METODI SPETTRALI PER L'EQ. DEL CALORE

La principale caratteristica dei metodi spettrali, che li rende diversi dagli elementi finiti, è che essi utilizzano polinomi globali sul dominio computazionale anziché polinomi a tratti. In particolare grazie alla scelta di polinomi opportuni ed alla regolarità della funzione incognita possono raggiungere velocità di convergenza superiori a quelle degli usuali approcci agli elementi finiti e differenze finite.

Allo scopo di illustrare in modo diretto i metodi spettrali consideriamo l'eq. del calore in una variabile di spazio

$$u_t = u_{xx}$$

$$u(x,0) = u_0(x)$$

e condizioni al bordo periodiche, ossia

$$u(0,t) = u(L,t).$$

Inoltre ha u_0 t.c. $u_0(0) = u_0(L)$.

Possiamo rappresentare la soluzione in serie di Fourier

$$(1)\quad \begin{cases} \hat{u}_k(t) = \dfrac{1}{2\pi}\displaystyle\int_0^{2\pi} u(x,t)\, e^{-ikx}\, dx \\ u(x,t) = \displaystyle\sum_{k=-\infty}^{\infty}{}' \hat{u}_k(t)\, e^{ikx} \end{cases}$$

dove abbiamo scelto l'intervallo $[0, 2\pi]$ per la rappresentazione ed u periodica sullo stesso intervallo. Analogamente, basta riscalare le serie (1) ottenuta la rappresentazione sull'intervallo $[0, L]$ dove al posto di x nei termini esponenziali compare il fattore $\frac{2\pi}{L}x$.

Utilizzando la (1) ed il fatto che

$$u_{xx} = \sum_{k=-\infty}^{\infty}{}' \hat{u}_k(t) \cdot \frac{d^2}{dx^2} e^{ikx}$$

$$\frac{d^2}{dx^2} e^{ikx} = (ik)^2 e^{ikx} = -k^2 e^{ikx}$$

otteniamo

$$\sum_{k=-\infty}^{\infty}{}' \frac{d\hat{u}_k}{dt} e^{ikx} = -\sum_{k=-\infty}^{\infty}{}' \hat{u}_k k^2 e^{ikx}.$$

Identificando termine a termine delle espressioni appena trovate si ha

$$(*)\quad \begin{cases} \dfrac{d\hat{u}_k}{dt} = -k^2 \hat{u}_k \quad , \quad -\infty < k < \infty \\ \hat{u}_k(0) = \hat{u}_{0k} \end{cases}$$

La soluzione si calcola facilmente ed è data da

$$\hat{u}_k(t) = \hat{u}_{0k}\, e^{-k^2 t} \quad , \quad \forall k$$

che sostituita nell'espressione originale della serie di Fourier porta alla nota soluzione analitica

$$(2)\quad u(x,t) = \sum_{k=-\infty}^{\infty} \hat{u}_{0k}\, e^{-k^2 t}\, e^{ikx} .$$

Numericamente non sempre è possibile utilizzare direttamente l'espressione (2) in quanto bisogna essere in grado di sommare la serie. In generale dunque si approssima (2) con la serie troncata ad N termini

$$P_N u = \sum_{k=-N/2}^{N/2-1} \hat{u}_{0k}\, e^{-k^2 t}\, e^{ikx} .$$

Si noti che l'operatore P_N che opera il troncamento è un operatore di proiezione dallo spazio a punto divergente in cui si trova la funzione u ad uno spazio di dimensione finita generato dalla base dei

polinomi trigonometrici

$$\left\{ e^{ikx} \right\} \quad , \quad k = -\frac{N}{2}, \dots, \frac{N}{2} - 1 .$$

In generale dunque data una funzione u periodica su $[0, 2\pi]$ e quadrato integrabile nello stesso intervallo ossia $u \in L^2([0,2\pi])$ allora indicha con

$$u_N = P_N u$$

ed ne rappresentazione nello spazio di polinomi generati da $\left\{ \phi_k \right\}$, $k = -\frac{N}{2}, \dots, \frac{N}{2} - 1$,

con $\phi_k = e^{ikx}$. Indicheremo tale spazio con $\mathbb{P}^N$.

Dunque $P_N : L^2([0,2\pi]) \to \mathbb{P}^N$. Introduciamo inoltre il prodotto interno in $L^2([0,2\pi])$ definito da

$$\langle u, v \rangle = \frac{1}{2\pi} \int_0^{2\pi} u(x)\, v(x)\, dx .$$

Allora quindi (1) è equivalente a

$$\hat{u}_k(t) = \langle u, \phi_k \rangle$$

ed il problema (*) risulta

$$\langle u_t, \phi_K \rangle = \langle u_{xx}, \phi_K \rangle$$

analoga alla forma debole vista per gli elementi finiti.
Il metodo numerico visto è detto di Fourier-Galerkin
e può essere riscritto tutto l'operatore di proiezione
come

$$\begin{cases} \dfrac{\partial u_N}{\partial t} = \dfrac{\partial^2 u_N}{\partial x^2} \\ u_N(x,0) = u_{0N}(x) \end{cases}$$

da cui eseguendo il prodotto interno si ottiene

$$\begin{cases} \dfrac{d(\hat{u}_N)_K}{dt} = -K^2 (\hat{u}_N)_K \qquad -N/2 \leq K \leq N/2 - 1 \\ (\hat{u}_N)_K(0) = (\hat{u}_{0N})_K \end{cases}$$

che risolto esattamente fornisce

$$(3) \qquad u_N(x,t) = \sum_{K=-N/2}^{N/2-1} (\hat{u}_{0N})_K \, e^{-K^2 t} \, e^{iKx}$$

che coincide esattamente con il troncamento ai primi N
termini della soluzione esatta (2) del problema.
Questo risultato è vero per tutte le equazioni lineari a
coefficienti costanti

CENNI SULLA CONVERGENZA SPETTRALE

Indichiamo con f una generica funzione di $L^2([0,2\pi])$ e con $P_N f$ la sua serie di Fourier troncata.

Vogliamo dimostrare che $P_N f \to f$ per $N \to \infty$ in una norma opportuna.

Ad esempio in norma L^2, ossia che

$$\int_0^{2\pi} |f - P_N f|^2 dx \to 0 \qquad \text{per } N \to \infty.$$

Vale l'<u>identità di Parseval</u>

$$\int_0^{2\pi} |f|^2 dx = \sum_{k=-\infty}^{\infty} |\hat{f}_k|^2$$

Da questa possiamo stimare l'errore in norma L^2

$$\int_0^{2\pi} |f - P_N f|^2 dx = \int_0^{2\pi} \Big| \sum_{\substack{k<-N/2 \\ k>\frac{N}{2}-1}} \hat{f}_k e^{ikx} \Big|^2 dx = \sum_{\substack{k<-N/2 \\ k>\frac{N}{2}-1}} |\hat{f}_k|^2$$

Per l'errore puntuale invece

$$|f(x_0) - P_N f(x_0)| = \Big| \sum_{\substack{k<-N/2 \\ k>\frac{N}{2}-1}} \hat{f}_k e^{ikx} \Big| \le \sum_{\substack{k<-N/2 \\ k>\frac{N}{2}-1}} |\hat{f}_k|$$

Dunque in entrambi i casi l'errore dipende dal comportamento dei coefficienti di Fourier al crescere di $|k|$.

Osserviamo che se la funzione è regolare allora è possibile ottenere stime molto migliori. Infatti se f è derivabile con continuità abbiamo integrando per parti

$$\hat{f}_k = \frac{1}{2\pi}\int_0^{2\pi} f(x)\, e^{-ikx}\, dx = -\frac{1}{2\pi i k}\left(f(2\pi)-f(0)\right) +$$

$$+\frac{1}{2\pi i k}\int_0^{2\pi} f'(x)\, e^{-ikx}$$

Passando ai valori assoluti e ricordando che f è periodica si ha

$$|\hat{f}_k| \le \frac{1}{2\pi|k|}\int_0^{2\pi}|f'(x)|\,dx = O\left(\frac{1}{|k|}\right)$$

ossia i coefficienti decadono linearmente rispetto a $|k|$.

Se ora il polinomio f derivabile con continuità p volte, iterando il procedimento otteniamo

$$|\hat{f}_k| \le \frac{1}{2\pi|k|^p}\int_0^{2\pi}|f^{(p)}(x)|\,dx = O\left(\frac{1}{|k|^p}\right)$$

Dove x è funzione e' infinitamente derivabile allora i coefficienti di Fourier decadono più rapidamente di qualsiasi potenza di $|k|$. Questo comporta che l'errore commesso dalla serie di Fourier troncata decade più rapidamente di qualsiasi potenza di N. Si parla in questo caso di <u>convergenza esponenziale</u> o di <u>accuratezza spettrale</u>.

<u>OSS</u>: In pratica i coefficienti di Fourier sono calcolati tramite la trasformata di Fourier discreta (DFT) che corrisponde a

$$\tilde{u}_k = \frac{1}{N} \sum_{j=0}^{N} u(x_j)\, e^{-ik\frac{2\pi}{N}j}$$

con $x_j = \frac{2\pi}{N} j$.

Tale trasformata per calcolare N coefficienti $\tilde{u}_k$ comporterebbe il costo proibitivo di N^2. Tramite l'algoritmo della Fast Fourier Transform (FFT) il costo può essere ridotto a $N \log_2 N$.

ESTENSIONE A PROBLEMI NON LINEARI

Consideriamo l'equazione di Burgers viscosa

$$(1)\quad \begin{cases} u_t + \left(\frac{u^2}{2}\right)_x = u_{xx} \\ u(0,t) = u(2\pi,t) \\ u(x,0) = u_0(x) \end{cases}$$

Ripercorriamo in particolare i passi che servono per la costruzione di un metodo di tipo Fourier-Galerkin.

Per i coefficienti di Fourier osserviamo innanzitutto che se indichiamo con

$$h = \left(\frac{u^2}{2}\right)_x$$

abbiamo

$$h = \left(\frac{u^2}{2}\right)_x = \frac{d}{dx}\left[\frac{1}{2}\sum_{\ell=-\infty}^{\infty} \hat{u}_\ell e^{i\ell x} \cdot \sum_{m=-\infty}^{\infty} \hat{u}_m e^{imx}\right]$$

$$= \frac{d}{dx}\left[\frac{1}{2}\sum_{\substack{k=-\infty \\ k=\ell+m}}^{\infty}\ \sum_{m=-\infty}^{\infty} \hat{u}_{k-m}\,\hat{u}_m\, e^{ikx}\right]$$

$$= \frac{1}{2} \sum_{k=-\infty}^{\infty}{}' k \sum_{m=-\infty}^{\infty}{}' \hat{u}_{k-m} \hat{u}_m e^{ikx}$$

$$= \sum_{k=-\infty}^{\infty}{}' \hat{h}_k e^{ikx}$$

con $$\hat{h}_k = \frac{1}{2} k \sum_{m=-\infty}^{\infty}{}' \hat{u}_{k-m} \hat{u}_m$$

ossia è rappresentabile da una convoluzione discreta.

In Fourier abbiamo dunque

$$(2) \quad \begin{cases} \dfrac{d\hat{u}_k}{dt} + \hat{h}_k = -k^2 \hat{u}_k & -\infty < k < \infty \\ \hat{u}_k(0) = \hat{u}_{0k} \end{cases}$$

Al solito per ottenere un metodo numerico dobbiamo considerare le serie troncate ad un numero finito di termini. In questo caso però non troncheremo solo la rappresentazione di u ma anche la rappresentazione di h in quanto la convoluzione discreta deve anch'essa essere proiettata sullo spazio finito dimensionale.

Tale ulteriore approssimazione, che non analizzeremo in dettaglio, non altera comunque le proprietà di accuratezza del metodo.

In breve, il metodo di Fourier-Galerkin diventa

$$(3)\quad \begin{cases} \dfrac{\partial \bar{u}_N}{\partial t} + P_N \dfrac{\partial}{\partial x}\left(\dfrac{\bar{u}_N^2}{2}\right) = \dfrac{\partial^2 \bar{u}_N}{\partial x^2} \\ \bar{u}_N(0,t) = \bar{u}_N(2\pi,t) \\ \bar{u}_N(x,0) = P_N u_0(x) \end{cases}$$

Si noti che dovendo per l'appunto proiettare $\dfrac{\partial}{\partial x}\left(\dfrac{u_N^2}{2}\right)$ sullo spazio di polinomi trigonometrici di grado N tramite P_N la soluzione $\bar{u}_N \neq u_N = P_N u$, ossia non coincide con la serie troncata della sol. esatta.

Il metodo infine risulta

$$\frac{d(\hat{\bar{u}}_N)_k}{dt} + (\hat{\bar{h}}_N)_k = -k^2 (\hat{\bar{u}}_N)_k \,, \quad -\frac{N}{2} \le k \le \frac{N}{2} - 1$$

con $$(\hat{\bar{h}}_N)_k = \frac{1}{2} k \sum_{m=-\frac{N}{2}}^{\frac{N}{2}-1} \hat{u}_{k-m}\, \hat{u}_m \qquad (*)$$

dove si è supposto $\hat{u}_\ell = 0$ per $\ell > \frac{N}{2} - 1$, $\ell < -\frac{N}{2}$.

Si noti che $(\hat{\hat{h}}_N)_k$ è una somma di convoluzione che può essere risolta efficacemente tramite l'uso algoritmico della FFT.

In sostanza l'uso della FFT in questo caso equivale ad operare nello spazio fisico con il calcolo di $(\hat{\hat{h}}_N)_k$ impiegando un metodo pseudo-spettrale.

Osserviamo infatti che se avessimo calcolato

$$h = u\, u_x = \sum_{\ell=-\infty}^{\infty} \hat{u}_\ell \sum_{m=-\infty}^{\infty} m \hat{u}_m e^{i(\ell+m)x}$$

$$= \sum_{k=-\infty}^{\infty} \hat{u}_{k-m}\, m\, \hat{u}_m\, e^{ikx} = \sum_{k=-\infty}^{\infty} \hat{\hat{h}}_k\, e^{ikx}$$

con $\hat{\hat{h}}_k = \sum_{m=-\infty}^{\infty} \hat{u}_{k-m}\, m\, \hat{u}_m$ diverso da $\hat{h}_k$.

Tale scelta originerà comunque un metodo equivalente sia in termini di costo computazionale che di accuratezza.

CENNI SUI METODI AI VOLUMI FINITI

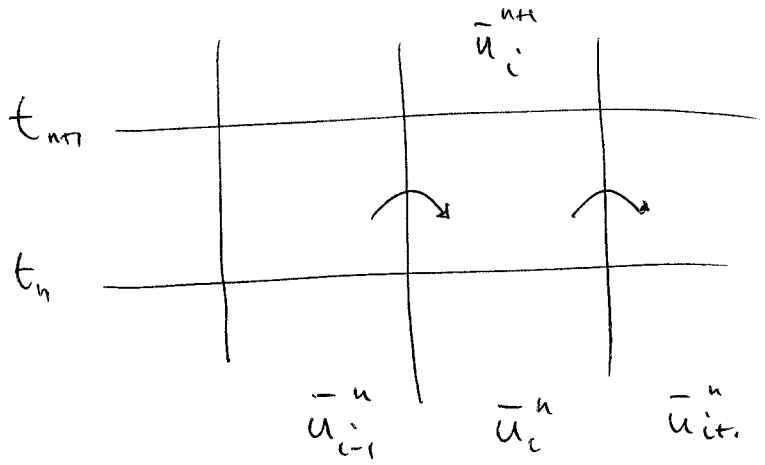
Definiamo innanzitutto il concetto di "media di cella".

$$C_i = \left]x_{i-1/2}, x_{i+1/2}\right[\quad \text{cella } i$$

$$\bar{U}_i^n \approx \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} u(x,t_n)\,dx = \frac{1}{\Delta x}\int_{C_i} u(x,t_n)\,dx \quad \text{(media di cella)}$$

Consideriamo ora l'equazione

$$u_t + f(u)_x = 0 \quad \text{(legge cons. scalare)}$$



Integrando la legge di conservazione sulla cella C_i si ha

$$\frac{d}{dt}\int_{C_i} u(x,t)\,dx = f\left(u\left(x_{i-1/2},t\right)\right) - f\left(u\left(x_{i+1/2},t\right)\right)$$

Vogliamo approssimare la media di cella al tempo $n+1$ in funzione delle medie di cella al tempo n. Possiamo approssimare la derivata temporale con un metodo Runge-Kutta per cui dovremo definire il <u>flusso numerico</u> $f(u(x_{i\pm1/2},t))$ in funzione delle medie di cella $\bar{u}_i$, $\bar{u}_{i\pm1}$, ...

Alternativamente si integra anche la variabile temporale.

Integro il tempo da t_n a t_{n+1} e ottengo

$$\int_{C_i} u(x,t_{n+1})\,dx - \int_{C_i} u(x,t_n)\,dx = \int_{t_n}^{t_{n+1}} f(u(x_{i-\frac{1}{2}},t))\,dt -$$

$$- \int_{t_n}^{t_{n+1}} f(u(x_{i+\frac{1}{2}},t))\,dt$$

Dividendo per Δx si ha

$$\frac{1}{\Delta x}\int_{C_i} u(x,t_{n+1})\,dx = \frac{1}{\Delta x}\int_{C_i} u(x,t_n)\,dx -$$

$$- \frac{1}{\Delta x}\left[\int_{t_n}^{t_{n+1}} f(u(x_{i+\frac{1}{2}},t))\,dt - \int_{t_n}^{t_{n+1}} f(u(x_{i+\frac{1}{2}},t))\,dt\right]$$

$$\boxed{\bar{u}_i^{n+1} = \bar{u}_i^n - \frac{\Delta t}{\Delta x}\left[F_{i+\frac{1}{2}}^n - F_{i-\frac{1}{2}}^n\right]}$$

$$F_{i+\frac{1}{2}}^n = \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(u(x_{i+\frac{1}{2}},t))\,dt$$

In generale calcolerò $F_{i+\frac{1}{2}}$ in modo approssimato, detto flusso numerico a partire dalle medie di cella.

Dato che per un sistema iperbolico la soluzione si propaga con velocità finita è ragionevole assumere che

$$F^n_{i+1/2} = F\left(\bar{u}^n_i, \bar{u}^n_{i+1}\right)$$

dove la funzione F caratterizza il flusso numerico.

La forma di F caratterizza il metodo numerico, in questo caso basato su uno "stencil" a tre punti.

Ossevano che

$$\Delta x \sum_{i=1}^{N} \bar{u}_i^{n+1} = \Delta x \sum_{i=1}^{N} \bar{u}_i^n - \frac{\Delta t}{\Delta x}\left(F^n_{N+1/2} - F^n_{1/2}\right)$$

la somma delle differenze dei flussi si cancella ad eccezione dei valori al bordo, dove dovrà applicare opportune condizioni.

Il metodo può anche essere rappresentato come metodo alle <u>differenze finite</u>

$$\frac{\bar{u}_i^{n+1} - \bar{u}_i^n}{\Delta t} + \frac{F^n_{i+1/2} - F^n_{i-1/2}}{\Delta x} = 0 .$$

<u>ESEMPIO</u>: Otteniamo dapprima i corrispettivi flussi utilizzando i metodi già visti come upwind, $L \times F$, $L \times W$

(<u>esercizio</u> scriverli per l'eq. $u_t + a u_x = 0$)