Roberto Ferretti

ALCUNI ARGOMENTI COMPLEMENTARI DEL CORSO DI ANALISI NUMERICA

Dispensa per il corso di
"Analisi Numerica"
Dipartimento di Matematica, Universita' di Roma Tre
1998

1. Analisi di sensibilitá nella soluzione di sistemi lineari.

Consideriamo innanzitutto la situazione in cui in un sistema

$$Ax = b \tag{1.1}$$

venga perturbato della quantitá δb il vettore b dei termini noti. La soluzione del sistema perturbato soddisferá

$$A(x + \delta x) = b + \delta b \tag{1.2}$$

e la grandezza di interesse é l'errore relativo sulla soluzione x. Definito il numero di condizionamento della matrice (nonsingolare) A dato da $K(A) = \|A\| \, \|A^{-1}\|$, si ha il seguente

Teorema 1.1. *Dato il sistema* (1.2)*, in cui* A*,* x *e* b *soddisfino* (1.1)*, l'errore relativo sulla soluzione é maggiorato da*

$$\frac{\|\delta x\|}{\|x\|} \leq K(A) \frac{\|\delta b\|}{\|b\|}.$$

Dim. Da (1.1), (1.2) si ottiene

$$A\delta x = \delta b$$

da cui, utilizzando le proprietá delle norme matriciali,

$$\|\delta x\| \leq \|A^{-1}\| \|\delta b\|. \tag{1.3}$$

Moltiplicando ambo i membri per $\|b\|/\|x\|$ ed usando la maggiorazione $\|b\| \leq \|A\| \, \|x\|$, si ha infine da (1.3):

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\| \, \|A^{-1}\| \frac{\|\delta b\|}{\|b\|}.$$

■

Nel caso piú generale, in cui venga perturbata anche la matrice A, il sistema perturbato si scriverá come

$$(A + \delta A)(x + \delta x) = b + \delta b \tag{1.4}$$

ed la corrispondente stima dell'errore é data da:

Teorema 1.2. *Dato il sistema* (1.4)*, in cui* A*,* x *e* b *soddisfino* (1.1)*, e supponendo che* $\|\delta A\| < 1/\|A^{-1}\|$*, l'errore relativo sulla soluzione é maggiorato da*

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{K(A)}{1 - K(A)\frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right).$$

Dim. Operando come nella dimostrazione precedente, e ricordando che per l'ipotesi fatta su δA la matrice $A + \delta A = A(I + A^{-1}\delta A)$ é nonsingolare, si ottiene

$$\delta x = (A + \delta A)^{-1}(\delta b - \delta A x)$$

da cui si ha immediatamente

$$\|\delta x\| \leq \|(A + \delta A)^{-1}\|(\|\delta b\| + \|\delta A\| \, \|x\|)$$

e quindi, dividendo ambo i membri per $\|x\|$ e moltiplicando e dividendo il secondo per $\|A\|$:

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\| \, \|(A + \delta A)^{-1}\| \left(\frac{\|\delta b\|}{\|A\| \, \|x\|} + \frac{\|\delta A\|}{\|A\|} \right) \leq \|A\| \, \|(A + \delta A)^{-1}\| \left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right). \tag{1.5}$$

La quantitá che resta da stimare in (1.5) é

$$\|(A + \delta A)^{-1}\| = \|(A(I + A^{-1}\delta A))^{-1}\| \leq \|A^{-1}\| \, \|(I + A^{-1}\delta A)^{-1}\|. \tag{1.6}$$

La funzione di matrice che appare nell'ultimo membro di (1.6) si puó scrivere in serie di potenze (con punto inziale la matrice I) come

$$(I + A^{-1}\delta A)^{-1} = \sum_{k=0}^{\infty} (-1)^k (A^{-1}\delta A)^k. \tag{1.7}$$

La convergenza della serie geometrica (1.7) é ancora conseguenza della ipotesi fatta su δA (infatti da questa ipotesi segue che $\|A^{-1}\delta A\| < 1$). Da (1.7) si ottiene poi

$$\|(I + A^{-1}\delta A)^{-1}\| \leq \sum_{k=0}^{\infty} \|A^{-1}\|^k \|\delta A\|^k = \frac{1}{1 - \|A^{-1}\| \, \|\delta A\|} = \frac{1}{1 - K(A)\frac{\|\delta A\|}{\|A\|}}$$

che insieme con (1.5) e (1.6) prova il teorema.

■

Osservazione. Si noti che questo risultato é intrinseco, e non dipende dall'algoritmo con cui si risolve il sistema lineare. A seconda poi della stabilitá dei diversi algoritmi nei confronti degli errori di arrotondamento, andrebbero introdotti ulteriori termini di errore (questa volta dipendenti dall'algoritmo usato).

2. I metodi di rilassamento.

Nel caso in cui la matrice A sia definita positiva, in metodo di Gauss–Seidel si puó interpretare come un metodo di minimizzazione che utilizzi la strategia del *rilassamento.* Tale strategia consiste, in poche parole, nel minimizzare una funzione rispetto ad una

variabile per volta, e per questa classe di algoritmi la convergenza é garantita da un teorema generale che enunceremo nel seguito. La funzione che ci interessa nel caso dei sistemi lineari é la forma quadratica

$$f(x) = \frac{1}{2}(Ax, x) - (b, x). \tag{2.1}$$

Tale funzione sará dotata di minimo unico se $A > 0$, il punto di minimo essendo dato dalla soluzione del sistema lineare $Ax = b$.

Per semplicitá, scriviamo in metodo di Gauss–Seidel per un sistema 2×2 nella forma generica:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 = b_1 \\ a_{21}x_1 + a_{22}x_2 = b_2. \end{cases} \tag{2.2}$$

Alla generica iterazione k occorrerá dapprima aggiornare la variabile x_1 secondo la formula

$$x_1^{(k)} = \frac{1}{a_{11}}(b_1 - a_{12}x_2^{(k-1)}). \tag{2.3}$$

Dopo questo aggiornamento, ci si puó figurare di essere passati dal punto $(x_1^{(k-1)}, x_2^{(k-1)})$ al punto $(x_1^{(k)}, x_2^{(k-1)})$. In quest'ultimo punto il gradiente di f sará dato da

$$\nabla f(x_1^{(k)}, x_2^{(k-1)}) = \begin{pmatrix} a_{11}x_1^{(k)} + a_{12}x_2^{(k-1)} - b_1 \\ a_{21}x_1^{(k)} + a_{22}x_2^{(k-1)} - b_2 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{a_{21}}{a_{11}}(b_1 - a_{12}x_2^{(k-1)}) + a_{22}x_2^{(k-1)} - b_2 \end{pmatrix} \tag{2.3}$$

In seguito, occorrerá aggiornare x_2 come:

$$x_2^{(k)} = \frac{1}{a_{22}}(b_2 - a_{21}x_1^{(k)}). \tag{2.4}$$

Ci si puó figurare ora di essere passati dal punto $(x_1^{(k)}, x_2^{(k-1)})$ al punto $(x_1^{(k)}, x_2^{(k)})$, dove il gradiente di f vale

$$\nabla f(x_1^{(k)}, x_2^{(k)}) = \begin{pmatrix} a_{11}x_1^{(k)} + a_{12}x_2^{(k)} - b_1 \\ a_{21}x_1^{(k)} + a_{22}x_2^{(k)} - b_2 \end{pmatrix} = \begin{pmatrix} a_{11}x_1^{(k)} + \frac{a_{12}}{a_{22}}(b_2 - a_{21}x_1^{(k)}) - b_1 \\ 0 \end{pmatrix}. \tag{2.5}$$

Da (2.3), (2.5) si puó osservare che dopo l'aggiornamento di una variabile x_i il gradiente di f é ortogonale alla direzione della variabile stessa, il che corrisponde a dire che il valore aggiornato di x_i é quello che corrisponde al minimo di f con gli $n - 1$ vincoli

$$x_j = x_j^{(k-1)} \quad (j \neq i). \tag{2.6}$$

Nel caso in cui A non sia definita positiva ma sia comunque simmetrica, e quindi $Ax - b$ si puó ancora interpretare come il gradiente della funzione $f(x)$ data da (2.1), il valore aggiornato di x_i corrisponderá in generale all'unico punto stazionario di f con i vincoli (2.6). Come é noto, in questo caso la convergenza si dimostra sotto ipotesi piú forti (dominanza diagonale stretta).

La formulazione generale di un metodo di rilassamento in R^n, per una funzione f che non sia del tipo (2.1), é nella forma

$$x_j^{(k)} = \underset{t}{\text{argmin}}\ f(x_1^{(k)}, \ldots, x_{j-1}^{(k)}, t, x_{j+1}^{(k-1)}, \ldots, x_n^{(k-1)}) \tag{2.7}$$

(per j *in sequenza* da 1 a n, $k \geq 1$) e la convergenza é assicurata dal seguente

Teorema 2.1. *Si supponga $f \in C^1(R^n)$, f strettamente convessa, $\lim_{\|x\|\to\infty} f(x) = +\infty$. Allora la successione $x^{(k)}$ definita da (2.7) converge all'unico punto di minimo di f su R^n.* ■

Piuttosto che dare la dimostrazione del teorema (che é piuttosto laboriosa) si preferisce fornire un controesempio che mostra come la ipotesi di differenziabilitá sia fondamentale. Si consideri la funzione

$$f(x_1, x_2) = x_1^2 + x_2^2 - 2(x_1 + x_2) + 2|x_1 - x_2| \tag{2.8}$$

che é strettamente convessa ed ha minimo (come é facile verificare) nel punto $(1, 1)$. Nell'origine, $f(0, 0) = 0$, mentre $f(x_1, 0) = x_1^2 - 2x_1 + 2|x_1| \geq 0$ e $f(0, x_2) = x_2^2 - 2x_2 + 2|x_2| \geq 0$. Quindi il punto $(0, 0)$ é di minimo sia rispetto alla variabile x_1 che a x_2, ma non é il minimo assoluto (questo é dovuto al fatto che si tratta di un punto di non differenziabilitá). Il grafico dell'insieme di livello $\{x : f(x) \leq 0\}$ é mostrato nella figura 1.

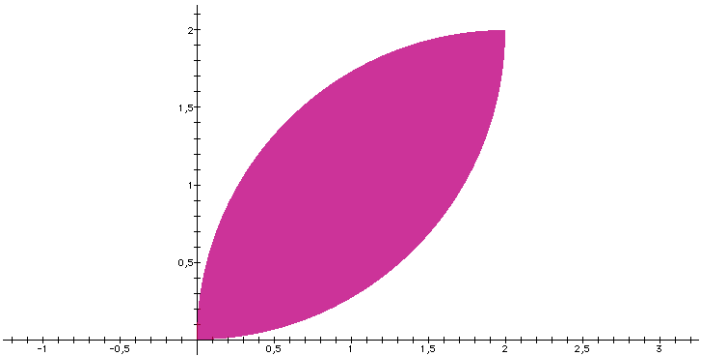

Figura 1. Controesempio sulla convergenza dei metodi di rilassamento.

Esercizi

1. Formulare un metodo di tipo Gauss–Seidel per sistemi di equazioni nonlineari. Quale condizione é l'equivalente della condizione di dominanza diagonale di A nel caso nonlineare?
2. Fornire un esempio (anche 2×2) di sistema lineare in cui la convergenza di un metodo iterativo (Jacobi o Gauss–Seidel) dipenda da come sono ordinate le equazioni.

3. Sui metodi di punto fisso per equazioni nonlineari.

Una maniera piuttosto generale di ottenere da una equazione

$$f(x) = 0 \tag{3.1}$$

uno schema iterativo in forma di punto fisso, consiste nel definirlo come

$$x_{k+1} = g(x_k) = x_k + \alpha f(x_k), \tag{3.2}$$

tentando poi di determinare un valore della costante α in modo che lo schema sia convergente il piú velocemente possibile. Dato per buono che esista una radice $\bar{x}$ della equazione (3.1), il valore ottimale di α (per ragioni che si vedranno) é $\alpha_{opt} = -1/f'(\bar{x})$. Di conseguenza lo schema (3.2) diviene

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(\bar{x})} \tag{3.3}$$

(si puó notare che $\bar{x}$ é ovviamente un punto fisso anche per questa iterazione). La convenienza di questa scelta proviene da un risultato del tutto generale:

Teorema 3.1. *Dato lo schema iterativo*

$$x_{k+1} = g(x_k)$$

con $g \in C^2$, se lo schema é convergente a $\bar{x}$ e $g'(\bar{x}) = 0$ (con $\bar{x}$ soluzione di (3.1)), allora la convergenza $x_k \to \bar{x}$ é quadratica.

Dim. Notiamo intanto che in queste ipotesi la funzione g é una contrazione (almeno localmente), visto che nell'intorno di $\bar{x}$ viene sicuramente verificata la condizione $|g'(x)| \leq L < 1$.

Dalla formula di ricorrenza dello schema si ha poi:

$$x_{k+1} - \bar{x} = g(x_k) - \bar{x} = g(x_k) - g(\bar{x}).$$

Sviluppando g con il suo polinomio di Taylor di centro $\bar{x}$, e tenendo conto del fatto che il differenziale primo si annulla nel nostro caso, si ottiene

$$x_{k+1} - \bar{x} = \frac{1}{2} g''(\xi_k)(x_k - \bar{x})^2 \tag{3.4}$$

con ξ_k compreso tra x_k e $\bar{x}$. Passando ai moduli,

$$|x_{k+1} - \bar{x}| = \frac{1}{2} |g''(\xi_k)| \, |x_k - \bar{x}|^2 \tag{3.5}$$

e dalla (locale) limitatezza di $g''(\cdot)$ si ottiene la tesi.

■

Osservazione. Si puó usare immediatamente questo risultato per dimostrare la convergenza quadratica del metodo di Newton in R^1, dato che per questo metodo si ha

$$g'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2}$$

e quindi $g'(\bar{x}) = 0$. In generale, peró, in metodi che non calcolino la derivata prima non sará possibile scegliere esattamente in questo modo la costante α in (3.2). Si tratta quindi di darne, con metodi euristici, la migliore approssimazione possibile (ricordando che in ogni caso, anche se la convergenza non é quadratica, un valore piú basso del coefficiente di contrazione porterá ad una convergenza piú veloce). Notiamo anche che per il metodo di Newton, l'esistenza della radice $\bar{x}$, assieme con (3.5), implica la convergenza della successione x_k. Infatti, dalla (3.5) si ha

$$|x_{k+1} - \bar{x}| \leq \frac{1}{2} \max_{[\bar{x}-\varepsilon, \bar{x}+\varepsilon]} |g''(x)| \, |x_k - \bar{x}|^2 \leq$$

$$\leq \frac{1}{2} \max_{[\bar{x}-\varepsilon, \bar{x}+\varepsilon]} |g''(x)| \varepsilon \, |x_k - \bar{x}|.$$

Basta perció scegliere ε in modo tale che

$$\frac{1}{2} \max_{[\bar{x}-\varepsilon, \bar{x}+\varepsilon]} |g''(x)| \varepsilon < L < 1$$

perche' la successione x_k converga a $\bar{x}$ (sempreché, naturalmente, $|x_0 - \bar{x}| < \varepsilon$).

Esercizi

1. Dare una versione n–dimensionale della equazione di punto fisso (3.2) e della relativa condizione di convergenza.
2. Dimostrare che se una radice $\bar{x}$ ha molteplicitá $p > 1$, il metodo di Newton non converge quadraticamente, ma che converge quadraticamente il metodo definito da

$$g(x) = x - p\frac{f(x)}{f'(x)}.$$

4. Ottimizzazione: metodi di ricerca unidimensionale.

Nella ricerca del minimo di una funzione $f : R \to R$ si puó naturalmente applicare un qualsiasi algoritmo di soluzione di equazioni nonlineari alla equazione $f'(x) = 0$. In pratica, peró, questo richiede il calcolo della derivata prima, e, nel caso del metodo di Newton, anche della derivata seconda. Per evitare questo inconveniente, su possono utilizzare metodi senza derivate, come il metodo di bisezione che si descrive di seguito. Tale metodo non é il piú efficiente tra i metodi di ricerca unidimensionale, ma é sufficiente per le applicazioni usuali.

Si supponga che la $f(x)$ sia continua e *unimodale* nell'intervallo $[a_0, b_0]$, ovvero che esista un punto $x^* \in [a_0, b_0]$ tale che:

$$x_1 < x_2 < x^* \Rightarrow f(x_1) > f(x_2)$$

$$x^* < x_1 < x_2 \Rightarrow f(x_1) < f(x_2).$$

Questa ipotesi implica che $f(x^*) = \min_{[a_0,b_0]} f(x)$, ed in particolare é soddisfatta se f é convessa. Si pone ora $k = 0$ e si opera iterativamente secondo i passi che seguono.

1. Poni:
$$c_k = \frac{a_k + b_k}{2}$$
2. Se é soddisfatta la condizione di arresto, STOP. Altrimenti poni:
$$d_k = \frac{a_k + c_k}{2} \ , \quad e_k = \frac{c_k + b_k}{2}$$
3. Se $f(a_k) = \min[f(a_k), f(b_k), f(c_k), f(d_k), f(e_k)]$, poni
$$a_{k+1} = a_k \ , \quad b_{k+1} = d_k$$
incrementa k e vai a 1.
4. Se $f(b_k) = \min[f(a_k), f(b_k), f(c_k), f(d_k), f(e_k)]$, poni
$$a_{k+1} = e_k \ , \quad b_{k+1} = b_k$$
incrementa k e vai a 1.
5. Se $f(c_k) = \min[f(a_k), f(b_k), f(c_k), f(d_k), f(e_k)]$, poni
$$a_{k+1} = d_k \ , \quad b_{k+1} = e_k \ , \quad c_{k+1} = c_k$$
incrementa k e vai a 2.
6. Se $f(d_k) = \min[f(a_k), f(b_k), f(c_k), f(d_k), f(e_k)]$, poni
$$a_{k+1} = a_k \ , \quad b_{k+1} = c_k \ , \quad c_{k+1} = d_k$$
incrementa k e vai a 2.
7. Poni
$$a_{k+1} = c_k \ , \quad b_{k+1} = b_k \ , \quad c_{k+1} = e_k$$
incrementa k e vai a 2.

Ad ogni passo, l'intervallo $[a_{k+1}, b_{k+1}]$ contiene il punto di minimo x^*. Per dimostrarlo, supponiamo ad esempio che il nodo a cui corrisponde il valore minimo sia c_k. Se per assurdo il punto x^* fosse esterno all'intervallo $[d_k, e_k]$, diciamo ad esempio $x^* \in [e_k, b_k]$,

poiché $f(x^*) \leq f(c_k)$, si dovrebbe avere $f(c_k) < f(e_k)$, $f(e_k) > f(x^*)$ e $f(x^*) < f(b_k)$, contro la ipotesi di unimodalitá.

Ovviamente, $b_{k+1} - a_{k+1} \leq (b_k - a_k)/2$ e quindi il metodo ha convergenza lineare. Ogni iterazione riduce l'errore della metá e richiede normalmente di valutare la funzione f due volte (nei punti ad un quarto e tre quarti dell'intervallo di ricerca, sempreché il nodo di minimo non sia in un estremo, nel qual caso l'intervallo di ricerca viene ridotto ad un quarto e la funzione viene valutata tre volte). La tipica condizione di arresto é del tipo $b_k - a_k < \varepsilon$, ovvero $k > \log_2(b_0 - a_0) - \log_2 \varepsilon$.

5. Metodi di gradiente con ricerca parziale.

Dato un metodo di gradiente (o, piú generalmente, di discesa) nella forma

$$x_{k+1} = x_k - \beta_k \nabla f(x_k), \tag{5.1}$$

ci si pone il problema di dare delle strategie generali per la determinazione degli scalari β_k. A tale scopo, definiamo intanto la famiglia di funzioni $\phi_k : R \to R$ data da:

$$\phi_k(\beta) := f(x_k - \beta \nabla f(x_k)). \tag{5.2}$$

Osserviamo che $\phi_k(0) = f(x_k)$ e che $\phi'(0) = -\|\nabla f(x_k)\|^2 < 0$. Supporremo costantemente nel seguito che f sia inferiormente limitata e che il gradiente ∇f sia un campo vettoriale lipschitziano con costante di Lipschitz (rispetto alla norma euclidea):

$$L := \sup_{R^n} \left\| \frac{\partial}{\partial x} \nabla f(x) \right\| = \sup_{R^n} \|H(x)\|$$

dove $\|\cdot\|$ indica la norma matriciale associata alla norma euclidea di R^n, e $H(x)$ indica la matrice Hessiana di f.

Le strategie tipiche per la scelta dei β_k sono di tre tipi:

a) Passo fisso. In questo caso $\beta_k \equiv \beta$ costante;

b) Ricerca esatta. Si sceglie β_k in modo che si abbia

$$\phi_k(\beta_k) = \min_{\beta} \phi_k(\beta)$$

applicando un metodo di ricerca unidimensionale, ad esempio il metodo di bisezione precedentemente descritto;

c) Ricerca parziale (criterio di Goldstein). In questo caso β_k é determinato (in modo non univoco) dalla condizione

$$f(x_k) + \sigma_1 \beta_k \phi_k'(0) \leq \phi_k(\beta_k) \leq f(x_k) + \sigma_2 \beta_k \phi_k'(0) \tag{5.3}$$

dove $0 < \sigma_2 < \sigma_1 < 1$. La strategia (5.3) é ben definita se f é inferiormente limitata. La sua interpretazione grafica é data nella figura 2, che mostra la funzione $\phi(\beta) -$

$\phi(0)$ insieme con le rette $\phi'(0)\beta$, $\sigma_1\phi'(0)\beta$ e $\sigma_2\phi'(0)\beta$ (in neretto l'intervallo di valori accetabili per β).

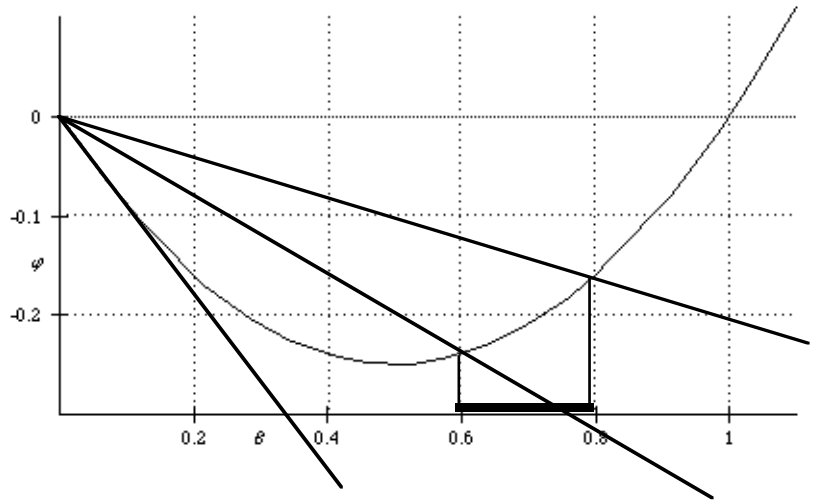

Figura 2. Interpretazione grafica del criterio di ricerca parziale.

Poiché si richiede comunque che $f(x_{k+1}) < f(x_k)$, si intuisce che la scelta che porta alla convergenza piú veloce (in termini di iterazioni) é la seconda. In pratica, peró, la convergenza resta lineare in tutti e tre i casi e la applicazione, ad ogni iterazione, di un metodo di minimizzazione unidimensionale puó appesantire notevolmente l'algoritmo (a meno che f non sia quadratica, nel qual caso β_k puó essere espresso esplicitamente come mostrato nel paragrafo 6).

Passiamo ora a dimostrare la convergenza per il terzo algoritmo.

Teorema 5.1. *Sia dato lo schema iterativo* (5.1) *con la strategia di ricerca parziale* (5.3). *Allora tutte le sottosuccessioni convergenti di* $\{x_k\}$ *convergono a punti stazionari di* f. *In particolare, se* f *é coercitiva e strettamente convessa,* $x_k \to x^*$.

Dim. Dimostriamo intanto che la successione β_k é limitata dal basso. La disuguaglianza di sinistra in (5.3) si puó riscrivere

$$\phi_k(\beta_k) - \phi_k(0) - \sigma_1\beta_k\phi_k'(0) \geq 0. \tag{5.4}$$

D'altra parte, sviluppando ϕ_k nell'intorno di $\beta = 0$, si ha

$$\phi_k(\beta_k) - \phi_k(0) - \sigma_1\beta_k\phi_k'(0) = (1-\sigma_1)\beta_k\phi_k'(0) + \frac{1}{2}\beta_k^2\phi_k''(\xi_k) \tag{5.5}$$

con $\xi_k \in [0, \beta_k]$. Da (5.4), (5.5) si ottiene

$$\beta_k\left[(1-\sigma_1)\phi_k'(0) + \frac{1}{2}\beta_k\phi_k''(\xi_k)\right] \geq 0. \tag{5.6}$$

La esistenza di soluzioni per (5.6), che segue dalla limitatezza inferiore di f, richiede che $\phi_k''(\xi_k) > 0$. Allora,

$$\beta_k \geq -\frac{2(1-\sigma_1)\phi_k'(0)}{\phi_k''(\xi_k)}$$

Ma per la lipschitzianitá di ∇f si ha (come é facile verificare) $|\phi_k''(\xi_k)| \leq L\|\nabla f(x_k)\|^2$ e quindi:

$$\beta_k \geq \frac{2(1-\sigma_1)}{L}.$$

Passiamo ora a dimostrare il risultato principale. Poiché $f(x_{k+1}) < f(x_k)$ e la f é limitata dal basso, ne segue che la successione $f(x_k)$ converge, e quindi $f(x_{k+1}) - f(x_k) \to 0$. D'altro canto, per una sottosuccessione (che indicheremo ancora con x_k) convergente ad un punto $\bar{x}$, si ha la catena di disuguaglianze

$$\sigma_1\beta_k\phi_k'(0) \leq f(x_{k+1}) - f(x_k) \leq \sigma_2\beta_k\phi_k'(0) \tag{5.7}$$

che puó essere soddisfatta per $k \to \infty$ solo se $\phi_k'(0) \to 0$, che equivale a dire $\|\nabla f(x_k)\| \to 0$. Per continuitá, $\|\nabla f(\bar{x})\| = 0$. Se questo punto stazionario é unico, si ottiene infine la convergenza di tutta la successione a $\bar{x}$.

■

Osservazione. Si puó notare che tutta la dimostrazione funziona anche se si sceglie, al posto della direzione di ricerca $-\nabla f(x_k)$, una generica direzione (di discesa) d_k che soddisfi la condizione

$$\frac{(d_k, \nabla f(x_k))}{\|d_k\|\,\|\nabla f(x_k)\|} \leq -c < 0. \tag{5.8}$$

Questo permette di trattare come metodo di discesa anche il metodo di Newton, quando si applichi a funzioni con Hessiana uniformemente definita positiva. Infatti la direzione $d_k = -H(x_k)^{-1}\nabla f(x_k)$ soddisfa (5.8), avendosi

$$\frac{(d_k, \nabla f(x_k))}{\|d_k\|\,\|\nabla f(x_k)\|} \leq -\frac{\lambda\|\nabla f(x_k)\|^2}{\Lambda\|\nabla f(x_k)\|^2} = -\frac{\lambda}{\Lambda}$$

(dove λ e Λ sono rispettivamente minimo e massimo autovalore di $H(x_k)^{-1}$) e di conseguenza il metodo di Newton si puó riscrivere, introducendo un parametro di ricerca unidimensionale, come

$$x_{k+1} = x_k - \beta_k H(x_k)^{-1}\nabla f(x_k). \tag{5.9}$$

Sotto queste ipotesi, lo schema (5.9) converge quindi globalmente sia in ricerca parziale che esatta (in quest'ultimo caso, ci si attende ovviamente che $\beta_k \to 1$, cioé che asintoticamente la scelta ottima sia proprio quella che corrisponde al metodo di Newton "puro").

Esercizi

1. Interpretare il metodo di Jacobi per la soluzione di sistemi lineari come metodo di discesa. A cosa corrisponde la condizione di dominanza diagonale?

2. Scrivere un algoritmo che implementi il metodi di ricerca parziale di Goldstein.
3. Dimostrare che per β sufficientemente piccolo (eventualmente, darne una stima esplicita) il metodo di gradiente con passo fisso converge.

6. Metodi di direzioni coniugate e gradiente coniugato.

Data una matrice simmetrica definita positiva A, definiamo k direzioni $d_0, \ldots, d_{k-1}$ *coniugate* rispetto ad A se soddisfano le condizioni

$$(Ad_i, d_j) = (d_i, Ad_j) = 0 \quad (i \neq j)$$

$$(Ad_i, d_i) = (d_i, Ad_i) \neq 0 \quad (i = 0, \ldots, k-1).$$

Se $d_i \in R^n$, l'esistenza di n direzioni coniugate linearmente indipendenti é assicurata da un noto teorema di agebra lineare (processo di ortogonalizzazione di Gram–Schmidt).

Supponiamo ora che la funzione f da minimizzare abbia la forma

$$f(x) = \frac{1}{2}(Ax, x) - (b, x) \tag{6.1}$$

con A definita positiva. In un generico algoritmo di minimizzazione iterativo, nella forma

$$x_{k+1} = x_k + \beta_k d_k \tag{6.2}$$

é in questo caso possibile esprimere esplicitamente il valore di β che corrisponde alla minimizzazione unidimensionale esatta. Infatti, appplicando la definizione (5.2) (omettiamo l'indice k per semplicitá):

$$\phi(\beta) = \frac{1}{2}(x + \beta d)^t A(x + \beta d) - b^t(x + \beta d) =$$

$$= \frac{1}{2}x^t Ax + \beta x^t Ad + \frac{1}{2}\beta^2 d^t Ad - b^t x - \beta b^t d =$$

$$= \phi(0) + \frac{1}{2}\beta^2(Ad, d) + \beta(\nabla f(x), d). \tag{6.3}$$

Annullando la derivata di ϕ si ottiene

$$\phi'(\beta) = \beta(Ad, d) + (\nabla f(x), d) = 0$$

che fornisce per β_k il valore

$$\beta_k = -\frac{(\nabla f(x_k), d_k)}{(Ad_k, d_k)} \tag{6.4}$$

Supponiamo ora che le prime n direzioni $d_0, \ldots, d_{n-1}$ in (6.3) siano linearmente indipendenti. Fissato x_0, definiamo la famiglia di varietá lineari

$$\pi_k = \left\{ x \in R^n : x = x_0 + \sum_{i=0}^{k-1} \alpha_i d_i \ , \quad \alpha_i \in R \right\}. \tag{6.5}$$

E' chiaro che si tratta di varietá di dimensione crescente, e che $\pi_n \equiv R^n$. Dimostriamo ora che se nello schema (6.2) si utilizzano come direzioni di spostamento direzioni coniugate rispetto ad A, si ottiene una convergenza al minimo in un numero finito di passi:

Teorema 6.1. *Sia data la funzione* (6.1)*, con A definita positiva. Allora, se nello schema (6.2) le direzioni $d_0, \ldots, d_{n-1}$ sono coniugate rispetto ad A, ed i β_k sono definiti da* (6.4)*, per ogni scelta di x_0 si ha $f(x_k) = \min_{\pi_k} f(x)$, ed in particolare $x_n = x^* = A^{-1}b$.*

Dim. Operando per induzione, occorre per prima cosa dimostrare che

$$f(x_0 + \beta_0 d_0) = \min_{\pi_1} f(x),$$

ma questo é ovviamente garantito, qualunque sia d_0, dalla scelta (6.4) per β. Si tratta di provare poi che, se $\nabla f(x_k)$ é ortogonale alla varietá π_k, allora $\nabla f(x_{k+1})$ é ortogonale alla varietá π_{k+1}, ovvero

$$(\nabla f(x_{k+1}), \bar{x} - x_0) = 0 \tag{6.6}$$

per ogni $\bar{x} \in \pi_{k+1}$. Osserviamo intanto che

$$\nabla f(x_{k+1}) = A(x_k + \beta_k d_k) - b = \nabla f(x_k) + \beta_k A d_k. \tag{6.7}$$

Poi, scritto $\bar{x}$ come $\bar{x} = x_0 + \bar{\alpha}_0 d_0 + \ldots + \bar{\alpha}_k d_k$, il primo membro di (6.6) diviene

$$(\nabla f(x_k) + \beta_k A d_k, \bar{\alpha}_0 d_0 + \ldots + \bar{\alpha}_k d_k) =$$

$$= (\nabla f(x_k), \bar{\alpha}_0 d_0 + \ldots + \bar{\alpha}_{k-1} d_{k-1}) + \bar{\alpha}_k (\nabla f(x_k), d_k) + \beta_k \sum_{i=0}^{k} \bar{\alpha}_i (A d_k, d_i) =$$

$$= \bar{\alpha}_k [(\nabla f(x_k), d_k) + \beta_k (A d_k, d_k)] \tag{6.8}$$

dove l'ultimo passaggio é ottenuto applicando la ipotesi di ortogonalitá tra $\nabla f(x_k)$ e π_k, e la definizione di direzioni coniugate. E' poi immediato osservare che l'ultimo membro di (6.8) é nullo in virtú della scelta fatta per β_k.

■

Un metodo iterativo del tipo appena visto va sotto il nome di *metodo di direzioni coniugate* (o CD nella letteratura anglosassone). Nella pratica, peró, il caso di gran lunga piú utilizzato é quello in cui le direzioni di spostamento non sono ottenute all'inizio, ma in forma ricorrente nel corso dell'algoritmo, ad esempio secondo lo schema

$$d_0 = \nabla f(x_0),$$

$$d_k = \nabla f(x_k) + \frac{\|\nabla f(x_k)\|^2}{\|\nabla f(x_{k-1})\|^2} d_{k-1} \ , \quad (k \geq 1) \tag{6.9}$$

che corrisponde al cosiddetto metodo di gradiente coniugato (CG) di Fletcher–Reeves. Si puó dimostrare, in modo concettualmente semplice anche se abbastanza tecnico, che le direzioni generate da (6.9) per una forma quadratica sono coniugate.

Questa strategia di generazione delle direzioni coniugate ha due importanti vantaggi. Da una parte, permette di memorizzare una sola direzione per volta, cosa cruciale in problemi di grandi dimensioni. Dall'altra, la sua definizione permette di applicarla anche a funzioni che non siano nella forma (6.1). In questo caso, peró, le direzioni di ricerca possono non essere particolarmente efficienti, a meno che non venga effettuata periodicamente (ad esempio dopo n iterazioni) una reinizializzazione dell'algoritmo consistente nell'applicare un passo di gradiente puro. Questo espediente si rende necessario anche nel caso di funzioni quadratiche, quando (a causa per esempio della dimensione alta) l'accumulo degli errori di troncamento renda necessario di effettuare piú delle n iterazioni teoricamente sufficienti per raggiungere il minimo. Si dimostra che la convergenza dell'algoritmo definito da (6.2), (6.9) per funzioni non quadratiche (in questo caso β_k non si ottiene da (6.4) ma da una ricerca unidimensionale) é sopralineare.

7. Metodo di penalizzazione per problemi vincolati.

Consideriamo un problema di minimizzazione vincolata con vincoli di disuguaglianza:

$$f(x^*) = \min_{x \in S} f(x) \tag{7.1}$$

$$S = \{x \in R^n : g_i(x) \leq 0 \quad (i = 1, \ldots, m)\}. \tag{7.2}$$

Ricordiamo che se le funzioni g_i sono convesse, anche S lo é, e che eventuali vincoli di uguaglianza possono essere trattati come doppi vincoli di disuguaglianza.

Suponendo per semplicitá che f sia definita su tutto R^n, un metodo molto semplice di approssimare la soluzione di (7.1), (7.2) é il cosiddetto *metodo di penalizzazione*, che si basa sulle considerazioni che seguono. Definiamo una funzione (di penalizzazione) $H(x)$ continua (in realtá, per le ragioni che si vedranno, si impone che sia anche derivabile con continuitá) e tale che:

$$H(x) = 0 \quad (x \in S) \quad , \qquad H(x) > 0 \quad (x \notin S). \tag{7.3}$$

Se H fosse definita $+\infty$ fuori di S, la minimizzazione vincolata di f equivarrebbe alla minimizzazione libera di $f + H$. Questa scelta non é peró realizzabile in pratica, e la funzione penalizzata f_ε é normalmente definita da

$$f_\varepsilon(x) = f(x) + \frac{1}{\varepsilon} H(x). \tag{7.4}$$

Si intuisce che per $\varepsilon \to 0$ il comportamento delle soluzioni (non vincolate) del problema penalizzato si avvicini a quello delle soluzioni di (7.1), (7.2). Piú precisamente, si ha il seguente

Teorema 7.1. *Indicato con x_ε un punto di minimo in R^n per (7.4), con H che soddisfa (7.3), allora le sottosuccessioni convergenti di x_ε convergono, per $\varepsilon \to 0$, a soluzioni di*

(7.1), (7.2). *In particolare, se f, g_i e H sono convesse, tutta la successione x_ε converge a x^*.*

Dim. Presa una sottosuccessione convergente di minimi $x_{\varepsilon_k} \to \bar{x}$, si ha

$$f(x_{\varepsilon_k}) \leq f_{\varepsilon_k}(x_{\varepsilon_k}) \leq \min_S f_{\varepsilon_k}(x) = f(x^*). \tag{7.5}$$

La prima disuguaglianza in (7.5) discende dalla definizione di f_ε, mentre la seconda dal fatto che x_{ε_k} é il minimo di f_{ε_k} su tutto R^n. L'uguaglianza tra gli ultimi due membri di (7.5) discende dal fatto che $H(x) = 0$ in S. Passando ora al limite per $x_{\varepsilon_k} \to \bar{x}$ si ha

$$f(\bar{x}) \leq f(x^*).$$

Il teorema é quindi dimostrato, una volta che si verifichi che $\bar{x} \in S$. Da secondo e quarto membro di (7.5) si ha

$$\frac{1}{\varepsilon_k} H(x_{\varepsilon_k}) \leq f(x^*) - f(x_{\varepsilon_k}) \tag{7.6}$$

e passando al limite si vede che il limite del primo membro, che é nonnegativo, é anche minore o uguale di $f(x^*) - f(\bar{x})$. Ma perché il primo membro abbia limite finito é necessario che $H(\bar{x}) = 0$, ovvero che $\bar{x} \in S$.

■

Osservazione. Si noti che il vincolo viene rispettato solo in senso asintotico. Niente garantisce infatti che la generica soluzione x_ε sia in S.

Una scelta abbastanza naturale per la funzione di penalizzazione H, che ne garantisce la differenziabilitá e, nel caso le g_i siano convesse, anche la convessitá, é:

$$H(x) = \sum_{i=1}^{m} [g_i(x)^+]^2. \tag{7.7}$$

Nel caso di vincoli di uguaglianza nella forma $g_i(x) = 0$ $(i = 1, \ldots, m)$, la funzione di penalizzazione puó essere scelta nella forma

$$H(x) = \sum_{i=1}^{m} g_i(x)^2 \tag{7.8}$$

ma in questo caso H non sará convessa se non nel caso di vincoli lineari. L'utilitá di considerare termini di penalizzazione differenziabili dipende naturalmente dalla posibilitá di applicare i metodi di minimizzazione iterativi visti in precedenza.

Si potrebbe pensare che per ottenere buone approssimazioni sia necessario usare valori molto piccoli di ε. In realtá questa scelta introduce un forte malcondizionamento nel problema. Per vederlo, supponiamo di dover risolvere per penalizzazione il problema in R^2:

$$f(x^*) = \min \frac{1}{2}(x_1^2 + x_2^2)$$

con il vincolo di uguaglianza

$$x_1 = 1.$$

La funzione penalizzata f_ε, con il termine di penalizzazione (7.8), ha la forma

$$f_\varepsilon(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2) + \frac{1}{\varepsilon}(x_1 - 1)^2$$

la cui matrice hessiana é

$$\begin{pmatrix} 1 + \frac{2}{\varepsilon} & 0 \\ 0 & 1 \end{pmatrix},$$

da cui si vede che valori piccoli di ε causano un aumento incontrollato del numero di condizionamento. In queste condizioni, il miglioramento di accuratezza che proviene dalla penalizzazione piú precisa puó essere completamente annullato dal rallentamento di convergenza nel metodo di minimizzazione.

8. Forma di Newton del polinomio interpolatore.

La forma di Lagrange del polinomio interpolatore relativo a $n+1$ nodi fissati ha l'inconveniente di richiedere il ricalcolo completo di tutti i polinomi $L_i(x)$ della base ogniqualvolta si aggiunge un nodo (questa é una operazione che puó rendersi necessaria per migliorare la qualitá della approssimazione). Fermo restando che il polinomio interpolatore é unico, si puó cercarne una forma che non richieda questo ricalcolo. Piú precisamente, cerchiamo un polinomio che sia nella forma

$$\Pi_n(x) = \Pi_{0\cdots n}(x) = \Pi_{0\cdots n-1}(x) + g(x). \tag{8.1}$$

In (8.1), $\Pi_{i\cdots j}(x)$ indica il polinomio interpolatore relativo ai nodi $x_i, \ldots, x_j$, mentre $g(x)$ é un polinomio di grado n, che si annulla in tutti i nodi $x_0, \ldots, x_{n-1}$. Si puó facilmente verificare che (8.1) é un polinomio di grado n, che coincide con $\Pi_{0\cdots n-1}(x)$ (e quindi con $f(x)$) nei nodi $x_0, \ldots, x_{n-1}$. $g(x)$ deve necessariamente avere la struttura

$$g(x) = f[x_0, \ldots, x_n](x - x_0)\cdots(x - x_{n-1}) \tag{8.2}$$

dove $f[x_0, \ldots, x_n]$ indica una costante (che coincide con il coefficiente del termine di grado massimo di $\Pi_{0\cdots n}$) da determinare in modo che $\Pi_n(x_n) = f(x_n)$. Con questa simbologia, il polinomio assume quindi la forma

$$\Pi_n(x) = f[x_0] + f[x_0, x_1](x - x_0) + \cdots + f[x_0, \ldots, x_n](x - x_0)\cdots(x - x_{n-1}). \tag{8.3}$$

La determinazione delle costanti $f[\cdots]$ si puó effettuare tramite il risultato seguente.

Teorema 8.1. *I coefficienti* $f[x_0], \ldots, f[x_0, \ldots, x_n]$ *di* (8.3) *soddisfano la relazione ricorrente*

$$f[x_0, \ldots, x_k] = \frac{f[x_1, \ldots, x_k] - f[x_0, \ldots, x_{k-1}]}{x_k - x_0}, \tag{8.4a}$$

$$f[x_0] = f(x_0). \tag{8.4b}$$

Dim. Dimostriamo per prima cosa che

$$\Pi_k(x) = \frac{(x - x_0)\Pi_{1\dots k}(x) - (x - x_k)\Pi_{0\dots k-1}(x)}{x_k - x_0}. \tag{8.5}$$

In effetti, calcolando il secondo membro di (8.5) nei nodi $x_1, \dots, x_{k-1}$ si ottiene

$$\frac{(x_j - x_0)\Pi_{1\dots k}(x_j) - (x_j - x_n)\Pi_{0\dots k-1}(x_j)}{x_k - x_0} = \frac{(x_j - x_0)f(x_j) - (x_j - x_k)f(x_j)}{x_k - x_0} = f(x_j)$$

mentre nel nodo x_0 si ha:

$$-\frac{(x_0 - x_k)\Pi_{0\dots k-1}(x_0)}{x_k - x_0} = f(x_0)$$

ed analogamente nel nodo x_k si ottiene:

$$\frac{(x_k - x_0)\Pi_{1\dots k}(x_k)}{x_k - x_0} = f(x_k).$$

D'altra parte il secondo membro di (8.5) é un polinomio di grado k e coincide quindi necessariamente con il polinomio interpolatore a primo membro.

Uguagliando ora i coefficienti dei termini di grado massimo di Π_k nelle forme (8.3) e (8.5), si ha

$$f[x_0, \dots, x_k] = \frac{f[x_1, \dots, x_k] - f[x_0, \dots, x_{k-1}]}{x_k - x_0},$$

che coincide con la (8.4*a*). Inoltre, per $k = 0$ si ha $f[x_0] = \Pi_k(x) = f(x_0)$ e quindi la (8.4*b*).

■

Si noti che la forma di Newton del polinomio interpolatore conserva il vantaggio, tipico della formula di Lagrange, di non richiedere ne' che i nodi siano equidistanti, ne' che siano ordinati.

9. Convergenza delle formule di Newton–Cotes composite.

Le pesanti ipotesi necessarie alla convergenza delle interpolazioni a nodi equidistanti rendono problematico il loro utilizzo anche nel calcolo di integrali. In effetti, anche nel caso di funzioni analitiche, il fatto che la formula di quadratura presenti pesi negativi quando il grado del polinomio interpolatore supera 7, causa una instabilitá verso gli errori di troncamento che ne rende comunque impossibile l'applicazione.

Supponiamo infatti che i dati $f(x_i)$ siano affetti da errori di troncamento δ_i, e che si abbia $|\delta_i| < \delta$. La differenza tra il valore corretto della quadratura numerica e quello affetto dagli errori di arrotondamento é

$$\left| \sum_k \alpha_k f(x_k) - \sum_k \alpha_k (f(x_k) + \delta_k) \right| = \left| \sum_k \alpha_k \delta_k \right| \leq \delta \sum_k |\alpha_k|. \tag{9.1}$$

Nonostante (come é noto) la somma dei pesi sia costante, non é invece vero che la *somma dei moduli* che appare all'ultimo membro di (9.1) sia uniformemente limitata al crescere del numero di nodi. Tale considerazione spinge verso l'uso di formule di Newton–Cotes composite (o generalizzate) le quali convergono sotto ipotesi alquanto piú generali, come é mostrato dal teorema seguente. Supporremo che la formula di quadratura sia nella forma

$$I_{n,m}(f) = \sum_{j=0}^{m-1} \sum_{k=0}^{n} \alpha_k^{(j)} f(x_k^{(j)}) \tag{9.2}$$

in cui l'indice j si riferisce al sottointervallo considerato, k al nodo all'interno del j–mo sottointervallo (si noti che $x_n^{(j)} = x_0^{(j+1)}$). Porremo inoltre $H_j = x_n^{(j)} - x_0^{(j)}$ e con $H = \max_j H_j$.

Teorema 9.1. *Sia f Riemann–integrabile in $[a, b]$. Allora, se la formula di quadratura composita* (9.2) *é esatta sulle costanti, si ha*

$$\lim_{H \to 0} I_{n,m}(f) = \int_a^b f(x)dx.$$

Dim. Ricordiamo che se (9.2) é esatta sulle costanti, allora $\sum_k \alpha_k^{(j)} = H_j$. D'altra parte, posto $h_j = H_j/n$, é anche noto che $\alpha_k^{(j)} = h_j w_k$, con le costanti w_k definite attraverso una formula di Newton–Cotes con distanza unitaria tra i nodi (in questo caso, $\sum_k w_k = n$). Riscrivendo allora (9.2) tramite queste costanti, si ottiene:

$$I_{n,m}(f) = \sum_{j=0}^{m-1} \frac{H_j}{n} \sum_{k=0}^{n} w_k f(x_k^{(j)}) =$$

$$= \frac{1}{n} \sum_{k=0}^{n} w_k \sum_{j=0}^{m-1} H_j f(x_k^{(j)}) \tag{9.3}$$

Nella sommatoria piú interna all'ultimo membro di (9.3) si puó riconoscere una somma integrale della f su $[a, b]$. Per $H \to 0$, si ha quindi

$$I_{n,m}(f) \to \frac{1}{n} \sum_{k=0}^{n} w_k \int_a^b f(x)dx = \int_a^b f(x)dx$$

■

10. Convergenza dei metodi ad un passo per EDO lineari.

Si supponga di approssimare il sistema di equazioni differenziali ordinarie lineari

$$\begin{cases} y'(x) = Ay(x) \\ y(x_0) = y_0 \end{cases} \tag{10.1}$$

con $y_0, y \in R^n$ ed A matrice reale $n \times n$ costante. Scriviamo la approssimazione ad un passo di (10.1) nella forma

$$\begin{cases} u_{j+1} = A_h u_j \\ u_0 = y_0 \end{cases} \tag{10.2}$$

in cui A_h é la matrice risultante dal tipo di approssimazione scelto (ad esempio, $A_h = I + hA$ per il metodo di Eulero). La forma (10.2) é giustificata dal fatto che tutti i metodi usuali ad un passo conservano la linearitá (forniscono approssimazioni lineari come (10.2) a partire da equazioni lineari). Lo schema (10.2) é *consistente* (rispettivamente, *consistente con ordine* p) se si ha:

$$y(x_{j+1}) = A_h y(x_j) + h\tau(x_j, y(x_j); h) \tag{10.3}$$

con $\|\tau(x_j, y(x_j); h)\| \to 0$ per $h \to 0$ (rispettivamente, $\|\tau(x_j, y(x_j); h)\| \leq Ch^p$). E' chiaro che, poiché per la soluzione esatta si ha

$$y(x_{j+1}) = e^{hA} y(x_j), \tag{10.4}$$

se lo schema (10.2) é esplicito perché sia consistente si deve avere $A_h = I + hA + o(h)$, e perché sia consistente con ordine p occorre che A_h sia lo sviluppo di Taylor di ordine p di e^{hA} (si puó verificare facilmente che questo accade, ad esempio, per tutti gli schemi di Runge–Kutta espliciti).

Fissando $x = M$ ed effettuando un numero di passi $k = M/h$ (supponiamo per semplicitá che questo sia un numero intero), definiamo *stabile* uno schema per cui si abbia, per una qualche costante C_0 indipendente da h,

$$\|A_h\|^{M/h} \leq C_0. \tag{10.5}$$

La sufficienza delle condizioni di stabilitá e consistenza é conseguenza del teorema che segue.

Teorema 10.1. *Se lo schema* (10.2) *é consistente con ordine* $p \geq 1$ *e stabile, allora*

$$\|u_{M/h} - y(M)\| \leq Ch^p.$$

Dim. Definita per ricorrenza la successione

$$u^*_{j+1} = A_h y_j, \tag{10.6}$$

possiamo dare per l'errore al passo generico $k = M/h$ la maggiorazione

$$\begin{aligned}\|e_k\| \leq \|y(x_k) - u^*_k\| + \|u^*_k - u_k\| \leq h\|\tau(x_{k-1}, y(x_{k-1}); h) + \|A_h\|\|e_{k-1}\| \leq \\ \leq Ch^{p+1}(1 + \|A_h\|) + \|A_h\|^2\|e_{k-2}\| \leq \\ \leq Ch^{p+1}(1 + \|A_h\| + \cdots + \|A_h\|^{k-1}) + \|A_h\|^k\|e_0\|. \end{aligned} \tag{10.7}$$

Ricordando ora che $e_0 = 0$, otteniamo

$$\|e_k\| \leq \begin{cases} Ch^{p+1}\frac{\|A_h\|^k - 1}{\|A_h\| - 1} & \text{se } \|A_h\| > 1 \\ Ch^{p+1}k & \text{se } \|A_h\| \leq 1. \end{cases} \tag{10.8}$$

Tenendo conto che $k = M/h$, nel primo caso la condizione di stabilitá garantisce che il numeratore sia limitato, mentre (per consistenza) il denominatore é un infinitesimo del primo ordine ed il teorema é quindi dimostrato. Nel secondo caso, il risultato si ottiene direttamente sostituendo l'espressione di k in (10.8).

■

Osservazione. In effetti la situazione si semplifica ulteriormente con schemi espliciti. Nel caso di schemi consistenti si ha $A_h = I + hA + o(h)$ e quindi:

$$\|A_h\| \leq 1 + h(\|A\| + o(1)) \leq 1 + h(\|A\| + \varepsilon)$$

per $h < h_\varepsilon$ (dipendente da ε), da cui:

$$\|A_h\|^{M/h} \leq [1 + h(\|A\| + \varepsilon)]^{M/h} \leq e^{M(\|A\| + \varepsilon)} \tag{10.9}$$

e quindi la consistenza implica anche la stabilitá.

11. Stabilitá assoluta di metodi a piú passi: un esempio.

Diamo qui un esempio di studio della stabilitá (normale ed assoluta) di un metodo a piú passi, il metodo del punto medio. Questo modo di procedere é del tutto generale e puó essere applicato, in effetti, a tutti gli schemi di tipo multistep. Dato il problema modello scalare

$$\begin{cases} y'(x) = -ay(x) \\ y(x_0) = y_0 \end{cases} \tag{11.1}$$

con $y_0, y \in R$ ed $a > 0$, la approssimazione per punto medio di (10.1) ha la forma

$$u_{j+1} = u_{j-1} - 2hau_j \tag{11.2}$$

in cui si é supposto di avere a disposizione opportuni valori u_0, u_1. Con un procedimento del tutto simile a quello che permette di ricondurre equazioni differenziali ordinarie di ordine maggiore di 1 a sistemi di primo ordine, mettiano (11.2) nella forma:

$$\begin{pmatrix} u_j \\ u_{j+1} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -2ah \end{pmatrix} \begin{pmatrix} u_{j-1} \\ u_j \end{pmatrix} \tag{11.3}$$

che é nella forma (analoga alla (10.2))

$$U_{j+1} = A_h U_j \tag{11.4}$$

con la condizione iniziale $U_1 = (u_0, u_1)^t$. La soluzione di (11.4) é ovviamente

$$U_{j+1} = A_h^j U_1.$$

La stabilitá dello schema (11.3) si dimostra come nella (10.9). Infatti, riferendosi ad esempio alla norma l^∞ sui vettori, si ha:

$$\|A_h\| = 1 + 2ah$$

da cui

$$\|A_h\|^{M/h} \leq [1 + 2ah]^{M/h} \leq e^{2aM}. \tag{11.5}$$

Per analizzare la stabilitá assoluta, occorre invece esaminare gli autovalori di A_h. L'equazione caratteristica é:

$$P(\lambda) = \lambda(\lambda + 2ah) - 1 = \lambda^2 + 2ah\lambda - 1 = 0$$

le cui soluzioni sono $\lambda_1 = -ah - \sqrt{1 + a^2h^2}$, $\lambda_2 = -ah + \sqrt{1 + a^2h^2}$. Si noti che $|\lambda_2| < 1$, $\lambda_2 = 1 - ah + 1/2a^2h^2 + o(h^2) = e^{-ah} + o(h^2)$ (questo é l'autovalore interessante ai fini della approssimazione di (11.1)), mentre $|\lambda_1| > 1$. Scrivendo $A_h = T^{-1}\Lambda T$, con $\Lambda = diag(\lambda_1, \lambda_2)$, si ha

$$U_{j+1} = A_h^{j+1} U_1 = T^{-1}\Lambda^{j+1} T U_1 = T^{-1} diag(\lambda_1^{j+1}, \lambda_2^{j+1}) T U_1$$

da cui si vede che la componente della soluzione relativa all'autovalore λ_1 diverge (ad h fissato) per $j \to \infty$, ed é quindi violata la condizione di stabilitá assoluta.

In conclusione (e questo é vero per tutte le approssimazioni a piú passi), la operazione di approssimare una equazione differenziale di primo ordine con un sistema alle differenze di ordine superiore introduce in generale autovalori che non sono caratteristici del problema originale. Alcuni di questi autovalori convergeranno al valore corretto e^{-ah}, mentre altri potranno causare, per x grande, fenomeni di instabilitá, anche quando la equazione originale non ne presenta.

12. Calcolo del polinomio caratteristico per matrici simmetriche tridiagonali.

Data una matrice simmetrica tridiagonale,

$$A = \begin{pmatrix} a_1 & b_1 & & & 0 \\ b_1 & a_2 & b_2 & & \\ & \ddots & \ddots & \ddots & \\ & & b_{n-2} & a_{n-1} & b_{n-1} \\ 0 & & & b_{n-1} & a_n \end{pmatrix}, \tag{12.1}$$

é possibile calcolarne il polinomio caratteristico in forma ricorrente con complessitá $O(n)$ (ricordiamo che, nel caso generale, il calcolo del polinomio caratteristico secondo la sua definizione ha una complessitá pari a quella del calcolo di un determinante, ovvero $O(n^3)$).

Premettiamo, senza dimostrarlo, un risultato sugli autovalori di matrici con questa struttura. Questo é solo uno dei molti risultati qualitativi e di localizzazione disponibili sugli autovalori di matrici del tipo (12.1).

Proposizione 12.1. *Gli autovalori della matrice* (12.1) *sono reali e semplici.*

■

Per calcolare effettivamente il polinomio caratteristico, partiamo dal minore principale formato dall'unico elemento a_1, il cui polinomio caratteristico é ovviamente

$$P_1(\lambda) = a_1 - \lambda, \tag{12.2}$$

ed aggiungiamo ad ogni passo una riga ed una colonna. Gli unici elementi non nulli che vengono aggiunti al passo k–esimo sono a_k, b_{k-1} ed il suo simmetrico. Di conseguenza, supponendo di conoscere i polinomi caratteristici dei minori fino al passo $k-1$, potremo scrivere il polinomio caratteristico al passo k come

$$P_k(\lambda) = (a_k - \lambda)P_{k-1} - b_{k-1}^2 P_{k-2}. \tag{12.3}$$

Assegnato λ, la relazione di ricorrenza (12.2)–(12.3) permette di calcolare il valore tutti i polinomi caratteristici dei minori considerati (ed infine di $P(\lambda) = P_n(\lambda)$), a patto di porre convenzionalmente $P_0(\lambda) = 1$.

Nonostante da questa relazione non si ottengano in modo semplice ne' i coefficienti ne' le derivate del polinomio caratteristico, il fatto che le radici siano semplici assicura che $P(\lambda)$ cambia segno in ogni radice e si puó quindi pensare di risolvere l'equazione caratteristica per bisezione (preceduta, ovviamente, da una fase di tabulazione necessaria a separare le singole radici). L'intervallo iniziale di ricerca puó essere dato ad esempio da $[-\|A\|, \|A\|]$; ricordiamo infatti che una qualsiasi norma di una matrice A ne maggiora il raggio spettrale. Altrimenti, si possono utilizzare le stime di Gershgorin fatte mediante somma sulle righe o sulle colonne.

13. Equazioni a Derivate Parziali: questioni generali.

Per illustrare in modo semplice i metodi numerici per Equazioni a Derivate Parziali, utilizzeremo esempi lineari, a coefficienti costanti ed in una dimensione spaziale. In genere, si suppone che i problemi che si incontrano in questi modelli semplificati siano abbastanza rappresentativi di quelli presentati da problemi piú complessi (ció, in effetti, é vero solo in parte).

Equazione del trasporto

Iniziamo con la cosiddetta equazione del trasporto, una cui forma sufficientemente generale é

$$\begin{cases} u_t(x,t) + a u_x(x,t) = f(x,t) & \text{in } \Omega = (0,1) \\ u(x,0) = u_0(x). \end{cases} \tag{13.1}$$

Per fissare le idee, supponiamo ad esempio $a > 0$. A (13.1) andrebbero aggiunte condizioni al bordo (compatibili con il dato iniziale u_0) che, per i motivi che si vedranno, possono essere imposte solo in uno dei due punti sulla frontiera a seconda del segno di a. L'equazione (13.1) é di tipo iperbolico; questa in effetti é l'unica possibilitá per equazioni del primo ordine.

Il nome di equazione del trasporto deriva da una formula di rappresentazione delle soluzioni che deriveremo ora euristicamente. Per semplicitá, consideriamo la versione omogenea di (13.1) e calcoliamo i vari termini che vi compaiono in un punto $x = \bar{x} + at$ (con $\bar{x}$ fissato):

$$\begin{cases} u_t(\bar{x} + at, t) + a u_x(\bar{x} + at, t) = 0 & \text{in } \Omega = (0,1) \\ u(\bar{x}, 0) = u_0(\bar{x}). \end{cases} \tag{13.2}$$

Si puó riconoscere che l'equazione differenziale in (13.2) equivale a

$$\frac{d}{dt} u(\bar{x} + at, t) = 0. \tag{13.3}$$

Per calcolare la soluzione in (x,t) si puó dunque integrare la (13.3) tra 0 e t, ponendo $\bar{x} = x - at$. Se $x - at > 0$, integrando si ottiene

$$u(x,t) - u(\bar{x}, 0) = 0$$

e quindi

$$u(x,t) = u_0(x - at). \tag{13.4}$$

Se $x - at < 0$, $u_0(\bar{x})$ non é definita e si deve quindi integrare partendo dal tempo t_0 tale che $x - at_0 = 0$ (ovvero $t_0 = x/a$), assegnando nel contempo la condizione al bordo $u(0,t) = c$ nel punto $x = 0$. Si ottiene cosí

$$u(x,t) - u(0,t_0) = 0$$

ovvero

$$u(x,t) = c. \tag{13.5}$$

Il risultato é quindi dato da (13.4) se $x - at > 0$, da (13.5) altrimenti, e la condizione al bordo viene assegnata solo in $x = 0$ (se fosse stato $a < 0$, la condizione al bordo sarebbe stata imposta nell'altro estremo). Naturalmente nel caso in cui l'equazione del trasporto si ponga su tutto R, la soluzione sará data sempre da (13.4). Inoltre, con semplici adattamenti si puó vedere che nel caso non omogeneo (13.1) la soluzione si puó rappresentare come

$$u(x,t) = u_0(x - at) + \int_0^t f(x - as)ds. \tag{13.6}$$

Si osserva quindi che la soluzione si propaga lungo le rette $x = at + cost$ del piano x–t (come mostrato in figura 3), conservando la sua regolaritá (questo é un fatto comune a tutte le equazioni iperboliche lineari).

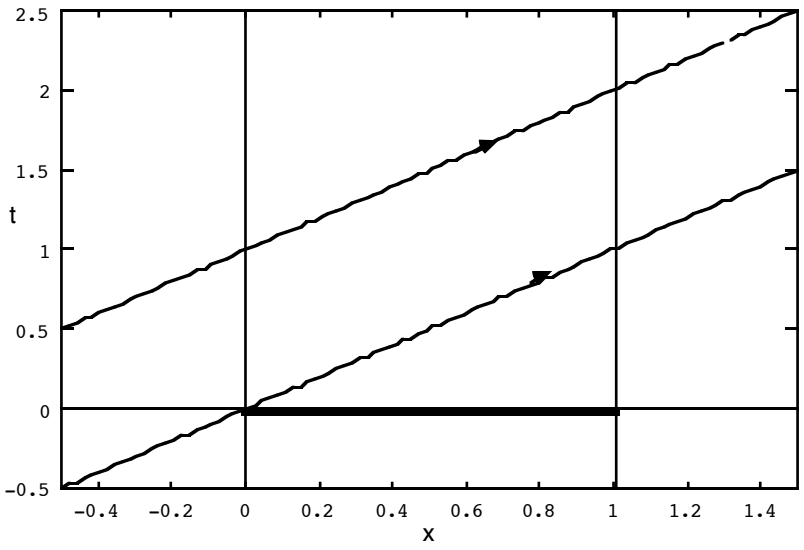

Figura 3. Propagazione della soluzione lungo le caratteristiche.

Equazione del calore

La equazione del calore in $(0,1) \subset R$ (con condizioni di Dirichlet) é nella forma

$$\begin{cases} u_t(x,t) - au_{xx}(x,t) = f(x,t) \quad \text{in } \Omega = (0,1) \\ u(x,0) = u_0(x) \\ u(0,t) = u(1,t) = 0. \end{cases} \tag{13.7}$$

Ricordiamo che, se $a > 0$, (13.7) é ben posta solo per $t > 0$, e viceversa. Le condizioni di Dirichlet sono poco significative dal punto di vista fisico, ma piú semplici da trattare analiticamente e numericamente.

E' noto che, quando la (13.7) é posta su tutto R ed omogenea ($f \equiv 0$), la sua soluzione si puó rappresentare nella forma

$$u(x,t) = \frac{1}{\sqrt{4\pi a t}} \int_R e^{\frac{(x-\xi)^2}{4at}} u_0(\xi) d\xi \tag{13.8}$$

che si ottiene formalmente con metodi di trasformata di Fourier. Su un limitato, una analisi analoga si puó condurre con la serie di Fourier; il risultato é in questo caso meno esplicito, ma, come vedremo, puó essere utilizzato a sua volta per l'approssimazione numerica della soluzione. In ogni caso la soluzione, che (13.8) permette di definire per ogni dato iniziale $u_0 \in L^2$, diviene analitica per tempi (positivi) arbitrariamente piccoli.

Per semplicitá facciamo l'ipotesi che, invece delle condizioni di Dirichlet, la soluzione soddisfi condizioni di periodicitá tipo $u(0,t) = u(1,t)$, e scriviamo la serie di Fourier (rispetto alla sola x) della soluzione $u(x,t)$ come:

$$u(x,t) = \frac{a_0(t)}{2} + \sum_{n=1}^{\infty} a_n(t) \cos n2\pi x + b_n(t) \sin n2\pi x. \tag{13.9}$$

Nella (13.9), si é trattato il tempo t come un parametro (questa scelta ricade nella filosofia della semidiscretizzazione, che sará esposta in forma generale piú avanti). In effetti, si sarebbe potuto sviluppare in serie di Fourier anche rispetto a t se si fosse operato su un intervallo limtato di tipo $[0,T]$. Possiamo ora scrivere formalmente in serie di Fourier le derivate della (13.9):

$$u_t(x,t) = \frac{\dot{a}_0(t)}{2} + \sum_{n=1}^{\infty} \dot{a}_n(t) \cos n2\pi x + \dot{b}_n(t) \sin n2\pi x, \tag{13.10}$$

$$u_x(x,t) = 2\pi \sum_{n=1}^{\infty} -n a_n(t) \sin n2\pi x + n b_n(t) \cos n2\pi x, \tag{13.11}$$

$$u_{xx}(x,t) = 4\pi^2 \sum_{n=1}^{\infty} -n^2 a_n(t) \cos n2\pi x - n^2 b_n(t) \sin n2\pi x. \tag{13.12}$$

Scrivendo in serie di Fourier anche il termine noto come

$$f(x,t) = \frac{f_0(t)}{2} + \sum_{n=1}^{\infty} f_n(t) \cos n2\pi x + g_n(t) \sin n2\pi x,$$

si puó imporre che questi sviluppi formali soddisfino l'equazione (13.7):

$$u_t - a u_{xx} - f = \frac{\dot{a}_0(t) - f_0(t)}{2} + \sum_{n=1}^{\infty} \left[\dot{a}_n(t) + 4\pi^2 a n^2 a_n(t) - f_n(t)\right] \cos n2\pi x +$$

$$+ \left[\dot{b}_n(t) + 4\pi^2 a n^2 b_n(t) - g_n(t)\right] \sin n2\pi x = 0. \tag{13.13}$$

Dalla (13.13) si ricava il sistema di (infinite) equazioni differenziali ordinarie disaccoppiate

$$\begin{cases} \dot{a}_0(t) = f_0(t) \\ \dot{a}_n(t) = -4\pi^2 an^2 a_n(t) + f_n(t) \\ \dot{b}_n(t) = -4\pi^2 an^2 b_n(t) + g_n(t) \end{cases} \tag{13.14}$$

con le condizioni iniziali $a_k(0) = \alpha_k$, $b_k(0) = \beta_k$ dove α_k, β_k sono i coefficienti di Fourier del dato iniziale u_0.

E' immediato osservare che i coefficienti di Fourier dati dalla (13.14) descrivono una funzione sufficientemente regolare da giustificare il procedimento formale che si é seguito. Inoltre, essendo il sistema (13.14) a variabili disaccoppiate, ogni coefficiente di Fourier della soluzione puó essere calcolato indipendentemente.

E' possibile approssimare il troncamento di ordine N della serie integrando numericamente in avanti il sistema differenziale formato dalle prime $2N + 1$ equazioni. Si noti che, nel caso questa operazione si effettui con uno schema esplicito, l'aumento di ordine della serie (e quindi di dimensioni del sistema) causa delle serie limitazioni nel passo di tempo. Infatti gli autovalori della matrice del sistema, anche se tutti negativi, hanno modulo $|\lambda_n| = 4\pi^2 an^2$ e di conseguenza il passo di tempo per un metodo esplicito deve essere dell'ordine di $1/N^2$ se N é l'ordine della serie troncata. Notiamo anche che la convergenza di una tale approssimazione numerica per $N \to \infty$ é conseguenza ovvia della uguaglianza di Parseval e dei risultati di convergenza delle approssimazioni per equazioni differenziali ordinarie. Naturalmente, lo stesso procedimento si potrebbe seguire con diversi sviluppi in serie di funzioni ortogonali (ad esempio polinomi di Legendre o Chebyshev).

Equazione di Poisson

La equazione di Poisson in $(0, 1)$, con condizioni di Dirichlet, ha invece la forma

$$\begin{cases} -u_{xx}(x) = f(x) \quad \text{in } \Omega = (0, 1) \\ u(0) = u(1) = 0 \end{cases} \tag{13.15}$$

ed é una equazione stazionaria (non compaiono la variabile t né le derivate rispetto a t), di tipo ellittico. La (13.15) é ben posta in senso classico se $f \in C^{0,\alpha}(0, 1)$, ed in questo caso $u \in C^{2,\alpha}(0, 1)$. Piú generalmente, si puó definire una soluzione di (13.15) in senso debole come minimo (sullo spazio $H_0^1(0, 1)$ delle funzioni con derivata prima u_x di quadrato sommabile in $(0, 1)$ e tali che $u(0) = u(1) = 0$) del funzionale

$$J(u) = \int_0^1 \left(\frac{u_x(x)^2}{2} - f(x)u(x) \right) dx \tag{13.16}$$

ed in questo caso il problema é ben posto a patto che il termine noto f sia in $L^2(0, 1)$. E' chiaro come la formulazione (13.16) sia piú generale della (13.15), in quanto richiede una regolaritá minore per la funzione u.

Anche per l'equazione di Poisson la soluzione si puó rappresentare in serie di Fourier. Anche in questo caso faremo l'ipotesi semplificatice di avere condizioni periodiche. Con le

stesse posizioni fatte in precedenza (ma supponendo ora i coefficienti di Fourier a_n, b_n, f_n e g_n indipendenti da t), si ha

$$u_{xx}(x) = 4\pi^2 \sum_{n=1}^{\infty} -n^2 a_n \cos n2\pi x - n^2 b_n \sin n2\pi x, \tag{13.17}$$

da cui:

$$f - u_{xx} = \frac{f_0}{2} + \sum_{n=1}^{\infty} [f_n - 4\pi^2 n^2 a_n] \cos n2\pi x + [g_n - 4\pi^2 n^2 b_n] \sin n2\pi x, \tag{13.17}$$

che da' il sistema di equazioni (ancora disaccoppiate)

$$\begin{cases} 0 = f_0 \\ a_n = -\dfrac{1}{4\pi^2 n^2} f_n \\ b_n = -\dfrac{1}{4\pi^2 n^2} g_n. \end{cases} \tag{13.18}$$

La prima equazione in (13.18) esprime semplicemente una condizione di buona posizione caratteristica dei problemi periodici. Dalle altre é possibile ricavare i coefficienti di Fourier della soluzione a partire da quelli del termine noto. Anche in questo caso la serie troncata si puó calcolare numericamente, in modo peraltro banale.

Esercizi

1. Dimostrare che, per una soluzione periodica, il funzionale (13.16) visto come funzione dei coefficienti di Fourier della soluzione, é quadratico, strettamente convesso ed ha un unico minimo dato da (13.18).

14. Semidiscretizzazione di Equazioni a Derivate Parziali.

Gli schemi di approssimazione che verranno considerati appartengono alla classe degli schemi alle differenze: si sostituiscono le derivate spaziali con opportuni rapporti incrementali, mentre la discretizzazione in tempo (se si tratta di equazioni evolutive) viene effettuata in un passo successivo, secondo lo schema:

Problema continuo in x e t

$\downarrow$

Problema discreto in x, continuo in t

$\downarrow$

Problema discreto in x e t

Questo modo di procedere permette di studiare in un primo tempo la convergenza della approssimazione semi–discreta (ovvero discreta solo in x) utilizzando poi la teoria giá costruita per le approssimazioni di sistemi di Equazioni Differenziali Ordinarie per studiare la convergenza del secondo passo della discretizzazione. A differenza di quello che succede per le Equazioni Differenziali Ordinarie, nella approssimazione di Equazioni a Derivate Parziali é d'uso indicare con h il parametro di discretizzazione spaziale e con k quello di discretizzazione temporale, ed a questa convenzione ci atterremo nel seguito.

La costruzione di schemi alle differenze richiede di stabilire un certo numero N di nodi $x_1, \ldots, x_N$ nell'insieme Ω (tipicamente, anche se non necessariamente, si tratta di nodi equidistanti e, in piú dimensioni, allineati secondo una griglia ortogonale). In corrispondenza, la soluzione approssimata viene data da un vettore $U = (u_1 \cdots u_N)^t$, dipendente dal tempo per i problemi evolutivi, in cui ogni variabile u_k approssima il valore $u(x_k)$ (o $u(x_k, t)$ nel caso di problemi evolutivi). La forma tipica della approssimazione semidiscreta é quindi

$$\begin{cases} \dot{u}_j(t) = f(u_1(t), \ldots, u_N(t), t) \quad (j = 1, \ldots, N) \\ u_j(0) = u_0(x_j) \end{cases} \tag{14.1}$$

mentre per equazioni stazionarie (in cui la discretizzazione va fatta esclusivamente rispetto alle derivate spaziali) si ottiene la forma

$$f(u_1(t), \ldots, u_N(t)) = 0 \quad (j = 1, \ldots, N). \tag{14.2}$$

Nel primo caso, si dovrá ulteriormente integrare numericamente il sistema differenziale (14.1), mentre nel secondo andrá risolto il sistema di equazioni (14.2). Aggiungiamo anche che negli schemi semplici che considereremo, la discretizzazione conserva la linearitá, ovvero da una equazione lineare come quelle che abbiamo considerato si ottengono schemi lineari (f é una funzione lineare dei suoi argomenti). Va tenuto presente che nei problemi di interesse pratico la dimensione N puó essere molto grande (dell'ordine delle decine o anche centinaia di migliaia) e divengono quindi cruciali le considerazioni sulla complessitá e sulla efficienza degli algoritmi.

Per studiare la convergenza delle approssimazioni semidiscrete, definiamo intanto *consistente* uno schema se soddisfa la condizione

$$\dot{u}(x_j, t) = f(u(x_1, t), \ldots, u(x_N, t), t) + \tau(x_j, t, h) \tag{14.3}$$

con $\|\tau(t, h)\| = \|(\tau(x_1, t, h) \cdots \tau(x_N, t, h))^t\| \to 0$ per $h \to 0$ ([1]). Inoltre, scritto lo schema nella forma compatta

$$\dot{U}(t) = A_h U(t), \tag{14.4}$$

diremo *stabile* uno schema che al variare di h verifichi la condizione

$$\|e^{tA_h}\| \leq K(t) \tag{14.5}$$

([1]) In effetti, come si vedrá dalla dimostrazione del teorema di Lax, é necessaria anche la ipotesi tecnica, normalmente verificata, che al variare di h la norma dell'errore di consistenza resti maggiorata da una funzione localmente sommabile rispetto a t

con $K(t)$ localmente limitata in $[0, \infty)$. Uno schema si dirá infine *convergente* se

$$u_j(t) \to u(x_j, t) \quad (h \to 0) \tag{14.6}$$

Il risultato principale della analisi di convergenza é il cosiddetto teorema di Lax–Richtmeyer:

Teorema 14.1. *Se la approssimazione* (14.4) *é consistente, allora é convergente se e solo se é stabile.*

Dim. Dimostreremo solo la sufficienza della ipotesi di stabilitá per la convergenza. La implicazione opposta (meno interessante per i nostri scopi) viene dimostrata utilizzando il teorema di Banach–Steinhaus, o principio di limitatezza uniforme.

Definiamo quindi il vettore

$$W(t) := \begin{pmatrix} u(x_1, t) \\ \vdots \\ u(x_N, t) \end{pmatrix}$$

i cui elementi sono i valori della soluzione esatta nei nodi. La condizione di consistenza (14.3) si puó riscrivere

$$\dot{W}(t) = A_h W(t) + \tau(t, h)$$

che insieme alla (14.4) da'

$$\dot{W}(t) - \dot{U}(t) = A_h(W(t) - U(t)) + \tau(t, h). \tag{14.7}$$

dove l'errore di consistenza τ va inteso come vettore. Si puó risolvere esplicitamente la (14.7) nella forma

$$(W - U)(t) = (W - U)(0) + \int_0^t e^{(t-s)A_h} \tau(s, h) ds$$

da cui, tenendo conto del fatto che $U(0) = W(0)$ e della condizione di stabilitá, si ottiene per il vettore degli errori la stima

$$\|(W - U)(t)\| \leq \int_0^t K(t - s) \|\tau(s, h)\| ds$$

che, dal teorema della convergenza dominata, assicura che $\|(W - U)(t)\| \to 0$

■

Osservazione. Si puó notare che, se lo schema é consistente con ordine p, ovvero $\|\tau(t, h)\| \leq C(t)h^p$, allora anche l'errore ha lo stesso ordine di convergenza.

15. Schema "upwind" di primo ordine per l'equazione del trasporto.

Torniamo ora all'equazione del trasporto (13.1), supponendo per comoditá $f(x) \equiv 0$, $a > 0$. Definendo $x_j = (j-1)h$ (con $h = 1/(N-1)$), consideriamo due diverse discretizzazioni della derivata $u_x(x_j, t)$:

$$u_x(x_j,t) \approx \begin{cases} 1/h(u(x_j,t) - u(x_{j-1},t)) & \text{(rapporto incrementale all'indietro)} \\ 1/h(u(x_{j+1},t) - u(x_j,t)) & \text{(rapporto incrementale in avanti).} \end{cases} \tag{15.1}$$

Lo schema semidiscreto risultante, scritto ad esempio per le differenze all'indietro, é

$$\begin{cases} \dot{u}_j = -a/h(u_j - u_{j-1}) & (j = 2, \ldots, N) \\ \dot{u}_j = 0 & (j = 1) \\ u_j(0) = u_0(x_j) \end{cases} \tag{15.2}$$

in cui si é ipotizzata una condizione di compatibilitá $u_0(0) = u(0,t) = c$.

Come vedremo, la prima delle scelte in (15.1) da' luogo ad uno schema convergente, ma non la seconda. In entrambi i casi, peró, lo schema che si ottiene é consistente: per le differenze all'indietro si ha infatti (almeno se la soluzione é sufficientemente regolare):

$$\dot{u}(x_j) = -au_x(x_j) = -\frac{a}{h}(u(x_j) - u(x_{j-1})) + o(1) \tag{15.3}$$

e ad una analoga conclusione si arriva nel caso delle differenze in avanti. Diverso é invece il comportamento dei due schemi per quanto riguarda la stabilitá. Nella forma compatta (14.4), la matrice A_h ha la struttura

$$A_h = \begin{pmatrix} 0 & 0 & & & 0 \\ a/h & -a/h & 0 & & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & & a/h & -a/h & 0 \\ 0 & & 0 & a/h & -a/h \end{pmatrix} \tag{15.4}$$

per il metodo con differenze all'indietro, e

$$A_h = \begin{pmatrix} 0 & 0 & & & 0 \\ 0 & a/h & -a/h & & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & & 0 & a/h & -a/h \\ 0 & & 0 & 0 & a/h \end{pmatrix} \tag{15.5}$$

per le differenze in avanti. Senza entrare nel dettaglio della analisi di stabilitá, notiamo che in entrambi i casi c'é un autovalore nullo; peró, mentre nel primo caso i restanti autovalori di A_h valgono $\lambda_2 = \cdots = \lambda_N = -a/h$ (ed é quindi ragionevole pensare che la norma della matrice esponenziale e^{tA_h} resti limitata), nel secondo valgono $\lambda_2 = \cdots = \lambda_N = a/h$ ed é noto che in questa situazione $\|e^{tA_h}\|$ non resta limitata e quindi lo schema non é

stabile. Nel primo caso (detto schema "upwind" o "First Order Upwind") si ha uno schema convergente, nel secondo no.

Naturalmente anche in questo caso vale la osservazione giá fatta in precedenza: uno schema semidiscreto stabile ma con autovalori di modulo crescente, se discretizzato in tempo con uno schema esplicito, richiede passi che siano abbastanza piccoli da garantire la stabilitá assoluta. Nello schema upwind, supponendo di discretizzare in tempo con un metodo di Eulero esplicito, si ha:

$$u_{j,m+1} = u_{j,m} - a\frac{k}{h}(u_{j,m} - u_{j-1,m}) \tag{15.6}$$

(in cui $u_{j,m}$ approssima il valore $u(x_j, t_m) = u((j-1)h, mk)$) e la condizione di stabilitá assoluta $k < 2/|\lambda_N|$ implica che il cosiddetto *numero di Courant* k/h deve soddisfare la limitazione

$$\frac{k}{h} < \frac{2}{a}. \tag{15.7}$$

D'altra parte, una volta dimostrata la convergenza della approssimazione semidiscreta, dalla teoria delle approssimazioni ad un passo per Equazioni Differenziali Ordinarie segue anche la convergenza dello schema completamente discreto (15.6).

La forma (15.6) permette di dare una diversa interpretazione della mancata convergenza dello schema con differenze in avanti. Notiamo che il valore della soluzione approssimata in (x_j, t_m) dipende dal valore in (x_j, t_{m-1}) ed (x_{j-1}, t_{m-1}); a loro volta i valori approssimati in questi nodi sono esprimibili in funzione di valori al passo $m-2$ e cosí via. Tornando fino all'istante iniziale si vede che il valore $u_{j,m}$ dipende dai valori $u_{j-m,0}, \ldots, u_{j,0}$. Il relativo insieme di nodi va sotto il nome di *dominio di dipendenza discreto*. D'altra parte é anche evidente che non si puó sperare di convergere alla soluzione corretta se questo dominio di dipendenza discreto non contiene il punto $x_j - at_m$ da cui parte la caratteristica che porta il valore corretto della soluzione in (x_j, t_m) (per analogia, il punto $x_j - at_m$ costituisce il dominio di dipendenza continuo). La condizione $x_{j-m} < x_j - at_m$ implica ulteriormente che

$$\frac{k}{h} < \frac{1}{a}. \tag{15.8}$$

Condizioni del tipo della (15.8) sono dette anche condizioni di Courant–Friedrichs–Lewy o CFL. Esse indicano la relazione tra i passi di discretizzazione che permette nei problemi evolutivi la inclusione del dominio di dipendenza continuo in quello discreto (condizione questa necessaria per la convergenza).

Una analisi simile mostrerebbe poi che nel caso dello schema con differenze in avanti il dominio di dipendenza discreto é costituito dai nodi $x_j, \ldots, x_{j+m}$, ed evidentemente in questo caso nessun valore del numero di Courant puó assicurare che tale insieme di nodi contenga il punto $x_j - at_m$. Una rappresentazione grafica della condizione sui domini di dipendenza per entrambi gli schemi é data in figura 4.

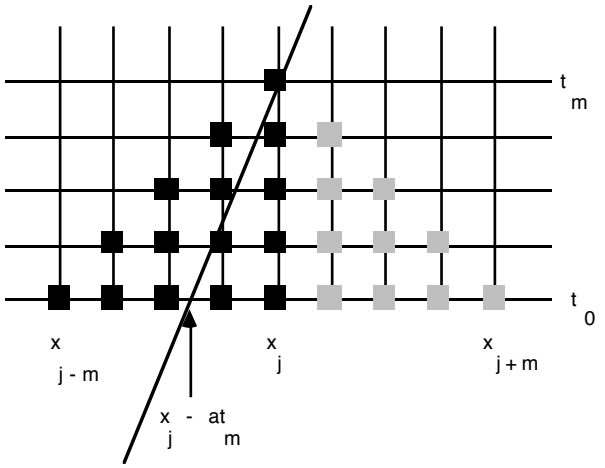

Figura 4. Domini di dipendenza per gli schemi a differenze all'idietro ed in avanti.

Esercizi

1. Riformulare lo schema upwind in modo da trattare problemi non omogenei.
2. Trovare la soluzione stazionaria (ovvero l'autovettore) associata all'autovalore nullo di A_h. Verificare che questo tipo di soluzione si presenta anche nel continuo.

16. Schema a differenze centrate per l'equazione del calore.

Riprendiamo ora in esame la equazione del calore (13.7), ancora per comoditá nella versione omogenea ($f \equiv 0$). Definendo ora $x_j = jh$ con $h = 1/(N+1)$ in modo che gli N nodi $x_1 = h, \ldots, x_N = 1 - h$ siano tutti interni all'intervallo $(0,1)$, consideriamo la discretizzazione della derivata seconda $u_{xx}(x_j, t)$ basata sul rapporto incrementale secondo:

$$u_{xx}(x_j,t) \approx \frac{1}{h}\left(\frac{u(x_{j+1},t)-u(x_j,t)}{h} - \frac{u(x_j,t)-u(x_{j-1},t)}{h}\right) =$$

$$= \frac{1}{h^2}(u(x_{j-1},t) - 2u(x_j,t) + u(x_{j+1},t)). \tag{16.1}$$

Ne risulta lo schema semidiscreto:

$$\begin{cases} \dot{u}_j = a/h^2(u_{j-1} - 2u_j + u_{j+1}) & (j = 2, \ldots, N-1) \\ \dot{u}_j = a/h^2(-2u_j + u_{j+1}) & (j = 1) \\ \dot{u}_j = a/h^2(u_{j-1} - 2u_j) & (j = N) \\ u_j(0) = u_0(x_j) \end{cases} \tag{16.2}$$

dove l'espressione delle differenze seconde agli estremi deriva dall'aver imposto la condizione di Dirichlet al bordo.

Verifichiamo ora che (16.1) é una approssimazione consistente della derivata seconda. Data $u \in C^2$, dalla formula di Taylor centrata in x_j si ha:

$$u(x_{j-1},t) = u(x_j,t) - hu_x(x_j,t) + \frac{h^2}{2}u_{xx}(\xi_-,t)$$

$$u(x_{j+1},t) = u(x_j,t) + hu_x(x_j,t) + \frac{h^2}{2}u_{xx}(\xi_+,t)$$

dove $\xi_- \in [x_{j-1},x_j]$, $\xi_+ \in [x_j,x_{j+1}]$. Sommando membro a membro e dividendo per h^2 si ottiene

$$\frac{1}{h^2}(u(x_{j-1},t) - 2u(x_j,t) + u(x_{j+1},t)) = \frac{1}{2}(u_{xx}(\xi_-,t) + u_{xx}(\xi_+,t)) \tag{16.3}$$

e, per la continuitá di u_{xx}, il secondo membro di (16.3) tende a $u_{xx}(x_j)$.

Per quanto riguarda la stabilitá dello schema, la matrice A_h ha la struttura

$$A_h = \frac{a}{h^2}\begin{pmatrix} -2 & 1 & 0 & & 0 \\ 1 & -2 & 1 & & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & & 1 & -2 & 1 \\ 0 & & 0 & 1 & -2 \end{pmatrix} = \frac{a}{h^2}A. \tag{16.4}$$

Si noti che tale matrice é simmetrica ed ha quindi autovalori reali e distinti. Applicando il teorema di Gershgorin, si ottiene per gli autovalori di A la stima $-4 \leq \lambda_i(A) \leq 0$. Dimostriamo ora che non ci sono autovalori nulli. Infatti, se cosí fosse, il determinante di A dovrebbe essere nullo, mentre sviluppandolo tramite i minori principali A_k di dimensione crescente (analogamente a quanto si é visto per il calcolo del polinomio caratteristico di una matrice tridiagonale) si ha la formula di ricorrenza

$$\begin{cases} \det A_0 = 1\ , \quad \det A_1 = -2 \\ \det A_k = -2\det A_{k-1} - \det A_{k-2} \end{cases}$$

da cui si ottiene (ad esempio per induzione) $\det A_k = (-1)^k(k+1)$ ed infine $\det A = (-1)^N(N+1)$. Di conseguenza si ha per gli autovalori di A_h la stima $-4a/h^2 \leq \lambda_i(A_h) < 0$.

Osserviamo ora dalle proprietá della matrice A_h discende che, per ogni $t \geq 0$, la matrice e^{tA_h} é simmetrica, definita positiva e con autovalori $\lambda_i(e^{tA_h}) = e^{t\lambda_i(A_h)} \leq 1$. Ció implica che nella norma euclidea si ha $\|e^{tA_h}\| < 1$ e di conseguenza la stabilitá e la convergenza dello schema.

In pratica, peró, anche in questo caso la discretizzazione in tempo pone serie limitazioni per ottenere la stabilitá assoluta. Si puó dimostrare infatti che il piú grande autovalore (in modulo) di A_h é dell'ordine di $1/h^2$. Ne segue che il passo di discretizzazione temporale per uno schema esplicito deve essere dell'ordine di h^2. Questa condizione é piuttosto limitante e fa sí che in pratica nei problemi parabolici (almeno oltre un certo grado di precisione) sia d'obbligo utilizzare schemi impliciti.

Per quanto riguarda i domini di dipendenza continuo e discreto, abbiamo giá visto che la soluzione esatta dipende, per tempi arbitrariamente piccoli, dal valore di u_0 in punti arbitrariamente lontani. Nel caso discreto, applicando considerazioni analoghe a quelle viste per lo schema upwind, si vede che la soluzione nel punto (x_j, t_m) dipende dai valori $u_{j-m,0}, \ldots, u_{j+m,0}$, come mostrato in figura 5.

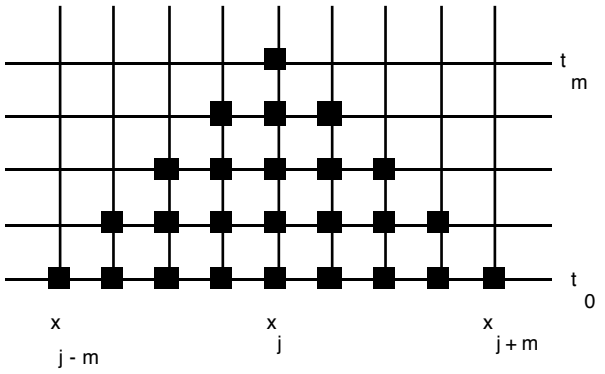

Figura 5. Dominio di dipendenza discreto per lo schema a differenze centrate.

In questo caso, é abbastanza intuitivo che al tendere a zero dei passi di discretizzazione il dominio di dipendenza discreto non debba restare limitato. La condizione (necessaria) risultante é perció $k = o(h)$. Questa condizione deve essere rispettata anche nel caso di schemi impliciti; per schemi espliciti é a sua volta implicata dalla condizione di stabilitá assoluta vista in precedenza.

Esercizi

1. Riformulare lo schema a differenze centrate in modo da trattare problemi non omogenei.
2. Estendere (con una opportuna numerazione dei nodi) l'approssimazione con rapporti incrementali secondi all'operatore laplaciano in R^2 (o R^n). Cosa si puó dire della localizzazione degli autovalori?
3. Adattare la approssimazione per differenze centrate al problema di Neumann (ovvero con le condizioni al bordo $u_x(0) = u_x(1) = 0$). Verificare che in questo caso A_h ha un autovalore nullo e dire a che tipo di soluzione stazionaria é associato.

17. Schema a differenze centrate per l'equazione di Poisson.

Consideriamo ora l'equazione di Poisson (13.15). Anche in questo caso ne prenderemo in considerazione una approssimazione per differenze centrate, la cui costruzione é stata

giá vista nel paragrafo precedente; trattandosi di un problema stazionario, comunque, la teoria generale della convergenza va opportunamente modificata.

Discretizzando quindi la derivata seconda u_{xx} come in (16.1), si ottiene lo schema:

$$\begin{cases} -1/h^2(u_{j-1} - 2u_j + u_{j+1}) = f(x_j) & (j = 2, \ldots, N-1) \\ -1/h^2(-2u_j + u_{j+1}) = f(x_j) & (j = 1) \\ -1/h^2(u_{j-1} - 2u_j) = f(x_j) & (j = N). \end{cases} \tag{17.1}$$

Che questa discretizzazione sia consistente é giá stato dimostrato nel paragrafo precedente. Tuttavia, come nel caso evolutivo, c'é comunque bisogno di una opportuna condizione di stabilitá senza la quale puó non esserci convergenza. Riscriviamo quindi lo schema (17.1) nella forma compatta

$$-A_h U = F \tag{17.2}$$

dove $U = (u_1 \cdots u_N)^t$, $F = (f(x_1) \cdots f(x_N))^t$ e A_h é definita da (16.4). Si noti che essendo A_h una matrice nonsingolare, il problema discreto é ben posto. Definito ora il vettore $W = (u(x_1) \cdots u(x_N))^t$, per la consistenza si ha

$$-A_h W = F + \tau(h) \tag{17.3}$$

dove ancora l'errore di consistenza é da interpretarsi come un vettore. Da (17.2) e (17.3) si ha:

$$A_h(W - U) = \tau(h)$$

da cui infine si ottiene per il vettore degli errori nei nodi la maggiorazione

$$\|W - U\| \leq \|A_h^{-1}\| \|\tau(h)\|. \tag{17.4}$$

La condizione di stabilitá dello schema, che ne garantisce la convergenza se $\|\tau(h)\| \to 0$, é la limitatezza (uniforme rispetto ad h) della norma $\|A_h^{-1}\|$. D'altra parte, si puó dimostrare che l'autovalore di modulo minimo $\lambda_1(A_h)$ tende, per $h \to 0$, ad un limite finito e non nullo, che é proprio il primo autovalore del laplaciano. Di conseguenza, essendo A_h^{-1} una matrice simmetrica e definita negativa, nella norma euclidea si avrá:

$$\|A_h^{-1}\| \leq \frac{1}{\lambda_1(A_h)},$$

il che garantisce la convergenza dello schema.

Esercizi

1. Verificare che in presenza di condizioni di Neumann (quindi con un autovalore nullo per A_h) lo schema (17.1)–(17.2) non é ben posto per tutti i termini noti. Caratterizzare i termini noti per cui il problema ammette soluzione.
2. Modificare lo schema in modo da inserire termini di ordine zero nell'equazione (ad esempio considerando $-\Delta u + u$ al posto di $-\Delta u$). Dire come va adattata la teoria della convergenza e studiare le proprietá della matrice A_h modificata.