8 Elementi di differenze finite per equazioni differenziali[1]

Prof. Marco Vianello - Dipartimento di Matematica, Università di Padova

aggiornamento: 9 gennaio 2020

8.1 Problemi ai valori iniziali

- dato un problema ai valori iniziali $y' = f(t,y)$, $t \in [t_0, t_f]$; $y(t_0) = a$, con f di classe C^1 tale che $|\partial f/\partial y| \leq L$, si dimostri che la "legge di propagazione" dell'errore $e_n = |y_n - y(t_n)|$ dei metodi di *Eulero esplicito*

$$y_{n+1} = y_n + hf(t_n, y_n) \ , \ \ 0 \leq n \leq N-1$$

ed *Eulero implicito*

$$y_{n+1} = y_n + hf(t_{n+1}, y_{n+1}) \ , \ \ 0 \leq n \leq N-1$$

su una discretizzazione a passo costante $h = (t_f - t_0)/N$, $t_n = t_0 + nh$, è del tipo

$$e_{n+1} \leq \alpha e_n + \delta_{n+1} \ , \ \ \delta_{n+1} = \delta_{n+1}(h) = \mathcal{O}(h^2)$$

dove $\alpha = \alpha(h) = (1 + hL)$ nel caso puramente Lipschitziano, oppure $\alpha = 1$ nel caso dissipativo ($-L \leq \partial f/\partial y \leq 0$) senza vincoli sul passo per Eulero implicito e con il vincolo $h \leq 2/L$ per Eulero esplicito
(traccia: per Eulero esplicito si usi la sequenza ausiliaria $u_n = y(t_n) + hf(t_n, y(t_n))$ e la scrittura $y_{n+1} - y(t_{n+1}) = y_{n+1} - u_n + u_n - y(t_{n+1})$, $\delta_{n+1}(h) = |u_n - y(t_{n+1})|$, ricorrendo poi al teorema del valor medio per il primo addendo e alla formula di Taylor per il secondo ...; analogamente per Eulero implicito con la sequenza ausiliaria $u_{n+1} = y(t_{n+1}) + hf(t_{n+1}, y(t_{n+1}))$, ...)

- si deduca dall'esercizio precedente che l'errore globale dei metodi di Eulero è

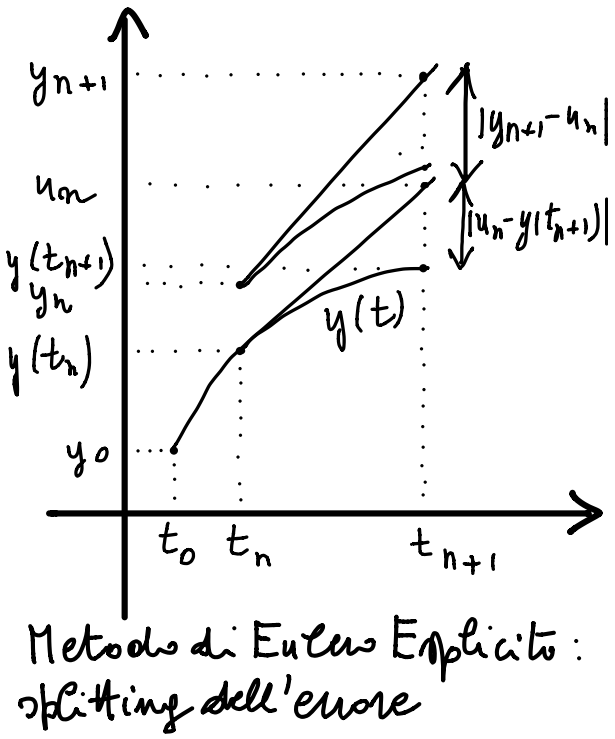
$$\max_{0 \leq n \leq N} e_n \leq \alpha^N e_0 + \max_{1 \leq n \leq N} \{\delta_n\} \sum_{n=0}^{N-1} \alpha^n \leq c_1 e_0 + c_2 h \ ,$$

stimando le costanti c_1 e c_2 nei casi Lipschitziano e dissipativo

- si estenda l'analisi di convergenza del metodo di Eulero esplicito al caso di un sistema $m \times m$ di ODEs, con $f(t, \cdot)$ campo vettoriale Lipschitziano di costante L in una norma vettoriale

- $*$ la nozione di dissipatività si estende al caso di un campo vettoriale nella forma $(f(t,y_2) - f(t,y_1), y_2 - y_1) \leq 0$, dove (u,v) indica il prodotto scalare euclideo

[1] argomenti e quesiti contrassegnati da * sono più impegnativi



di due vettori u, v; a cosa corrisponde la dissipatività nel caso di un sistema differenziale lineare $y' = A(t)y + g(t)$? Si dimostri che il metodo di Eulero implicito per un sistema $m \times m$ di ODEs con campo vettoriale f dissipativo soddisfa la stessa legge di propagazione dell'errore del caso scalare (in norma euclidea), senza vincoli sul passo.
(sugg.: partendo da $y_{n+1}-y(t_{n+1}) = y_n-y(t_n)+h[f(t_{n+1}, y_{n+1})-f(t_{n+1}, y(t_{n+1})]+ u_{n+1}-y(t_{n+1})$, dove $\|u_{n+1}-y(t_{n+1})\|_2 = \mathcal{O}(h^2)$, si faccia a sinistra e a destra il prodotto scalare per il vettore $y_{n+1}-y(t_{n+1})$, ...)

- $*$ dopo aver individuato un opportuno modello di "metodo perturbato" che fornisca una sequenza $\tilde{y}_n \approx y_n$ per Eulero esplicito ed implicito tenendo conto degli errori introdotti ad ogni passo, si studi la stabilità dei due metodi nei casi Lipschtiziano e dissipativo, ottenendo una "legge di propagazione" degli errori e una stima dell'effetto globale di tali errori. Ci si aspetta che $\max |\tilde{y}_n-y(t_n)| \to 0$ per $h \to 0$?

- il *metodo di Crank-Nicolson* (o trapezoidale)

$$y_{n+1} = y_n + (h/2)[f(t_n, y_n) + f(t_{n+1}, y_{n+1})]$$

ha ordine di approssimazione locale $\delta_{n+1}(h) = \mathcal{O}(h^3)$ per $f \in C^2$ (traccia: il metodo si ricava applicando alla rappresentazione integrale $y(t_{n+1}) - y(t_n) = \int_{t_n}^{t_{n+1}} y'(t)\,dt$ la formula di quadratura del trapezio)

- dato il problema test

$$y' = \lambda y + b(t)\ ,\ \ t > 0\ ;\ \ y(0) = y_0$$

dove $g \in C[0, +\infty)$ e $\lambda \in \mathbb{C}$, $Re\lambda \leq 0$ (problema "stiff"), qual'è l'effetto sulla soluzione di un errore $\varepsilon_0 = |y_0 - \tilde{y}_0|$ sul dato iniziale? si verifichi poi che la propagazione di un errore ε_0 sul dato iniziale per una discretizzazione a passo costante $h > 0$ della semiretta è del tipo

$$\varepsilon_n = (\phi(h\lambda))^n \varepsilon_0\ ,\ \ n > 0$$

per i metodi di Eulero (esplicito ed implicito) e per il metodo di Crank-Nicolson. Quale è la regione di stabilità di ciascun metodo, ovvero $\{z \in \mathbb{C} :\ \phi(z) \leq 1\}$? Quale dei tre metodi è stabile senza vincoli sul passo?

- $*$ si estenda l'analisi dell'esercizio precedente al caso del sistema test

$$y' = Ay + b(t)\ ,\ \ t > 0\ ;\ \ y(0) = y_0$$

dove $y(t), y_0, b(t) \in \mathbb{R}^m$ e $A \in R^{m\times m}$ è una matrice costante diagonalizzabile con autovalori di parte reale non positiva (sistema "stiff")
(traccia: lavorando nella base di autovettori di A ...)

8.2 Problemi ai valori al contorno

- dato il problema ai valori al contorno

$$u''(x) - cu(x) = f(x)\ ,\ \ x \in (a, b)\ ;\ \ u(a) = u(b) = 0$$

dove c è una costante positiva, si assuma che la soluzione u sia di classe $C^4[a, b]$ e si consideri una discretizzazione dell'intervallo a passo costante $h = (b-a)/(n+1)$, con nodi $x_i = a+ih$, $0 \leq i \leq n+1$. Tramite l'approssimazione della derivata seconda ottenuta con le *differenze centrate*

$$u''(x) \approx \delta_h^2 u(x) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2}$$

si vede che il vettore $\{u(x_i)\}_{1\leq i\leq n}$ soddisfa un sistema lineare del tipo

$$A\{u(x_i)\} = \{f(x_i)\} + \{\varepsilon_i\}$$

dove $A \in \mathbb{R}^{n\times n}$ è una matrice *tridiagonale* con $-2h^{-2}-c$ sulla diagonale principale e h^{-2} sulle due diagonali adiacenti, e $|\varepsilon_i| \le Mh^2$ (sugg.: si utilizzi la formula di Taylor di centro x_i in $x_i - h = x_{i-1}$ e $x_i + h = x_{i+1}$). La matrice A risulta *simmetrica* e *definita negativa*, con autovalori nell'intervallo $[-(c+4h^{-2}), -c]$ (si usi il teorema di Gershgorin per la localizzazione degli autovalori, ovvero: se λ è autovalore di $A \in \mathbb{C}^{m\times m}$ allora $\lambda \in \bigcup_{i=1}^{m} C[a_{ii}, r_i]$, $r_i = \sum_{j\neq i} |a_{ij}|$, dove $C[a,r]$ è il cerchio complesso chiuso di centro $a \in \mathbb{C}$ e raggio r).

Detto $\{u_i\}$ il vettore soluzione del sistema lineare

$$A\{u_i\} = \{f(x_i)\}$$

si provi che vale la stima

$$\frac{\|\{u_i\} - \{u(x_i)\}\|_2}{\sqrt{n}} = \mathcal{O}(h^2)$$

(sugg.: si ha che $\|A^{-1}\|_2 = \rho(A^{-1}) \le 1/c$, ...; si osservi che $\|\{u(x_i)\}\|_2/\sqrt{n} \approx \|u\|_{L^2(a,b)}/\sqrt{b-a}$). Il sistema può essere agevolmente risolto con il metodo di eliminazione di Gauss, che in questo caso ha complessità $\mathcal{O}(n)$ (perché?) Come va modificato il sistema per condizioni al contorno del tipo $u(a) = \alpha$, $u(b) = \beta$?

- nel caso di $c = 0$ (*equazione di Poisson* unidimensionale) si può ancora far vedere (non richiesto) che A è definita negativa, perché i suoi autovalori appartengono all'intervallo $[-4h^{-2}, -\delta]$, con un opportuno $\delta > 0$ indipendente da h

- $*$ dato il problema ai valori al contorno

$$\Delta u(P) - cu(P) = f(P)\ , \quad P = (x,y) \in \Omega = (a,b)\times(c,d)\ ; \quad u|_{\partial\Omega} \equiv 0$$

dove $\Delta = \partial/\partial x^2 + \partial/\partial y^2$ è l'operatore laplaciano e c è una costante positiva, si assuma che la soluzione u sia di classe $C^4(\overline{\Omega})$ e si consideri una discretizzazione del rettangolo con passo $h = (b-a)/n$ nella direzione x e $k = (d-c)/m$ nella direzione y, con nodi $P_{ij} = (x_i, y_j)$, $x_i = a + ih$, $0 \le i \le n+1$, $y_j = c + jk$, $0 \le j \le m+1$. Discretizzando l'operatore di Laplace tramite lo schema alle differenze "a croce"

$$\delta^2_{h,k} u(P) = \delta^2_{x,h} u(P) + \delta^2_{y,k} u(P) \approx \Delta u(P)$$

si vede che il vettore $\{u(P_{ij})\}_{1\le i\le n, 1\le j\le m}$ soddisfa un sistema lineare del tipo

$$A\{u(P_{ij})\} = \{f(P_{ij})\} + \{\varepsilon_{ij}\}$$

dove $A \in \mathbb{R}^{nm\times nm}$ e $|\varepsilon_{ij}| \le M_1 h^2 + M_2 k^2$.

La struttura della matrice dipende dalla numerazione dei nodi: utilizzando l'ordinamento *lessicografico* delle coppie (i,j), si vede che la matrice risulta *simmetrica*, *tridiagonale a blocchi* con una diagonale di blocchi $n \times n$ che sono matrici tridiagonali con $-c - 2(h^{-2} + k^{-2})$ sulla diagonale principale e h^{-2} sulle

due diagonali adiacenti, e due diagonali adiacenti di blocchi $n \times n$ che sono matrici diagonali con k^{-2} sulla diagonale principale. Inoltre $\sigma(A) \subset [-(c + 4(h^{-2}+k^{-2})), -c]$, quindi A è *definita negativa* (questo è vero (di. non richiesta) anche per $c = 0$, *equazione di Poisson* bidimensionale). Detto $\{u_{ij}\}$ il vettore soluzione del sistema lineare

$$A\{u_{ij}\} = \{f(P_{ij})\}$$

si provi che vale la stima

$$\frac{\|\{u_{ij}\} - \{u(P_{ij})\}\|_2}{\sqrt{nm}} = \mathcal{O}(h^2) + \mathcal{O}(k^2)$$

(si osservi che $\|\{u(P_{ij})\}\|_2/\sqrt{nm} \approx \|u\|_{L^2(\Omega)}/\sqrt{area(\Omega)}$). In questo caso il metodo di eliminazione di Gauss non è conveniente (perché?), essendo A fortemente *sparsa* (su ogni riga ci sono al massimo 5 elementi non nulli) e tendenzialmente di grande dimensione sono più adatti metodi iterativi, opportunamente precondizionati visto che A è *mal condizionata*, $k_2(A) = \mathcal{O}(h^{-2} + k^{-2})$ (sugg.: si usi il teorema di Gershgorin per stimare il condizionamento di A)

8.3 Problemi evolutivi: equazione del calore

- *metodo delle linee* per l'equazione del calore: dato il problema evolutivo alle derivate parziali con condizioni iniziali e al contorno

$$\frac{\partial u}{\partial t}(P,t) = \sigma \Delta u(P,t) + g(P,t) \ , \quad (P,t) \in \Omega \times (0, +\infty)$$

$$u(P,0) = u_0(P) \ , \quad u(P,t)|_{P \in \partial\Omega} \equiv 0$$

nel caso unidimensionale con $P = x \in \Omega = (a,b)$ o bidimensionale con $P = (x,y) \in \Omega = (a,b) \times (c,d)$, la discretizzazione nelle variabili spaziali tramite $\Delta u \approx \delta^2 u$, posto rispettivamente $y(t) = \{u_i(t)\} \approx \{u(x_i,t)\}$ o $y(t) = \{u_{ij}(t)\} \approx \{u(P_{ij},t)\}$, porta ad un sistema di equazioni differenziali ordinarie nel tempo

$$y' = Ay + b(t) \ , \quad t > 0 \ ; \quad y(0) = y_0$$

dove A è la matrice di discretizzazione del Laplaciano vista sopra (tridiagonale o tridiagonale a blocchi), scalata con il fattore $\sigma > 0$ (coefficiente di diffusione). Si mostri che il metodo di Eulero esplicito è stabile con un passo temporale dell'ordine del quadrato dei passi spaziali, mentre i metodi di Eulero implicito e di Crank-Nicolson sono incondizionatamente stabili.
(sugg.: essendo la matrice A simmetrica e definita negativa si tratta di un sistema stiff, ...)

- si osservi che le matrici (simmetriche definite positive) dei sistemi lineari ad ogni passo dei metodi di Eulero implicito e di Crank-Nicolson sono molto meglio

condizionate della matrice A; infatti, considerando ad esempio Eulero implicito, detto Δt il passo di discretizzazione nel tempo, si ha che

$$k_2(I - \Delta t A) = \frac{1 + \Delta t|\lambda_{max}(A)|}{1 + \Delta t|\lambda_{min}(A)|} \approx \Delta t|\lambda_{max}(A)| \ll \frac{|\lambda_{max}(A)|}{|\lambda_{min}(A)|} = k_2(A)$$

- la soluzione dell'equazione del calore discretizzata nello spazio col metodo delle linee si può scrivere esplicitamente usando l'esponenziale di matrice come

$$y(t) = \exp(tA)y_0 + \int_0^t \exp((t-s)A)\, b(s)\, ds \ ,$$

dove $\exp(B) = \sum_{j\geq 0} B^j/j!$ con B matrice quadrata. Nel caso in cui $g(P,t) = g(P)$ è indipendente dal tempo, si ha $b(t) \equiv b$ vettore costante e quindi

$$y(t) = \exp(tA)y_0 + (\exp(tA) - I)A^{-1}b \to -A^{-1}b \ , \quad t \to +\infty \ ,$$

cioè la soluzione dell'equazione (discretizzata) del calore ha uno stato stazionario asintotico che corrisponde alla soluzione dell'equazione (discretizzata) di Poisson con termine noto $-g$ (si può dimostrare (NR) che questa proprietà vale anche per la soluzione $u(P,t)$). Nel ricavare la rappresentazione esponenziale della soluzione si usa il fatto che $d/ds \exp(sA) = A\exp(sA)$ (perché?) e di conseguenza $\int \exp(sA)ds = A^{-1}\exp(sA)$.

La rappresentazione esponenziale della soluzione è alla base di una famiglia di solutori, i cosiddetti *integratori esponenziali*, che invece di discretizzare il sistema stiff approssimano direttamente gli operatori esponenziali coinvolti.