta)(1 - \theta) + u_H\theta\eta + u_Q(1 - \theta)\eta$  combinazione convessa dei 4 punti. Perdiamo di accuratezza ma questa tecnica puó facilmente essere generata per condiz. di Neumann.

L'idea di questo metodo quindi 1'e di trovare una equazione per ogni punto fantasma cos1'i da bilanciare il numero di equazioni e numero di incognite. Se volessi ottenere alto ordine dovrei considerare interpolazione biliniare. Riassumendo Stencil 5 punti, secondo ordine per problemi con C.D. solo per la soluzione  $u$ , primo ordine se Neumann. Stencil 9 punti secondo ordine per  $u$  e per  $\Delta u$ , con C.D. e C.N.

N.B.: In realtá quando si ha a che fare con metodi di grande dimensione si utilizzano metodi iterativi, es. Jacobi o G-S.

## ESERCIZIO.



Usare il metodo ADI. Tempo finale  $T = 1600$ ,  $N_t = 160$ ,  $N_x = N_y = 15$ .

## ESERCIZIO 1:

$$u_{xx} + u_{yy} = f(x, y),$$

$$\text{con } f(x, y) = -(\cos(x + y) + \cos(x - y))$$

$$u(0, y) = \cos y, \quad u(\pi, y) = -\cos y, \quad 0 \leq y \leq \pi/2,$$

$$u(x, 0) = \cos x, \quad u(x, \pi/2) = 0, \quad 0 \leq y \leq \pi,$$

Usare  $h = \pi/5$ ,  $k = \pi/10$ , e comparare i risultati con la soluzione esatta  $u(x, y) = \cos(x) \cos(y)$ .

ESERCIZIO 2:  $f(x, y) = (x^2 + x^2)e^{xy}$ ,  $0 \leq x \leq 2$ ,  $0 \leq y \leq 1$ .

$$u(0, y) = 1, \quad u(2, y) = e^{2y} \quad 0 \leq y \leq 1,$$

$$u(x, 0) = 1, \quad u(x, 1) = e^x \quad 0 \leq x \leq 2,$$

$h = 0.2$ ,  $h = 0.1$  e comparare con la soluzione esatta  $u(x, y) = e^{xy}$ .

# Lezioni Analisi Numerica II modulo

Sebastiano Boscarino

Dipartimento di Matematica e Informatica  
Università di Catania, ITALY

May 7, 2020

## EQUAZIONE DELLE ONDE

L' equazione delle onde in fisica conosciuta anche come equazione di De-lambert del tipo (in una dimensione):

$$u_{tt} - c^2 u_{xx} = 0.$$

Questa equazione in piú dimensione (3D) non é di facile soluzione. Al contrario in una dimensione é piú semplice dato che possiamo scriverla nella forma

$$\left( \frac{\partial}{\partial t} + c \frac{\partial}{\partial x} \right) \left( \frac{\partial}{\partial t} - c \frac{\partial}{\partial x} \right) u = 0$$

Adesso se  $u$  é una funzione che soddisfa una delle due equazioni allora automaticamente soddisfa l'altra.

Quindi per iniziare studiamo l'equazione piú semplice delle onde (LINEAR ADVECTION EQUATION):

$$u_t + cu_x = 0, \quad u(x, 0) = u_0. \tag{1}$$

dove  $u(x, t) \in \mathbb{R}$ ,  $c > 0$ ,  $t > 0$ ,  $x \in \mathbb{R}$ .

Risolviamo il seguente problema. Consideriamo la derivata totale di  $u$ :

$$\frac{du}{dt} = \frac{\partial u}{\partial t} + \frac{\partial x}{\partial t} \frac{\partial u}{\partial x},$$

quindi imponendo  $\frac{\partial x}{\partial t} = c$ , per cui ho

$$u_t + cu_x = \frac{du}{dt} = \frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = 0.$$

ovvero

$$\frac{du}{dt} = 0, \quad u(x, 0) = u_0 \tag{2}$$

quindi (2) è detta forma caratteristica dell'equazione (1) e ci dice che la  $u$  è costante sulle curve caratteristiche, definite da

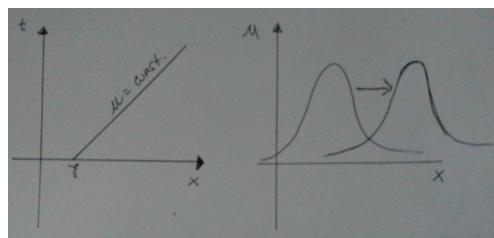
$$\frac{dx}{dt} = c, \quad x(0) = \xi, \quad t > 0,$$

e  $c > 0$  detta velocità dell'onda. Sono curve  $(x(t), t)$  nel piano  $(x, t)$ , e sono in particolare ( $c$  costante) delle rette  $x(t) = \xi + ct$ ,  $t > 0$ , (la pendenza della retta è data da  $dt/dx = 1/c$ ), e se  $c \rightarrow 0$  la curva tende a diventare verticale.

Quindi la soluzione  $u$  di (1) si mantiene costante lungo le linee caratteristiche, quindi  $u = u_0(\xi)$ , soluzione parametrica di (1), quindi per avere la soluzione in funzione di  $x$  e  $t$ , basta eliminare il parametro  $\xi$ :

$$u(x, t) = u_0(x - ct).$$

Ovvero la soluzione di (1) non è altro che una traslazione della soluzione iniziale, quindi il profilo dell'onda viaggia rimanendo inalterato.



Generalizzando

$$u_t + c(x, t)u_x = g(x, t)$$

la velocità non è costante. Prima riscriviamo l'equazione in forma caratteristica

$$\frac{du}{dt} = g(x, t), \quad \frac{dx}{dt} = c(x, t)$$

risolviamo l'equazione caratteristica

$$\frac{dx}{dt} = c(x, t), \quad x(\xi, 0) = \xi$$

una volta risolta questa equazione differenziale ordinaria, i.e.  $x(t) = \xi + c(x, t)t$  (soluzione parametrica), allora ho

$$\frac{du(x(t), t)}{dt} = g(x(t), t)$$

sostituendo la  $x(t)$  ho un'altra eq. diff. ordinaria per la  $u$ , otterremo una  $u = \tilde{u}(\xi, t)$  in funzione di  $\xi$  e  $t$ . In generale non è facile trovare una funzione di  $\xi$  in funzione di  $x$  e  $t$ , ovvero è molto difficile trovare la funzione inversa. Ma se supponiamo che la funzione  $c$  è lipschiziana negli argomenti allora il prob. ai valori iniziali delle caratteristiche ci da una e una sola soluzione, ovvero per ogni valore iniziale  $\forall \xi$  ci sarà una sola soluzione ovvero che le curve caratteristiche nel piano  $(t, x)$  non si possono mai intersecare.



Quindi risolviamo eq. differenziali ordinarie disaccoppiate, prima risolvo il problema per le caratteristiche e poi la soluzione. Questa è una equazione semi-lineare, ovvero lineare nella  $u$  e i coefficienti dipendono dalla  $u$ .

Il passo successivo è se la  $c$  dipende dalla  $u$ , i.e.  $c(u)$ , (detta equazione quasi-lineare). Il caso di una equazione quasi-lineare 1'e l'equazione di Burgers

$$u_t + uu_x = 0, \quad u(x, 0) = u_0(x).$$

Attraverso questa equazione si può vedere che le soluzioni ( dette forti) esistono fino a certi tempi e dopo non esistono più.

Studiamo l'equ. in forma caratteristica:

$$\frac{du}{dt} = 0, \quad \frac{dx}{dt} = u, \quad u(x, 0) = u_0(x).$$

in particolare dato che  $\frac{du}{dt} = 0$ ,  $u$  è costante lungo le linee caratteristiche e dal fatto che  $\frac{dx}{dt} = u = \text{cost.}$  quindi sia:

$$u = u_0(\xi), \quad \text{e} \quad x = \xi + u_0(\xi)t,$$

soluzione del prob.  $\frac{du}{dt} = 0$ , con  $u$  costante lungo le caratteristiche. Adesso vorrei una soluzione  $u(x, t)$  e non in forms parametrica. Basta eliminare la  $\xi$  dall'equazione  $x = \xi + u_0(\xi)t$  ma è di tipo non-lineare.



Per risolvere chiamiamo:  $F_t(\xi) = \xi + u_0(\xi)t$ , ma dobbiamo trovare la funzione inversa  $F_t^{-1}$ , quindi tutto dipende dall'invertibilità della funzione  $F_t$ .

Per avere la funzione inversa ci serve la condizione di monotonicità per la funzione  $F_t$ . Se  $F_t$  è regolare per studiare la monotonicità basta vedere la derivata prima:

$$\frac{dF_t}{d\xi} = 1 + u'_0(\xi)t \quad (3)$$

e al tempo  $t = 0$  segue:  $\frac{dF_t}{d\xi} > 0$  (ovvero monotonamente crescente). Vediamo se rimaner monotonamente crescente  $t > 0$ .

Se  $1 + u'_0(\xi)t > 0 \forall \xi \in \mathbb{R}$  quindi  $F_t(\xi)$  è monotonamente crescente, ovvero si può invertire e quindi la soluzione si può esprimere come

$$u = u_0(\xi) = u_0(\xi(x, t)) = \text{invertito} = \tilde{u}(x, t).$$

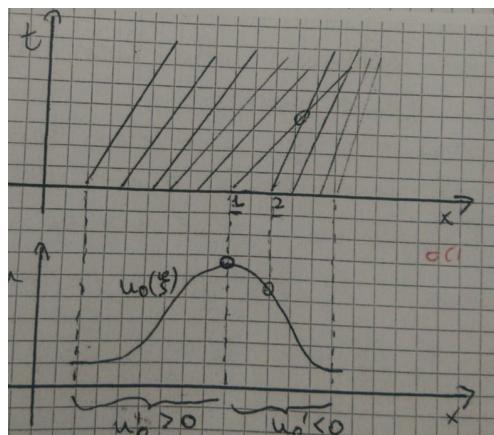
Quindi da (3) se  $u'_0(\xi) \geq 0$  segue  $1 + u'_0(\xi)t > 0, \forall t > 0$ . Se esiste un  $\xi$  tale che  $u'_0(\xi) < 0$ , allora esiste  $t_\xi^*$  tale che

$$\frac{dF_t}{d\xi} < 0, \quad \forall t > t_\xi^*,$$

Cosa vuol dire questa ultima relazione?



Cosa succede alle curve caratteristiche? Le caratteristiche hanno la pendenza che dipende da  $u_0(\xi)$ ,



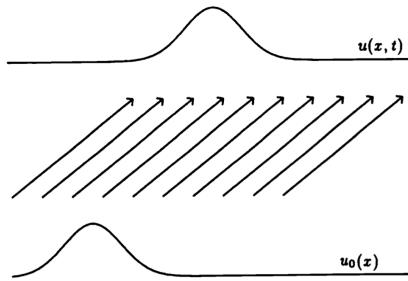
Tanto maggiore è la velocità  $u'_0(\xi)$  (zona con maggior pendenza) caratteristiche con maggiore pendenza (schiacciate), al contrario nella zona bassa, avrà caratteristiche di minore pendenza. Ci sarà un punto in cui si intersecheranno queste caratteristiche e in quel punto che valore ha la  $u$ ? Se guardo i due punti 1 e 2 le due caratteristiche si incontrano in un punto (lo stesso punto dello spazio-tempo) ovvero la funzione  $u$  avrebbe due valori! Ovvero ad uno stesso valore della  $x$  (ordinate) due valori della  $\xi$  (ascisse)! Le due caratteristiche si intersecano in una regione dove  $u'_0 < 0$ , e la soluzione  $u$  qui non è più monodroma (un solo valore)!



Se considero l'equazione

$$u_t + cu_x = 0$$

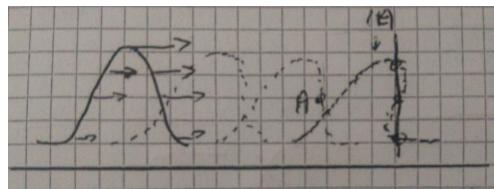
la soluzione corrisponde ad una oda che viene traslata.



mentre l'equazione di Burgers

$$u_t + uu_x = 0$$

il profilo si muoverà con una velocità che è proporzionale all'altezza del profilo stesso (punto più alto velocità maggiore di uno basso), cioè avrò una parte del profilo che si allunga mentre un'altra parte che si accorcia (si deforma)

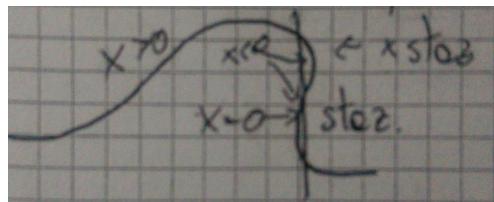


Ci sarà un momento in cui un punto del profilo presenterà un flesso verticale. Se superiamo quell'istante il profilo diventa a più valori.

Cosa vuol dire che il profilo presenta un flesso verticale? Quando la funzione cessa di essere invertibile per un certo valore di  $\xi$ , ci' quando la derivata si annulla  $\frac{dF_t}{d\xi} = 1 + u'_0(\xi)t = 0$ . Se io aumento il valore  $\xi$  (il parametro) sto percorrendo il profilo, e aumentando  $\xi$  aumento anche  $x$ , ma ad un certo punto aumentando  $\xi$  la  $x$  può decrescere.

Se  $A$  è il punto in cui si annulla la derivata di  $F_t$ , ovvero cond. di stazionarietà, allora:

$$1 + u'_0(\xi) = 0, \quad t = -\frac{1}{u'_0(\xi)},$$



questo tempo è positivo quando  $u'_0 < 0$ , ovvero la zona che crea l'accavallamento.

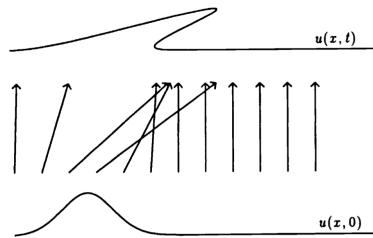
L'espressione precedente ci dice anche il tempo in cui ottengo la soluzione di stazionarietà, allora per ottenere il flesso verticale si calcola il tempo critico (minimo tempo della condizione di stazionarità):

$$t_c = \min_{u'_0(\xi) < 0} -\frac{1}{u'_0(\xi)}.$$

Allora se  $t < t_c$  esiste la soluzione "forte" dell'equazione:

$$\frac{du}{dt} = u_t + uu_x = 0, \quad u_0(x, 0) = u_0.$$

Se  $t > t_c$  non esiste soluzione "classica" (ovvero non esiste soluzione ad un sol valore!).



Allora cosa fare in quest'ultimo caso?

Esempio 1):

$$u_t + uu_x = 0, \quad u(x, 0) = 1 + \cos(\pi x), \quad x \in [-1, 1], \quad (4)$$

linee caratteristiche  $x(t) = \xi + tu_0(\xi)$ , con  $u_0(\xi) = 1 + \cos(\pi\xi)$ . Allora con  $\xi$  parametro la seguente rappresentazione parametrica

$$x(t) = \xi + tu_0(\xi), \quad u = u_0(\xi)$$

rappresenta una soluzione del sistema

$$\frac{dx}{dt} = u, \quad \frac{du(x(t), t)}{dt} = 0$$

e quindi del problema (4).

Esempio 2):

$$u(x, 0) = \begin{cases} 1, & x \leq 0 \\ 1-x, & 0 \leq x \leq 1, \\ 0, & x \geq 1. \end{cases}$$

$$x(t) = \xi + u_0(\xi)t = \begin{cases} \xi + t, & \xi \leq 0 \\ \xi + (1-\xi)t, & 0 \leq \xi \leq 1, \\ \xi, & \xi \geq 1. \end{cases}$$

le linee caratteristiche non si intersecano solo se  $t < 1$ .

Una soluzione per risolvere il problema delle multisoluzioni é rinunciare alla regolaritá della soluzione, ovvero formazione di discontinuitá ovvero per  $t > t_c$  si forma una discontinuitá, e in questo caso si parla di soluzione debole dell'equazione

$$u_t + uu_x = 0, \quad u(x, 0) = 0,$$

Scriviamo questa equazione in forma equivalente

$$u_t + \left( \frac{u^2}{2} \right)_x = 0, \quad u(x, 0) = 0,$$

detta in forma *conservativa*. Integriamo questa equazione in un'intervallo  $[a, b]$

$$\int_a^b u_t dx + \int_a^b \left( \frac{1}{2} u^2 \right)_x dx = 0,$$

$$\frac{d}{dt} \int_a^b u(x, t) dx + \frac{1}{2} (u^2(b, t) - u^2(a, t)) = 0,$$

se abbiamo condizioni periodiche, di periodo  $b - a$  allora abbiamo:

$$\frac{d}{dt} \int_a^b u(x, t) dx = 0$$

la soluzione  $u$  si conserva.



Supponiamo adesso che

$$\lim_{a \rightarrow -\infty} u(a, t) = \lim_{b \rightarrow \infty} u(b, t),$$

allora

$$\frac{d}{dt} \int_{-\infty}^{\infty} u(x, t) dx = 0$$

si conserva per tutto  $\mathbb{R}$ , e questo é il motivo per cui é detta forma conservativa.

La struttura di una legge di conservazione scalare ha la forma:

$$u_t + f(u)_x = 0, \quad u(x, 0) = u_0(x), \quad (5)$$

$f$  é detta funzione di flusso, e questa é un'equazione di bilancio. Per vederla meglio da un punti di vista fisico. Integriamo l'equazione:

$$\frac{d}{dt} \int_a^b u(x, t) dx = \underbrace{(f(a, t) - f(b, t))}_{\text{quantita' di flusso che entra e che esce}}, \quad (6)$$

le due equazioni sono la stessa cosa? Osserviamo

- ▶ Se integro la (5) ottengo la (6);
- ▶ Ma dato la (6) ottengo la (5);

Devo fare due ipotesi:



- 1) Se la  $u$  é derivabile rispetto a  $t$  allora posso portare la derivata dentro il segno di integrale quindi ho  $\int_a^b u_t$ .
- 2) Se la  $f$  é derivabile rispetto a  $x$ , allora per il teorema fondamentale del calcolo integregale:

$$(f(a, t) - f(b, t)) = \int_a^b f(u)_x dx$$

allora se valgono le due condizioni

$$\int_a^b (u_t + f(u)_x) dx = 0$$

e voglio che questa relazione sia valida per ogni intervallo  $[a, b]$  e allora posso dire che dalla (6) ottengo la (5).

Dalla precedente "osservazioni" osservo che la forma (6) é piú debole (detta forma debole) rispetto alle (5) (cioé meno restrittiva e ammette piú soluzioni, anche se meno regolari).

## Soluzioni deboli

. Se moltiplichiamo (5) per una funzione  $\phi(x, t)$  a supporto compatto (funzioni diverso da zero in un insieme compatto e al di fuori nulle) otteniamo:

$$\int_0^\infty dt \int_{-\infty}^\infty (u_t \phi + f(u)_x \phi) dx = 0,$$

applichiamo la formula per integrazione per parti:

$$(u\phi)_t = u_t \phi + u\phi_t, \quad (f\phi)_x = f_x \phi + f\phi_x$$

$\phi$  supporto compatto quindi  $t \rightarrow \infty$  e per  $x \rightarrow \pm\infty$  la  $\phi = 0$ ,

$$\int_0^\infty \int_{-\infty}^\infty ((u\phi)_t - u\phi_t + (f\phi)_x - f\phi_x) dx dt = 0,$$

$$\int_0^\infty \int_{-\infty}^\infty (\phi_t u + \phi_x f(u)) dx dt = - \int_{-\infty}^\infty \phi(x, 0) u(x, 0) dx$$

da questa relazione vediamo che compare la funzione  $u$  e non le sue derivate, per cui se la  $u$  non é regolare, discontinua per esempio, l'ultima relazione ha ancora senso, in quanto tutte le derivate sono scaricate sulla  $\phi$  che é una funzione regolare. Quindi questa é la formulazione debole ed é piú generale della (5).

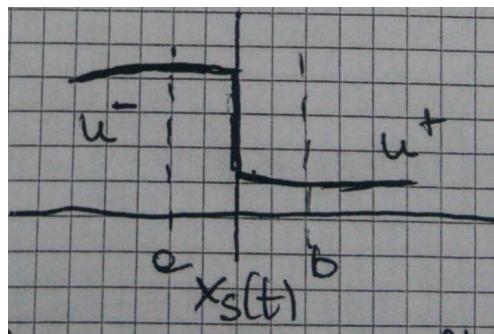
Tornando all'equazione di Burgers, scritta in forma conservativa:

$$u_t + \left( \frac{u^2}{2} \right)_x = 0, \quad u(x, 0) = u_0(x),$$

Allora

$$\frac{d}{dt} \left( \int_a^b u dx \right) = f_a - f_b,$$

con  $f_a = f(u(a, t), t)$ , facciamo vedere che ci può essere una soluzione discontinua che si propaga e supponiamo che in  $x_s(t)$  sia la posizione di tale discontinuità al tempo  $t$ , e consideriamo l'intervallo  $[a, b]$  dove tale punto di discontinuità è contenuta.



Supponiamo che la funzione  $u$  sia regolare a tratti, ovvero abbia la forma:

$$u(x, 0) = \begin{cases} u^-(x, t), & x < x_s(t), \\ u^+(x, t), & x > x_s(t). \end{cases}$$

Allora abbiamo

$$\frac{d}{dt} \left( \int_a^{x_s(t)} u dx + \int_{x_s(t)}^b u dx \right) = f_a - f_b,$$

adesso utilizzando la seguente proprietà

$$\frac{d}{dt} \int_0^{x(t)} g(z, t) dz = \int_0^{x(t)} g_t dx + g(x(t), t)x_t,$$

da questa relazione segue

$$\int_a^{x_s(t)} u_t^- dx + u^-(x_s(t), t)\dot{x}_s + \int_{x_s(t)}^b u_t^+ dx - u^+(x_s(t), t)\dot{x}_s = f_a - f_b,$$

facendo il limite di  $a \rightarrow x_s^-$   $b \rightarrow x_s^+$ , we get

$$(u^-(x_s(t), t)\dot{x}_s - u^+(x_s(t), t))\dot{x}_s = f(u^-(x_s(t), t)) - f(u^+(x_s(t), t)),$$

la discontinuità è legata a questa relazione (cioè alla velocità). In modo compatto possiamo scrivere:

$$[[u]]\dot{x}_s = [[f]], \quad \forall h(u) : [[h]] = h(u^-(x_s(t), t)) - h(u^+(x_s(t), t))$$

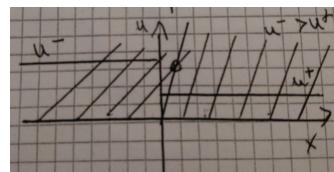
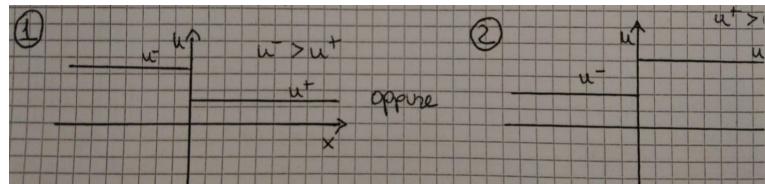
detta *Condizione di Salto di Rankine-Hugoniot*, e lega la velocità con cui si muove la discontinuità ai salti della "densità"  $u$  e ai salti del flusso  $f$  ed è una conseguenza dell'equazione in forma debole.  $[[\cdot]]$  indica il salto di una certa quantità attraverso la discontinuità.

Queste condizioni però non garantiscono ancora l'unicità della soluzione. Per mostrare ciò consideriamo Eq. di Burgers

$$u_t + \left(\frac{u^2}{2}\right)_x = 0,$$

consideriamo le condiz. iniziali:

$$u(x, 0) = \begin{cases} u^-, & x < 0, \\ u^+, & x \geq 0. \end{cases}$$



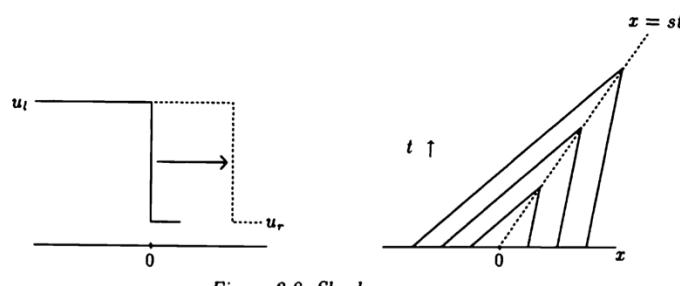
$u^- > u^+$  l'inclinazione è maggiore nella zona  $u^-$ , minore in  $u^+$ . La velocità della discontinuità in questo caso vale:

$$\dot{x}_s = \frac{[[f]]}{[[u]]} = \frac{\frac{1}{2}((u^+)^2 - (u^-)^2)}{u^+ - u^-} = \frac{1}{2}(u^+ + u^-)$$

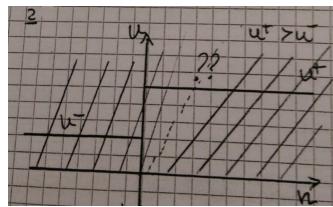
ovvero la discontinuità si propaga con una pendenza che è la media aritmetica tra le due pendenze, quindi alla fine la  $u$  è:

$$u = \begin{cases} u^+, & x > \dot{x}_s t, \\ u^-, & x < \dot{x}_s t. \end{cases}$$

con  $\dot{x}_s = \frac{1}{2}(u^+ + u^-)$  notiamo che in questo caso le caratteristiche vanno dentro lo shock.



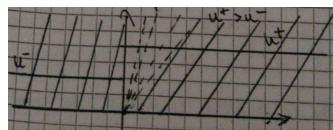
Nel caso 2.  $u^- < u^+$  che soluzione abbiamo? In questo caso ci sono infinite soluzioni deboli, qui ne presentiamo due:



Una soluzione può essere sempre:

$$u = \begin{cases} u^+, & x > \dot{x}_s t, \\ u^-, & x < \dot{x}_s t. \end{cases}$$

questa soluzione soddisfa l'equazione di B. come la precedente (notiamo che in questo caso le caratteristiche vanno fuori lo shock). Un'altra soluzione di questo caso può essere la seguente:



Considero le caratteristiche che partono dall'origine:  $x = ct$ , con  $c$  costante e queste rette hanno deriviamo da

$$\frac{dx}{dt} = c, \quad x(0) = 0,$$

e dalla forma caratteristica:  $\frac{du}{dt} = 0$ ,  $su$ :  $\frac{dx}{dt} = u$ .

Allora se  $c := u$ , allora la pendenza delle caratteristiche va da  $u^- \leq c \leq u^+$ , allora un'altra soluzione è:

$$u = \begin{cases} u^+, & x > u^+ t, \\ \frac{x}{t}, & u^- t < x < u^+ t, \\ u^-, & x < u^- t. \end{cases}$$

ONDE RAREFATTE. Per verificare che soddisfa la soluzione allora abbiamo:

$$u_t = -\frac{x}{t^2}, \quad u_x = \frac{1}{t},$$

$$u_t + uu_x = -\frac{x}{t^2} + \frac{x}{t} \cdot \frac{1}{t} = 0.$$

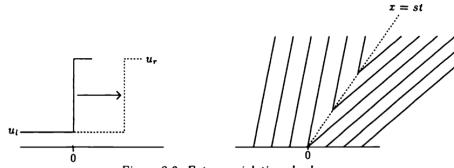


Figure 3.9. Entropy-violating shock.

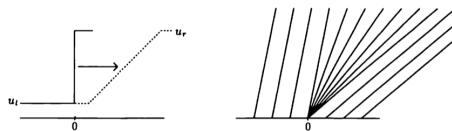


Figure 3.10. Rarefaction wave.

Cosa sta succedendo? Perché abbiamo così tante soluzioni?



Andiamo a rivedere il modello e consideriamo quelle che vengono chiamate SOLUZIONI VISCOSE.

Considero un problema che dipende da un parametro  $\varepsilon > 0$ . Nel nostro caso il nostro problema è:

$$u_t + uu_x = \varepsilon u_{xx}$$

dove abbiamo aggiunto un termine di diffusione. Se  $\varepsilon \rightarrow 0$  riottremiamo il prob. di B. L'equazione è un'equazione detta di convezione-diffusione (convezione non-lineare).

Caso Semplice. Eq. di convezione diffusione lineare è del tipo:

$$u_t + cu_x = \varepsilon u_{xx}, \quad \varepsilon > 0 \quad \text{coefficiente di diffusione.}$$

Abbiamo visto che il profilo della diffusione si "allarga", quindi se ho un gradino la diffusione non fa altro che smussare il gradino.



mentre il pezzo di convezione non fa altro che traslare il profilo di una certa quantità  $ct$ . Quindi avendo entrambi i termini quindi il profilo iniziale si sposta mentre la diffusione lo "sbrodola" per soluzioni regolari, mentre per c.i. a gradino si sposta la soluzione e si smussa agli angoli del gradino.



La diffusione ha un'effetto regolarizzante quindi se ho un prob. conv-diff con condizione iniziale anche discontinua avrò una soluzione per ogni tempo  $\in \mathbb{C}^\infty$ .

Caso di equazione di Burgers viscosa.

$$u_t + uu_x = \varepsilon u_{xx}, \quad u(x, 0) = u_0(x).$$

Caso non-lineare il problema è più delicato. Dal punto di vista teorico si può dimostrare che l'equaz. precedente ammette soluzioni regolari  $\forall t$ . Quindi per ogni  $\varepsilon$  esisterà una soluzione dell'eq.  $u_\varepsilon(x, t)$ .

Chiamo  $u$  soluzione di viscosità se

$$u(x, t) = \lim_{\varepsilon \rightarrow 0} u_\varepsilon(x, t)$$

il parametro  $\varepsilon$  detto di perturbazione singolare perché appena lo consideriamo cambia il carattere dell'equazione da iperbolica a parabolica.

Quando  $\varepsilon \rightarrow 0$ , si può studiare il profilo dell'onda viaggiante, ovvero una soluzione che non dipende separatamente da  $x$  e da  $t$ , ma da un unico parametro  $\xi = x - ct$ . Quindi si cerca una soluzione del tipo:

$$u = \phi(\xi) = \phi(x - ct),$$

e sostituendola nella soluzione si ha  $-c\phi' + \phi\phi' = \varepsilon\phi''$ , quindi si passa da una pde ad una ode (unico parametro di derivazione è  $\xi$ ). ◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏷ ⏸ ⏹ ⏺

Un'altra osservazione è che:

$$\phi\phi' = \frac{1}{2} \frac{d\phi^2}{d\xi}$$

allora

$$\frac{d}{d\xi} \left( \frac{1}{2}\phi^2 - c\phi - \varepsilon\phi' \right) = 0,$$

ovvero

$$Q := \frac{1}{2}\phi^2 - c\phi - \varepsilon\phi' = Cost.$$

Supponiamo che all'infinito la  $\phi$  sia costante (per tempi lunghi), quindi se

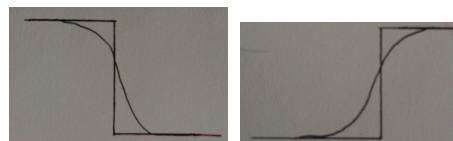
$$\phi_1 \text{ valore di } \phi \text{ a } \infty,$$

$$\phi_2 \text{ valore di } \phi \text{ a } -\infty,$$

allora

$$\frac{1}{2}\phi_1^2 - c\phi_1 = \frac{1}{2}\phi_2^2 - c\phi_2 := Q.$$

Nel nostro caso cerco profili a forma di S:



Chiaramente in tali profili, caso 1) diffusione smussa e la non linearità rende più ripida la soluzione che convergerà per  $\varepsilon \rightarrow 0$ . Nel caso 2) non potremo trovarlo mai tale profilo perché la diffusione tenderà a essere più "sbrodolata" e si allunga ... Per cui posso cercare solo il primo profilo caso 1).

Per cui so che all'infinito ( $\pm$ ) deve fare:  $Q$ , e siccome so che deve essere un'onda viaggiante e so pure che  $Q$  è costante. Da

$$-c\phi' + \phi\phi'' - \varepsilon\phi''' = 0$$

è un'equazione di secondo grado, ed ha un'integrale primo, ovvero  $Q$  costante.

La derivata prima deve fare zero, perché il profilo diventa piatto, per cui vale:

$$c = \frac{1}{2}(\phi_1 + \phi_2).$$

quest'ultima relazione ci dice che la velocità dell'onda viaggiante è uguale alla media aritmetica dei valori a  $\pm\infty$ , (ricordiamo che la velocità dello shock è uguale alla media dei due valori a sinistra e a destra della discontinuità). Comunque niente di sorprendente perché l'equazione di viscosità è sempre una legge di conservazione  $f = u^2/2 - \varepsilon u_x$ .

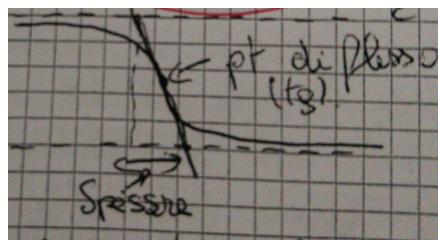
Se si risolve l'equazione differenziale, si prova che la soluzione ha la forma:

$$\phi = \frac{\phi_1 + \phi_2}{2} + \frac{\phi_1 - \phi_2}{2} \tanh(\alpha\xi + \beta),$$

Quando  $\xi \rightarrow \pm\infty \tanh(\dots) = \pm 1$ ,  $\beta = \text{cost}$ . che dipende dalle condizioni iniziali mentre  $\alpha$  dipende da  $\varepsilon$  e vale  $\alpha = \frac{1}{L}$  con

$$L \propto \frac{\xi}{(\phi_1 - \phi_2)}$$

$L$  è lo spessore della tangente iperbolica. Tangente iperbolica ha la forma di:



Quindi più grande è il salto più piccolo lo spessore, e più piccolo  $\varepsilon$ ,  $\varepsilon \rightarrow 0$  la tanh tende a diventare un gradino (profilo più ripido).

La soluzione di viscosità è quella che restituisce l'unicità.  $\varepsilon \rightarrow 0$  soluz. di onda d'urto. Le caratteristiche entrano nella discontinuità. Le onde rarefatte sono soluzioni ma non le discontinuità uscenti. C'è una irreversibilità nelle soluzioni, ovvero non si può tornare indietro, in fisica questa irreversibilità è legata all'entropia. E questo irreversibilità è possibile descriverla da una funzione di entropia matematica.

**Condizione di Entropia.** Una discontinuità che si propaga con velocità

$$\dot{s} = \frac{f(u_2) - f(u_1)}{u_2 - u_1},$$

soddisfa la condizione di entropia se  $f(u_2) > s > f(u_1)$ . Per Burgers questa condizione di entropia si riduce a richiedere che se una discontinuità si propaga con velocità  $\dot{s}$  allora  $u_2 > u_1$ .

Caso  $u^+ = 1$  e  $u^- = 0$ ,  $\dot{s} = 1/2$ , e quindi la soluzione è

$$u(x, t) = \begin{cases} 1, & x \leq t/2, \\ 0, & x > t/2 \end{cases}$$

e questa soluzione soddisfa la condizione di entropia. E questa è l'unica soluzione debole del problema.

Caso  $u^+ = 0$ ,  $u^- = 1$  soluzione:

$$u(x, t) = \begin{cases} 0, & x \leq 0, \\ x/t, & 0 < x \leq t \\ 0, & x > t \end{cases}$$

# Lezioni Analisi Numerica II modulo

Sebastiano Boscarino

Dipartimento di Matematica e Informatica  
Università di Catania, ITALY

May 17, 2020

## Metodi Numerici

Risolviamo l'equazione di convezione lineare scalare:

$$u_t + cu_x = 0, \quad u(x, 0) = u_0.$$

$c > 0$ . Discretizziamo prima lo spazio (metodo delle linee)

$$\frac{dU_j}{dt} + c \frac{U_{j+1} - U_{j-1}}{h} = 0,$$

discretizzazione centrale del secondo ordine, con condizioni periodiche al contorno (esempio:  $u(0,t) = u(1,t) \forall t > 0$  with  $0 \leq x \leq 1$ ). In tempo utilizziamo Eulero-explicito per il tempo

$$U_j^{n+1} = U_j^n - c \frac{k}{h} \frac{(U_{j+1} - U_{j-1})}{h}.$$

Consistenza ordine 1 in tempo e ordine 2 in spazio.

Sia  $U_j(t) = (U_1(t), U_2(t), \dots, U_m(t))^T$ , con  $U_j(t) \approx u(x_j, t)$ . Caso periodico  $U_0(t) = U_{m+1}(t)$ .

Per  $2 \leq j \leq m$  abbiamo quindi una ODE:

$$\frac{dU_j}{dt} + c \frac{U_{j+1} - U_{j-1}}{h} = 0,$$

e puó essere riscritta come:

$$U'(t) = AU(t),$$

con

$$A = -c \frac{k}{h} \begin{pmatrix} 0 & 1 & 0 & \cdots & -1 \\ -1 & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 1 & \cdots & \cdots & -1 & 0 \end{pmatrix}$$

Notiamo che questa matrice é anti-simmetrica  $A^T = -A$  e i suoi autovalori devono essere puri e immaginari. Infatti sono

$$\lambda_p = -\frac{ic}{h} \sin(2\pi ph), \quad p = 1, 2, \dots, m+1$$

e autovettori

$$u_j^p = e^{2\pi ipjh}, \quad j = 1, 2, \dots, m+1$$

gli autovalori giaciono nell'asse immaginaria tra  $-ic/h$  e  $ic/h$ .

Analisi di stabilitá di Von-Newmann.

Sia  $u_j^n = \rho^n e^{ij\xi}$ , segue:

$$\rho = 1 - \frac{ck}{h} \frac{e^{ij\xi} - e^{-ij\xi}}{2} = 1 - \frac{ick}{a} \frac{e^{ij\xi} - e^{-ij\xi}}{2i}$$

$$\rho = 1 - \frac{ck}{h} i \sin(\xi)$$

$$|\rho|^2 = 1 + c^2 \frac{c^2}{h^2} i^2 \sin^2(\xi) \approx 1 + c^2 \frac{k^2}{c^2} > 1$$

Cioé il fattore di amplificazione é maggiore di 1. Cioé il metodo é incondizionatamente instabile.

Quindi il metodo di Eulero associato alla discretizzazione centrale è incondizionatamente instabile perché noi sappiamo che il metodo di Eulero è stabile se  $|1 + \lambda_p k| \leq 1$  e la regione di stabilità  $S$  è il cerchio di centro -1 e raggio 1. Per quanto scegliamo piccolo il rapporto  $k/h$  poiché l'autovalore  $\lambda_p$  è immaginario  $\lambda_p k$  non starà mai in  $S$ . Quindi il metodo è instabile per qualunque rapporto  $k/h$ .

Da Eulero abbiamo

$$U_j^{n+1} = C_k U_j^n$$

Qui  $C_k = I + kA$ . Allora noi abbiamo:

$$|1 + k\lambda_p|^2 \leq 1 + (ck/h)^2,$$

per ogni  $p$  e quindi posto  $k = h^2$

$$|1 + k\lambda_p|^2 \leq 1 + c^2 h^2 = 1 + c^2 k$$

La condizione di Von-Neumann ci dice quindi che

$$\|C_k\|^2 \leq 1 + \mathcal{O}(\Delta t) = 1 + c^2 k$$

e se  $nk \leq T$ , abbiamo

$$\|(I + kA)^n\| \leq (1 + c^2 k)^n / 2 \leq e^{c^2 T / 2}$$

quindi mostra l'uniforme limitatezza di  $\|C_k^n\|$  (in norma 2) necessaria per la stabilità. Per cui  $e^{c^2 T / 2}$  non diverge se la quantità  $c^2 T$  è finita.

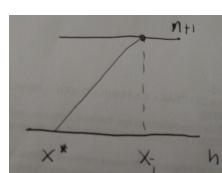
Per la convergenza vogliamo la consistenza  $h, k \rightarrow 0$ , dalla relazione precedente  $k/h^2$  non deve divergere allora:  $k = \mathcal{O}(h^2)$  e uno scaling di questo tipo l'abbiamo visto per il metodo di Eulero nel caso dell'equazione del calore, chiaramente tale restrizione era ovvia per i metodi utilizzati.

Abbiamo visto che E.E. non va, ma se applico un RK-2, mi aspetto una discretizzazione del secondo ordine, ma invece ottengo  $k < Ch^{4/3}$ , c'è qualcosa che non va, forse l'approccio delle linee non è il miglior metodo per la discretizzazione dell'equazione? Proviamo una discretizzazione globale, ovvero un metodo completamente discreto.

Voglio trovare la soluzione al punto  $x_j$  al tempo  $n + 1$ . Scriviamo quindi l'equazione in forma caratteristica

$$\frac{du}{dt} = 0, \quad \frac{dx}{dt} = c.$$

Quindi se voglio sapere questa soluzione traccio la caratteristica all'indietro ( $c > 0$ ), e incotra l'asse delle  $x$  in un certo punto  $x^*$



Ma non ho  $x^*$ , per ottenerlo uso l'interplazione lineare per esempio tra i punti  $(x_{j-1}, u_{j-1})$ ,  $(x_j, u_j)$ , ottengo

$$u_j^{n+1} = u_j^n \frac{x^* - x_{j-1}}{\Delta x} + u_{j-1}^n \frac{x^* - x_{j-1}}{\Delta x},$$

mentre  $x^* = x_j - ck$ , sostituendo

$$\frac{u_j^{n+1} - u_j^n}{k} + c \frac{(u_j^n - u_{j-1}^n)}{h} = 0,$$

Con

$$u_t \approx \frac{u_j^{n+1} - u_j^n}{k}, \quad u_x \approx \frac{(u_j^n - u_{j-1}^n)}{h}$$

Se la  $c < 0$  allora la avremmo

$$u_x \approx \frac{(u_{j+1}^n - u_j^n)}{h}$$

Questo metodo é detto metodo upwind, e si vede che la consistenza del metodo globalmente é 1. Proviamo adesso che il metodo é stabile solo se il piede della caratteristica cade nell'intervallo  $[x_{j-1}, x_j]$ .

Considero sempre il modo di Fourier

$$u_j^n = \rho^n e^{ijh\xi}$$

allora sostituendo ottengo

$$\rho = 1 - c \frac{k}{h} (1 - e^{-ih\xi}) = (1 - a) + ae^{-ih\xi} = (1 - a) + a(\cos(\theta) - i \sin(\theta)),$$

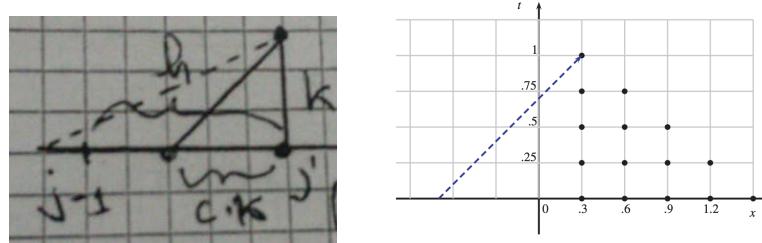
posto  $a = ck/h$ , e  $h\xi = \theta$  abbiamo quindi

$$|\rho|^2 = (1 - a + a\cos(\theta))^2 + a^2 \sin^2(\theta) = 1 - 4a(1 - a)\sin\left(\frac{1}{2}\phi\right),$$

sfruttando che  $(1 - \cos(\phi)) = 2\sin^2(\frac{1}{2}\phi)$ ,  $\sin(\phi) = 2\sin(\frac{1}{2}\phi)\cos(\frac{1}{2}\phi)$ .  
Ovvero  $|\rho|^2 \leq 1$  se  $0 \leq a \leq 1$ ,

$$0 \leq c \frac{k}{h} \leq 1.$$

Abbiamo supposto che  $c > 0$ , ma se usassimo gli stessi punti non avrei invece più stabilità con  $c < 0$ . La pendenza è  $1/c$



distanza percorsa in un tempo  $k$ . Quindi la distanza percorsa dalla caratteristica in un tempo  $k$  deve essere più piccola del passo spaziale  $h$ . Se la caratteristica cade fuori dall'intervallo allora il metodo diventerà instabile. Questa condizione viene chiamata condizione CFL (Courant-Friedrich-Levy). C.N. per la stabilità è che il dominio di dipendenza analitico sia interno al dominio di dipendenza numerico.

Introduciamo un operatore definito come:

$$L_{h,k}u = \frac{u(x, t+k) - u(x, t)}{k} + c \frac{u(x, t) - u(x-h, t)}{h},$$

facciamo lo sviluppo di Taylor

$$u(x+h, t) = u(x) + hu_x + \frac{1}{2}h^2u_{xx} + \mathcal{O}(h^3)$$

$$u(x, t+k) = u(x) + ku_t + \frac{1}{2}k^2u_{tt} + \mathcal{O}(k^3)$$

quindi abbiamo:

$$(L_{h,k} - L)u = \frac{1}{2}u_{tt}k - \frac{c}{2}u_{xx}h + \mathcal{O}(h^2) + \mathcal{O}(k^2)$$

metodo del primo ordine.

## Metodo Lax-Friedrichs

$$u_j^{n+1} = \frac{u_{j+1}^n + u_{j-1}^n}{2} - c \frac{k}{2h} (u_{j+1}^n - u_{j-1}^n),$$

Allora consideriamo l'operatore differenziale

$$L_{h,k}u = \frac{u(x, t+k) - \frac{u(x+h, t) - u(x-h, t)}{2}}{k} + c \frac{u(x+h, t) - u(x-h, t)}{2h},$$

La media

$$\text{media} = u + \frac{1}{2} h^2 u_{xx} + \mathcal{O}(h^4)$$

$$D.C. = hu_x + \mathcal{O}(h^3)$$

sostituendo:

$$L_{h,k}u(x, t) = \frac{u + ku_t + \frac{1}{2}u_{tt}k^2 + \mathcal{O}(k^3) - (u + \frac{1}{2}h^2u_{xx} + \mathcal{O}(h^4))}{k} + c(u_x + \mathcal{O}(h^2))$$

$$L_{h,k}u = \underbrace{u_t + cu_x}_{LU=0} + \underbrace{\frac{1}{2}ku_{tt} - \frac{h^2}{k}u_{xx} + \mathcal{O}(k^2, h^2)}_{d_{h,k}}$$

con  $d_{h,k}$  errore di discretizzazione del metodo. Se  $h \rightarrow 0$ ,  $k \rightarrow 0$ ,  $k/h = \mu = \text{const}$ , e, se  $u$  é soluzione di  $u_t + cu_x = 0$ .

Qui i termini di consistenza sono un po' strane, il metodo é consistente se e solo se  $h^2$  tende a zero piú velocemente di  $k$ . In tal caso é del primo ordine nello spazio e nel tempo.

Facendo un'analisi di stabilitá  $u_j^n = \rho^n e^{ijh\xi}$  abbiamo:

$$\rho = \frac{e^{ih\xi} + e^{-ih\xi}}{2} - \frac{ick}{h} \frac{e^{ih\xi} - e^{-ih\xi}}{2i} = \cos(\theta) - c \frac{k}{h} i \sin(\xi),$$

$$|\rho|^2 = \cos^2(\theta) + a^2 \sin^2(\xi) = 1 - (1 - a^2) \sin^2(\xi),$$

$$1 - a^2 \geq 0, \quad a^2 \leq 1, \quad -1 \leq \frac{ck}{h} \leq 1.$$

non c'è nessuna restrizione sul segno di  $c$ .

Ci sono metodi di alto ordine?

## Metodo do Lax-Wendroff

Sia

$$u(x, t + k) = u + u_t k + \frac{1}{2} u_{tt} k^2 + \mathcal{O}(k^3)$$

sfrutto l'equazione delle onde  $u_t + cu_x = 0$ :

$$u_{tt} = -c - u_{xt} = -cu_{tx} = c^2 u_{xx},$$

sostituendo abbiamo:

$$u(x, t + k) = u - kcu_x + \frac{c^2}{2} u_{xx} k^2 + \mathcal{O}(k^3)$$

vado a discretizzare lo spazio

$$u_j^{n+1} = u_j^n - kc \frac{u_{j+1}^n - u_{j-1}^n}{2h} + \frac{c^2}{2} k^2 \left( \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} \right).$$

Qual é l'ordine di discretizzazione che stiamo commettendo? Se  $L_{h,k}$  é l'operatore iperbolico discretizzato e  $L$  é l'operatore iperbolico continuo, abbiamo:

$$(L_{h,k} - L)u = \underbrace{\frac{1}{2} (u_{tt} - c^2 u_{xx})}_= k + \frac{1}{6} u_{ttt} k^2 + \frac{c}{6} u_{xxx} h^2 + \mathcal{O}(k^3) + \mathcal{O}(h^4)$$

metodo del secondo ordine globale.

## Analisi di Stabilitá.

Facendo sempre gli stessi calcoli abbiamo:

$$\rho = 1 + a^2(\cos(\theta) - 1) - ia \sin(\theta)$$

da cui

$$|\rho|^2 = 1 - a^2(1 - a^2)(1 - \cos(\theta))^2.$$

Quindi il metodo é stabile se e solo se

$$|c| \frac{k}{h} \leq 1.$$

## EQUAZIONE MODIFICATA

Un buon comportamento qualitativo degli schemi numerici può essere meglio capito con l'uso della equazione modificata associata agli scheme numerici. Abbiamo visto che:

$$L_{h,k}^{up} u = u_t + cu_x + \frac{1}{2} u_{tt} k - \frac{c}{2} u_{xx} h + \mathcal{O}(k^2, h^2)$$

$$L_{h,k}^{LxF} u = u_t + cu_x + \frac{1}{2} ku_{tt} - \frac{h^2}{k} u_{xx} + \mathcal{O}(k^2, h^2)$$

$$L_{h,k}^{LW} = u_t + cu_x + \frac{1}{2} \underbrace{(u_{tt} - c^2 u_{xx})}_{=0} k + \frac{1}{6} u_{ttt} k^2 + \frac{c}{6} u_{xxx} h^2 + \mathcal{O}(k^3) + \mathcal{O}(h^4)$$

Dall'equazione  $L_{h,k}^{LxF}$  abbiamo l'errore di discretizzazione, segue che se noi lo applichiamo alla funzione  $u$  soddisfacente l'equazione:

$$u_t + cu_x = \frac{1}{2} k \left( \frac{h^2}{k^2} u_{xx} - u_{tt} \right)$$

avremo  $L_{h,k}^{LxF} = \mathcal{O}(k^2, h^2)$ . Allo stesso ordine di accuratezza la funzione  $u$  soddisfa l'equazione

$$u_t + cu_x = \nu_{LxF} u_{xx},$$



con

$$\nu_{LxF} = \frac{1}{2} k \left( \frac{h^2}{k^2} - c^2 \right) u_{xx} = \frac{ch}{2} \frac{1 - \lambda^2}{\lambda}$$

con  $\lambda = ck/h$  e  $u_{tt} = c^2 u_{xx}$ .

Questo significa che la soluzione numerica del metodo di LxF applicato all'equazione  $u_t + cu_x = 0$  approssima ad un più alto ordine di accuratezza la soluzione dell'equ. di convezione e diffusione detta Equaz. Modificata, associata al metodo di LxF.

Il comportamento dello schema applicato all'equazione di convezione può essere qualitativamente descritto da questa equazione EM. In particolare osserviamo che

$$\mu_{LxF} \geq 0, \quad |\lambda| \leq 1, \quad \text{cond. CFL}$$

Quindi la condizione di stabilità per il metodo di LxF corrisponde alla condizione di buona posizione dell'equazione modificata.

Analogamente M. Up.

$$L_{hk} u = u_t + cu_x + ch(\lambda - 1)u_{xx} + \mathcal{O}(h^2 + k^2),$$



Eq. modificata del metodo upwind

$$u_t + cu_x = ch(1 - \lambda)u_{xx}$$

con  $\mu_{up} = ch(1 - \lambda) \geq 0$ , segue  $|\lambda| \leq 1$ .

Facciamo il rapporto

$$\frac{\mu_{up}}{\mu_{LxF}} = \frac{\lambda}{1 + \lambda} \in [0, 1/2],$$

caso limite  $\lambda = 1$ , e  $\mu_{up} \leq \mu_{LxF}$ , quindi entrambi gli schemi sono dissipativi, il coefficiente di diffusione è dell'ordine di  $h$  e scompare per  $\lambda = 1$ , ma il metodo upwind è meno dissipativo del metodo LxF.

$$L_{h,k}^{LW} = u_t + cu_x + \frac{1}{2}(u_{tt} - c^2 u_{xx})k + \frac{1}{6}(u_{ttt}k^2 + ch^2 u_{xxx}) + \mathcal{O}(k^3) + \mathcal{O}(h^4)$$

equazione modificata è

$$u_t + cu_x = \frac{1}{2}(c^2 u_{xx} - u_{tt})k + \frac{1}{6}(ch^2 u_{xxx} - u_{ttt}k^2)$$

allora da questa equazione uno ha

$$u_t = -cu_x + \frac{1}{2}(c^2 u_{xx} - u_{tt})k + \mathcal{O}(k^2)$$

quindi differenziando in tempo

$$u_{tt} = c^2 u_{xx} - \frac{1}{2}(c^3 u_{xxx} + u_{ttt})k + \mathcal{O}(k^2), \quad u_{ttt} = -c^3 u_{xxx} + \mathcal{O}(k)$$

ovvero segue

$$u_{ttt} + c^3 u_{xxx} = \mathcal{O}(k), \quad u_{tt} - c^2 u_{xx} = \mathcal{O}(k^2)$$

quindi abbiamo che per l'equazione modifiata segue:

$$u_t + cu_x = \frac{1}{6}(ch^2 - k^2 c^3)u_{xxx} + \mathcal{O}(k^3)$$

quindi  $u_t + cu_x = \mu_{LW} u_{xxx}$ , con  $\mu_{LW} = \frac{ch^2}{6}(\lambda^2 - 1)$ .

Questa equazione modificata é detta essere di carattere dispersivo, poiché piccole perturbazioni viaggiano con velocitá che dipende dalla frequenza. Quindi cerchiamo soluzioni di onde elementari viaggianti della forma

$$u(x, t) = \rho \exp(i(\kappa x - \omega t))$$

con  $i$  unitá immaginaria e  $\kappa$  é il numero d'onda. Questa é soluzione dell'E.M. se

$$-i\omega + i\kappa = -i\mu_{LW}k^3$$

cioé se  $\omega = c\kappa + \mu_{LW}k^3$ .

Il rapporto  $\omega/\kappa$  é detto velocitá di fase e la derivata  $v_g(\kappa) = \partial\omega/\partial\kappa$  é detta velocitá di gruppo e rappresenta la velocitá di propagazione di un pacchetto d'onda centrato al numero d'onda  $\kappa$ . La conseguenza del carattere dispersivo é che un profilo iniziale non viaggia imperturbato poiché le sue componenti di Fourier si muovono a velocitá differenti.

Caso LW  $\partial\omega/\partial\kappa = c + 3\mu_{LW}\kappa^3$ , ovvero finché le frequenze non sono grandi la velocitá di propagazione é  $c$ , ma per tempi lunghi  $\partial\omega/\partial\kappa \neq c$  la velocitá ha una dipendenza da  $k$ . Qual é la conseguenza nel nostro caso del metodo LW? Mentre il metodo di LW é piú accurato il comportamento dispersivo diventa drammatico nel caso di condizioni iniziali discontinue e delle oscillazioni appaiono nei profili.

## Schema Upwind per sistemi

Dall'equazione scalare  $u_t + cu_x = 0$  possiamo generalizzarla al sistema lineare al primo ordine:

$$u_t + Au_x = 0, \quad u(x, 0) = u_0(x),$$

con  $u : \mathbb{R} \times R \rightarrow \mathbb{R}^n$  e  $A \in \mathbb{R}^{n \times n}$  matrice costante. Il sistema é detto iperbolico se  $A$  é diagonalizzabile con autovalori reali, cosí che possiamo decomporla come

$$A = R\Lambda R^{-1}$$

con  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$  é una matrice diagonale di autovalori e  $R = [r_1, r_2, \dots, r_n]$  é la matrice degli autovettori destri. Notiamo che  $AR = R\Lambda$ , cioé

$$Ar_p = \lambda_p r_p, \quad \text{per } p = 1, 2, \dots, n$$

Il sistema é detto strettamente iperbolico se gli autovalori sono distinti.

Esprimendo adesso il vettore  $u$  come una combinazione lineare degli autovettori di  $A$ ,  $u = Rv$ ,  $v \in \mathbb{R}^n$  allora sostituendo nel sistema abbiamo:

$$v_t + \Lambda v_x = 0,$$

con  $v = R^{-1}u$  dette variabili caratteristiche. Ovvero le equazioni son tutte disaccopiate,

$$(v_p)_t + \lambda_p(v_p)_x = 0$$

con soluzione (eq. lineare di convezione)  $v_p(x, t) = v_p(x - \lambda_p t, 0)$ , il dato iniziale é dato da

$$v(x, 0) = R^{-1}u_0(x)$$

soluzione del sistema originale allora:  $u(x, t) = Rv(x, t)$ , ovvero  $u(x, t) = \sum_{p=1}^n v_p(x, t)r_p$  alla fine

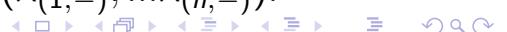
$$u(x, t) = \sum_{p=1}^n v_p(x - \lambda_p t, 0)r_p.$$

Adesso applichiamo il metodo upwind a ogni equazione scalare:

$$v_j^{n+1} = v_j^n - \frac{k}{h} (\Lambda_+(v_j^n - v_{j-1}^n) + \Lambda_-(v_{j+1}^n - v_j^n))$$

dove

$$\Lambda_+ = \text{diag}(\lambda_{(1,+)}, \dots, \lambda_{(n,+)}), \quad \Lambda_- = \text{diag}(\lambda_{(1,-)}, \dots, \lambda_{(n,-)}).$$



Usando la trasformazione  $u = Rv$ , abbiamo lo schema upwind per il sistema:

$$u_j^{n+1} = u_j^n - \frac{k}{h} (A_+(u_j^n - u_{j-1}^n) + A_-(v_{j+1}^n - v_j^n)),$$

con  $A_+ = R\Lambda_+R^{-1}$ ,  $A_- = R\Lambda_-R^{-1}$ .

La restrizione é che per tutti gli autovalori vale la relazione:

$$|\lambda_\ell| \frac{k}{h} \leq 1, \quad \ell = 1, \dots, n.$$

Questa condizione puó essere scritta come

$$\rho(A) \frac{k}{h} \leq 1,$$

dove  $\rho(A) = \max_{1 \leq j \leq n} |\lambda_j(A)|$ , raggio spettrale della matrice  $A$ .



# Lezioni Analisi Numerica II modulo

Sebastiano Boscarino

Dipartimento di Matematica e Informatica  
Università di Catania, ITALY

May 28, 2020

## Metodi Conservativi

Consideriamo l'equazione

$$u_t + f(u)_x = 0.$$

Se  $f(u)$  è non lineare, allora della discontinuità anche con C.I. regolari si può sviluppare in un tempo finito. Per soluzioni regolari questo sistema è equivalente a

$$u_t + Au_x = 0,$$

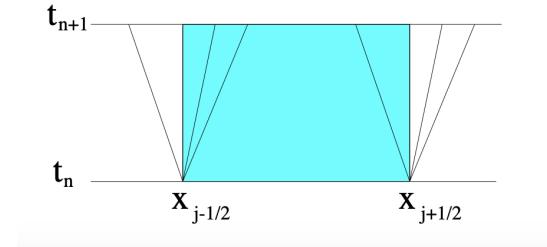
dove  $A = df/du := f'(u)$  detta matrice Jacobiana. Comunque non appena si sviluppano discontinuità le due forme non sono equivalenti. Per esempio, E. B.  $u_t + uu_x = 0$  e la sua forma conservativa  $u_t + (u^2/2)_x = 0$ .

Quindi una discontinuità si forma e si muove con una certa velocità  $s$  data dalla condizione di salto di R. H.

Inoltre la forma conservativa garantisce che  $\int_{-\infty}^{+\infty} u dt = const.$  con condizioni periodiche.

Quindi vogliamo che tali proprietà persistono anche a livello discreto , producendo la corretta velocità dipropagazione della discontinuità e sia conservativo.

Allora consideriamo un intervallo  $[a, b]$  e lo dividiamo in un numero  $N$  di celle  $I_j = [x_{j-1/2}, x_{j+1/2}]$  centrate in  $x_j$ ,  $j = 1, \dots, N$ . Quindi integriamo l'equazione  $u_t + f(u)_x = 0$  nella cella  $I_j$



$$\int_{x_{j-1/2}}^{x_{j-1/2}} u_t dx + \int_{x_{j-1/2}}^{x_{j-1/2}} f(u)_x = 0,$$

dividiamo per  $\Delta x$

$$\frac{d}{dt} \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j-1/2}} u dx + \frac{1}{\Delta x} [f(u(x_{j-1/2}, t)) - f(u(x_{j+1/2}, t))] = 0,$$

Poniamo

$$\bar{u}_j := \langle u \rangle_j = \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j-1/2}} u(x, t) dx$$

detta media di cella, abbiamo

$$\frac{d\bar{u}_j}{dt} + \frac{1}{\Delta x} [f(u(x_{j+1/2}, t)) - f(u(x_{j-1/2}, t))] = 0, \quad (1)$$



Quindi la forma precedente suggerisce di trovare un metodo numerico della forma:

$$\frac{d\bar{u}_j}{dt} = -\frac{1}{\Delta x} [F_{j+1/2} - F_{j-1/2}],$$

Questa espressione non è un metodo numerico, quindi si devono mettere in relazione i valori puntuali della  $u(x, t)$  con le medie di cella. In questo modo la  $F_{j+1/2}$  sarà una funzione di  $\bar{u}_j$  e  $\bar{u}_{j+1}$ ,

$$f(u(x_{j+1/2}, t)) \approx F_{j+1/2} = F(\bar{u}_j, \bar{u}_{j+1})$$

Quindi schemi di questa forma con la scelta del flusso numerico  $F(\cdot, \cdot)$  è uno schema semi-discreto. Quindi possiamo risolvere una discretizzazione temporale per il sistema di ODEs. Questo ci da una ampia scelta per il flusso numerico e la discretizzazione temporale (metodo delle linee).

Adesso se andiamo a integrare direttamente in tempo la (1) abbiamo:

$$\bar{u}_j^{n+1} = \bar{u}_j^n - \frac{1}{\Delta x} \int_{t_n}^{t_{n+1}} [f(u(x_{j+1/2}, t)) - f(u(x_{j-1/2}, t))] dt,$$

abbiamo uno schema full-discrete

$$\bar{u}_j^{n+1} = \bar{u}_j^n - \frac{\Delta t}{\Delta x} [F_{j+1/2}^n - F_{j-1/2}^n]$$

con

$$F_{j+1/2}^n = \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} [f(u(x_{j+1/2}, t)) - f(u(x_{j-1/2}, t))] dt$$



Il caso piú semplice é

$$F_{j+1/2}^n = F(\bar{u}_j^n, \bar{u}_{j+1}^n).$$

Gli schemi visti in precedenza (Upwind, LxF, LW) sono metodi conservativi con un opportuna scelta di  $F_{j+1/2}^n$ .

Per esempio il metodo upwind applicando all'equazione  $u_t + f(u)_x = 0$  con  $f'(u) > 0$ , sarà

$$\bar{u}_j^{n+1} = \bar{u}_j^n - \frac{\Delta t}{\Delta x} [f(\bar{u}_j^n) - f(\bar{u}_{j-1}^n)]$$

con

$$F_{j+1/2}^n = f(\bar{u}_j^n)$$

Per il metodo di LxF abbiamo

$$\bar{u}_j^{n+1} = \frac{1}{2} (\bar{u}_{j+1}^n + \bar{u}_{j-1}^n) - \frac{\Delta t}{2\Delta x} [f(\bar{u}_{j+1}^n) - f(\bar{u}_{j-1}^n)]$$

Allora aggiungo e sottraggo  $\bar{u}_j^n$  abbiamo:

$$\begin{aligned} \bar{u}_j^{n+1} &= \frac{1}{2} (\bar{u}_{j+1}^n - 2\bar{u}_j^n + \bar{u}_{j-1}^n) - \frac{\Delta t}{2\Delta x} [f_{j+1} - f_{j-1}] \\ &= \bar{u}_j^n - \frac{\Delta t}{2\Delta x} \left( -\frac{\Delta x}{\Delta t} (\bar{u}_{j+1}^n - 2\bar{u}_j^n + \bar{u}_{j-1}^n) + [f_{j+1} - f_{j-1}] \right) \end{aligned}$$



$$= \bar{u}_j^n - \frac{\Delta t}{2\Delta x} \left( (f_{j+1} - f_j + f_j - f_{j-1}) - \frac{\Delta x}{\Delta t} (\bar{u}_{j+1}^n - 2\bar{u}_j^n + \bar{u}_{j-1}^n) \right)$$

Allora con

$$F_{j+1/2}^n = \frac{f(\bar{u}_j^n) + f(\bar{u}_{j+1}^n)}{2} - \frac{\Delta t}{2\Delta x} (\bar{u}_{j+1}^n - \bar{u}_j^n) := F(\bar{u}_j^n, \bar{u}_{j+1}^n)$$

e

$$F_{j-1/2}^n = \frac{f(\bar{u}_j^n) + f(\bar{u}_{j-1}^n)}{2} - \frac{\Delta t}{2\Delta x} (\bar{u}_j^n - \bar{u}_{j-1}^n) := F(\bar{u}_{j-1}^n, \bar{u}_j^n)$$

Ovvero

$$\bar{u}_j^{n+1} = \bar{u}_j^n - \frac{\Delta t}{\Delta x} [F_{j+1/2}^n - F_{j-1/2}^n].$$

Caso schema LW. E' basato sullo sviluppo di Taylor

$$u(x, t+k) = u + ku_t + \frac{1}{2}k^2u_{tt} + \mathcal{O}(k^3)$$

da

$$\begin{aligned} u_{tt} &= -f(u)_{tx} = -\left(\frac{df}{du} \frac{du}{dt}\right)_x = -\left(\frac{df}{du} \left(-\frac{df}{dx}\right)\right)_x = \left(A \frac{df}{dx}\right)_x = (Af(u)_x)_x, \\ &= u - k(f(u) + \frac{1}{2}kA f(u)_x)_x + \mathcal{O}(k^3). \end{aligned}$$



Quindi trascurando i termini di ordine alto, possiamo scrivere lo schema in forma conservativo come:

$$\bar{u}_j^{n+1} = \bar{u}_j^n - \frac{\Delta t}{2\Delta x} [f(\bar{u}_{j+1}^n) - f(\bar{u}_{j-1}^n)] + \frac{\Delta t^2}{2} \left[ \frac{Af(u)_x|_{j+1/2} - Af(u)_x|_{j-1/2}}{\Delta x} \right].$$

Valutiamo  $Au_{j+1/2}$

$$\begin{aligned} & \frac{1}{2\Delta x} (Af(u)_x|_{j+1/2} - Af(u)_x|_{j-1/2}) = \\ & = \frac{1}{2\Delta x^2} (A_{j+1/2}(f(u)_{j+1} - f(u)_j) - A_{j-1/2}(f(u)_j - f(u)_{j-1})) \end{aligned}$$

questa é una discrittizzazione del secondo ordine (vedere equaz. del calore).

Allora sostituendo abbiamo:

$$F_{j+1/2}^n = \frac{f(\bar{u}_j^n) + f(\bar{u}_{j+1}^n)}{2} - \frac{\Delta t}{2\Delta x} A_{j+1/2}(f(\bar{u}^n)_{j+1} - f(\bar{u}^n)_j) := F(\bar{u}_j^n, \bar{u}_{j+1}^n)$$

ovvero

$$\bar{u}_j^{n+1} = \bar{u}_j^n - \frac{\Delta t}{\Delta x} [F_{j+1/2}^n - F_{j-1/2}^n].$$



**Proprietá di conservazione.** Consideriamo condizioni periodiche al bordo  $[a, b]$  e abbiamo  $\frac{d}{dt} \int_a^b u(x, t) dx = 0$  poiché  $u(a, t) = u(b, t)$ . Dividiamo l'intervallo in un numero  $N$  di sottointervalli,  $(b-a)/N$ . Se utilizziamo un metodo conservativo:

$$\bar{u}_j^{n+1} = \bar{u}_j^n - \frac{\Delta t}{\Delta x} [F_{j+1/2}^n - F_{j-1/2}^n],$$

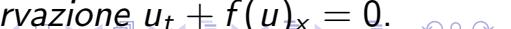
allora tale proprietá é mantenuta? Sommiamo su tutte le celle  $N$ ,

$$\begin{aligned} \sum_{j=1}^N \bar{u}_j^{n+1} &= \sum_{j=1}^N \bar{u}_j^n = \\ -\frac{\Delta t}{\Delta x} [F_{N+1/2}^n - F_{N-1/2}^n + F_{N-1/2}^n - F_{N-3/2}^n + \cdots + F_{3/2}^n - F_{1/2}^n] &= \sum_{j=1}^N \bar{u}_j^n, \end{aligned}$$

perché tutti termini intermedi sono cancellati e  $F_{N+1/2}^n = F_{1/2}^n$  per le condizioni di periodicitá  $F_{N+1/2}^n = F(\bar{u}_N, \bar{u}_{N+1}) = F(\bar{u}_0, \bar{u}_1) = F_{1/2}^n$ , e poiché  $\bar{u}_0 = \bar{u}_N$  e  $\bar{u}_{N+1} = \bar{u}_1$ .

La consistenza inoltre ci suggerisce che lo schema conservativo ci da la corretta velocitá di propagazione della discontinuitá e inoltre abbiamo in generale il Teorema di LW:

*Se un sistema discreto converge a una funzione  $u(x, t)$ , allora questa funzione é una soluzione debole della legge di conservazione  $u_t + f(u)_x = 0$ .*



Per uno schema conservativo per essere consistente bisogna richiedere:

$$L_\Delta u \rightarrow 0, \quad \text{as} \quad h, k \rightarrow 0, \quad h/k \quad \text{fixed.}$$

La condizione

$$F(v, v) = f(v), \quad \forall v \in \mathbb{R}^n$$

insieme a qualche assunzione di regolarità per la  $F$  (per esempio  $F$  Lipschitz continua in ciascuno argomento) assicura la consistenza dello schema

$$\bar{u}_{j+1}^n = \bar{u}_j^n - \frac{1}{\Delta x} [F_{j+1/2} - F_{j-1/2}],$$

con

$$F_{j+1/2}^n = F(\bar{u}_j^n, \bar{u}_{j+1}^n).$$

Infatti abbiamo

$$L_\Delta u(x, t) = \frac{u(x, t+k) - u(x, t)}{k} + \frac{F(u(x, t), u(x+h, t)) - F(u(x-h, t), u(x, t))}{h}$$

allora facciamo vedere che se:  $F(v, v) = f(v)$  allora vale la consistenza.

Usando espansione di Taylor, e assumendo la differenzialità di  $F$ ,

$$F(u(x, t), u(x+h, t)) = F(u, u) + hF'(u_1, u_2) + \mathcal{O}(h^2)$$

con  $F'(u_1, u_2) = \partial F / \partial u_2 \cdot \partial u / \partial x$  e

$$F(u(x-h, t), u(x, t)) = F(u, u) - hF'(u_1, u_2) + \mathcal{O}(h^2)$$

con  $F'(u_1, u_2) = \partial F / \partial u_1 \cdot \partial u / \partial x$  allora

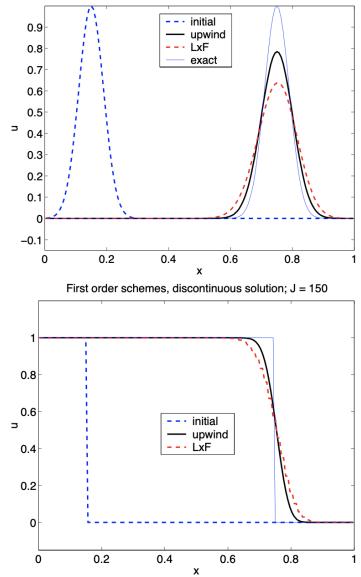
$$L_\Delta u = u_t + \left( \frac{\partial F}{\partial u_1} + \frac{\partial F}{\partial u_2} \right) u_x + \mathcal{O}(h, k)$$

Allora da  $F(v, v) = f(v)$  abbiamo

$$\frac{\partial F}{\partial u_1} + \frac{\partial F}{\partial u_2} = \frac{\partial f}{\partial u}$$

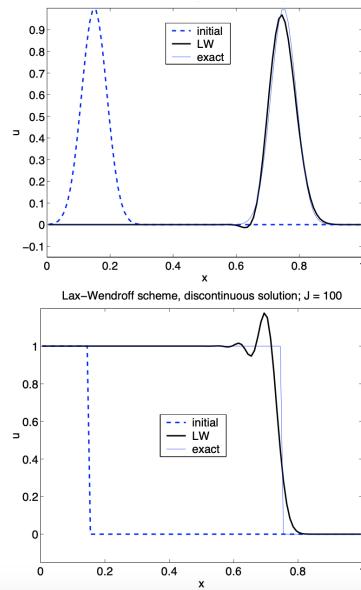
poiché  $u$  soddisfa l'equazione di partenza e quindi  $u_t + f'(u)u_x = 0$  allora segue  $L_\Delta = \mathcal{O}(h, k)$  per soluzioni regolari.

## Esempio. Upwind LxF



Dissipazione schemi primo ordine applicato al caso lineare. Top: Soluzione regolare. Bottom: Soluzione regolare,  $c = 1$ ,  $T_f = 0.6$ .  $N = 150$ ,  $CFL = 0.6$ .

## Esempio. LW



Dispersione schema secondo ordine applicato al caso lineare. Top: Soluzione regolare. Bottom: Soluzione regolare,  $c = 1$ ,  $T_f = 0.6$ .  $N = 150$ ,  $CFL = 0.6$ .

## Condizione di Entropia

E' ben noto che soluzioni deboli per leggi di conservazioni non sono uniche, anche nel caso scalare. L'unicitá puó essere recuperata se imponiamo una condizione aggiuntiva. Questa condizione aggiuntiva é detta condizione di entropia. Cosa é una entropia matematica?

Funzione di entropia puó essere definita sia per sistemi che eq. scalari per leggi di conservazione. Essa é una funzione convessa della variabile  $u$  che soddisfa una equazione addizionale di legge di conservazione per soluzioni regolari. Allora:

*Una funzione convessa  $\mu(u)$  é un'entropia associata all'equazione  $u_t + f(u)_x = 0$  se esiste una funzione  $\psi(u)$ , chiamato flusso di entropia. tale che per tutte le soluzioni regolari della equazione  $u_t + f(u)_x = 0$  la seguente equazione é soddisfatta*

$$\eta(u)_t + \psi(u)_x = 0,$$

Quando un sistema ammette entropia? Per soluzioni regolari la precedente equazione puó essere scritta come:

$$\eta'(u)u_t + \psi'(u)u_x = 0.$$

Relazione di compatibilitá con  $u_t + f'(u)u_x = 0$ , richiede che  $\psi'(u) = \eta'(u)f'(u)$ . Questa infatti é la condizione che é usata per costruire il flusso di entropia. Allora nel caso scalare  $\psi(u)$  é primitiva di  $\eta'(u)f'(u)$ .



Cosa garantisce l'esistenza di una funzione di entropia? L'esistenza di una funzione convessa di entropia é una proprietá molto importante per sistemi iperbolici di leggi di conservazione, perché se un sistema possiede una funzione convessa di entropia allora é simmetrizzabile, ovvero, esiste una trasformazione inversa del campo delle variabili tale che nelle nuove variabili il sistema é simmetrico.

Esempio. E. di B.

$$u_t + uu_x = 0, \quad u_t + (u^2/2)_x = 0,$$

moltiplichiamo la prima per  $u$  abbiamo

$$(\frac{1}{2}u^2)_t + (\frac{1}{3}u^3)_x = 0$$

consideriamo  $\eta(u) = \frac{1}{2}u^2$ ,  $\psi(u) = \frac{1}{3}u^3$ . Cosa succede nel caso discontinuo? Scegliamo funzione a gradino con  $u_L = 1$  e  $u_R = 0$ , come condizione iniziale. Il salto é 1 e la velocitá di propagazione per E-B. é,  $\dot{s} = 1/2$  (applicato a  $u_t + f(u)_x = 0$ ).

Mentre applichiamo all'equazione  $(\frac{1}{2}u^2)_t + (\frac{1}{3}u^3)_x = 0$ , abbiamo  $\dot{s}[\eta] = [\psi]$  allora

$$\dot{s} = \frac{[\psi]}{[\eta]} = \frac{2}{3}$$

ovvero sono diverse le velocitá caso discontinuo (invece sono le stesse nel caso soluzioni regolari, provare).



In generale partendo da una funzione di entropia convessa  $\eta(u)$  trovo  $\psi(u)$ , ma ho infinite coppie  $(\eta(u), \psi(u))$  che soddisfala relazione  $\eta_t + \psi_x = 0$ , come faccio a rispristinare l'unicità?

Un metodo per ripristinare l'unicità della soluzione per un'equazione scalare o sistema è di considerare la soluzione debole dell'equazione o del sistema come il limite di una sequenza di soluzioni dell'equazione regolarizzata,

$$u_t^\varepsilon + f(u)^\varepsilon_x = \varepsilon u_{xx}^\varepsilon$$

ovvero  $\lim_{\varepsilon \rightarrow 0} u^\varepsilon = u$ , per ogni  $\varepsilon > 0$ , tale soluzione  $u$  è detta soluzione viscosa dell'equazione  $u_t + f(u)_x = 0$ , con  $u^\varepsilon$  soluzione regolare dell'equazione di conv-diff.

Adesso consideriamo  $u_t + f(u)_x = \varepsilon u_{xx}$  e moltiplichiamo per  $\eta'(u)$  allora

$$\eta'(u)u_t + \eta'(u)f'(u)u_x = \varepsilon\eta'(u)u_{xx}$$

$$\eta'(u)u_t + \psi'(u)u_x = \varepsilon\eta'(u)u_{xx}$$

quindi

$$\eta(u)_t + \psi(u)_x = \varepsilon\eta'(u)u_{xx}$$

integriamo rispetto a  $x$ :

$$\begin{aligned} \frac{d}{dt} \int_a^b \eta(u)dx + \int_a^b \psi_x dx &= \varepsilon \int_a^b \eta'(u)u_{xx} dx \\ \frac{d}{dt} \int_a^b \eta(u)dx + [\psi_b - \psi_a] &= \varepsilon [\eta'(u)u_x]_a^b - \varepsilon \int_a^b \eta''(u)u_x^2 dx \end{aligned}$$

Se  $[a, b]$  intervallo finito e condizioni periodiche  $\varepsilon[\eta'(u)u_x]_a^b = 0$  se  $[a, b]$  è piccolo  $\varepsilon[\eta'(u)u_x]_a^b = 0$  è trascurabile rispetto a  $\varepsilon \int_a^b \eta''(u)u_x^2 dx$ , inoltre questa ultima quantità è sempre negativa perché  $\eta$  connessa  $\eta'' > 0$  quindi varrà sempre la diseguaglianza:

$$\eta(u)_t + \psi(u)_x \leq 0$$

qualunque sia  $\varepsilon$  e qualunque sia la soluzione viscosa. Tale diseguaglianza è detta diseguaglianza di Entropia.

La precedente relazione può essere integrata in senso debole

$$\int_0^T \int_a^b (\eta(u)\phi_t + \psi(u)\phi_x) dx dt + \int_a^b (\eta(u_0(x))\phi(x, 0)) dx \geq 0$$

con  $\phi(x, t)$  funzione test regolare a supporto compatto (nulla in  $x = a$ ,  $x = b$  e in  $t = T$ ).

Quindi se  $u$  é una soluzione di viscositá allora é una soluzione di entropia. Inoltre, si dimostra che se la disuguaglianza di entropia vale  $\forall \eta(u)$  convessa allora la soluzione che sto cercando é la soluzione di viscositá, ovvero ho l'unicitá (nel caso scalare é vero ma ancora é una questione aperta per i sistemi). Per alcuni sistemi di rilevanza fisica che ammettono un'entropia, l'unicitá della soluzione entropica, puó essere provata.

**Condizione di entropia discreta** Da

$$\eta(u)_t + \psi(u)_x \leq 0,$$

Alcuni schemi numerici possiedono una coppia di flusso di entropia che ci aiuta a selezionare la corretta soluzione di entropia.

La condizione discreta di entropia puó essere scritta come:

$$\eta(u_j^{n+1}) \leq \eta(u_j^n) - \frac{\Delta t}{\Delta x} [\Psi(\bar{u}_j^n, \bar{u}_{j+1}^n) - \Psi(\bar{u}_{j-1}^n, \bar{u}_j^n)],$$

dove  $\Psi(\bar{u}_r, \bar{u}) = \psi(u)$ . Ricordiamo che il teorema di L-W asserisce che: se una soluzione numerica dello schema conservativo applicato all'equazione scalare converge alla funzione  $u(x, t)$  allora la soluzione é una soluzione debole dell'equazione. Se lo schema possiede una disuguaglianza discreta di entropia, allora é garantito che la funzione  $u(x, t)$  é la soluzione entropica del problema.



## Il problema di Riemann

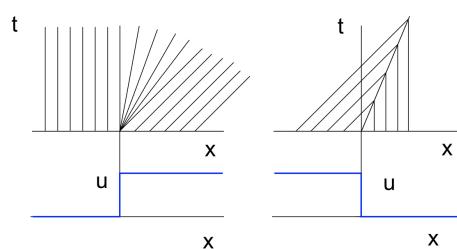
Come possiamo generalizzare il metodo upwind per equazioni non lineari di leggi di conservazione? Un metodo molto popolare é il metodo di Godunov. Questo metodo é basato sulla risoluzione del problema di Riemann.

Il problema di Riemann é un problema ai valori iniziali con dati iniziali costanti a tratti:

$$u_t + f(u)_x = 0,$$

$$u(x, 0) = \begin{cases} u_l, & x \leq 0, \\ u_r, & x > 0 \end{cases}$$

Per il caso scalare il problema di Riemann puó essere risolto esplicitamente. Per esempio per il caso  $f(u) = u^2/2$  (E.B.) noi abbiamo i due seguenti casi:



Di alcuni sistemi iperbolici di leggi di conservazione il problema di Riemann é ben conosciuto e risolto, come nel caso della gas dinamica. Ma in molti altri casi o non si conosce o é molto costoso da risolvere, e in questi casi si cerca di risolvere con degli risolutori approssimati di Riemann, o di schemi che non richiedono la risoluzione del problema di Riemann.

Per il momento noi supponiamo di conoscere la soluzione del problema Riemann

**schema di Godunov.** Assumiamo di conoscere una approssimazione della media di cella al tempo  $t^n$ ,  $\{\bar{u}_j^n\}$  e che la soluzione sia una funzione costanti a tratti

$$u(x, t^n) \approx \sum_j \bar{u}_j^n \chi_j(x)$$

dove  $\chi(x)_j = 1$ , per  $x \in [x_{j-1/2}, x_{j+1/2}]$ , e 0 altrove. Per brevi tempi il campo vettoriale  $u(x, t)$  sarà la soluzione di diversi problemi di Riemann (in ciascun intervallino) centrato in  $x_{j+1/2}$ . Integriamo la legge di conservazione nella cella  $I_j \times [t_n, t_{n+1}]$

$$\bar{u}_j^{n+1} = \bar{u}_j^n - \frac{1}{\Delta x} \int_{t_n}^{t^{n+1}} [f(u(x_{j+1/2}, t)) - f(u(x_{j-1/2}, t))] dt.$$

Adesso se il ventaglio di Riemann non interagisce (che si ottiene se  $\Delta t$  soddisfa la condizione CFL) allora la funzione  $u(x_{j+1/2})$  può essere ottenuta dalla soluzione del problema di Riemann con stati  $\bar{u}_j$  e  $\bar{u}_{j+1}$  attraverso l'interfaccia  $x_{j+1/2}$ :

$$u(x_{j+1/2}) = u^*(\bar{u}_j, \bar{u}_{j+1}).$$

Questa quantitá non dipende dal tempo, e quindi abbiamo

$$\bar{u}_j^{n+1} = \bar{u}_j^n - \frac{\Delta t}{\Delta x} [f(u^*(\bar{u}_j, \bar{u}_{j+1})) - f(u^*(\bar{u}_{j-1}, \bar{u}_j))] dt.$$

Se la funzione  $u(x, t^n)$  é realmente una funzione costante a tratti, allora la precedente equazione da la corretta media della soluzione al tempo  $t^{n+1}$ . Quando applichiamo il metodo di Godunov al caso lineare allora si riduce al metodo upwind. Il metodo di Godunov é un metodo al primo ordine in spazio e tempo.

Tale schema inoltre soddisfa la disuguaglianza discreta di entropia cioé

$$\eta(u_j^{n+1}) \leq \eta(u_j^n) - \frac{\Delta t}{\Delta x} [\Psi(\bar{u}_j^n, \bar{u}_{j+1}^n) - \Psi(\bar{u}_{j-1}^n, \bar{u}_j^n)].$$

## Ricostruzione non lineare e schemi di alto ordine

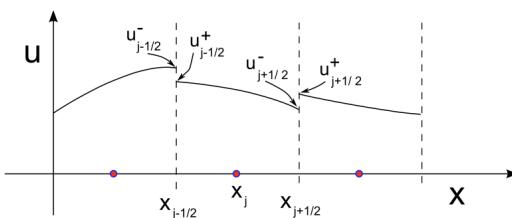
Abbiamo visto che la soluzione soddisfa un' equazione del tipo

$$\frac{d\bar{u}_j}{dt} + \frac{f(u(x_{j+1/2}, t)) - f(u(x_{j-1/2}, t))}{\Delta x} = 0,$$

dove  $\bar{u}_j = \frac{1}{\Delta x} \int_{I_j} u(x, t) dt$  con  $I_j = [x_{j-1/2}, x_{j+1/2}]$ . Un primo ordine schema semidiscreto può essere ottenuto usando  $F_{j+1/2} = F(\bar{u}_j, \bar{u}_{j+1})$  al posto di  $f(u(x_{j+1/2}, t))$ :

$$\frac{d\bar{u}_j}{dt} = -\frac{F(\bar{u}_j, \bar{u}_{j+1}) - F(\bar{u}_{j-1}, \bar{u}_j)}{\Delta x},$$

Uno schema basato su questa formula da uno schema al primo ordine. Schemi di alto ordine sono ottenuti usando ricostruzione polinomiali a tratti in ogni cella (caso costante a tratti è Godunov ma è del primo ordine) e valutare il flusso numerico sui due lati dell'interfaccia.



Questi sono ottenute come segue.

Date le medie di cella  $\{\bar{u}_j\}$ , valutiamo la seguente ricostruzione lineare a tratti

$$L(x) = \sum_j L_j(x) \chi_j(x),$$

con

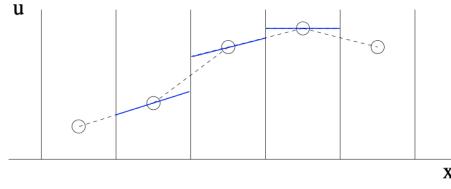
$$L_j(x) = \bar{u}_j + u'_j(x - x_j)$$

La quantità  $u'_j$  è una approssimazione al primo ordine della derivata spaziale del profilo  $u(x)$  in  $x_j$ .

L'approssimazione numerica della derivata prima è molto importante poiché le proprietà di accuratezza e TVD dello schema dipende da essa. Per esempio l'utilizzo della discretizzazione alle differenze centrali per la derivata, nella ricostruzione lineare a tratti produrrà delle oscillazioni spurie.

Per prevenire queste oscillazioni spurie, la derivata deve essere ricostruita attraverso un limitatore di flusso accettabile. Il più semplice limitatore è chiamato limitatore *minmod* definito come:

$$MinMod(a, b) = \begin{cases} a \text{ if } |a| \leq |b| \text{ and } ab > 0 \\ b \text{ if } |a| > |b| \text{ and } ab > 0 \\ 0 \text{ if } ab \leq 0 \end{cases} \quad (2)$$



Il limitatore minmod é molto robusto, ma ha svantaggio di degradare al primo ordine l'accuratezza dello schema vicino agli estremi locali. Una volta che il profilo é ricostruito allora la funzione agli estremi della cella é valutata come

$$u_{j+1/2}^- = L_j(x_{j+1/2}), \quad u_{j+1/2}^+ = L_{j+1}(x_{j+1/2})$$

MINMOD-limitatore:

$$\sigma_j = minmod \left( \frac{u_j - u_{j-1}}{h}, \frac{u_{j+1} - u_j}{h} \right)$$

Se  $\sigma_j = 0$  metodo upwind!

Se  $\sigma_j = \frac{u_{j+1} - u_j}{h}$  schema LW.

Metodi di risoluzione di Alto ordine

$$\frac{d\bar{u}_j}{dt} = -\frac{F_{j+1/2} - F_{j-1/2}}{h}$$

indicatore di regolaritá:

$$\theta_j = \frac{u_j - u_{j-1}}{u_{j+1} - u_j} \begin{cases} \approx 1 & \text{smooth,} \\ & \text{lontano da 1} \\ & \text{vicino shock} \end{cases} \quad (3)$$

Flusso:  $F_{j+1/2} = f_{j+1/2}^{low} + \phi(\theta_j)(f_{j+1/2}^{low} - f_{j-1/2}^{high})$

Condizioni TVD per la  $\phi(\theta)$ :

$$\theta \leq \phi(\theta) \leq 2\theta \quad (0 \leq \theta \leq 1), \quad 1 \leq \phi(\theta) \leq \theta \quad (1 \leq \theta \leq 2), \quad 1 \leq \phi(\theta) \leq 2 \quad (\theta \geq 2)$$

Secondo ordine  $\phi(\theta) = 1$ ,  $\phi$

Limitatori popolari:

$$Superbee : \phi(\theta) = \max(0, \min(1, 2\theta), \min(\theta, 2)), \quad Van Leer : \phi(\theta) = \frac{|\theta| + \theta}{1 + |\theta|}$$

Quando studiamo le proprietà di stabilità per un metodo numerico uno di solito verifica se per la soluzione numerica, le stesse proprietà della soluzione esatta si verificano.

Quindi per una soluzione entropica di una legge di conservazione soddisfacente le condizioni di entropia per la coppia  $(\eta, \psi)$ , verifica una serie di proprietà. Per esempio la TVD e la monotonicità.

Per la monotonicità abbiamo: Ogni coppia di soluzioni deboli  $u(x, t)$ ,  $v(x, t)$  con  $u_0(x) \geq v_0(x)$ ,  $\forall x$ , soddisfa:

$$v(x, t) \geq u(x, t), \forall x, t$$

Quindi nel costruire schemi numerici per leggi di conservazione si deve preservare tale proprietà.

Caso discreto della monotonicità

$$v_j^n \geq u_j^n, \rightarrow v_j^{n+1} \geq u_j^{n+1}.$$

Una semplice condizione sufficiente per provare la monotonicità è la seguente. Sia dato un metodo numerico definito dalla seguente funzione iterativa  $H$ :

$$u_j^{n+1} = H_j(u^n),$$

Allora se  $\forall j$ ,  $H_j(u)$  è una funzione non decrescente di tutti gli argomenti, allora è chiaro che:  $v_j^n \geq u_j^n \rightarrow H_j(v^n) \geq H_j(u^n)$ , e quindi lo schema è monotone.



La variazione totale (TV) di una funzione di valori reali è definita come segue

$$TV(u) = \sup \sum_{j=1}^N |v(\xi_{j-1}) - v(\xi_j)|$$

dove il sup è preso su tutte le possibili suddivisioni  $\xi_0, \xi_1, \dots, \xi_N$  di  $[a, b]$ . La variazione totale di una funzione è la misura del comportamento delle sue oscillazioni.

Proprietà TVD (Total variation diminishing)

$$TV(u(\cdot, t_2)) \leq TV(u(\cdot, t_1)), \quad t_2 \geq t_1,$$

dove  $\forall v \in L^1(a, b)$ ,  $TV(v)$  denota la variazione totale.

TVD significa che la totale quantità di oscillazioni decresce in tempo.

Caso discreto allora uno schema è TVD se

$$TV(u^{n+1}) \leq TV(u^n)$$

dove  $TV(u^n) = \sum_j |u_{j+1}^n - u_j^n|$ .



E' possibile provare che uno schema monotone é anche TVD.

E' facile provare che lo schema di Lax-Friedrichs é monotone.

Si puó provare che uno schema monotone é al massimo del primo ordine di accuratezza.

Per questa ragione si cercano schemi TVD che non sono monotonici ma che possono essere di alto ordine di accuratezza.

La nozione di flusso numerico monotone é molto importante nella derivazione di schemi di alto ordine. Un flusso monotone é un flusso associato ad uno schema monotone.

Proviamo che, se consideriamo un flusso monotone della forma

$$F_{j+1/2} = F(u_j, u_{j+1}),$$

se una condizione CFL é soddisfatta, e se  $F$  é non decrescente nel primo argomento e non crescente nel secondo argomento il corrispondente schema é monotono.

Infatti:

$$\begin{aligned} H_j(u) &= u_j - \frac{k}{h}(F(u_j, u_{j+1}) - F(u_{j-1}, u_j)); \\ \frac{\partial H_j}{\partial u_j} &= 1 - \frac{k}{h}\left(\frac{\partial F}{\partial u^{(1)}}(u_j, u_{j+1}) - \frac{\partial F}{\partial u^{(2)}}(u_{j-1}, u_j)\right); \\ \frac{\partial H_j}{\partial u_{j+1}} &= -\frac{k}{h}\frac{\partial F}{\partial u^{(2)}}(u_{j+1}, u_j) \geq 0; \\ \frac{\partial H_j}{\partial u_{j-1}} &= \frac{k}{h}\frac{\partial F}{\partial u^{(1)}}(u_{j-1}, u_j) \geq 0; \\ \frac{\partial H_j}{\partial u_\ell} &= 0 \quad \ell \neq j-1, j, j+1. \end{aligned}$$

Per la prima derivata il segno bisogna studiarlo separatamente. Per esempio upwind e Lax-Friedrichs soddisfano la condizione  $F(\uparrow, \downarrow)$ . Per upwind  $F(u_j, u_{j+1}) = f(u_j)$ , la condizione  $\frac{\partial H_j}{\partial u_j} \geq 0$  coincide con la CFL, mentre per Lax-Friedrichs  $\frac{\partial H_j}{\partial u_j} = 0$  e le altre due condizioni corrispondono alla CFL.

## Ricostruzione ENO (Essentially non oscillatory)

Ricostruzioni di alto ordine hanno lo scopo di fornire accuratezza di alto ordine in spazio. Questo step é molto importante per costruire metodi *shock capturing* perché ricostruzioni naive di alto ordine introducono oscillazioni spurie. Una ricostruzione che soddisfa questa proprietà é la ricostruzione ENO.

L'obiettivo é il seguente: assumiamo che esiste una funzione regolare  $u(x)$  e conosciamo solo le sue medie di cella  $\{\bar{u}_j\}$ . Allora noi vogliamo costruire in ogni cella  $I_j$  un polinomio  $p_j$  di un dato ordine  $m - 1$ , i.e.  $p_j \in \mathbb{P}^{(m-1)}$

$$p_j(x) = u(x) + \mathcal{O}(\Delta x^m).$$

In particolare siamo interessati a valutare il polinomio nei bordi cella:

$$u_{j+1/2}^- = p_j(x_{j+1/2}), \quad u_{j-1/2}^+ = p_j(x_{j-1/2})$$

Tale polinomio é costruito come segue. Prendiamo  $m$  celle adiacenti che includono la cella  $I_j$ . Denotiamo queste celle:  $j - r, j - r + 1, \dots, j, \dots, j + s$  con  $r + s + 1 = m$ ,  $r, s \geq 0$ . Quindi imponiamo che:

$$\langle p_j \rangle_\ell = \bar{u}_\ell := \frac{1}{\Delta x} \int_{x_{\ell-1/2}}^{x_{\ell+1/2}} u(x) dx, \quad \ell = j - r, \dots, j + s. \quad (4)$$

Queste  $m$  condizioni indipendenti determinano in modo univoco un polinomio di grado  $m - 1$ .



Facciamo vedere che questo polinomio é di grado  $m - 1$ . Definendo:

$$U(x) = \int_a^x u(\tau) d\tau$$

una primitiva di  $u(x)$ . L'estremo  $a$  non é rilevante. Scegliamolo tale che  $a = x_{j_a-1/2}$ . All'estremo destro abbiamo allora:

$$U(x_{i+1/2}) = \Delta x \sum_{j=j_a}^i \bar{u}_j.$$

Sia ora  $P_j(x) \in \mathbb{P}^m$  e sia  $P_j(x_{i+1/2}) = U(x_{i+1/2})$ ,  $i = j - r - 1, \dots, j + s$ , questi  $m + 1$  condizioni determinano unicamente  $P_j \in \mathbb{P}^m$ , e dalla teoria dell'interpolazione abbiamo:

$$P_j(x) = U(x) + \mathcal{O}(\Delta x^{m+1}),$$

e quindi

$$p_j(x) = P'_j(x) = U'(x) + \mathcal{O}(\Delta x^m) = u(x) + \mathcal{O}(\Delta x^m).$$

Il polinomio  $p(x)$  quindi soddisfa la condizione

$$p_j(x) = u(x) + \mathcal{O}(\Delta x^m),$$

e (4).



Ci sono  $m$  polinomi di questo tipo. Per esempio di grado 2 uno puó scegliere come celle:

$$j-2, j-1, j \quad j-1, j, j+1, \quad j, j+1, j+2.$$

Quale sceglieremo per la ricostruzione? L'idea ENO é la seguente. Prendiamo la cella  $I_j$  e costruiamo una funzione lineare tra i punti  $(x_{j-1/2}, U(x_{j-1/2}))$ ,  $(x_{j+1/2}, U(x_{j+1/2}))$ :

$$P_1(x) = U_{j-1/2} + U[x_{j-1/2}, x_{j+1/2}](x - x_{j-1/2})$$

con

$$U[x_{j-1/2}, x_{j+1/2}] = \frac{U(x_{j+1/2}) - U(x_{j-1/2})}{x_{j+1/2} - x_{j-1/2}} = \bar{u}_j,$$

Questa frazione rappresenta la media nella cella  $j$ -esima della  $u$ . Inoltre da  $p = P'$  abbiamo  $p_1(x) = \bar{u}_j$ .

Aggiungendo un altro punto a sinistra  $(x_{j-3/2}, U(x_{j-3/2}))$  oppure a destra  $(x_{j+3/2}, U(x_{j+3/2}))$  otteniamo:

$$R(x) = P_1(x) + U[x_{j-3/2}, x_{j-1/2}, x_{j+1/2}](x - x_{j-1/2})(x - x_{j+1/2})$$

oppure a destra

$$R(x) = P_1(x) + U[x_{j-1/2}, x_{j+1/2}, x_{j+3/2}](x - x_{j-1/2})(x - x_{j+1/2}).$$

Quindi quale delle due ricostruzioni devo scegliere? Quale stencil? Scelgo quello che é meno oscillante, cioé quello con la derivata seconda piú piccola. La piú piccola differenza divisa inoltre implica che la funzione é piú regolare in quello stencil allora prendiamo quello stencil con il piú piccolo valore assoluto, ovvero:

$$|U[x_{j-3/2}, x_{j-1/2}, x_{j+1/2}]| \leq |U[x_{j-1/2}, x_{j+1/2}, x_{j+3/2}]|$$

allora prendiamo come stencil  $S_3 = \{x_{j-3/2}, x_{j-1/2}, x_{j+1/2}\}$  altrimenti  $S_3 = \{x_{j-1/2}, x_{j+1/2}, x_{j+3/2}\}$

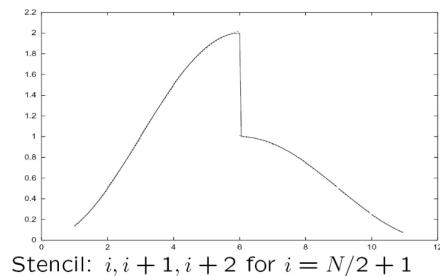
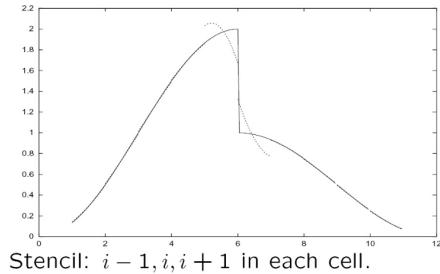
Notiamo che se vale ultimo stencil abbiamo:

$$U[x_{j-1/2}, x_{j+1/2}, x_{j+3/2}] = \frac{U[x_{j+1/2}, x_{j+3/2}] - U[x_{j-1/2}, x_{j+1/2}]}{2\Delta x} = \frac{\bar{u}_{j+1} - \bar{u}_j}{2\Delta x}$$

cioé non é necessario usare la funzione  $U$ .

Chiaramente possiamo ripetere la procedura introducendo altri punti allo stencil, a sinistra o a destra.

Supponiamo adesso che la funzione ha una discontinuità attraverso il bordo di una cella. La ricosistruzione  $R(x)$  nella parte alta è ottenuta usando una parabola nella cella  $I_j$ , prendendo i punti  $j - 1, j, j + 1$ , mentre quello in basso è presa utilizzando una procedura di tipo ENO. Tale procedura sceglie lo stencil basato sullo stencil  $j, j + 1, j + 2$ , ovvero lo stencil alla destra della discontinuità.



Allora l'effetto di questa procedura è di usare lo stencil che usa la parte più smooth della funzione nella ricostruzione.

Quindi per un dato grado  $m - 1$  ci sono abbiammo detto  $m$  possibilità di scelta di stencils. Per ogni uno di loro è possibile scegliere due set di coefficienti  $\{c_{ri}\}$   $\{\tilde{c}_{ri}\}$ , per valutare  $u_{j+1/2}^-$   $u_{j-1/2}^+$  come una combinazione lineare di medie di celle sullo stencil. Questi coefficienti sono valutati una sola volta e poi usati.

Uno volta scelto lo stencil, dalla procedura ENO, uno sa quale set di coefficienti  $c_{ri}$  usare. Le espressioni per ogni scelta dello stencil sono

$$u_{j+1/2}^- = \sum_{i=0}^{m-1} c_{ri} \bar{u}_{j-r+1}, \quad u_{j-1/2}^+ = \sum_{i=0}^{m-1} \tilde{c}_{ri} \bar{u}_{j-r+1}.$$

Nella ricostruzione ENO si scegli uno stencil con  $m$  nodi per ricostruire un polinomio di grado  $m - 1$ , i.e.  $m = 3$  per polinomio con grado  $m - 1 = 2$  e per avere un'accuratezza di ordine  $\mathcal{O}(h^m)$  nella cella  $I_j$ . Comunque ci vogliono in totale  $2m - 1$  punti ( $m = 5$ , i.e.  $x_{i-2}, x_{i-1}, x_i, x_{i+1}, x_{i+2}$ ). Ma con tutti questi punti è possibile avere una accuratezza maggiore, WENO. Sia caso  $m = 3$ , e usiamo una parabola per ricostruire la funzione  $u(x)$  nella cella  $I_j$ . Sia  $q_k$  tale parabola ottenuta ottenuta sulla celle  $k - 1, k, k + 1$ , i.e.  $q_k(x)$  è ottenuta dall'imporre:

$$\langle q_k \rangle_\ell = \bar{u}_\ell, \quad \ell = k - 1, k, k + 1.$$

Quindi per il polinomio  $p_2 \in \mathbb{P}^2$  possiamo usare le parbole  $q_{j-1}, q_j, q_{j+1}$ . Ogni scelta da' un'accuratezza del terzo ordine. Noi inoltre possiamo scegliere una combinazione convessa di  $q_k$

$$p_j = w_{-1}^j q_{j-1} + w_0^j q_j + w_1^j q_{j+1},$$

con  $w_{-1}^j + w_0^j + w_1^j = 1$   $w_\ell^j \geq 0$ ,  $\ell = -1, 0, 1$ . Ogni combinazione del genere da' una approssimazione del terzo ordine. Come scegliamo i pesi  $w_\ell^j$ ?

- ▶ Nella regione di regolarità della  $u(x)$  il valore dei pesi sono scelti tale tale che di avere una ricostruzione della funzione in qualche particolare punto alto ordine di accuratezza. Tipicamente noi vogliamo alto ordine in  $x_j + h/2, x_j - h/2$ . (Notiamo che due gradi di libertà in più abbiamo un quinto ordine di accuratezza nel punto  $x_{j+1/2}$ , i.e.  $q_{j-1}, (j-2, j-1, j), q_j, (j-1, j, j+1), q_{j+1}, (j, j+1, j+2)$ , totale 5 punti.) Denotiamo  $C_{-1}^+, C_0^+, C_1^+$ , delle costanti che forniscono l'accuratezza di alto ordine nel punto  $x_{j+1/2}$ :

$$p_j(x_{j+1/2}) = \sum_{-1}^1 C_k^+ q_{j+1/2} = u(x_{j+1/2}) + \mathcal{O}(h^5)$$

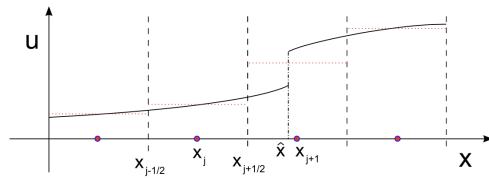
analogamente:

$$p_j(x_{j-1/2}) = \sum_{-1}^1 C_k^- q_{j-1/2} = u(x_{j-1/2}) + \mathcal{O}(h^5)$$

I valori di queste costanti sono state valutate e sono:

$$C_1^+ = C_{-1}^- = 3/10, \quad C_0^+ = C_0^- = 3/5, \quad C_{-1}^+ = C_1^- = 1/10,$$

- ▶ Nella regione vicino la discontinuità, dovremmo fare uso solo dei valori delle celle medie che appartengono alla parte regolare del profilo. Consideriamo la figura.



Supponiamo che la funzione  $u(x)$  ha la discontinuità in  $\hat{x} \in I_{j+1}$ . Allora per ricostruire la funzione nella cella  $I_j$  a uno piacerebbe fare uno solo della parabola  $q_{j-1}$ , ovvero con dei pesi:

$$w_{-1}^j \approx 1, w_0^j \approx 0, w_1^j \approx 0.$$

Questo è possibile ottenerlo facendo dipendere i pesi dalla regolarità della funzione nella cella corrispondente. Nella procedura WENO questo è ottenuto prendendo:

$$\alpha_j = \frac{C_k}{\beta_k^j + \epsilon}, \quad k = -1, 0, 1 \quad w_k^j = \frac{a_k^j}{\sum_\ell \alpha_\ell^j}.$$

$\beta_k$  sono detti indicatori di regolarità e sono usati per misurare la regolarità oppure no della funzione:

$$\beta_k^j = \sum_{\ell=1}^2 \int_{x_{j-1/2}}^{x_{j+1/2}} h^{2\ell-1} \left( \frac{d^\ell q_{j+k}(x)}{dx^\ell} \right), \quad k = -1, 0, 1$$

Queste integrazioni sono state calcolate esplicitamente per esempio caso parbole abbiamo:

$$\beta_{-1} = \frac{13}{12}(\bar{u}_{j-2} - 2\bar{u}_{j-1} + \bar{u}_j)^2 + \frac{1}{4}(\bar{u}_{j-2} - 4\bar{u}_{j-1} + 3\bar{u}_j)^2.$$

Quindi con tre parabole abbiamo un'accuratezza al quinto ordine nelle zone regolari e un terzo ordine in vicinanza della discontinuità.

### Scheme alle differenze finite conservativi

Schemi alle differenze non si considera la media di cella della funzione, ma si considerano i valori puntuali della  $u(x)$ , i.e.  $u_i \approx u(x_i)$ .

Consideriamo allora il sistema:  $u_t + f(u)_x = 0$  e scriviamo:

$$f(u(x))_x = \frac{\hat{f}(u(x + h/2)) - \hat{f}(u(x - h/2))}{h}.$$

La relazione tra  $f$  e  $\hat{f}$  è la seguente.

Se consideriamo la cella media:

$$\bar{u}(x) = \frac{1}{h} \int_{x-h/2}^{x+h/2} u(\xi) d\xi.$$

differenziamo rispetto a  $x$ :

$$\bar{u}_x = \frac{(u(x + h/2) - u(x - h/2))}{h}.$$

Quindi la relazione tra  $f$  e  $\hat{f}$  é la stessa tra  $\bar{u}$  e  $u$ , ovvero la funzione  $f$  é la cella media di  $\hat{f}$ . Quindi questo ci suggerisce un modo per valutare la funzione flusso.

Allora negli schemi a differenze finite la tecnica qui é di valutare la funzione flusso in  $x_j$  e poi ricostruire in  $x_{j+1/2}$ , ovvero valutare  $\hat{f}(u(x_{j+1/2}))$  da  $f(u(x_j))$ . Ma la ricostruzione in  $x_{j+1/2}$  puó essere discontinua, quale valore usare?

Una risposta in generale puó essere data se si considera uno split della funzione  $f$ :

$$f(u) = f^+(u) + f^-(u),$$

con le condizioni (per flussi scalari) che

$$\frac{df^+(u)}{du} \geq 0 \quad \frac{df^-(u)}{du} \leq 0.$$

Infatti se una funzione  $f$  puó essere scritta come splitting di due funzioni con quelle proprietá allora

$$F(u, v) = f^+(u) + f^-(v),$$

noi definiamo un flusso monotone consistente.



Questo é il caso per esempio del flusso locale di Lax-Friedrichs (conosciuto anche come metodo di Rusanov):

$$F(u, v) = \frac{1}{2}(f(u) + \alpha u) + \frac{1}{2}(f(v) - \alpha v), \quad \alpha = \max_w |f'(w)|.$$

### procedura schemi conservativi alle differenze finite



$$\frac{du_j}{dt} = -\frac{1}{h}[\hat{F}_{j+1/2} - \hat{F}_{j-1/2}]$$

con

$$\hat{F}_{j+1/2} = \hat{f}^+(u_{j+1/2}^-) + \hat{f}^-(u_{j+1/2}^+)$$

► dove  $\hat{f}^+(u_{j+1/2}^-)$  é ottenuta come segue:

- 1) Valutiamo  $f^+(u_\ell)$  e interpretiamola come la cella media di  $\hat{f}^+$ ,
- 2) eseguiamo la ricostruzione di  $\hat{f}^+$  nella cella  $j$  e valutiamola in  $x_{j+1/2}$ .
- 3) Analogamente per  $\hat{f}^-(u_{j+1/2}^+)$ .

Osserviamo che per metodi ai volumi finiti si possono usare griglie non uniformi, mentre per metodi alle differenze finite si possono usare solo griglie uniformi.



# Integrazione temporale: metodi SSP-RK

In generale sistemi iperbolici non sono stiff, quindi non è necessario usare metodi di tipo implicito per l'integrazione temporale.

Quando si costruiscono schemi per leggi iperboliche bisogna stare attenti ad avere schemi, in particolare di alto ordine che evitano oscillazioni spurie vicino alle discontinuità o shocks. Questo è possibile non solo utilizzando una buona discretizzazione spaziale ma anche temporale.

Soluzione di legge di conservazione scalare e equazioni con sorgenti dissipative possiedono qualche norma che decresce nel tempo. Vorremo che tale proprietà si mantenesse a livello discreto.

Quindi sia  $U^n$  un vettore di soluzioni (per esempio ottenuto da un metodo delle linee) diamo la seguente definizione:

*Una sequenza  $\{U^n\}$  è detta essere fortemente stabile in una data norma  $\|\cdot\|$  se  $\|U^{n+1}\| \leq \|U^n\|$  per ogni  $n \geq 0$ .*

Le norme più comunemente usate sono la TV e la norma infinito. Uno schema numerico che mantiene tale proprietà è detto *Strong stability preserving (SSP)*

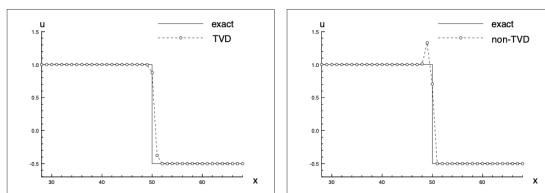


Legge di conservazione  $u_t = -f(u)_x$ , method of lines

$$U' = L(U)$$

Caso Burger's equation  $u_t + (u^2/2)_x = 0$ , initial condition

$$u(x, 0) = \begin{cases} 1, & x \leq 0, \\ -0.5 & x > 0 \end{cases}$$



Time discretization: SSP2 RK

$$\begin{aligned} U^{(1)} &= U^n + \Delta t L(U^{(n)}), \\ U^{n+1} &= \frac{1}{2} U^n + \frac{1}{2} U^{(1)} + \frac{1}{2} \Delta t L(U^{(1)}) \end{aligned} \quad (5)$$

Non SSP:

$$\begin{aligned} U^{(1)} &= U^n - 20 \Delta t L(U^{(n)}), \\ U^{n+1} &= U^n + \frac{41}{40} U^{(1)} - \frac{1}{40} \Delta t L(U^{(1)}) \end{aligned} \quad (6)$$



Notiamo che da (5) abbiamo, sostituendo  $U^{(1)}$  in  $U^{n+1}$ :

$$U^{n+1} = \frac{1}{2}U^n + \frac{1}{2}(U^n + \Delta t L(U^{(n)})) + \frac{1}{2}\Delta t L(U^{(1)})$$

segue Heun method nella rappresentazione di RK Butcher

$$U^{n+1} = U^n + \frac{1}{2}\Delta t \left( L(U^{(n)}) + L(U^{(1)}) \right).$$

Si puó provare che, innanzitutto schemi di alto ordine SSP, sono tutti esplicativi, e un generico schema esplicito SSP sono riscritti come:

$$\begin{aligned} U^{(0)} &= U^n \\ U^{(i)} &= \sum_{k=0}^{i-1} \left( \alpha_{ik} U^{(k)} + \Delta t \beta_{ik} L(U^{(k)}) \right), \quad i = 1, \dots, s \\ u^{n+1} &= U^{(s)} \end{aligned}$$

dove  $\alpha_{ik} \geq 0$ , e  $\alpha_{ik} = 0$  solo se  $\beta_{ik} = 0$ . Osserviamo che per consistenza si ha  $\sum_{k=0}^{i-1} \alpha_{ik} = 1$ . Segue anche che se lo schema numerico puó essere scritto in questa forma con coefficienti non negativi  $\beta_{ik}$ , allora puó essere visto come una combinazione convessa di passi di Eulero esplicito con passi temporale  $(\beta_{ik}/\alpha_{ik})\Delta t$ .

Una conseguenza di questo fatto é: se il metodo di Eulero é SSP per  $\Delta t \leq \Delta t^*$ , allora lo schema esplicito RK é anche SSP per  $\Delta t \leq c\Delta t^*$  con  $c = \min_{ik}(\alpha_{ik}/\beta_{ik})$ ,  $c$  é una misura dell'efficienza del metodo quindi 1'e opportuno avere  $c$  la piú grande possibile.

Questa rappresentazione di schemi RK, che non é unica, puó essere convertita in uno standard tableau di Butcher (Runge-Kutta esplicativi).