Appunti del corso di Analisi Numerica

Roberto Ferretti

12 dicembre 2017

ROMA TRE
UNIVERSITÀ DEGLI STUDI

Indice

1 Sistemi di equazioni lineari e nonlineari

Nel caso generale, il problema che si pone è di risolvere un sistema di equazioni della forma

$$F(x) = 0. \quad (F : \mathbb{R}^n \to \mathbb{R}^n) \tag{1.1}$$

Come è noto, non esistono risultati di esistenza e molteplicità di soluzioni per questo problema, meno che in situazioni particolari (teorema degli zeri in $\mathbb{R}$, teoremi di punto fisso in spazi metrici completi).

Nel caso in cui F sia lineare, il sistema si scrive nella forma estesa

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n \end{cases} \tag{1.2}$$

o in forma compatta $Ax = b$. In questo caso è noto che il sistema è univocamente risolubile a patto che la matrice A sia nonsingolare, ed esistono algoritmi risolutivi che possono essere completati in un numero finito di operazioni.

Le strategie più comuni di soluzione numerica di questo problema sono di tre tipi: metodi diretti, iterativi e di minimizzazione.

Metodi diretti – Si applicano solo al problema lineare (1.2). Consentono di arrivare alla soluzione con un numero finito di operazioni, a patto di operare *in aritmetica esatta*. In genere, si ritengono convenienti per sistemi lineari pieni e relativamente di piccole dimensioni. Occupano più memoria degli altri metodi e soffrono di una maggiore instabilità.

Metodi iterativi – Si possono applicare sia al problema lineare che a quello nolineare. Generano successioni che, sotto opportune ipotesi, convergono ad una soluzione del sistema. In generale la soluzione non viene raggiunta in un numero finito di passi, quindi nei sistemi lineari possono essere meno precisi (a parità di numero di operazioni), a meno che non vengano applicati a problemi sparsi. Occupano però meno memoria e sono più stabili dei metodi diretti.

Metodi di minimizzazione – Sono metodi iterativi che sotto ipotesi opportune convergono al minimo di una funzione $f : \mathbb{R}^n \to \mathbb{R}$. Se si può trovare una funzione f tale che $F(x) = \nabla f(x)$, e se f ha minimo (nel caso dei sistemi lineari, se A è definita positiva si può porre $f(x) = 1/2(Ax, x) - (b, x)$), si possono applicare alla soluzione di (1.1) o (1.2). Una maniera standard di applicarli consiste nel porre $f(x) = F(x)^t F(x)$, il che corrisponde alla

strategia del *minimo residuo*. I metodi di minimizzazione hanno tutte le caratteristiche dei metodi iterativi; sono particolarmente efficienti se F è già nella forma di un gradiente. Se invece si minimizza il residuo, allora sia a causa della eventuale mancanza di convessità di f, sia a causa di un peggioramento del condizionamento, possono essere meno competitivi.

Esempio: sistemi lineari Anche nel caso dei sistemi lineari, in cui la soluzione si fornisce in modo esplicito, può non essere ovvio il fatto che i metodi per calcolarla possano portare a complessità computazionali molto diverse. Se nel caso dei sistemi di due equazioni in due incognite tutti i metodi più usuali (sostituzione, eliminazione, Cramer) calcolano la soluzione nello stesso modo, già per un sistema di tre equazioni in tre incognite la situazione è molto diversa.

Infatti, con il metodo di Cramer ogni incognita viene calcolata come

$$x_k = \frac{\Delta_k}{\Delta} \quad (k = 1, 2, 3)$$

dove Δ_k e Δ sono determinanti di matrici di ordine 3. Sviluppando questi quattro determinanti, si deve sommare in ognuno il contributo di 6 termini, ognuno prodotto di 3 elementi della matrice. Per ognuno di questi termini servono dunque 2 moltiplicazioni ed una somma. Al tutto vanno aggiunte tre divisioni con cui vengono calcolati i valori delle x_k, per un totale di

$$4 \cdot 6 \cdot (2 + 1) + 3 = 75$$

operazioni in virgola mobile.

Nel caso invece del metodo di eliminazione (Gauss–Jordan), si tratta di portare il sistema nella forma diagonale

$$\begin{cases} \alpha_1 x_1 = \beta_1 \\ \alpha_2 x_2 = \beta_2 \\ \alpha_3 x_3 = \beta_3 \end{cases}$$

eliminando la $k-$esima variabile da tutte le equazioni meno la $k-$esima mediante operazioni di combinazione lineare tra equazioni. Poiché le variabili sono tre ed ognuna va eliminata da due equazioni, queste combinazioni lineari vanno effettuate 6 volte. Ognuna di loro richiede una divisione per calcolare il coefficiente della combinazione lineare e 3 prodotti più 3 somme per ricalcolare i tre termini non nulli della nuova equazione. Aggiungendo le 3 divisioni con cui sono calcolati i valori delle incognite, si ottiene un totale di

$$6 \cdot (1 + 3 + 3) + 3 = 45$$

operazioni in virgola mobile, contro le 75 del metodo di Cramer.

Un terzo modo di risolvere il sistema è di portarlo, sempre mediante combinazioni lineari di righe, nella forma triangolare

$$\begin{cases} \alpha_{11}x_1 + \alpha_{12}x_2 + \alpha_{13}x_3 = \beta_1 \\ \alpha_{22}x_2 + \alpha_{23}x_3 = \beta_2 \\ \alpha_{33}x_3 = \beta_3 \end{cases}$$

calcolando poi il valore delle incognite a partire dall'ultima, procedendo all'indietro. In questo caso al sistema triangolare si arriva con solo tre operazioni di eliminazione, dopodiché x_3 si calcola con una divisione, x_2 con un prodotto, una somma ed una divisione ed x_1 con due prodotti, due somme ed una divisione, per un totale di

$$3 \cdot (1 + 3 + 3) + 1 + 3 + 5 = 30$$

operazioni in virgola mobile, contro le 75 del metodo di Cramer e le 45 del metodo di Gauss–Jordan. Il divario di complessità diviene ancora più vistoso in dimensione maggiore.

Esempio: equazioni scalari nonlineari Premendo ripetutamente il tasto del coseno su una calcolatrice da tasca, si ottiene una successione convergente. Il limite di questa successione è chiaramente il numero reale che coincide con il suo coseno, ovvero la soluzione della equazione (di punto fisso)

$$x = \cos x.$$

La convergenza della successione, ottenuta per ricorrenza,

$$x_{k+1} = \cos x_k$$

deriva dal fatto che il secondo membro è una *contrazione*, almeno in un opportuno intorno della soluzione.

1.1 I metodi diretti

La logica generale dei metodi diretti è di risolvere il sistema lineare (1.2) mediante manipolazioni algebriche che portino alla soluzione esatta (in aritmetica esatta) in un numero finito di passi. Nei casi più semplici, questa operazione si effettua riducendo il sistema alla forma triangolare.

Infatti, dato un sistema triangolare

$$\begin{cases} \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1n}x_n = \beta_1 \\ \alpha_{22}x_2 + \cdots + \alpha_{2n}x_n = \beta_2 \\ \vdots \\ \alpha_{nn}x_n = \beta_n \end{cases} \tag{1.3}$$

la sua soluzione si puo' calcolare tramite le cosiddette *sostituzioni all'indietro* come

$$x_n = \frac{\beta_n}{\alpha_{nn}}$$

$$x_k = \frac{1}{\alpha_{kk}} \left(\beta_k - \sum_{j=k+1}^{n} \alpha_{kj} x_j \right) \quad (k = n-1, \ldots, 1) \tag{1.4}$$

in cui il valore di una incognita viene ottenuto sulla base di quelle (successive) già calcolate.

1.1.1 Il metodo di eliminazione di Gauss

Il metodo di eliminazione di Gauss è in genere noto dai corsi precedenti. Ne richiamiamo comunque brevemente la filosofia.

La logica del metodo di eliminazione è di riportare un generico sistema quadrato nella forma (1.2) alla forma del sistema triangolare (1.3). Per mettere in evidenza i vari passi di eliminazione delle variabili, riscriviamo il sistema come $A^{(1)}x = b^{(1)}$, o per esteso:

$$\begin{cases} a_{11}^{(1)} x_1 + a_{12}^{(1)} x_2 + \cdots + a_{1n}^{(1)} x_n = b_1^{(1)} \\ a_{21}^{(1)} x_1 + a_{22}^{(1)} x_2 + \cdots + a_{2n}^{(1)} x_n = b_2^{(1)} \\ \vdots \\ a_{n1}^{(1)} x_1 + a_{n2}^{(1)} x_2 + \cdots + a_{nn}^{(1)} x_n = b_n^{(1)} \end{cases} \tag{1.5}$$

Si parte dalla eliminazione della variabile x_1 sottraendo alla riga k–esima la prima moltiplicata per il cosiddetto *moltiplicatore* $m_{k1} = \frac{a_{k1}^{(1)}}{a_{11}^{(1)}}$. Alla fine di $n-1$ combinazioni lineari così costruite, il sistema sarà nella forma

$$\begin{cases} a_{11}^{(1)} x_1 + a_{12}^{(1)} x_2 + \cdots + a_{1n}^{(1)} x_n = b_1^{(1)} \\ \qquad\quad a_{22}^{(2)} x_2 + \cdots + a_{2n}^{(2)} x_n = b_2^{(2)} \\ \qquad\quad \vdots \\ \qquad\quad a_{n2}^{(2)} x_2 + \cdots + a_{nn}^{(2)} x_n = b_n^{(2)} \end{cases} \tag{1.6}$$

e la variabile x_1 sarà presente solo nella prima equazione. Si riparte quindi dalla variabile x_2 con la stessa modalità (meno che per il fatto che le combinazioni lineari si effettuano sulle righe dalla terza in poi con i moltiplicatori $m_{ki} = \frac{a_{ki}^{(i)}}{a_{ii}^{(i)}}$), e così via. Il risultato è una sequenza di sistemi equivalenti

$$A^{(1)}x = b^{(1)},$$

$$A^{(2)}x = b^{(2)},$$

$$\vdots$$

$$A^{(n)}x = b^{(n)},$$

l'ultimo dei quali è un sistema triangolare

$$\begin{cases} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \cdots + a_{1n}^{(1)}x_n = b_1^{(1)} \\ a_{22}^{(2)}x_2 + \cdots + a_{2n}^{(2)}x_n = b_2^{(2)} \\ \vdots \\ a_{nn}^{(n)}x_n = b_n^{(n)} \end{cases} \tag{1.7}$$

che è nella forma (1.3) e può essere risolto per sostituzioni all'indietro.

In tutto questo procedimento, si è implicitamente supposto che all'i–esimo passo l'elemento (detto *pivot*) $a_{ii}^{(i)}$ sia non nullo. Questo non è necessariamente vero; è vero però che se la matrice A è nonsingolare *almeno uno* tra gli elementi $a_{ki}^{(i)}$ (con $k \geq i$) è non nullo. L'algoritmo viene perciò definito a meno di permutazioni tra la riga i–esima ed una riga successiva.

Stabilità E' abbastanza intuibile che nel metodo di eliminazione la operazione di combinazione lineare tra righe porti a lungo andare alla perdita di cifre significative per sottrazione, e questo effetto porta alla impossibilità di applicare il metodo di eliminazione in dimensione alta.

Moltiplicatori di modulo molto grande causano a loro volta una ulteriore amplificazione di questo errore, che si traduce nel fatto che il sistema che si ottiene dal procedimento di eliminazione è solo approssimativamente triangolare. D'altra parte, nella operazione di prodotto della riga i–esima per il moltiplicatore, gli elementi $a_{ij}^{(i)}$ per $j < i$, se non nulli, si propagano con modulo dipendente dal moltiplicatore. Per aumentare la stabilità dell'algoritmo l'idea è quindi quella di scegliere il pivot in modo che i moltiplicatori risultanti siano "piccoli".

Nella strategia di *pivoting parziale*, che è basata su permutazioni delle righe, al passo i–esimo di eliminazione viene portata in i–esima posizione l'equazione j–esima (con $j \geq i$), dove

$$|a_{ji}^{(i)}| = \max_{k \geq i} |a_{ki}^{(i)}|.$$

Nella strategia di *pivoting totale*, basata su permutazioni sia di righe che di colonne, al passo i–esimo di eliminazione viene portato in posizione di pivot l'elemento $a_{jl}^{(i)}$ (con $j, l \geq i$) per cui si abbia

$$|a_{jl}^{(i)}| = \max_{k,h \geq i} |a_{kh}^{(i)}|.$$

Questo significa portare in i–esima posizione l'equazione j–esima, e scambiare li variabili x_i e x_l (operazione quest'ultima di cui va tenuta opportunamente memoria).

Complessità Per prima cosa calcoliamo la complessità della soluzione del sistema triangolare (1.3). In questa situazione, da (1.4) si ottiene che il calcolo della k–esima variabile richiede $n-k$ somme ed altrettanti prodotti. Sommando tutti questi contributi si ottiene una complessità globale di

$$2+4+6+\cdots+2(n-1) = 2O\left(\frac{n^2}{2}\right) = O(n^2)$$

operazioni in virgola mobile.

Nella fase di triangolarizzazione del sistema, d'altra parte, per eliminare la variabile k–esima si effettuano $(n-k)^2$ prodotti ed altrettante somme. La complessità globale è quindi di

$$2(n-1)^2+2(n-2)^2+\cdots+8+2 = 2O\left(\frac{n^3}{3}\right) = O\left(\frac{2n^3}{3}\right)$$

operazioni, ed è quindi la complessità asintoticamente prevalente.

Per quanto riguarda l'operazione di pivoting, la sua complessità è data dal numero di confronti necessari a determinare il nuovo pivot (mentre la complessità della operazione di scambio di righe/colonne è sempre lineare). Per ogni eliminazione, tale numero è lineare in n per il pivoting parziale e quadratico per il pivoting totale. Ne risulta un incremento di complessità che è quadratico (e quindi asintoticamente trascurabile) per il pivoting parziale, e cubico (quindi *non* trascurabile asintoticamente) per il pivoting totale.

Risultati fondamentali

- Esistenza della soluzione per il metodo di eliminazione

  Teorema 1.1 *Se la matrice A del sistema (1.5) è nonsingolare, esiste una permutazione delle equazioni per cui questo algoritmo può essere completato. Il vettore che si ottiene dalla sostituzione all'indietro (1.4) è la unica soluzione del sistema (1.5).*

  Dim. Basta osservare che ogni passo del processo di eliminazione trasforma il sistema in un sistema equivalente. D'altra parte, questa equivalenza non sussisterebbe se non fosse possibile trovare ad ogni passo

un pivot non nullo (infatti, in questo caso da un sistema ben posto se ne sarebbe ottenuto un altro senza soluzione o con infinite soluzioni).

■

- Necessità delle permutazioni di righe

  Teorema 1.2 *Se tutti i minori principali di A sono nonsingolari, in particolare se la matrice A è a diagonale dominante o definita positiva, allora l'algoritmo di eliminazione può essere completato senza permutazione delle righe.*

1.1.2 La fattorizzazione LU

Dal metodo di eliminazione discende la possibilità di fattorizzare (a meno di permutazioni di righe) la matrice A nel prodotto di due matrici triangolari, L triangolare inferiore e U triangolare superiore. Si può quindi scrivere, in modo chiaramente non unico,

$$A = LU$$

e porre il sistema (1.5) nella forma

$$Ax = LUx = b.$$

Introducendo una variabile ausiliaria z, la soluzione del sistema (1.5) si ottiene quindi dalla successiva soluzione dei due sistemi triangolari $Lz = b$ e $Ux = z$.

Nel caso in cui sia necessaria una permutazione P di righe, o si utilizzi una fattorizzazione pivotata, si ha

$$PA = LU$$

ed i due sistemi triangolari da risolvere sono $Lz = Pb$ e $Ux = z$. Di questa permutazione di righe occorre perciò tenere memoria nel caso in cui si risolvano più sistemi lineari con la stessa matrice A ma con diversi termini noti.

Un modo per calcolare la fattorizzazione LU della matrice A sarà dato nel Teorema 1.3. Un'altra possibilità è di utilizzare la formula di prodotto tra L e U,

$$a_{ij} = \sum_k l_{ik} u_{kj}$$

con le condizioni

$$l_{ik} = 0 \quad (k > i)$$

$$u_{kj} = 0 \quad (k > j)$$
$$l_{ii} = 1$$

(quest'ultima condizione porta ad un risultato univocamente determinato, che coincide con la cosiddetta *fattorizzazione di Doolittle* per la quale si ha $U = A^{(n)}$). Partendo dalla prima riga di A si ha:

$$a_{1j} = l_{11}u_{1j} = u_{1j}$$

da cui si ottiene $u_{1j} = a_{1j}$ e quindi tutta la prima riga di U. Passando poi alla prima colonna di A:

$$a_{i1} = l_{i1}u_{11}$$

e poiché l'elemento u_{11} è stato già calcolato in precedenza, si ottiene per $i \geq 2$:

$$l_{i1} = \frac{a_{i1}}{u_{11}}.$$

Dalla seconda riga di A si ha:

$$a_{2j} = l_{21}u_{1j} + l_{22}u_{2j}$$

e quindi, per $j \geq 2$,

$$u_{2j} = a_{2j} - l_{21}u_{1j},$$

mentre considerando la seconda colonna di A si ha analogamente

$$a_{i2} = l_{i1}u_{12} + l_{i2}u_{22}$$

da cui si ottiene per $i \geq 3$:

$$l_{i2} = \frac{1}{u_{22}}(a_{i2} - l_{i1}u_{12}).$$

L'algoritmo continua in questo modo, alternativamente ottenendo per $j \geq p$ dalla p–esima riga di A:

$$u_{pj} = a_{pj} - \sum_{k<p} l_{pk}u_{kj}, \tag{1.8}$$

e per $i > q$ dalla q–esima colonna di A:

$$l_{iq} = \frac{1}{u_{qq}}\left(a_{iq} - \sum_{k<q} l_{ik}u_{kq}\right). \tag{1.9}$$

La necessità di riordinare le righe della matrice A appare nel caso in cui in (1.9) compaia un pivot u_{qq} nullo. Esistono varianti pivotate di questo algoritmo.

Analogamente a quanto si è fatto per la fattorizzazione di Doolittle ponendo $l_{ii} = 1$, si possono ottenere altri tipi di fattorizzazione fissando in modo diverso gli elementi sulle diagonali. Con la scelta $u_{ii} = 1$ si ottiene la cosiddetta *fattorizzazione di Crout*, mentre scegliendo $l_{ii} = u_{ii}$ (si dimostra che ciò è possibile nelle matrici definite positive) si ottiene la *fattorizzazione di Cholesky*, che pone A nella forma

$$A = LL^t.$$

Con un procedimento del tutto analogo a quanto fatto per la fattorizzazione di Doolittle, si ottiene dalla p–esima riga di A:

$$l_{pp} = \left(a_{pp} - \sum_{k<p} l_{pk}^2 \right)^{\frac{1}{2}}, \tag{1.10}$$

e per $i > p$:

$$l_{ip} = \frac{1}{l_{pp}} \left(a_{ip} - \sum_{k<p} l_{ik} l_{pk} \right). \tag{1.11}$$

In questo caso non serve utilizzare le colonne di A, vista la sua simmetria.

La fattorizzazione LU permette anche di calcolare determinante ed inversa della matrice A. Per quanto riguarda il determinante, ricordando che in una matrice triangolare esso è dato dal prodotto degli elementi sulla diagonale, e che, se $A = LU$, allora $\det A = \det L \det U$, si ha immediatamente

$$\det A = \prod_k u_{kk}.$$

Per quanto riguarda l'inversa A^{-1}, si può notare che la sua colonna i–sima η_i è soluzione del sistema lineare

$$A\eta_i = e_i. \tag{1.12}$$

La matrice A^{-1} si può calcolare quindi risolvendo gli n sistemi lineari (1.12) per $i = 1, \ldots, n$. La soluzione di ognuno di questi sistemi, d'altra parte, ha a sua volta complessità quadratica se A è stata fattorizzata, nel qual caso la complessità risultante per il calcolo dell'inversa è cubica.

Stabilità Nella fattorizzazione LU, trattandosi in sostanza di un riarrangiamento delle stesse operazioni che si effettuano nel metodo di eliminazione, si hanno caratteristiche di stabilità simili a quelle del metodo di Gauss. La situazione cambia nella fattorizzazione di Cholesky: in questo caso da (1.10) si ottiene

$$l_{pp} = \left(a_{pp} - \sum_{k<p} l_{pk}^2 \right)^{\frac{1}{2}} \leq a_{pp}^{\frac{1}{2}}, \tag{1.13}$$

che mostra che gli elementi sulla diagonale del fattore triangolare (e di conseguenza *tutti* gli elementi della matrice) crescono, rispetto agli elementi di A, più lentamente di quanto non accada nella fattorizzazione LU. Questa caratteristica permette al metodo di Cholesky di essere applicabile in dimensione considerevolmente più alta rispetto al metodo di Gauss o alla fattorizzazione LU.

Complessità Il numero di operazioni in virgola mobile necessarie alla fattorizzazione della matrice si ottiene sommando i contributi di (1.8) ed (1.9). Nel triangolo superiore il calcolo di ognuno degli $n - p + 1$ elementi della p–esima riga di U richiede $(p - 1)$ prodotti ed altrettante somme; si effettuano quindi

$$2\Big(0 \cdot n + 1 \cdot (n-1) + 2 \cdot (n-2) + \cdots + (n-1) \cdot (n-(n-1))\Big) =$$

$$= 2n(1 + 2 + \cdots + (n-1)) - 2(1 + 4 + \cdots + (n-1)^2) =$$

$$= 2O\left(\frac{n^3}{2}\right) - 2O\left(\frac{n^3}{3}\right) = O\left(\frac{n^3}{3}\right)$$

operazioni (come si può facilmente verificare espandendo i prodotti della prima riga). Poiché (asintoticamente) lo stesso numero di operazioni viene effettuato per calcolare L dal triangolo inferiore di A, possiamo concludere che il costo totale della fattorizzazione è lo stesso del metodo di eliminazione di Gauss. Per la fattorizzazione di Cholesky si tratta invece di calcolare solo uno dei due fattori triangolari, e perciò la complessità scende a $O(n^3/3)$ operazioni.

D'altra parte, occorre ricordare che nel metodo di eliminazione questo costo è richiesto anche quando *pur restando fissa la matrice* A, viene cambiato il vettore dei termini noti. In questa eventualità, nel metodo di fattorizzazione LU non occorre fattorizzare di nuovo la matrice ed il costo computazionale resta solo quello della soluzione dei due sistemi triangolari (ovvero $O(2n^2)$). Nel caso di sistemi con matrice definita positiva, il metodo di Cholesky è invece conveniente anche nella soluzione di un singolo sistema lineare.

Infine il calcolo dell'inversa, richiedendo prima la fattorizzazione e poi la soluzione di $2n$ sistemi triangolari, ha una complessità globale di $O(8n^3/3)$ operazioni in virgola mobile.

Risultati fondamentali

- Esistenza della fattorizzazione LU

Teorema 1.3 *Se la matrice A del sistema (1.5) è nonsingolare, esiste una permutazione P delle equazioni che permette di scrivere $PA = LU$ con L triangolare inferiore, U triangolare superiore (tale permutazione coincide con quella che permette l'esecuzione dell'algoritmo di eliminazione). Inoltre, gli elementi della matrice L coincidono con i moltiplicatori, e più precisamente*

$$l_{ki} = \begin{cases} 1 & se\ k = i \\ m_{ki} = \frac{a_{ki}^{(i)}}{a_{ii}^{(i)}} & se\ k > i \end{cases} \tag{1.14}$$

Dim. Osserviamo che nel Metodo di Eliminazione (in assenza di permutazioni di righe) tutte le trasformazioni che si effettuano sulla matrice A sono operazioni sulle righe, consistenti nel rimpiazzare una riga con la sua combinazione lineare con le righe precedenti. In termini di matrici trasformanti (vedi §A.1), l'eliminazione di una generica variabile x_i equivale a ottenere $A^{(i+1)} = T_i A^{(i)}$ moltiplicando a sinistra $A^{(i)}$ per una matrice di trasformazione

$$T_i = \begin{pmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & 1 & & \\ & & -m_{i+1,i} & & \\ & & \vdots & \ddots & \\ 0 & & -m_{ni} & & 1 \end{pmatrix}$$

dove gli elementi m_{ki} per $k > i$ sono i moltiplicatori già definiti in precedenza. Il prodotto di tutte queste trasformazioni è ancora una matrice triangolare inferiore che indicheremo con $\Lambda = T_{n-1} T_{n-2} \cdots T_1$. Si ha quindi, supponendo di avere già effettuato la corretta permutazione di righe ed indicando con U la matrice del sistema triangolarizzato:

$$\Lambda A^{(1)} = A^{(n)} = U$$

e quindi

$$A = A^{(1)} = \Lambda^{-1} A^{(n)} = LU$$

in cui si è posto $L = \Lambda^{-1}$. Questa ultima matrice è triangolare inferiore in quanto inversa di una matrice triangolare inferiore. Per quanto riguarda la seconda parte dell'enunciato, ponendo

$$m_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ m_{k+1,k} \\ \vdots \\ m_{nk} \end{pmatrix}$$

si ha $T_k = I - m_k e_k^t$, mentre $T_k^{-1} = I + m_k e_k^t$. Infatti,

$$(I - m_k e_k^t)(I + m_k e_k^t) = I + m_k e_k^t - m_k e_k^t - m_k e_k^t m_k e_k^t = I$$

poiché l'ultimo termine nello sviluppo del prodotto è nullo, essendo nullo il prodotto scalare $e_k^t m_k$. Si ha quindi

$$L = \Lambda^{-1} = T_1^{-1} T_2^{-1} \cdots T_{n-1}^{-1} =$$

$$= (I + m_1 e_1^t)(I + m_2 e_2^t) \cdots (I + m_{n-1} e_{n-1}^t).$$

Per dimostrare la (1.14), verifichiamo per induzione che si ha

$$T_1^{-1} T_2^{-1} \cdots T_k^{-1} = I + m_1 e_1^t + m_2 e_2^t + \cdots + m_k e_k^t. \tag{1.15}$$

Questa forma vale sicuramente per $k = 1$, ed inoltre

$$(I + m_1 e_1^t + m_2 e_2^t + \cdots + m_k e_k^t)(I + m_{k+1} e_{k+1}^t) =$$

$$= I + m_1 e_1^t + m_2 e_2^t + \cdots + m_k e_k^t + m_{k+1} e_{k+1}^t +$$

$$+ m_1 e_1^t m_{k+1} e_{k+1}^t + \cdots + m_k e_k^t m_{k+1} e_{k+1}^t =$$

$$= I + m_1 e_1^t + m_2 e_2^t + \cdots + m_k e_k^t + m_{k+1} e_{k+1}^t$$

dove l'ultimo passaggio è motivato dal fatto che i prodotti scalari $e_1^t m_{k+1}, \ldots, e_k^t m_{k+1}$ sono tutti nulli. E' soddisfatta quindi la (1.15) e di conseguenza, come è facile verificare, la (1.14).

■

- Esistenza della fattorizzazione di Cholesky

Teorema 1.4 *Se la matrice A è definita positiva, allora si può fattorizzare nella forma (detta di Cholesky) $A = LL^t$ con L triangolare inferiore.*

1.1.3 La fattorizzazione QR

Un'altra tecnica importante di fattorizzazione di una matrice $n \times m$ (con $n \geq m$, e quindi *non necessariamente quadrata*) consiste nel decomporla nel prodotto QR in cui Q è una matrice ortogonale $n \times n$ ed R una matrice $n \times m$ in cui $r_{ij} = 0$ se $i > j$, in particolare triangolare superiore se $m = n$. Nel caso del sistema (1.5), dalla fattorizzazione QR della matrice A si può calcolare la soluzione del sistema risolvendo nell'ordine i sistemi lineari $Qz = b$ (che ha soluzione $z = Q^t b$) e $Rx = z$ (che è triangolare).

Un caso in cui questo tipo di fattorizzazione si rivela particolarmente utile è la soluzione del sistema delle equazioni normali che si pone nella approssimazione per errore quadratico minimo (vedi §5.1.4).

La tecnica di fattorizzazione QR più efficiente è tramite le matrici di Householder, che verranno trattate nella sezione dedicata agli autovalori.

Complessità La fattorizzazione QR ha complessità maggiore della fattorizzazione LU, il suo uso è quindi in genere limitato ai sistemi di equazioni normali, o comunque a sistemi particolarmente malcondizionati. Una analisi più dettagliata della complessità del metodo di Householder verrà fatta nella sezione sugli autovalori.

Risultati fondamentali

- Esistenza della fattorizzazione QR

  Teorema 1.5 *Se la matrice quadrata A è nonsingolare, esiste una matrice ortogonale Q che permette di scrivere $A = QR$ con R triangolare superiore.*

  Dim. Applicando il metodo di ortogonalizzazione di Gram–Schmidt alle colonne della matrice A, si ha che tutte le operazioni di sostituzione di una colonna con la combinazione lineare di colonne *precedenti* vengono rappresentate, in termini di matrici trasformanti (vedi §A.1), dal prodotto a destra per matrici triangolari superiori nonsingolari. Analogamente a quanto fatto per la fattorizzazione LU, possiamo quindi scrivere

  $$Q = AR^{-1}$$

  con R^{-1}, e di conseguenza anche R, triangolare superiore. Ne segue immediatamente che $A = QR$.

  ■

1.2 Metodi iterativi di punto fisso.

Dato il sistema di n equazioni, lineari o nonlineari,

$$F(x) = 0 \tag{1.16}$$

dove $F : \mathbb{R}^n \to \mathbb{R}^n$, si chiama *equazione di punto fisso* una sua formulazione equivalente nella forma

$$x = T(x). \tag{1.17}$$

Dalla forma (1.17), dato un punto iniziale $x^{(0)}$, è possibile definire per ricorrenza la successione

$$x^{(k+1)} = T(x^{(k)}). \tag{1.18}$$

Nel caso la successione $x^{(k)}$ converga ad un certo $\bar{x}$, si caratterizza la velocità di convergenza tramite il più grande esponente γ che permette di verificare la disuguaglianza

$$\|x^{(k+1)} - \bar{x}\| \leq C\|x^{(k)} - \bar{x}\|^{\gamma} \tag{1.19}$$

per una qualche costante C. Ad esempio, se $\gamma = 1$ si parla di convergenza lineare, se $1 < \gamma < 2$ di convergenza sopralineare, se $\gamma = 2$ di convergenza quadratica.

Risultati fondamentali

- Teorema delle contrazioni

  Teorema 1.6 *Se esiste un insieme chiuso $E \subseteq \mathbb{R}^n$ tale che $T(E) \subseteq E$ e che $\|T(x) - T(y)\| \leq L\|x - y\|$ per ogni $x, y \in E$ con $L < 1$, allora l'equazione (1.17) ha soluzione unica $\bar{x}$ in E, e se $x^{(0)} \in E$, si ha $\bar{x} = \lim_k x^{(k)}$, con $x^{(k)}$ definito da (1.18).*

- Convergenza dei metodi di sostituzioni successive nella forma (1.18)

  Teorema 1.7 *Se esiste una soluzione $\bar{x}$ di (1.17) ed un suo intorno sferico $U = B(\bar{x}, \rho) \subseteq \mathbb{R}^n$ tale che $\|T(x) - T(y)\| \leq L\|x - y\|$ per ogni $x, y \in U$ con $L < 1$, e se $x^{(0)} \in U$, allora $\bar{x} = \lim_k x^{(k)}$. La convergenza di $x^{(k)}$ verso $\bar{x}$ è (almeno) lineare, e più esattamente*

  $$\|x^{(k+1)} - \bar{x}\| \leq L\|x^{(k)} - \bar{x}\|. \tag{1.20}$$

Dim. Per il teorema delle contrazioni basta verificare che U sia un insieme invariante, ovvero che $T(U) \subseteq U$. In effetti, considerando un punto $x^{(k)} \in U$, tale quindi che $\|x^{(k)} - \bar{x}\| < \rho$, si ha

$$x^{(k+1)} - \bar{x} = T(x^{(k)}) - T(\bar{x})$$

e passando alle norme:

$$\|x^{(k+1)} - \bar{x}\| = \|T(x^{(k)}) - T(\bar{x})\| \leq L\|x^{(k)} - \bar{x}\| < L\rho$$

da cui si ottiene che anche $x^{(k+1)} \in U$.

■

1.2.1 Metodi iterativi per sistemi lineari

I metodi iterativi per sistemi lineari si basano su una forma generale del tipo:

$$x^{(k+1)} = T(x^{(k)}) = Bx^{(k)} + c \tag{1.21}$$

La matrice jacobiana della trasformazione è la matrice B (detta anche matrice di iterazione). Per darne una espressione esplicita nei vari casi, si partiziona la matrice A come $A = D + E + F$ dove

$$d_{ij} = \begin{cases} a_{ij} & \text{se } i = j \\ 0 & \text{se } i \neq j \end{cases}$$

$$e_{ij} = \begin{cases} a_{ij} & \text{se } i > j \\ 0 & \text{se } i \leq j \end{cases}$$

$$f_{ij} = \begin{cases} a_{ij} & \text{se } i < j \\ 0 & \text{se } i \geq j \end{cases}$$

Descriviamo di seguito tre metodi iterativi classici per sistemi lineari. La loro convergenza, ed in alcuni casi la loro stessa applicabilità dipendono in misura notevole da eventuali permutazioni di righe nel sistema (1.16).

Il metodo di Jacobi Consiste nel porre la iterazione (1.18) nella forma

$$x_j^{(k+1)} = \frac{1}{a_{jj}} \left(b_j - \sum_{i \neq j} a_{ji} x_i^{(k)} \right) \qquad (j = 1, \ldots, n) \tag{1.22}$$

che si ottiene esplicitando la variabile j–esima dalla j–esima equazione. Questo metodo è nella forma (1.21) quando si ponga

$$B_J = -D^{-1}(E + F), \quad c_J = D^{-1}b.$$

Il metodo di Gauss–Seidel Consiste nel modificare il metodo di Jacobi utilizzando nella stessa iterazione le variabili già aggiornate, ovvero:

$$x_j^{(k+1)} = \frac{1}{a_{jj}} \left(b_j - \sum_{i<j} a_{ji} x_i^{(k+1)} - \sum_{i>j} a_{ji} x_i^{(k)} \right) \qquad (j = 1, \ldots, n). \tag{1.23}$$

Questo procedimento è più "naturale" dal punto di vista della programmazione perché permette di lavorare sullo stesso vettore senza tenere memoria del risultato della iterazione precedente. Il metodo di Gauss–Seidel si porta nella forma (1.21) ponendo

$$B_{GS} = -(D+E)^{-1}F, \quad c_{GS} = (D+E)^{-1}b.$$

Il metodo SOR (Successive Over–Relaxation) E' una ulteriore modifica del metodo di Gauss–Seidel in cui, per accelerare la convergenza, si introduce un parametro (detto *parametro di rilassamento*) ω modificando lo schema nella forma:

$$x_j^{(k+1)} = (1-\omega)x_j^{(k)} + \omega x_{j,GS}^{(k+1)} \qquad (j = 1, \ldots, n). \tag{1.24}$$

in cui $x_{j,GS}^{(k+1)}$ è il secondo membro di (1.23). Si noti che il valore $\omega = 0$ corrisponde a non effettuare aggiornamenti, ed il valore $\omega = 1$ al metodo di Gauss–Seidel. Anche il metodo SOR può poi essere messo nella forma (1.21) ponendo

$$B_{SOR} = (D+\omega E)^{-1}[(1-\omega)D - \omega F], \quad c_{SOR} = \omega(D+\omega E)^{-1}b$$

Il metodo di Richardson In questo metodo, l'aggiornamento di $x^{(k)}$ viene fatto sommando un multiplo del vettore residuo $Ax-b$ del sistema, ottenendo quindi uno schema nella forma:

$$x^{(k+1)} = x^{(k)} - \beta \left(Ax^{(k)} - b \right) \tag{1.25}$$

con $\beta \in \mathbb{R}$. Nella tipica situazione di uso di questo metodo, la matrice A è definita positiva, ed in questo caso si suppone $\beta > 0$. Il metodo di Richardson può essere messo nella forma (1.21) senza ricorrere alla partizione di A utilizzata negli altri casi, ma ponendo invece

$$B_R = I - \beta A, \quad c_R = \beta b.$$

Criteri di arresto Il criterio più naturale e computazionalmente più economico è di arrestare le iterazioni quando la norma dell'aggiornamento, $\|x^{(k+1)} - x^{(k)}\|$, scenda al di sotto di una certa soglia ε. In questo caso, indicata con $\bar{x}$ la soluzione esatta, si può maggiorare esplicitamente l'errore $\|x^{(k)} - \bar{x}\|$ mediante la serie geometrica degli aggiornamenti:

$$\|x^{(k)} - \bar{x}\| \leq$$

$$\leq \|x^{(k+1)} - x^{(k)}\| + \|x^{(k+2)} - x^{(k+1)}\| + \|x^{(k+3)} - x^{(k+2)}\| + \cdots \leq$$

$$\leq \varepsilon + L\varepsilon + L^2\varepsilon + \cdots = \frac{\varepsilon}{1-L}.$$

Se invece è possibile maggiorare l'errore iniziale $\|x^{(0)} - \bar{x}\|$, allora

$$\|x^{(k)} - \bar{x}\| \leq L\|x^{(k-1)} - \bar{x}\| \leq \cdots \leq L^k\|x^{(0)} - \bar{x}\|.$$

Entrambe queste maggiorazioni possono però essere poco significative se la costante di contrazione è molto vicina ad 1. Inoltre, simili criteri non danno alcuna indicazione su quale sia l'accuratezza con cui sono soddisfatte le equazioni del sistema, ovvero su quale sia il suo *residuo*. Per questo motivo un corretto criterio di arresto dovrebbe tenere in conto anche il fatto che a sua volta il residuo $\|Ax - b\|$ del sistema sia al di sotto di una certa soglia.

Stabilità I metodi iterativi presentano caratteristiche di stabilità migliori di quelli diretti. Il fatto che la soluzione sia un attrattore (globale nel caso dei sistemi lineari) per la iterazione (1.18) fa si che la propagazione delle perturbazioni, compresi gli errori di troncamento, non aumenti il loro ordine di grandezza.

Complessità In tutti e tre i metodi iterativi presentati, le operazioni da efettuarsi per ogni iterazione sono legate alla costruzione delle sommatorie a secondo membro, quindi un prodotto ed una somma *per ogni elemento non nullo della matrice* A. Questo vuol dire che per ogni iterazione si effettuano $O(2n^2)$ operazioni in virgola mobile per problemi pieni e $O(2cn)$ per problemi sparsi (supponendo tipicamente che il numero di elementi non nulli della matrice A sia circa costante, e dell'ordine di c, per ogni riga). Per quanto riguarda l'implementazione dei test di arresto, la valutazione della norma dell'aggiornamento ha in ogni caso costo lineare; al contrario, il calcolo del residuo ha costo dell'ordine del numero di elementi non nulli di A, in particolare quadratico se A è una matrice piena.

Risultati fondamentali

- Convergenza dei metodi iterativi

  Teorema 1.8 *La successione $x^{(k)}$ definita da (1.21) converge alla unica soluzione del sistema lineare (1.2) per ogni $x^{(0)} \in \mathbb{R}^n$, se e solo se il raggio spettrale della matrice di iterazione soddisfa la condizione*

  $$\rho(B) < 1.$$

- Condizione sufficiente di convergenza del metodo di Jacobi

  Teorema 1.9 *Se la matrice A è strettamente dominante diagonale, la successione $x^{(k)}$ definita da (1.22) converge alla unica soluzione del sistema lineare (1.2) per ogni $x^{(0)} \in \mathbb{R}^n$.*

  Dim. Si tratta di verificare che il metodo iterativo (1.22) è contrattivo, in una norma opportuna, su tutto $\mathbb{R}^n$. Poiché la costante di Lipschitz della trasformazione T è la norma della sua jacobiana, dimostriamo che la matrice jacobiana J_T del secondo membro (che è una matrice costante) soddisfa

  $$\|J_T\|_\infty < 1.$$

  Infatti,

  $$\frac{\partial T_j}{\partial x_i} = \begin{cases} -\dfrac{a_{ji}}{a_{jj}} & \text{se } i \neq j \\ 0 & \text{se } i = j \end{cases}$$

  e quindi, ricordando la definizione della norma $\|\cdot\|_\infty$ sulle matrici, si ottiene

  $$\|J_T\|_\infty = \max_j \sum_{i \neq j} \frac{|a_{ji}|}{|a_{jj}|}.$$

  Sotto l'ipotesi di dominanza diagonale stretta, per ogni indice j, si ha

  $$\sum_{i \neq j} \frac{|a_{ji}|}{|a_{jj}|} < 1$$

  e di conseguenza,

  $$\|J_T\|_\infty < 1.$$

  ■

- Condizioni sufficienti di convergenza del metodo di Gauss–Seidel

Teorema 1.10 *Se la matrice A è strettamente dominante diagonale o definita positiva, la successione $x^{(k)}$ definita da (1.23) converge alla unica soluzione del sistema lineare (1.2) per ogni $x^{(0)} \in \mathbb{R}^n$.*

- Condizione sufficiente di convergenza del metodo di sovrarilassamento (SOR)

Teorema 1.11 *Se la matrice A è definita positiva, la successione $x^{(k)}$ definita da (1.24) converge alla unica soluzione del sistema lineare (1.2) per ogni $x^{(0)} \in \mathbb{R}^n$ e per ogni $\omega \in (0, 2)$.*

- Condizione sufficiente di convergenza del metodo di Richardson

Teorema 1.12 *Se la matrice A è definita positiva, esiste un valore β_0 tale che, per $0 < \beta < \beta_0$, la successione $x^{(k)}$ definita da (1.25) converge alla unica soluzione del sistema lineare (1.2) per ogni $x^{(0)} \in \mathbb{R}^n$.*

Dim. In questo caso la convergenza verrà dimostrata applicando il criterio più generale. Si tratta quindi di verificare che

$$\rho(B_R) = \rho(I - \beta A) < 1.$$

D'altra parte, se $\lambda_i(A)$ è un autovalore di A (necessariamente reale e positivo per l'ipotesi di positività della matrice), il corrispondente autovalore di B_R sarà

$$\lambda_i(I - \beta A) = 1 - \beta\lambda_i(A)$$

e quindi lo schema è convergente sotto la condizione

$$-1 < 1 - \beta\lambda_i(A) < 1.$$

Mentre per l'ipotesi di positività di β e λ_i la seconda disuguaglianza è sempre soddisfatta, dalla prima si ottiene

$$\beta\lambda_i(A) < 2,$$

che dovendo essere soddisfatta per ogni autovalore λ_i, fornisce la condizione di convergenza

$$\beta < \frac{2}{\max_i \lambda_i(A)} = \beta_0.$$

■

1.2.2 Metodi iterativi per equazioni nonlineari

A partire dall'equazione scalare

$$f(x) = 0 \tag{1.26}$$

si costruisce una famiglia di metodi che possono essere messi tutti (con esclusione dei metodi di bisezione, delle secanti e di Muller) nella forma

$$x_{k+1} = g(x_k) \tag{1.27}$$

con x_0 assegnato [1] (in genere, si deve supporre x_0 sufficientemente vicino alla radice $\bar{x}$).

Il metodo di bisezione Si suppone di avere a disposizione un intervallo $[a_0, b_0]$ ai cui estremi la funzione assuma valori di segno differente, ovvero $f(a_0)f(b_0) < 0$. Si opera partendo da $k = 0$ e costruendo una successione di approssimazioni c_k secondo i passi seguenti:

1. Poni:

$$c_k = \frac{a_k + b_k}{2}$$

2. Se $f(c_k) = 0$ o se è soddisfatta una opportuna condizione di arresto, STOP.

3. Se $f(a_k)f(c_k) < 0$, poni

$$a_{k+1} = a_k \ , \quad b_{k+1} = c_k$$

incrementa k e vai a 1.

4. Se $f(b_k)f(c_k) < 0$, poni

$$a_{k+1} = c_k \ , \quad b_{k+1} = b_k$$

incrementa k e vai a 1.

[1] In questo sottoparagrafo, non essendoci bisogno di esplicitare le singole componenti del vettore $x^{(k)}$, useremo la notazione semplificata x_k.

Il metodo delle sostituzioni successive Consiste nel porre in forma di punto fisso $x = g(x)$ l'equazione

$$f(x) = 0$$

e definire mediante la funzione g una iterazione del tipo (1.27). Tale operazione non si effettua naturalmente in modo univoco; alcune scelte classiche sono:

$$g(x) = x + \alpha f(x) \tag{1.28}$$

$$g(x) = x + \eta(x) f(x) \tag{1.29}$$

$$g(x) = x + G(f(x)) \tag{1.30}$$

in cui si tenta di determinare il parametro α o le funzioni $\eta(\cdot)$, $G(\cdot)$ in modo da rendere convergente lo schema. Perché le forme (1.28)–(1.30) siano equivalenti all'equazione originale si deve avere $\alpha \neq 0$ nel primo caso, $\eta(x) \neq 0$ e di segno costante nel secondo, $G(0) = 0$ e $G(t) \neq 0$ per $t \neq 0$ nel terzo. Inoltre, in questi casi si deve supporre normalmente che $\bar{x}$ sia una radice semplice: in caso contrario, infatti, $g'(\bar{x}) = 1$ e la funzione g non può essere una contrazione nell'intorno di $\bar{x}$.

Il metodo di Newton Nel metodo di Newton si pone

$$g(x) = x - \frac{f(x)}{f'(x)}. \tag{1.31}$$

Questa scelta equivale, intuitivamente, a definire x_{k+1} come lo zero della tangente al grafico di $f(x)$ nel punto $(x_k, f(x_k))$. Tale tangente ha infatti equazione

$$y - f(x_k) = f'(x_k)(x - x_k)$$

da cui imponendo il passaggio per il punto $(x_{k+1}, 0)$ si ottiene

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}. \tag{1.32}$$

Se la derivata $f'(x)$ non è nota esplicitamente o è troppo complessa da calcolare, si può costruire un metodo di Newton approssimato sostituendola con il rapporto incrementale in x_k:

$$x_{k+1} = x_k - \frac{h}{f(x_k + h) - f(x_k)} f(x_k).$$

Il metodo delle secanti Se in (1.32) si sotituisce il calcolo di $f'(x)$ con il calcolo del rapporto incrementale tra i punti $(x_{k-1}, f(x_{k-1}))$ e $(x_k, f(x_k))$ si ottiene un metodo (che non è più nella forma (1.27)), noto come metodo delle secanti:

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} f(x_k). \tag{1.33}$$

In questo caso l'interpretazione geometrica del metodo è di definire x_{k+1} come lo zero della secante passante per le due ultime coppie $(x_{k-1}, f(x_{k-1}))$ e $(x_k, f(x_k))$.

Il metodo delle corde Una ulteriore approssimazione si ottiene utilizzando in (1.33) un rapporto incrementale fisso, calcolato agli estremi di un intervallo (sufficientemente piccolo) $[a, b]$ in cui si localizza la radice. Il metodo che si ottiene in questo modo,

$$x_{k+1} = x_k - \frac{b - a}{f(b) - f(a)} f(x_k) \tag{1.34}$$

è noto come metodo delle corde ed è di nuovo nella forma (1.27).

Il metodo di Steffensen In questo metodo si sostituisce ancora la derivata $f'(x)$ con un rapporto incrementale, ma utilizzando $f(x_k)$ come incremento (se si sta convergendo ad una radice, si ha infatti $f(x_k) \to 0$). Ne risulta il metodo iterativo

$$x_{k+1} = x_k - \frac{f(x_k)^2}{f(x_k + f(x_k)) - f(x_k)}. \tag{1.35}$$

Il metodo di Newton cubico In questo caso il punto x_{k+1} si calcola ancora tramite gli zeri della approssimazione di Taylor centrata in x_k, ma di secondo grado (utilizzando quindi anche la derivata seconda di f). Il metodo è nella forma (1.27), e più precisamente:

$$x_{k+1} = x_k - \frac{f'(x_k) \pm \sqrt{f'(x_k)^2 - 2f(x_k)f''(x_k)}}{f''(x_k)}. \tag{1.36}$$

Tra le due radici del polinomio di Taylor, se reali, si sceglie quella più vicina ad x_k.

Il metodo di Muller Analogamente al metodo di Newton cubico, il metodo di Muller è una generalizzazione del metodo delle secanti in cui x_{k+1} viene calcolato azzerando il polinomio interpolatore di secondo grado (vedi sezione 5) passante per i punti $(x_{k-2}, f(x_{k-2}))$, $(x_{k-1}, f(x_{k-1}))$ e $(x_k, f(x_k))$. Posto

$$t_k = f[x_k, x_{k-1}] + (x_k - x_{k-1}) f[x_k, x_{k-1}, x_{k-2}]$$

la iterazione da effettuare è

$$x_{k+1} = x_k - \frac{2}{t_k \pm \sqrt{t_k^2 - 4 f(x_k) f[x_k, x_{k-1}, x_{k-2}]}} f(x_k) \tag{1.37}$$

in cui il segno al denominatore viene scelto in modo da ottenere, tra le due possibili determinazioni, quella più vicina ad x_k.

Criteri di arresto Per i metodi a convergenza lineare valgono le considerazioni fatte nel §1.2.1 a proposito dei metodi iterativi per i sistemi lineari. In particolare, per un metodo con una costante di contrazione L, allora se $|x_{k+1} - x_k| \leq \varepsilon$, si ha

$$|x_k - \bar{x}| \leq \frac{\varepsilon}{1 - L},$$

mentre conoscendo una maggiorazione di $|x_0 - \bar{x}|$ si può stimare l'errore come

$$|x_k - \bar{x}| \leq L|x_{k-1} - \bar{x}| \leq \cdots \leq L^k |x_0 - \bar{x}|.$$

Ovviamente, valgono ancora le considerazioni sul residuo, rappresentato in questo caso dal valore $|f(x_{k+1})|$, che normalmente si richiede sia al di sotto di una certa soglia. In questo caso, è geometricamente abbastanza intuitivo che se $|f'(\bar{x})| \gg 1$, piccoli errori sulla variabile x portino ad elevati residui.

Nel caso del metodo di Newton, la situazione è abbastanza diversa e per semplicità si può talvolta lavorare a numero di iterazioni fissato. Dalla (1.19) si ottiene

$$|x_k - \bar{x}| \leq C^k |x_0 - \bar{x}|^{2k}.$$

Se ad esempio si avesse $C = 1$ e l'errore iniziale $|x_0 - \bar{x}| \approx 0.1$, ad ogni iterazione del metodo il numero di cifre decimali esatte raddoppierebbe. Ciò vuol dire che l'errore di macchina verrebbe raggiunto in tre–quattro iterazioni in precisione semplice ed in quattro–cinque iterazioni in precisione doppia.

Stabilità Nel caso delle equazioni (o dei sistemi) nonlineari, la stabilità dei metodi iterativi è legata alla determinazione di un intorno in cui il metodo sia convergente, presentando tutti gli schemi più efficienti una convergenza di tipo locale. Questa considerazione porta all'uso preventivo di metodi più

lenti ma più robusti (ad esempio, la bisezione) in fase di separazione delle radici. Si noti che nei metodi delle secanti e di Steffensen, il rapporto incrementale che approssima $f'(x_k)$ viene effettuato (se lo schema converge) con incrementi infinitesimi, e si può quindi presentare un problema di perdita di precisione per sottrazione. In questa situazione può accadere che lo schema converga verso la soluzione, e una volta nel suo intorno continui ad avere leggere oscillazioni senza assestarsi su un valore definitivo.

Complessità Nel trattamento numerico delle equazioni scalari, la operazione che si considera critica è il calcolo di f (ricordiamo che questa funzione può essere non nota in forma esplicita, o comunque complessa da calcolare), ed a maggior ragione il calcolo delle derivate successive. I metodi di bisezione, delle corde, delle secanti e di Muller non richiedono il calcolo delle derivate, ed effettuano un solo calcolo della f ad ogni iterazione. Il metodo di Newton ha convergenza quadratica ma richiede il calcolo sia di f che di f'; se si sostituisce il calcolo di f' con un rapporto incrementale si perde la convergenza quadratica e si effettuano due valutazioni di f per iterazione. Analoga complessità si ha per il metodo di Steffensen che però ha convergenza quadratica. Il metodo di Newton cubico ha ordine elevato di convergenza ma richiede anche il calcolo di f'', insieme con forti ipotesi di regolarità.

Risultati fondamentali

- Convergenza del metodo di bisezione

  Teorema 1.13 *Se $f \in C^0([a_0, b_0])$ ed $f(a_0)f(b_0) < 0$, allora:*

  $$\lim_k a_k = \lim_k b_k = \lim_k c_k = \bar{x}$$

  dove $\bar{x}$ è una radice dell'equazione (1.26).

  Dim. Per come sono state costruite a_k e b_k, si ha $a_0 \leq a_k \leq b_0$ e $a_0 \leq b_k \leq b_0$; inoltre entrambe le successioni sono monotone (non decrescente a_k, non crescente b_k). Quindi le due successioni ammettono limite, e poiché

  $$b_k - a_k = \frac{b_0 - a_0}{2^k} \to 0,$$

  il loro limite $\bar{x}$ coincide (e, per confronto, coincide anche con il limite di c_k). Se poi supponiamo, per fissare le idee, che $f(a_k) < 0$, $f(b_k) > 0$, per la continuità di f e per il teorema di permanenza del segno si ha

  $$0 \leq \lim_k f(b_k) = f(\bar{x}) = \lim_k f(a_k) \leq 0$$

e quindi necessariamente $f(\bar{x}) = 0$.

■

- Convergenza dei metodi nella forma (1.27)

  Teorema 1.14 *Dato lo schema iterativo*

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

  con $g \in C^{m+1}$ (o prolungabile in una funzione $\tilde{g} \in C^{m+1}$) in un intorno di $\bar{x}$, se $g'(\bar{x}) = \cdots = g^{(m)}(\bar{x}) = 0$ (con $\bar{x}$ soluzione di (1.26)) ed x_0 è sufficientemente vicino a $\bar{x}$, allora $x_k \to \bar{x}$, e la convergenza ha ordine $m + 1$.

  Dim. Identifichiamo intanto g ed il suo eventuale prolungamento $\tilde{g}$. Notiamo che nelle ipotesi del teorema la funzione g è una contrazione (almeno localmente), visto che in un opportuno intorno di $\bar{x}$ viene sicuramente verificata la condizione $|g'(x)| \leq L < 1$. Questo assicura la convergenza dello schema applicando il Teorema 1.7, se l'approssimazione iniziale è sufficientemente vicina ad $\bar{x}$.

  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 i termini che vanno dal differenziale primo a quello m–esimo si annullano nel nostro caso, si ottiene

  $$x_{k+1} - \bar{x} = \frac{1}{(m+1)!} g^{(m+1)}(\xi_k)(x_k - \bar{x})^{m+1}$$

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

  $$|x_{k+1} - \bar{x}| = \frac{1}{(m+1)!} |g^{(m+1)}(\xi_k)| \, |x_k - \bar{x}|^{m+1} \tag{1.38}$$

  e dalla (locale) limitatezza di $g^{(m+1)}(\cdot)$ si ottiene la seconda parte della tesi.

  ■

- Convergenza locale del metodo di Newton

Teorema 1.15 *Se $f \in C^2$, $f'(\bar{x}) \neq 0$ ed x_0 è sufficientemente vicino a $\bar{x}$, allora la successione x_k definita dal metodo di Newton (1.32) converge con ordine quadratico alla soluzione $\bar{x}$.*

Dim. Daremo la dimostrazione nell'ipotesi supplementare che $f \in C^3$. In questo caso, basta notare che per il metodo di Newton si ha

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

Sotto le ipotesi fatte, $g \in C^2$ e $g'(\bar{x}) = 0$, e si può quindi applicare il Teorema 1.14 con $m = 1$.

■

- Convergenza monotona del metodo di Newton

Teorema 1.16 *Se $f \in C^1$, $f(\bar{x}) = 0$ ed una delle seguenti condizioni è soddisfatta:*

1. *$x_0 > \bar{x}$, f è crescente e convessa in $[\bar{x}, x_0]$;*
2. *$x_0 > \bar{x}$, f è decrescente e concava in $[\bar{x}, x_0]$;*
3. *$x_0 < \bar{x}$, f è crescente e concava in $[x_0, \bar{x}]$;*
4. *$x_0 < \bar{x}$, f è decrescente e convessa in $[x_0, \bar{x}]$;*

allora la successione x_k definita dal metodo di Newton (1.32) converge in modo monotono alla soluzione $\bar{x}$.

Dim. Dimostriamo il teorema nel caso 1, essendo analoghi gli altri. Se f è crescente, $f(x) > 0$ e $f'(x) > 0$ per $x > \bar{x}$ e, dalla (1.32), $x_{k+1} < x_k$. D'altra parte, x_{k+1} è lo zero della tangente al grafico di f in x_k, e poiché la funzione è convessa, si ha $f(x_{k+1}) > 0$ e quindi $x_{k+1} > \bar{x}$. Quindi la successione x_k è decrescente ed inferiormente limitata ed ammette un limite $\tilde{x} \geq \bar{x}$. Passando al limite per $k \to \infty$ nella (1.32) si ha poi:

$$\tilde{x} = \tilde{x} - \frac{f(\tilde{x})}{f'(\tilde{x})}$$

che può essere soddisfatta solo se $f(\tilde{x}) = 0$. Ma essendo la funzione f crescente nell'intervallo $[\bar{x}, x_0]$ deve necessariamente aversi $\tilde{x} = \bar{x}$.

■

- Convergenza del metodo delle secanti

  Teorema 1.17 *Se $f \in C^2$, $f'(\bar{x}) \neq 0$ ed x_0 e x_1 sono sufficientemente vicini a $\bar{x}$, allora o esiste un indice $\bar{k}$ finito tale che $x_{\bar{k}} = \bar{x}$, oppure la successione x_k definita dal metodo delle secanti (1.33) converge alla soluzione $\bar{x}$ con ordine $\gamma = (1+\sqrt{5})/2$.*

- Convergenza del metodo delle corde

  Teorema 1.18 *Se $f \in C^1$, $f'(\bar{x}) \neq 0$ ed a, b, x_0 sono sufficientemente vicini a $\bar{x}$, allora la successione x_k definita dal metodo delle corde (1.34) converge alla soluzione $\bar{x}$.*

  Dim. Per calcolare la costante di contrazione del metodo deriviamo g, ottenendo:

$$g'(x) = 1 - \frac{b-a}{f(b)-f(a)} f'(x) = 1 - \frac{f'(x)}{f'(\xi)} = \frac{f'(\xi) - f'(x)}{f'(\xi)} \qquad (1.39)$$

  in cui si è anche applicato il teorema di Lagrange con $\xi \in (a,b)$. Notiamo che $g'(x)$ dipende anche da a e b tramite ξ. Per la continuità di f', e se $f'(\bar{x}) \neq 0$, passando al limite per $a, b, x \to \bar{x}$ si ottiene

$$\lim_{a,b,x\to\bar{x}} g'(x) = 0,$$

  il che implica che esiste un intorno U di $\bar{x}$ tale che se $a, b, x \in U$ si ha

$$\sup_U |g'(x)| \leq L < 1$$

  e si può quindi dedurre la convergenza dal Teorema 1.7. Osserviamo che se $f \in C^2$, e poiché la radice $\bar{x}$ è semplice, è possibile determinare un intorno W di $\bar{x}$ in cui $|f''|$ è limitato dall'alto, e $|f'|$ limitato dal basso (da una valore strettamente positivo). Posto ora $a, b, x \in U \subseteq W$, si può anche scrivere più esplicitamente, applicando una seconda volta il teorema di Lagrange in (1.39),

$$\sup_U |g'(x)| \leq \frac{|\xi - x| \sup_W |f''|}{\inf_W |f'|} \leq \frac{|U| \sup_W |f''|}{\inf_W |f'|},$$

  e la condizione di contrattività si ottiene di nuovo prendendo un intervallo U di misura $|U|$ sufficientemente piccola.

  ■

- Convergenza del metodo di Steffensen

Teorema 1.19 *Se $f \in C^2$, $f'(\bar{x}) \neq 0$ ed x_0 è sufficientemente vicino a $\bar{x}$, allora o esiste un indice $\bar{k}$ finito tale che $x_{\bar{k}} = \bar{x}$, oppure la successione x_k definita dal metodo di Steffensen (1.35) converge con ordine quadratico alla soluzione $\bar{x}$.*

Dim. Si può applicare il teorema generale alla funzione di iterazione

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

Poiché $f(\bar{x}) = 0$, se esiste un indice $\bar{k}$ finito tale che $x_{\bar{k}} = \bar{x}$, il metodo non permette di definire oltre le iterazioni. Altrimenti, si ha $x_k \neq \bar{x}$ per ogni k e si tratta di verificare che g e g' sono entrambe prolungabili con continuità in $\bar{x}$ con i valori limite

$$\lim_{x \to \bar{x}} g(\bar{x}) = \bar{x}, \tag{1.40}$$

$$\lim_{x \to \bar{x}} g'(\bar{x}) = 0. \tag{1.41}$$

Notiamo subito che, per il teorema di Lagrange,

$$f(x + f(x)) - f(x) = f'(\xi) f(x) \tag{1.42}$$

con $\xi \to \bar{x}$ per $x \to \bar{x}$. Utilizzando (1.42) nell'espressione di $g(x)$, e ricordando che $\bar{x}$ è radice semplice, si ottiene immediatamente

$$\lim_{x \to \bar{x}} g(\bar{x}) = \lim_{x \to \bar{x}} \left(x - \frac{f(x)}{f'(\xi)} \right) = \bar{x}.$$

Per quanto riguarda la (1.41) si ha, con qualche passaggio:

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

$$- f(x)^2 (f'(x + f(x))(1 + f'(x)) - f'(x)) \Big]. \tag{1.43}$$

Utilizzando ora (1.42) in (1.43) si ottiene:

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

$$= 1 - \frac{2f'(x)f'(\xi) - (f'(x)^2(1+o(1)) + f'(x)o(1))}{f'(\xi)^2} \rightarrow$$

$$\rightarrow 1 - \frac{2f'(\bar{x})^2 - f'(\bar{x})^2}{f'(\bar{x})^2} = 0$$

in cui si è ancora supposto che la radice $\bar{x}$ sia semplice (cioè $f'(\bar{x}) \neq 0$). Omettiamo la dimostrazione della limitatezza di g''.

■

- Convergenza del metodo di Muller

Teorema 1.20 *Se $f \in C^3$, $f'(\bar{x}) \neq 0$ ed x_0, x_1 e x_2 sono sufficientemente vicini a $\bar{x}$, allora la successione x_k definita dal metodo di Muller (1.37) converge alla soluzione $\bar{x}$ con ordine γ dato dalla radice positiva della equazione*

$$\gamma^3 - \gamma^2 - \gamma - 1 = 0.$$

1.2.3 Metodi iterativi per sistemi nonlineari

Il metodo di Newton La versione n–dimensionale del metodo di Newton si basa (analogamente al caso unidimensionale) sulla linearizzazione locale del sistema di equazioni: si effettua lo sviluppo di Taylor (di primo ordine e punto iniziale $x^{(k)}$) della $F(x)$, e si impone che si annulli in $x^{(k+1)}$, ovvero

$$F(x^{(k)}) + J_F(x^{(k)})(x^{(k+1)} - x^{(k)}) = 0$$

(in cui J_F è la matrice Jacobiana della funzione F) che si può riscrivere in modo equivalente

$$J_F(x^{(k)})(x^{(k+1)} - x^{(k)}) = -F(x^{(k)}) \tag{1.44}$$

o anche, utilizzando formalmente l'inversa di J_F:

$$x^{(k+1)} = x^{(k)} - J_F(x^{(k)})^{-1}F(x^{(k)}). \tag{1.45}$$

Metodi di Newton approssimati Per evitare la soluzione ad ogni passo del sistema lineare in (1.44) o la inversione della matrice Jacobiana in (1.45) (operazione in ogni caso sconsigliabile per complessità ed instabilità) si può implementare il metodo di Newton in versioni approssimate ma di minore complessità computazionale. Un tipico esempio è quello in cui

$$J_F(x^{(\tilde{k})})(x^{(k+1)} - x^{(k)}) = -F(x^{(k)}) \tag{1.46}$$

in cui l'indice $\tilde{k}$ (e di conseguenza la Jacobiana) viene aggiornato solo una volta ogni m iterazioni, formalmente

$$\tilde{k} = \left[\frac{k}{m}\right] m$$

o in cui addirittura la Jacobiana viene calcolata una sola volta alla prima iterazione:

$$J_F(x^{(0)})(x^{(k+1)} - x^{(k)}) = -F(x^{(k)}), \tag{1.47}$$

situazione che rappresenta la versione n–dimensionale del metodo delle corde. In questi casi, il sistema lineare può essere risolto con minore complessità fattorizzando la matrice Jacobiana (una volta ogni m iterazioni nel primo caso, una volta per tutte nel secondo).

Risultati fondamentali

- Convergenza del metodo di Newton

  Teorema 1.21 *Se $F \in C^2$, $\det J_F(\bar{x}) \neq 0$ ed $x^{(0)}$ è sufficientemente vicino a $\bar{x}$, allora la successione $x^{(k)}$ definita dal metodo di Newton (1.45) converge con ordine quadratico alla soluzione $\bar{x}$.*

- Convergenza del metodo delle corde

  Teorema 1.22 *Se $F \in C^1$, $\det J_F(\bar{x}) \neq 0$ ed $x^{(0)}$ è sufficientemente vicino a $\bar{x}$, allora la successione $x^{(k)}$ definita dal metodo delle corde (1.47) converge alla soluzione $\bar{x}$.*

  Dim. Messo il metodo nella forma

  $$x^{(k+1)} = T\left(x^{(k)}\right),$$

  ricordiamo che T è una contrazione se la sua Jacobiana J_T soddisfa in una qualche norma la condizione $\|J_T(x^{(0)}, x)\| \leq L_T < 1$ (in cui si è esplicitata la dipendenza di T da $x^{(0)}$ che appare nella definizione (1.47)). In base al Teorema 1.7, perché il metodo sia localmente convergente a $\bar{x}$, basta che ciò avvenga in un intorno della soluzione. D'altra parte si ha, usando (1.47):

  $$J_T(x^{(0)}, x) = I - J_F(x^{(0)})^{-1} J_F(x),$$

e quindi, considerando che J_T è una funzione continua dei suoi argomenti, $\|J_T(x^{(0)}, x)\| \to 0$ per $x, x^{(0)} \to \bar{x}$. Ancora per continuità, è quindi possibile individuare un intorno U di $\bar{x}$ tale che, se $x, x^{(0)} \in U$,

$$\|J_T(x^{(0)}, x)\| \leq L_T < 1.$$

■

1.3 Metodi di minimizzazione

Questa strategia di soluzione dei sistemi è applicabile sia a sistemi ben posti che a sistemi sovradeterminati. Dato il sistema *non necessariamente quadrato*

$$F(x) = 0 \tag{1.48}$$

in cui supporremo $F : \mathbb{R}^n \to \mathbb{R}^m$, con $m \geq n$, definiamo $\bar{x}$ soluzione (nel senso del minimo residuo o dei minimi quadrati) se

$$\bar{x} \in \underset{x \in \mathbb{R}^n}{\text{argmin}}\ r(x) \tag{1.49}$$

in cui il residuo $r(x)$ è definito da $r(x) = F^t(x)F(x) = \|F(x)\|^2$ (nella norma euclidea). Se $r(\bar{x}) = 0$, e solo allora, $\bar{x}$ è soluzione di (1.48) in senso letterale, mentre se $r(\bar{x}) > 0$ (e questo avviene in generale nei sistemi sovradeterminati o comunque mal posti) si tratterà di una soluzione nel senso generalizzato della (1.49).

Nel caso dei sistemi lineari, si è visto che se A è simmetrica e definita positiva è possibile risolvere il sistema per minimizzazione senza passare per il residuo. Se A è una matrice generica nonsingolare, il calcolo del residuo fornisce la forma quadratica

$$r(x) = (Ax - b)^t(Ax - b) = x^t A^t Ax - 2b^t Ax + b^t b$$

la cui minimizzazione equivale a risolvere il sistema simmetrico $A^t Ax = A^t b$ (si noti che questo sistema in generale ha un condizionamento peggiore del sistema di partenza).

Per quanto riguarda gli algoritmi di minimizzazione, tutti di tipo iterativo, una veloce rassegna sarà data nel cap. 3.

Una tipica applicazione della soluzione di sistemi sovradeterminati nel senso dei minimi quadrati si vedrà poi a proposito della approssimazione ai minimi quadrati di funzioni (vedi §5.1.4).

1.4 Confronto fra i vari schemi

1.4.1 Sistemi lineari

La tabella riassume i dati di complessità ed occupazione di memoria degli schemi esaminati. Considerando che gli schemi iterativi hanno convergenza lineare, ne risulta che il loro uso è conveniente solo in caso di problemi sparsi e di grandi dimensioni, in cui questi schemi traggono vantaggio dal ridotto numero di elementi della matrice A (fa eccezione a questo quadro il problema di risolvere più sistemi lineari con identica matrice). In dimensione alta una ulteriore considerazione che spinge all'uso di metodi iterativi è la bassa stabilità di quelli diretti. Nella pratica, tuttavia, è più frequente che questi metodi siano metodi di minimizzazione (eventualmente a convergenza sopralineare), applicati alla funzione $f(x) = 1/2(Ax, x) - (b, x)$ se A è definita positiva o al residuo nel caso generico.

schema	complessità (probl. pieni)	complessità (probl. sparsi)	occupazione (probl. pieni)	occupazione (probl. sparsi)
MEG	$O\left(\frac{2n^3}{3}\right)$	$O\left(\frac{2n^3}{3}\right)$	$O(n^2)$	$O(n^2)$
LU	$O\left(\frac{2n^3}{3}\right)$ $O(2n^2)$ (*)	$O\left(\frac{2n^3}{3}\right)$ $O(2n^2)$ (*)	$O(n^2)$	$O(n^2)$
QR	$O\left(\frac{4n^3}{3}\right)$ $O(3n^2)$ (*)	$O\left(\frac{4n^3}{3}\right)$ $O(3n^2)$ (*)	$O(n^2)$	$O(n^2)$
iterat.	$O(n^2)$ per iter.	$O(n)$ per iter.	$O(n^2)$	$O(n)$

(*) in caso di più sistemi con la stessa matrice ma con termini noti diversi

1.4.2 Equazioni nonlineari

Per la soluzione approssimata di equazioni scalari, il metodo ritenuto più efficiente nella pratica è il metodo di Newton, se si conosce l'espressione esplicita della derivata. In caso opposto, a seconda della regolarità che si può ipotizzare, si possono utilizzare metodi a convergenza lineare o sopralineare che non richiedono il calcolo di f'. Occorre tenere presente che la maggior parte di questi metodi sono schemi a convergenza locale, e che vanno di regola preceduti (qualora non si conosca a priori una stima sufficientemente precisa della radice, o non si riesca a verficare le ipotesi per la convergenza monotona) da una fase di tabulazione e/o bisezione.

schema	complessità per iterazione	ordine di convergenza	regolarità
bisezione	calcolo $f(x)$	$\gamma = 1$	C^0
corde	calcolo $f(x)$	$\gamma = 1$	C^1
secanti	calcolo $f(x)$	$\gamma = \frac{1+\sqrt{5}}{2}$	C^2
Muller	calcolo $f(x)$	$\gamma \approx 1.84$	C^3
Newton	calcolo $f(x), f'(x)$	$\gamma = 2$	C^2
Steffensen	calcolo $f(x), f(x + f(x))$	$\gamma = 2$	C^2
Newton cubico	calcolo $f(x), f'(x), f''(x)$	$\gamma = 3$	C^3

1.5 Esercizi sperimentali

- Verificare sperimentalmente le differenze di velocità di convergenza tra i vari algoritmi iterativi per sistemi lineari. Verificare che tale differenza è più vistosa nel caso di sistemi lineari malcondizionati.

- Calcolare il residuo del sistema dopo la sua soluzione numerica (o all'arresto, in un metodo iterativo). Confrontare errore sulla soluzione e residuo, nel caso di sistemi più o meno malcondizionati.

- Dare un esempio che faccia vedere come nei metodi di Jacobi e Gauss–Seidel la convergenza può avvenire o meno a seconda di come sono ordinate le righe del sistema.

- Verificare che i metodi di Jacobi e Gauss–Seidel possono non convergere nel caso di matrici non dominanti diagonali. Dimostrare che in un sistema 2×2 (ma non in un sistema $n \times n$) uno dei due possibili ordinamenti delle righe dà luogo ad uno schema convergente, e non l'altro.

- Verificare sperimentalmente le differenze di velocità di convergenza tra i vari algoritmi iterativi per equazioni nonlineari. Verificare l'importanza della ipotesi di radice semplice e della regolarità ai fini della velocità di convergenza.

2 Calcolo di autovalori

Data una matrice quadrata A, calcolarne gli autovalori significa trovare i valori λ per cui si abbia

$$Ax = \lambda x. \tag{2.1}$$

Se una coppia (λ, x) soddisfa (2.1), λ si dice *autovalore* della matrice A ed x *autovettore* relativo all'autovalore λ. Si dimostra che gli autovalori sono soluzioni della *equazione caratteristica*

$$\det(A - \lambda I) = 0$$

(in cui il primo membro, che è un polinomio di grado n, si indica come *polinomio caratteristico*) e che per il teorema fondamentale dell'Algebra ne esistono esattamente n, se considerati nel piano complesso e con la loro molteplicità.

Ricordiamo anche che una matrice reale simmetrica, o più in generale una matrice hermitiana, ha autovalori reali semplici ed autovettori ortogonali; inoltre se (λ, x) è una coppia autovalore–autovettore relativa ad A, esisterà per l'inversa A^{-1} la coppia $(1/\lambda, x)$.

Le due classi principali di metodi per il calcolo approssimato di autovalori sono legate a due problemi principali:

Calcolo degli autovalori estremi – In questo caso i principali metodi che vengono utilizzati sono il metodo delle potenze e quello di Lanczos.

Calcolo di tutti gli autovalori – Questo secondo caso viene affrontato di regola con i cosiddetti *metodi di similitudine*, in cui dalla matrice di partenza si costruisce una successione di matrici simili (in genere tramite trasformazioni ortogonali), fino a portarla in forma triangolare superiore, diagonale o tridiagonale simmetrica.

2.1 Calcolo degli autovalori estremi

Per questo problema verrà esposto il solo algoritmo delle potenze (insieme con le sue varianti principali), nonostante il metodo di Lanczos venga considerato ampiamente più efficiente. Si può notare che, in linea di principio, dal calcolo di un autovalore λ_1 semplice (ad esempio l'autovalore di modulo massimo come nel metodo delle potenze) si può, con una operazione detta di *deflazione*, porre la matrice A nella forma diagonale a blocchi

$$\tilde{A} = T^{-1}AT = \begin{pmatrix} \lambda_1 & 0^t \\ 0 & \tilde{A}_{22} \end{pmatrix}$$

(dove 0 e 0^t indicano rispettivamente un blocco colonna ed uno riga di zeri), e riprendere la ricerca degli altri autovalori sulla sottomatrice $\tilde{A}_{22}$ con lo stesso metodo. Tale procedura è però instabile se iterata in dimensione alta, e nel caso si vogliano ottenere tutti gli autovalori di una matrice di grandi dimensioni si ricorre normalmente ai metodi di similitudine.

2.1.1 Metodo delle potenze e sue varianti

Si tratta di metodi che isolano un solo autovalore per volta, rispettivamente l'autovalore di modulo massimo, quello di modulo minimo o ancora quello più vicino ad un numero reale assegnato.

Metodo delle potenze Data la matrice quadrata A di ordine n, con autovalori che soddisfano la condizione

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|, \tag{2.2}$$

ed un vettore iniziale $z_0 \in \mathbb{R}^n$, si costruisce la successione di vettori z_k come

$$z_{k+1} = Az_k. \tag{2.3}$$

Si può osservare che $z_{k+1} = Az_k = A^2 z_{k-1} = \cdots = A^{k+1} z_0$, da cui il nome dell'algoritmo. Sotto l'ipotesi (2.2) si ha che i vettori z_k si allineano asintoticamente con la direzione di x_1 (autovettore relativo all'autovalore dominante λ_1), ed inoltre il cosiddetto *quoziente di Rayleigh* relativo a z_k,

$$\sigma_k = \frac{z_k^t A z_k}{z_k^t z_k} \tag{2.4}$$

converge a λ_1.

In pratica se $|\lambda_1| >> 1$, le componenti del vettore z_k possono divergere molto velocemente, e si pone quindi il problema di evitare l'overflow (analogamente, se $|\lambda_1| << 1$, occorre evitare l'underflow). Il metodo delle potenze viene quindi usualmente implementato in forma normalizzata:

$$y_k = \frac{\tilde{z}_k}{\|\tilde{z}_k\|}, \tag{2.5}$$

$$\tilde{z}_{k+1} = Ay_k, \tag{2.6}$$

e con il quoziente di Rayleigh che può essere dato ancora da (2.4) (sostituendo z_k con $\tilde{z}_k$), o ancora, supponendo di aver normalizzato rispetto alla norma euclidea, da

$$\sigma_k = \frac{y_k^t A y_k}{y_k^t y_k} = y_k^t \tilde{z}_{k+1}. \tag{2.7}$$

Metodo delle potenze inverse In questo caso l'autovalore che viene individuato è quello di modulo minimo. Ricordando che l'autovalore di modulo minimo di una matrice A è il reciproco dell'autovalore di modulo massimo di A^{-1}, si può formulare questo schema nella forma normalizzata

$$y_k = \frac{\tilde{z}_k}{\|\tilde{z}_k\|}, \tag{2.8}$$

$$\tilde{z}_{k+1} = A^{-1} y_k, \tag{2.9}$$

e con il quoziente di Rayleigh dato da

$$\sigma_k = \frac{y_k^t A^{-1} y_k}{y_k^t y_k} = y_k^t \tilde{z}_{k+1}. \tag{2.10}$$

In pratica, però, si può evitare l'inversione della matrice A calcolando $\tilde{z}_{k+1}$ come soluzione del sistema lineare

$$A\tilde{z}_{k+1} = y_k, \tag{2.11}$$

in cui la matrice A sia stata fattorizzata una volta per tutte all'inizio delle iterazioni. Viene così evitato il costo (e la instabilità) della operazione di inversione della matrice senza aumentare il costo di ogni iterazione. Naturalmente, nel caso del metodo delle potenze inverse, la condizione (2.2) va sostituita dalla

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_{n-1}| > |\lambda_n|, \tag{2.12}$$

ed il quoziente di Rayleigh σ_k converge a $1/\lambda_n$.

A sua volta il metodo delle potenze inverse può essere applicato con una traslazione dello spettro, sostituendo la matrice A con la matrice $A - \lambda I$, ovvero rimpiazzando (2.11) con

$$(A - \lambda I)\tilde{z}_{k+1} = y_k, \tag{2.13}$$

ed in questo caso l'autovalore che viene isolato è quello più vicino a λ. Più precisamente, se $\sigma_k \to \rho$, allora l'autovalore di A più vicino a λ vale $\lambda + 1/\rho$ (ovviamente, per garantire la convergenza, tale autovalore va supposto semplice).

Complessità L'operazione di complessità dominante nel metodo delle potenze è il prodotto matrice–vettore. Ogni iterazione richiede quindi un numero di operazioni di ordine $O(2n^2)$ per matrici piene e $O(cn)$ per matrici sparse. La convergenza dello schema è in generale lineare, ma diviene quadratica per matrici simmetriche. La situazione più favorevole all'uso del metodo delle potenze è perciò quella di matrice simmetrica sparsa.

Risultati fondamentali

- Convergenza del metodo delle potenze

Teorema 2.1 *Sia A una matrice reale $n \times n$. Se vale (2.2), allora il metodo delle potenze definito da (2.5), (2.6) e (2.7) converge, e più precisamente $y_k = c\,\mathrm{sgn}(\lambda_1)^k\, x_1 + o(1)$, $\sigma_k = \lambda_1 + o(1)$ (con c costante reale ed x_1 autovettore associato a λ_1).*

Dim. Dimostreremo il teorema nell'ipotesi supplementare che esista una base di autovettori $x_1, \ldots, x_n$ linearmente indipendenti. In questa base possiamo scrivere $\tilde{z}_0$ nella forma

$$\tilde{z}_0 = \sum_{i=1}^{n} \alpha_i x_i$$

e supporremo inoltre (questa ipotesi non è molto vincolante) che $\alpha_1 \neq 0$.

Tenendo conto della definizione di z_k e $\tilde{z}_k$, si può vedere che, se si pone $\tilde{z}_0 = z_0$, allora $y_k = z_k/\|z_k\|$. Si ha quindi:

$$z_k = A^k z_0 = \sum_{i=1}^{n} \alpha_i A^k x_i = \sum_{i=1}^{n} \alpha_i \lambda_i^k x_i =$$

$$= \lambda_1^k \left(\alpha_1 x_1 + \sum_{i=2}^{n} \alpha_i \left(\frac{\lambda_i}{\lambda_1} \right)^k x_i \right) = \lambda_1^k (\alpha_1 x_1 + o(1)) \qquad (2.14)$$

in cui l'ultimo passaggio è giustificato dal fatto che $|\lambda_i|/|\lambda_1| < 1$ e quindi le successive potenze di tutti questi rapporti tendono a zero. Dalla (2.14) si ha poi:

$$y_k = \frac{z_k}{\|z_k\|} = \left(\frac{\lambda_1}{|\lambda_1|} \right)^k \frac{\alpha_1 x_1 + o(1)}{|\alpha_1|\, \|x_1\|}$$

che corrisponde alla prima parte dell'enunciato. Inoltre,

$$\sigma_k = y_k^t A y_k = \left(\frac{\lambda_1}{|\lambda_1|} \right)^{2k} \frac{\alpha_1^2 \lambda_1 x_1^t x_1 + o(1)}{|\alpha_1|^2\, \|x_1\|^2} = \lambda_1 + o(1)$$

da cui la seconda parte dell'enunciato.

■

2.2 Metodi di similitudine

In questi metodi si operano trasformazioni successive di similitudine sulla matrice A per portarla (eventualmente in senso asintotico) verso una forma di cui si possano calcolare gli autovalori. Per motivi di buon condizionamemnto e per la maggior facilità di inversione le trasformazioni che si utilizzano sono in genere ortogonali, più precisamente *rotazioni piane* e *riflessioni*.

Se la matrice A non è simmetrica la sua forma finale sarà triangolare superiore; se simmetrica la forma finale può essere diagonale o tridiagonale simmetrica. Nei primi due casi gli autovalori compaiono sulla diagonale; affronteremo invece nel §2.2.1 il problema di calcolare gli autovalori nel terzo caso.

Rotazioni Nel caso delle rotazioni, l'interpretazione geometrica è quella di matrici che ruotano ogni vettore di un dato angolo θ intorno all'origine, su un piano coordinato $x_i - x_j$. Nel caso più semplice, quello di $\mathbb{R}^2$, hanno la forma

$$Q = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

mentre in generale su $\mathbb{R}^n$ la struttura della matrice è

$$Q = \begin{pmatrix}
1 & & & 0 & & & & 0 & & & & \\
& \ddots & & \vdots & & & & \vdots & & & & \\
& & 1 & 0 & & & & 0 & & & & \\
0 & \cdots & 0 & \cos\theta & 0 & \cdots & 0 & -\sin\theta & 0 & \cdots & 0 & \leftarrow i \\
& & & 0 & 1 & & & 0 & & & & \\
& & & \vdots & & \ddots & & \vdots & & & & \\
& & & 0 & & & 1 & 0 & & & & \\
0 & \cdots & 0 & \sin\theta & 0 & \cdots & 0 & \cos\theta & 0 & \cdots & 0 & \leftarrow j \\
& & & 0 & & & & 0 & 1 & & & \\
& & & \vdots & & & & \vdots & & \ddots & & \\
& & & 0 & & & & 0 & & & 1 &
\end{pmatrix} \tag{2.15}$$

in cui le righe e colonne diverse da quelle della matrice identità sono la i e la j; ciò corrisponde appunto ad effettuare una rotazione sul piano $x_i - x_j$, lasciando inalterate le altre componenti.

La tipica operazione che può essere effettuata tramite una rotazione è l'annullamento di una coppia di elementi simmetrici. Più precisamente, tramite una trasformazione di similitudine $\tilde{A} = Q^t A Q$ in cui A sia una matrice simmetrica e Q la matrice di rotazione (2.15) è possibile ottenere $\tilde{a}_{ij} = 0$

scegliendo θ in modo che

$$\tan\theta = \frac{a_{jj} - a_{ii} \pm \sqrt{(a_{jj} - a_{ii})^2 + 4a_{ij}^2}}{2a_{ij}}. \tag{2.16}$$

Infatti, restringendosi per semplicità al caso $n = 2$, e ponendo $s = \sin\theta$, $c = \cos\theta$, si ha

$$\tilde{A} = Q^t A Q = \begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix} \begin{pmatrix} c & -s \\ s & c \end{pmatrix}.$$

Gli elementi fuori diagonale di questa matrice valgono

$$\tilde{a}_{12} = \tilde{a}_{21} = -s^2 a_{12} + cs(a_{22} - a_{11}) + c^2 a_{12}$$

da cui, dividendo per c^2 ed imponendo $\tilde{a}_{12} = \tilde{a}_{21} = 0$ si ottiene la condizione in $t = \tan\theta$:

$$a_{12}t^2 - (a_{22} - a_{11})t - a_{12} = 0$$

la cui soluzione coincide, per $i = 1$ e $j = 2$, con (2.16).

Riflessioni Le riflessioni, dal punto di vista geometrico, sono matrici che trasformano un vettore nel suo speculare rispetto ad un dato piano. Indicato con $w = (w_1 \cdots w_m)^t$ uno dei due versori normali al piano, la matrice di riflessione relativa è data da

$$Q_m = I - 2ww^t = \begin{pmatrix} 1 - 2w_1^2 & -2w_1w_2 & \cdots & -2w_1w_m \\ -2w_2w_1 & 1 - 2w_2^2 & \cdots & -2w_2w_m \\ \vdots & \vdots & \ddots & \vdots \\ -2w_mw_1 & -2w_mw_2 & \cdots & 1 - 2w_m^2 \end{pmatrix}, \tag{2.17}$$

e si può verificare immediatamente che Q_m, oltre ad essere ortogonale, ha colonne di norma unitaria ed è una matrice simmetrica, da cui si ha che $Q_m^{-1} = Q_m^t = Q_m$. In genere le matrici di riflessione vengono usate su un sottospazio di dimensione minore di n, mediante trasformazioni nella forma

$$Q^{(k)} = \begin{pmatrix} I_k & 0 \\ 0 & Q_{n-k} \end{pmatrix} \tag{2.18}$$

in cui I_k è una matrice identità di dimensione k, Q_{n-k} è una riflessione di dimensione $m = n - k$. Si noti che, per la struttura di $Q^{(k)}$, il prodotto $Q^{(k)}A$ lascia inalterate le prime k righe di A, mentre il prodotto $(Q^{(k)}A)Q^{(k)}$ lascia inalterate le prime k colonne di $Q^{(k)}A$. La operazione tipica che si

può effettuare per riflessione è l'azzeramento di un certo numero di elementi di una colonna della matrice A (poiché una riflessione in dimensione $n-k$ è individuata da $n-k-1$ parametri liberi, ci si aspetta che questo sia il numero di elementi che è possibile azzerare mediante la trasformazione $Q^{(k)}AQ^{(k)}$). Più precisamente, dato un vettore colonna $x = (x_1 \cdots x_m)^t$, si tratta di determinare in funzione di x un versore w in modo che per la matrice Q_m definita da (2.17) lo trasformi nel vettore

$$\tilde{x} = Q_m x = \begin{pmatrix} \tilde{x}_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Si dimostra facilmente che ciò accade, ad esempio, se il versore w viene definito da

$$w = \frac{v}{\|v\|}, \quad v = x \pm \|x\| e_1 = \begin{pmatrix} x_1 \pm \|x\| \\ x_2 \\ \vdots \\ x_m \end{pmatrix} \tag{2.19}$$

(con $\|v\|^2 = 2(\|x\|^2 \pm x_1\|x\|)$). Infatti, in questo caso

$$\tilde{x} = x - 2ww^t x$$

e d'altra parte

$$w^t x = \frac{1}{\|v\|}(x_1^2 \pm x_1\|x\| + x_2^2 + \cdots + x_m^2) = \frac{\|v\|^2}{2\|v\|} = \frac{\|v\|}{2},$$

da cui si ottiene

$$\tilde{x} = x - 2w\frac{\|v\|}{2} = x - v = \begin{pmatrix} \pm\|x\| \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

In (2.19) la ambiguità del segno va risolta in modo da non poter causare l'annullamento della prima componente del vettore (detto *di Householder*) v, ovvero in modo che $\pm\|x\|$ abbia lo stesso segno di x_1. Questo permette anche di evitare che il valore $\|v\|$ a denominatore si annulli o comunque divenga troppo piccolo (cosa che renderebbe instabile l'operazione).

2.2.1 Metodo delle successioni di Sturm

Data una matrice simmetrica tridiagonale,

$$A = \begin{pmatrix} a_1 & b_1 & & & 0 \\ b_1 & a_2 & b_2 & & \\ & \ddots & \ddots & \ddots & \\ & & b_{n-2} & a_{n-1} & b_{n-1} \\ 0 & & & b_{n-1} & a_n \end{pmatrix}, \tag{2.20}$$

è possibile mediante questo metodo calcolarne il polinomio caratteristico in forma ricorrente con complessità lineare. Si tratta praticamente del solo caso in cui il calcolo degli autovalori viene effettuato risolvendo l'equazione caratteristica.

Ricordiamo che gli autovalori di matrici reali e simmetriche sono reali e semplici. Per calcolare effettivamente il polinomio caratteristico, partiamo dal minore principale formato dall'unico elemento a_1, il cui polinomio caratteristico è ovviamente

$$P_1(\lambda) = a_1 - \lambda, \tag{2.21}$$

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 ottenere il polinomio caratteristico al passo k, ad esempio sviluppando il determinante con i prodotti associati agli elementi della k–esima colonna, come

$$P_k(\lambda) = (a_k - \lambda)P_{k-1} - b_{k-1}^2 P_{k-2}. \tag{2.22}$$

Assegnato λ, la relazione di ricorrenza (2.21)–(2.22) 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 può quindi pensare di risolvere l'equazione caratteristica per bisezione, o precedendola con una fase di tabulazione per separare le singole radici, o utilizzando l'informazione supplementare data dal Teorema 2.2.

L'intervallo iniziale di ricerca può 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 (vedi §A.3) fatte mediante somma sulle righe o sulle colonne.

Complessità Il calcolo del polinomio caratteristico con questo metodo ha complessità $O(5n)$ (ricordiamo che, nel caso generale, il calcolo del polinomio caratteristico secondo la sua definizione ha una complessità pari a quella del calcolo di un determinante, ovvero $O(2n^3/3)$ oppure $O(n^3/3)$ passando rispettivamente per la fattorizzazione LU o di Cholesky).

Risultati fondamentali

- Proprietà delle successioni di Sturm

  Teorema 2.2 *Gli zeri di $P_k(\lambda)$ separano quelli di $P_{k+1}(\lambda)$.*

2.2.2 Metodo di Jacobi

In questo metodo si utilizzano rotazioni del tipo (2.15) per portare una matrice simmetrica in forma diagonale azzerando iterativamente gli elementi non nulli fuori della diagonale, secondo la iterazione

$$\begin{cases} A^{(0)} = A \\ A^{(k+1)} = {Q^{(k)}}^t A^{(k)} Q^{(k)} \end{cases} \tag{2.23}$$

Poichè azzerando una coppia di elementi si può far diventare diverso da zero un elemento precedentemente nullo, non si arriva in generale alla forma diagonale in un numero finito di passi. Il limite delle matrici $A^{(k)}$ è una matrice $\tilde{A} = diag(\lambda_1, \ldots, \lambda_n)$, mentre la matrice

$$\tilde{Q} = \lim_k \left(Q^{(0)} Q^{(1)} \cdots Q^{(k)} \right)$$

ha i rispettivi autovettori come colonne.

Nel metodo di Jacobi classico, al generico passo k–esimo si sceglie di azzerare la coppia di modulo massimo, ovvero si usa una matrice $Q^{(k)}$ nella forma (2.15) con i e $j > i$ scelti in modo che

$$|a_{ij}^{(k-1)}| \geq |a_{lm}^{(k-1)}| \tag{2.24}$$

per ogni l, $m > l$.

Complessità Ad ogni iterazione del metodo di Jacobi vengono modificate due righe e due colonne della matrice A; il numero di operazioni richieste da questa operazione è quindi $O(cn)$. Nel metodo "classico" la complessità dominante sarebbe quindi quella relativa alla determinazione della coppia da azzerare, operazione che richiede $O(n^2/2)$ confronti. Per questa ragione

il metodo di Jacobi viene spesso implementato nella versione *con soglia* in cui la coppia da azzerare viene scelta (in modo non univoco) in base ad una condizione del tipo

$$|a_{ij}^{(k-1)}| \geq s(k-1)$$

con una soglia $s(\cdot)$ opportunamente definita.

Risultati fondamentali

- Convergenza del metodo di Jacobi

  Teorema 2.3 *Sia A una matrice $n \times n$ reale simmetrica. Allora il metodo di Jacobi definito da (2.23), (2.24) converge, ovvero*

$$\lim_k A^{(k)} = diag(\lambda_1, \ldots, \lambda_n).$$

2.2.3 Metodo di Householder

Il metodo di Householder non porta di per se al calcolo degli autovalori, ma il suo scopo è quello di trasformare (mediante una successione di riflessioni) la matrice A in una matrice $\tilde{A}$ in forma di Hessemberg, ovvero tale che

$$\tilde{a}_{ij} = 0 \quad \text{per} \quad i > j+1.$$

Infatti, questa forma si utilizza come forma iniziale per il metodo QR, ed inoltre, poichè il metodo opera successive trasformazioni di similitudine nella forma

$$\begin{cases} A^{(1)} = A \\ A^{(k+1)} = Q^{(k)} A^{(k)} Q^{(k)} \\ \tilde{A} = A^{(n-1)} \end{cases} \tag{2.25}$$

in cui le matrici $Q^{(k)}$ hanno la struttura (2.18), e sono quindi simmetriche, se A è simmetrica lo è anche $\tilde{A}$, ed essendo anche in forma di Hessemberg è necessariamente tridiagonale. I suoi autovalori si possono quindi calcolare con il metodo delle successioni di Sturm.

In effetti, la trasformazione di una matrice nella forma di Hessemberg è possibile anche mediante rotazioni. Questa strategia dà luogo al cosiddetto *metodo di Givens*, il quale però presenta una complessità globale maggiore rispetto al metodo di Householder e non è quindi reputato competitivo.

In pratica, al passo k–esimo del metodo di Householder si ha

$$A^{(k)} = \begin{pmatrix} B^{(k)} & C^{(k)} \end{pmatrix}$$

in cui $B^{(k)}$ è una matrice $n \times (k-1)$ in forma di Hessemberg, e $C^{(k)}$ è una matrice $n \times (n-k+1)$, in generale piena e senza struttura. Il prodotto $Q^{(k)}A^{(k)}$ (con $Q^{(k)}$ data da (2.18)) lascia inalterate le prime k righe di $A^{(k)}$ e le prime $k-1$ colonne per quanto riguarda gli elementi dal $(k+1)$–esimo all'n–esimo (che sono tutti nulli), mentre è costruita in modo da azzerare gli elementi dal $(k+2)$–esimo all'n–esimo della k–esima colonna della matrice $A^{(k)}$, ovvero della prima colonna della matrice $C^{(k)}$. Per fare questa operazione, al passo k–esimo il vettore di Householder va definito come visto nel §2.2, ponendo

$$x = \begin{pmatrix} c^{(k)}_{k+1,1} \\ \vdots \\ c^{(k)}_{n,1} \end{pmatrix}.$$

D'altra parte, il prodotto $(Q^{(k)}A^{(k)})Q^{(k)}$ lascia inalterate le prime k colonne di $Q^{(k)}A^{(k)}$ ed in particolare le lascia nella forma di Hessemberg. Dopo $n-2$ trasformazioni tutta la matrice A è in forma di Hessemberg.

Una variante di questo algoritmo permette invece di porre la matrice A in forma triangolare superiore mediante prodotti *a sinistra* di matrici del tipo (2.18), mediante lo schema

$$\begin{cases} A^{(0)} = A \\ A^{(k+1)} = Q^{(k)}A^{(k)} \\ \tilde{A} = A^{(n-1)}. \end{cases} \tag{2.26}$$

In questo caso, la prima riflessione è in dimensione n ed azzera gli elementi dal secondo all'n–esimo della prima colonna, la seconda trasformazione lascia inalterata la prima riga e tutti gli elementi nulli della prima colonna ed azzera gli elementi dal terzo all'n–esimo della seconda colonna, e così via fino ad ottenere una matrice $A^{(n-1)} = R$ triangolare superiore. Per fare questa operazione il vettore di Householder va definito ponendo al passo k–esimo

$$x = \begin{pmatrix} a^{(k)}_{k+1,k+1} \\ \vdots \\ a^{(k)}_{n,k+1} \end{pmatrix}.$$

Ponendo poi

$$Q = Q^{(n-2)}Q^{(n-3)} \cdots Q^{(0)}$$

si ha $QA = R$, e quindi, poiché Q è ortogonale e simmetrica, $A = QR$.

Complessità Nel metodo di Householder, in effetti, le matrici $Q^{(k)}$ non vengono realmente costruite ad ogni passo; si può infatti esprimere direttamente in modo semplice il loro prodotto per $A^{(k)}$. A conti fatti, la complessità

sia per la riduzione a forma di Hessemberg che per la fattorizzazione QR è $O(4n^3/3)$. Nel primo caso, l'altra possibilità, offerta dal metodo di Givens, ha complessità maggiore. Nel secondo caso, la fattorizzazione QR si potrebbe effettuare mediante il processo di ortogonalizzazione di Gram–Schmidt, il quale però non è sufficientemente stabile se implementato in dimensione alta.

2.2.4 Metodo QR

Il metodo QR si applica a matrici con autovalori distinti, ma non necessariamente simmetriche. Il metodo si basa sulla iterazione

$$\begin{cases} A^{(0)} = A \\ A^{(k)} = Q^{(k)}R^{(k)} & \text{(fattorizzazione di } A^{(k)}) \\ A^{(k+1)} = R^{(k)}Q^{(k)} & \text{(definizione di } A^{(k+1)}). \end{cases} \tag{2.27}$$

Poichè dalla seconda riga di (2.27) si ha

$$Q^{(k)^t}A^{(k)} = R^{(k)},$$

dalla terza si ottiene

$$A^{(k+1)} = Q^{(k)^t}A^{(k)}Q^{(k)}$$

e quindi la successione $A^{(k)}$ è effettivamente costituita di matrici simili, ed inoltre se $A^{(k)}$ è simmetrica lo è anche $A^{(k+1)}$.

Si può dimostrare che se A è in forma di Hessemberg, tutte le matrici della successione $A^{(k)}$ sono nella stessa forma. E' quindi conveniente porre preventivamente in forma di Hessemberg la matrice A applicando il metodo di Householder in modo da effettuare ad ogni passo la fattorizzazione QR su di una matrice di Hessemberg, per la quale la fattorizzazione presenta minore complessità (infatti, la triangolarizzazione richiede di annullare un solo elemento per colonna).

Complessità La velocità di convergenza del metodo QR può essere bassa se ci sono autovalori molto vicini in modulo. Esistono opportune tecniche (di scalamento) per accelerare la convergenza in questi casi, e lo schema risultante può essere competitivo con il metodo delle successioni di Sturm nel caso di matrici simmetriche.

Risultati fondamentali

- Convergenza del metodo QR

Teorema 2.4 *Sia A una matrice reale $n \times n$. Se vale la condizione*

$$|\lambda_1| > |\lambda_2| > |\lambda_3| > \cdots > |\lambda_n| > 0, \tag{2.28}$$

allora la successione $A^{(k)}$ definita da (2.27) converge e si ha

$$\lim_{k\to\infty} A^{(k)} = U \tag{2.29}$$

con U matrice triangolare superiore (diagonale se A è simmetrica), e $\lambda_i = u_{ii} \quad (i = 1, \ldots, n)$.

2.3 Confronto fra i vari schemi

Ricapitoliamo intanto le possibilità di applicare i vari schemi nel calcolo di *tutti* gli autovalori:

- Matrici simmetriche tridiagonali: $\begin{cases} \text{Metodo di Jacobi} \\ \text{Metodo delle successioni di Sturm} \\ \text{Metodo } QR \end{cases}$
- Matrici simmetriche generiche: $\begin{cases} \text{Jacobi} \\ \text{Householder/Givens + } QR \\ \text{Householder/Givens + Sturm} \end{cases}$
- Matrici non simmetriche: Householder/Givens + QR

mentre, nel caso si vogliano calcolare solo *alcuni* autovalori, si possono utilizzare i metodi delle potenze (con le sue varianti, inclusa eventualmente la deflazione) e di Lanczos, che viene normalmente preferito per motivi di efficienza.

La trasformazione di una matrice nella forma di Hessemberg si effetua di preferenza con il metodo di Householder che, come si è già detto, ha complessità minore del metodo di Givens, almeno per matrici piene. L'uso di trasformazioni di rotazione tramite i metodi di Givens o di Jacobi può al contrario essere conveniente nel caso delle matrici sparse, in cui serva azzerare "pochi" elementi in modo più selettivo.

3 Problemi di minimizzazione libera

In questa sezione esamineremo algoritmi di risoluzione del problema

$$f(x^*) = \min_{x \in \mathbb{R}^n} f(x) \tag{3.1}$$

con $f \in C^1$ (per i metodi di tipo Newton $f \in C^2$). Ricordiamo che nel caso f sia strettamente convessa, se il punto di minimo in (3.1) esiste, allora è l'unico punto stazionario, e quindi questo problema equivale a trovare la soluzione del sistema nonlineare

$$\nabla f(x) = 0.$$

Se f è quadratica definita positiva, ovvero $f(x) = 1/2(Ax, x) - (b, x)$ con A definita positiva, allora il punto di minimo è soluzione del sistema lineare $Ax = b$.

La struttura generale di un metodo iterativo di minimizzazione è

$$x_{k+1} = x_k + \beta_k d_k, \tag{3.2}$$

dove $\beta_k \in \mathbb{R}$, $d_k \in \mathbb{R}^n$ ed x_0 è assegnato. Si tratta quindi di muoversi lungo la direzione d_k con passo β_k, entrambi da scegliersi opportunamente ad ogni passo, imponendo di regola che $f(x_{k+1}) < f(x_k)$ (questa condizione garantisce che la successione $\{x_k\}$ sia nell'insieme di sottolivello

$$\Sigma_0 = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\} \tag{3.3}$$

e permette, nel caso di funzioni coercitive, di ottenere una successione limitata). Usualmente si richiede che la direzione d_k sia una direzione di discesa, ovvero che $(d_k, \nabla f(x_k)) < 0$, ed in questo caso tipicamente $\beta_k \in \mathbb{R}^+$ (questa scelta porta ad indicare metodi del tipo (3.2) come *metodi di discesa*). Per evitare che le direzioni d_k possano diventare asintoticamente ortogonali al gradiente, cosa che potrebbe bloccare la convergenza dell'algoritmo, si impone normalmente la condizione più forte

$$\frac{(d_k, \nabla f(x_k))}{\|d_k\| \|\nabla f(x_k)\|} \leq -\cos\theta \tag{3.4}$$

con $\theta < \pi/2$ (ciò individua un cono di direzioni di semiapertura θ intorno alla direzione di massima discesa).

Criteri di arresto Analogamente a quanto visto nel caso dei sistemi lineari o delle equazioni scalari, anche nei problemi di minimizzazione il test di arresto tipico prevede che sia l'aggiornamento $\|x_{k+1} - x_k\|$, sia il residuo, rappresentato in questo caso da $\|\nabla f(x_{k+1})\|$, siano al di sotto di soglie date. Nel caso dei metodi di direzioni coniugate applicati a funzioni quadratiche (vedi §3.2.3), se si lavora in dimensione alta è comunque necessario calcolare il residuo alla fine delle n iterazioni, per verificare che l'accumulo di errori di arrotondamento non abbia danneggiato l'accuratezza della soluzione.

3.1 Strategie di scelta del passo

Definita la restrizione di f alla direzione d_k come

$$\phi_k(\beta) := f(x_k + \beta d_k),$$

le strategie più comuni per scegliere gli scalari β_k sono:

Ricerca esatta – Si sceglie β_k in modo che si abbia

$$\phi_k(\beta_k) = \min_{\beta} \phi_k(\beta) \tag{3.5}$$

applicando un metodo di ricerca unidimensionale, ad esempio il metodo di bisezione. Nel caso di funzioni quadratiche il minimo (3.5) si può calcolare esplicitamente.

Ricerca parziale – In questo caso β_k è determinato (in modo non univoco) da condizioni che garantiscono da un lato che la funzione decresca abbastanza tra x_k e x_{k+1}, dall'altro che l'algoritmo non si blocchi a causa di passi troppo piccoli. In genere queste condizioni sono costruite in modo da essere "facili" da soddisfare (per abbassare la complessità) e da includere le condizioni di spostamento ottimale (ad esempio, il passo che corrisponde alla ricerca unidimensionale esatta).

Passo fisso – In questo caso si sceglie $\beta_k \equiv \bar{\beta}$ (costante).

3.1.1 Ricerca esatta

La strategia di ricerca esatta è quella che permette di scegliere il punto x_{k+1} in modo che la funzione decresca il più possibile, quindi in un certo senso è la strategia che ci si aspetta converga più velocemente a parità di numero di iterazioni. Tuttavia questa maggiore velocità di convergenza si ottiene a prezzo di una maggiore complessità computazionale della singola iterazione,

ed in molti casi si ritiene che le strategie di ricerca parziale possano avere una efficienza globale maggiore.

Si osservi che in ricerca esatta il punto x_{k+1} è il minimo di $f(x)$ vincolato alla semiretta $x = x_k + \beta d_k$ al variare di $\beta > 0$. In funzioni convesse e differenziabili, tale minimo corrisponde all'unico punto in cui $\nabla f(x)$ è ortogonale alla semiretta. Si ha quindi

$$(\nabla f(x_{k+1}), d_k) = 0. \tag{3.6}$$

In generale la ricerca unidimensionale esatta è essa stessa un procedimento iterativo, a meno che la funzione non sia quadratica, nel qual caso il minimo si può calcolare esplicitamente. Consideriamo quindi separatamente i due casi.

Funzioni quadratiche Supponiamo inizialmente che la funzione f da minimizzare abbia la forma

$$f(x) = \frac{1}{2}(Ax, x) - (b, x) \tag{3.7}$$

con A definita positiva. In un generico algoritmo di minimizzazione iterativo, nella forma (3.2), è in questo caso possibile esprimere esplicitamente il valore di β che corrisponde alla minimizzazione unidimensionale esatta. Infatti, omettendo l'indice k per semplicità, si ha

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

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

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

Annullando la derivata di ϕ si ottiene

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

che fornisce per β_k il valore

$$\beta_k = -\frac{(\nabla f(x_k), d_k)}{(Ad_k, d_k)}. \tag{3.8}$$

E' immediato verificare che, se $d_k \neq 0$, il passo di ricerca β_k è nullo se e solo se $\nabla f(x_k) = 0$

Funzioni non quadratiche Nel caso di una funzione generica, si utilizzano algoritmi di minimizzazione unidimensionale il cui prototipo è il metodo di bisezione che si descrive di seguito. Tale metodo non è il più efficiente tra i metodi di ricerca unidimensionale, ma è sufficiente per le applicazioni usuali.

Si supponga che la $\phi(\beta)$ sia continua e unimodale nell'intervallo $[a_0, b_0]$ (se la ϕ è convessa e $\phi'(0) < 0$, un modo di scegliere questo intervallo iniziale può essere assegnando $a_0 = 0$, e b_0 tale che $\phi(b_0) > \phi(0)$). Si pone ora $j = 0$ e si opera iterativamente secondo i passi che seguono.

1. Poni:

$$c_j = \frac{a_j + b_j}{2}$$

2. Se è soddisfatta la condizione di arresto, STOP. Altrimenti poni:

$$d_j = \frac{a_j + c_j}{2} \ , \quad e_j = \frac{c_j + b_j}{2}$$

3. Se $\phi(a_j) = \min[\phi(a_j), \phi(b_j), \phi(c_j), \phi(d_j), \phi(e_j)]$, poni

$$a_{j+1} = a_j \ , \quad b_{j+1} = d_j$$

incrementa j e vai a 1.

4. Se $\phi(b_j) = \min[\phi(a_j), \phi(b_j), \phi(c_j), \phi(d_j), \phi(e_j)]$, poni

$$a_{j+1} = e_j \ , \quad b_{j+1} = b_j$$

incrementa j e vai a 1.

5. Se $\phi(c_j) = \min[\phi(a_j), \phi(b_j), \phi(c_j), \phi(d_j), \phi(e_j)]$, poni

$$a_{j+1} = d_j \ , \quad b_{j+1} = e_j \ , \quad c_{j+1} = c_j$$

incrementa j e vai a 2.

6. Se $\phi(d_j) = \min[\phi(a_j), \phi(b_j), \phi(c_j), \phi(d_j), \phi(e_j)]$, poni

$$a_{j+1} = a_j \ , \quad b_{j+1} = c_j \ , \quad c_{j+1} = d_j$$

incrementa j e vai a 2.

7. Poni

$$a_{j+1} = c_j \ , \quad b_{j+1} = b_j \ , \quad c_{j+1} = e_j$$

incrementa j e vai a 2.

Ad ogni passo, l'intervallo $[a_{j+1}, b_{j+1}]$ contiene il punto di minimo β^*. Per dimostrarlo, supponiamo ad esempio che il nodo a cui corrisponde il valore minimo sia c_j. Se per assurdo il punto β^* fosse esterno all'intervallo $[d_j, e_j]$, diciamo ad esempio $\beta^* \in [e_j, b_j]$, poiché $\phi(\beta^*) \leq \phi(c_j)$, si dovrebbe avere $\phi(c_j) < \phi(e_j)$, $\phi(e_j) > \phi(\beta^*)$ e $\phi(\beta^*) < \phi(b_j)$, contro la ipotesi di unimodalità.

Ovviamente, $b_{j+1} - a_{j+1} \leq (b_j - a_j)/2$ e quindi il metodo ha sostanzialmente convergenza lineare. La tipica condizione di arresto è del tipo $b_j - a_j < \varepsilon$, ovvero $k > \log_2(b_0 - a_0) - \log_2 \varepsilon$.

Ogni iterazione riduce l'errore della metà e richiede normalmente di valutare la funzione ϕ 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). A questo proposito, è importante notare che il valore della funzione negli altri punti utilizzati ad ogni passo non va realmente ricalcolato essendo disponibile dal passo precedente.

Si può notare che lo schema potrebbe funzionare anche calcolando la funzione in due punti interni anziché tre, e scartando ad ogni passo un solo sottointervallo di $[a_j, b_j]$ anziché due. Naturalmente, in questo caso i nodi vanno disposti opportunamente per permettere di riutilizzare ad un certo passo i valori della funzione calcolati al passo precedente. Una tipica strategia è quella della *sezione aurea* che porta ad uno schema un po' più efficiente della bisezione.

Complessità Nel caso di funzioni quadratiche il costo della ricerca unidimensionale è legato al calcolo del secondo membro di (3.8), ed in particolare (poiché il prodotto scalare ha complessità $O(n)$) di $\nabla f(x_k) = Ax_k - b$ e del prodotto Ad_k. La complessità di questi calcoli, in entrambi i casi, è dell'ordine del numero di elementi non nulli di A, quindi $O(n^2)$ per problemi pieni e $O(n)$ per problemi sparsi.

Nel caso di funzioni non quadratiche, la complessità dipende dalla precisione con cui si valuta il minimo di ϕ, ed è meno facile da determinare. Le prestazioni di molti metodi di discesa non dipendono però in modo critico dalla accuratezza di questa ricerca (considerazione che spinge, in caso di funzioni generiche, verso l'uso di strategie di ricerca parziale).

Risultati fondamentali

- Convergenza dei metodi di discesa in ricerca esatta

Teorema 3.1 *Sia $f(x) \in C^1$, strettamente convessa sull'insieme Σ_0 definito dalla (3.3), e la successione x_k sia generata tramite l'algoritmo (3.2). Si supponga*

i) che l'insieme Σ_0 sia compatto;

ii) che le direzioni d_k soddisfino (3.4) per $k \in \mathcal{I}$ (dove $\mathcal{I}$ è un insieme illimitato di indici;

iii) che per $k \in \mathcal{I}$, β_k sia ottenuto tramite ricerca esatta.

Allora la successione x_k converge all'unico punto x^* di minimo per f.

Dim. Senza perdita di generalità, possiamo ipotizzare che le direzioni d_k siano normalizzate, ovvero che $\|d_k\| = 1$.

Notiamo intanto che in ricerca esatta è sicuramente verificata la condizione $f(x_{k+1}) \leq f(x_k)$, che implica a sua volta che $x_k \in \Sigma_0$ per ogni k.

Supponiamo di estrarre dalla sottosuccessione di indici per cui (3.4) è soddisfatta, una ulteriore sottosuccessione $\{x_{k_l}\}$ convergente ad un punto $\bar{x}$ e tale che $d_{k_l} \to \bar{d}$ (una tale successione esiste certamente data la limitatezza di entrambe le successioni $\{x_k\}_{k\in\mathcal{I}}$ e $\{d_k\}_{k\in\mathcal{I}}$). Dalla continuità e differenziabilità di f deriva che $f(x_{k_l}) \to f(\bar{x})$ e che $\nabla f(x_{k_l}) \to \nabla f(\bar{x})$. Se si suppone per assurdo che $\nabla f(\bar{x}) \neq 0$, dalla (3.4) segue che $(\nabla f(\bar{x}), \bar{d}) < 0$ ed esiste uno scalare positivo $\bar{\beta}$ tale che

$$f(\bar{x} + \bar{\beta}\bar{d}) < f(\bar{x}). \tag{3.9}$$

Ora, dalla proprietà di monotonia $f(x_{k+1}) \leq f(x_k)$, si ottiene:

$$f(x_{k_{l+1}}) \leq f(x_{k_l+1}) = f(x_{k_l} + \beta_{k_l} d_{k_l}) \leq f(x_{k_l} + \bar{\beta} d_{k_l}) \tag{3.10}$$

in cui l'ultima disuguaglianza è motivata dal fatto che gli scalari β_{k_l} soddisfano la condizione di ricerca unidimensionale esatta. Passando ora al limite sui due termini estremi di (3.10), si ottiene:

$$f(\bar{x}) \leq f(\bar{x} + \bar{\beta}\bar{d})$$

ma questa relazione è in contrasto con (3.9), e da ciò si deduce che necessariamente $\nabla f(\bar{x}) = 0$. Essendo poi f convessa su Σ_0, esiste un unico punto di minimo x^* che è anche l'unico punto stazionario. Da ciò segue che tutta la successione $\{x_k\}$ converge a questo punto.

■

3.1.2 Ricerca parziale

In questa strategia i valori accettabili di β coprono un insieme relativamente "grande", caratterizzato, almeno nei due casi più comuni, da una doppia disuguaglianza. La prima di queste disuguaglianze impedisce di rallentare lo schema scegliendo valori di β troppo piccoli, la seconda assicura che la funzione f "decresca abbastanza" passando da x_k ad x_{k+1}. Queste due condizioni garantiscono a loro volta di soddisfare le condizioni di convergenza del Teorema 3.2. In quanto segue, si suppone che $\phi_k'(0) = (d_k, \nabla f(x_k)) < 0$.

Criterio di Armijo–Goldstein In questo caso β_k è determinato dalle condizioni

$$\phi_k(0) + \sigma_1\beta_k\phi_k'(0) \leq \phi_k(\beta_k) \leq \phi_k(0) + \sigma_2\beta_k\phi_k'(0)$$

che si possono riscrivere più esplicitamente come

$$f(x_k)+\sigma_1\beta_k(d_k, \nabla f(x_k)) \leq f(x_k+\beta_k d_k) \leq f(x_k)+\sigma_2\beta_k(d_k, \nabla f(x_k)) \quad (3.11)$$

dove $0 < \sigma_2 < \sigma_1 < 1$.

La ricerca di un opportuno valore di β può essere effettuata per bisezione, lavorando tra un valore $\beta = a_0$ che viola la prima disuguaglianza ed un valore $\beta = b_0$ che viola la seconda. Si pone $j = 0$, e si procede secondo l'algoritmo che segue.

1. Poni:

$$c_j = \frac{a_j + b_j}{2}$$

2. Se $\phi_k(0) + \sigma_1 c_j\phi_k'(0) > \phi_k(c_j)$, poni

$$a_{j+1} = c_j \ , \quad b_{j+1} = b_j$$

incrementa j e vai a 1.

3. Se $\phi_k(0) + \sigma_2 c_j\phi_k'(0) < \phi_k(c_j)$, poni

$$a_{j+1} = a_j \ , \quad b_{j+1} = c_j$$

incrementa j e vai a 1.

4. Poni $\beta_k = c_j$, STOP.

Una variante semplificata di questo algoritmo usa una sola delle due disuguaglianze:

$$f(x_k + \beta_k d_k) \leq f(x_k) + \sigma_2 \beta_k (d_k, \nabla f(x_k)) \tag{3.12}$$

(con $0 < \sigma_2 < 1$), cercando però di soddisfarla in modo il più stretto possibile. Posto $j = 0$ e scelto un valore a_0 che viola la disuguaglianza (3.12) ed una costante $\rho > 1$, si procede secondo l'algoritmo seguente.

1. Poni:

$$a_{j+1} = \rho a_j$$

2. Se $\phi_k(a_{j+1}) < \phi_k(0) + \sigma_2 a_{j+1} \phi_k'(0)$, incrementa j e vai a 1.

3. Poni $\beta_k = a_j$, STOP.

In questo modo, il passo β_k è limitato inferiormente dal fatto di non poter essere minore di $1/\rho$ volte il più piccolo passo che viola la disuguaglianza (3.12).

Criterio di Wolfe–Powell In questo caso β_k è determinato dalle condizioni

$$\phi_k'(\beta_k) \geq \sigma_1 \phi_k'(0)$$
$$\phi_k(\beta_k) \leq \phi_k(0) + \sigma_2 \beta_k \phi_k'(0)$$

la cui forma esplicita è

$$\begin{cases} (d_k, \nabla f(x_k + \beta_k d_k)) \geq \sigma_1 (d_k, \nabla f(x_k)) \\ f(x_k + \beta_k d_k) \leq f(x_k) + \sigma_2 \beta_k (d_k, \nabla f(x_k)) \end{cases} \tag{3.13}$$

dove ancora $0 < \sigma_2 < \sigma_1 < 1$.

Come per il criterio di Armijo–Goldstein, si può lavorare per bisezione, tra un valore che viola la prima disuguaglianza ed un valore che viola la seconda. A partire da $j = 0$, si procede secondo l'algoritmo:

1. Poni:

$$c_j = \frac{a_j + b_j}{2}$$

2. Se $\phi_k'(c_j) < \sigma_1 \phi_k'(0)$, poni

$$a_{j+1} = c_j \ , \quad b_{j+1} = b_j$$

incrementa j e vai a 1.

3. Se $\phi_k(0) + \sigma_2 c_j \phi_k'(0) < \phi_k(c_j)$, poni

$$a_{j+1} = a_j \ , \quad b_{j+1} = c_j$$

incrementa j e vai a 1.

4. Poni $\beta_k = c_j$, STOP.

Complessità Nella ricerca parziale va ricercato un compromesso tra numero di operazioni necessarie a determinare il passo e numero di iterazioni: allargare troppo l'intervallo $[\sigma_2, \sigma_1]$ facilita la determinazione di β_k ma rende più lenta la decrescita della funzione f. La letteratura riporta scelte efficienti per i valori σ_1 e σ_2.

Risultati fondamentali

- Convergenza dei metodi di discesa in ricerca parziale o a passo fisso

  Teorema 3.2 *Sia $f(x) \in C^1$, strettamente convessa sull'insieme Σ_0 definito dalla (3.3), e la successione x_k sia generata tramite l'algoritmo (3.2). Si supponga*

  i) che $f(x_{k+1}) < f(x_k)$;

  ii) che l'insieme Σ_0 sia compatto;

  iii) che le direzioni d_k soddisfino (3.4) per $k \in \mathcal{I}$ (dove $\mathcal{I}$ è un insieme illimitato di indici;

  iv) che per $k \in \mathcal{I}$ gli scalari β_k soddisfino, per qualche $\gamma, \underline{c}, \overline{c} > 0$ le condizioni

  $$\beta_k \|d_k\| \geq \underline{c} \, \|\nabla f(x_k)\|^\gamma, \tag{3.14}$$

  $$f(x_k + \beta_k d_k) \leq f(x_k) + \overline{c} \beta_k (\nabla f(x_k), d_k). \tag{3.15}$$

  Allora la successione x_k converge all'unico punto x^* di minimo per f.

  Dim. Notiamo intanto che la condizione $i)$ implica che $x_k \in \Sigma_0$ per ogni k. Inoltre, la successione $\{f(x_k)\}$ è monotona decrescente ed inferiormente limitata, ed ammette quindi limite.

Supponiamo di estrarre dalla sottosuccessione di indici per cui (3.4) è soddisfatta, una ulteriore sottosuccessione $\{x_{k_l}\}$ convergente ad un punto $\bar{x}$ (una tale successione esiste certamente data la limitatezza della successione $\{x_k\}$). Dalla continuità e differenziabilità di f deriva che $f(x_{k_l}) \to f(\bar{x})$ e che $\nabla f(x_{k_l}) \to \nabla f(\bar{x})$. Per la monotonia della successione $\{f(x_k)\}$ possiamo scrivere

$$f(x_{k_{l+1}}) \leq f(x_{k_l+1}) = f(x_{k_l} + \beta_{k_l} d_{k_l}) \tag{3.16}$$

ed applicando ulteriormente (3.4), (3.14) e (3.15):

$$\begin{aligned} f(x_{k_l} + \beta_{k_l} d_{k_l}) &\leq f(x_{k_l}) + \overline{c}\beta_{k_l}(\nabla f(x_{k_l}), d_{k_l}) \leq \\ &\leq f(x_{k_l}) - \overline{c}\beta_{k_l} \cos\theta \|\nabla f(x_{k_l})\| \|d_{k_l}\| \leq \\ &\leq f(x_{k_l}) - \overline{c}\,\underline{c} \cos\theta \|\nabla f(x_{k_l})\|^{1+\gamma}. \end{aligned} \tag{3.17}$$

Da (3.16), (3.17) segue quindi che

$$f(x_{k_{l+1}}) \leq f(x_{k_l}) - \overline{c}\,\underline{c} \cos\theta \|\nabla f(x_{k_l})\|^{1+\gamma},$$

e passando al limite sulla sottosuccessione $\{x_{k_l}\}$,

$$f(\bar{x}) \leq f(\bar{x}) - \overline{c}\,\underline{c} \cos\theta \|\nabla f(\bar{x})\|^{1+\gamma},$$

ma questa relazione può essere soddisfatta solo se $\nabla f(\bar{x}) = 0$. Essendo f convessa, esiste un unico punto di minimo x^* che è anche l'unico punto stazionario, e quindi $\bar{x} = x^*$. Poiché la successione $\{f(x_k)\}$ ammette limite, ne segue che necessariamente $f(x_k) \to \min f(x)$ e tutta la successione $\{x_k\}$ converge al punto x^*.

■

- Condizioni di convergenza per gli algoritmi di ricerca parziale

Teorema 3.3 *Se $f(x) \in C^2(\Sigma_0)$ ed è soddisfatta la condizione (3.4), allora sia il criterio di Armijo–Goldstein (3.11) che quello di Wolfe–Powell (3.13) soddisfano le condizioni (3.14), (3.15).*

Dim. Iniziamo osservando che la condizione (3.15) coincide esattamente con la disuguaglianza comune ad entrambe le strategie di ricerca,

$$\phi_k(\beta_k) \leq \phi_k(0) + \sigma_2 \beta_k \phi_k'(0),$$

una volta esplicitata la funzione ϕ_k e posto $\overline{c} = \sigma_2$.

Passiamo ora a considerare la condizione (3.14). Per quanto riguarda il metodo di Armijo–Goldstein, partiamo maggiorando il valore di $\phi_k(\beta_k)$ come

$$\phi_k(\beta_k) \leq \phi_k(0) + \beta_k \phi_k'(0) + \frac{1}{2}\beta_k^2 \max_\beta |\phi_k''(\beta)|.$$

Utilizzando questa maggiorazione nella prima disuguaglianza di (3.11), si ottiene

$$\phi_k(0) + \beta_k \phi_k'(0) + \frac{1}{2}\beta_k^2 \max_\beta |\phi_k''(\beta)| \geq \phi_k(0) + \sigma_1 \beta_k \phi_k'(0).$$

e quindi la condizione

$$(1-\sigma_1)\beta_k \phi_k'(0) + \frac{1}{2}\beta_k^2 \max_\beta |\phi_k''(\beta)| \geq 0. \tag{3.18}$$

D'altra parte, tenendo conto che $\phi_k' < 0$ e ϕ_k'' sono le derivate direzionali prima e seconda nella direzione d_k, si ha a fortiori

$$\phi_k'(0) \leq -\cos\theta \|d_k\| \|\nabla f(x_k)\|$$

$$\max_\beta |\phi_k''(\beta)| \leq \max_x \|H_f(x)\| \|d_k\|^2.$$

Sostituendo queste maggiorazioni nella (3.18) e risolvendo rispetto a β_k (ci si può ovviamente limitare ai valori positivi), si arriva alla stima

$$\beta_k \|d_k\| \geq \frac{2(1-\sigma_1)\cos\theta}{\max_x \|H_f(x)\|} \|\nabla f(x_k)\|.$$

che è nella forma (3.14). Per quanto riguarda il metodo di Wolfe–Powell, si parte maggiorando la derivata $\phi_k'(\beta_k)$ come

$$\phi_k'(\beta_k) \leq \phi_k'(0) + \beta_k \max_\beta |\phi_k''(\beta)| \leq \phi_k'(0) + \beta_k \max_x \|H_f(x)\| \|d_k\|^2.$$

Sostituendo questa stima nella prima delle (3.13), si ottiene

$$\phi_k'(0) + \beta_k \max_x \|H_f(x)\| \|d_k\|^2 \geq \sigma_1 \phi_k'(0).$$

Maggiorando $\phi_k'(0)$ come sopra e risolvendo rispetto a β_k si ha infine

$$\beta_k \|d_k\| \geq \frac{(1-\sigma_1)\cos\theta}{\max_x \|H_f(x)\|} \|\nabla f(x_k)\|,$$

ancora nella forma (3.14).

■

3.1.3 Passo fisso

Questa strategia ha il vantaggio della semplicità, e di non richiedere calcoli per la determinazione del passo. Tuttavia, per questa strada in generale si ottiene uno schema convergente solo quando $\bar{\beta}$ è sufficientemente piccolo, ed in caso di funzioni con matrice Hessiana malcondizionata la convergenza può essere lentissima. Una variante di questo metodo prevede passi non fissati ma *predeterminati*: una scelta classica è quella di porre $\beta_k = 1/k$ (o più in generale in modo che $\beta_k \to 0$, $\sum_k \beta_k$ divergente), scelta che permette di dimostrare la convergenza sotto ipotesi più generali, ma porta ad uno schema estremamente inefficiente.

Risultati fondamentali

- Condizioni di convergenza per il passo fisso

  Teorema 3.4 *Se $f(x) \in C^2(\Sigma_0)$ e la direzione d_k è scelta in modo che, per qualche $C > 0$ e $\gamma \geq 1$, si abbia*

  $$\|d_k\| \geq C\|\nabla f(x_k)\|^\gamma,$$

  allora il metodo a passo fisso soddisfa la condizione (3.14). Se inoltre è soddisfatta la condizione (3.4), allora esiste un valore β_M tale che per $\bar{\beta} < \beta_M$ il metodo a passo fisso soddisfa la condizione (3.15) in ogni insieme in cui $\|\nabla f(x)\|$ resta limitata (in particolare, nell'insieme di sottolivello Σ_0 definito nel Teorema 3.2).

  Dim. Osserviamo intanto cha la condizione (3.14) è banalmente verificata ponendo $\underline{c} = \bar{\beta}C$.

  Per quanto riguarda poi la (3.15), procedendo come nel Teorema 3.3, osserviamo che la disuguaglianza da verificare,

  $$\phi_k(\bar{\beta}) \leq \phi_k(0) + \overline{c}\bar{\beta}\phi_k'(0), \tag{3.19}$$

  è verificata a maggior ragione se si sostituisce alla ϕ_k la sua maggiorazione

  $$\phi_k(\bar{\beta}) \leq \phi_k(0) + \bar{\beta}\phi_k'(0) + \frac{1}{2}\bar{\beta}^2 \max_x \|H_f(x)\|\|d_k\|^2. \tag{3.20}$$

  Confrontando quindi i secondi membri di (3.19) e (3.20), si ha

  $$\phi_k(0) + \overline{c}\bar{\beta}\phi_k'(0) \leq \phi_k(0) + \bar{\beta}\phi_k'(0) + \frac{1}{2}\bar{\beta}^2 \max_x \|H_f(x)\|\|d_k\|^2, \tag{3.21}$$

ovvero, continuando a restringersi ai valori positivi di β,

$$\bar{\beta} \leq \frac{2(1-\overline{c})|\phi_k'(0)|}{\max_x \|H_f(x)\| \|d_k\|^2}. \tag{3.22}$$

D'altra parte, poiché

$$|\phi_k'(0)| \leq \|d_k\| \|\nabla f(x_k)\|,$$

il secondo membro di (3.22), usando anche la ipotesi su $\|d_k\|$, si può ulteriormente maggiorare ottenendo

$$\bar{\beta} \leq \frac{2(1-\overline{c})\|\nabla f(x_k)\|}{\max_x \|H_f(x)\| \|d_k\|} \leq \frac{2(1-\overline{c})\|\nabla f(x_k)\|^{1-\gamma}}{C \max_x \|H_f(x)\|},$$

che deve valere per ogni $k \geq 0$. Ricordando che $1-\gamma < 0$, la condizione di convergenza (3.15) è quindi soddisfatta per

$$\bar{\beta} \leq \beta_M = \frac{2(1-\overline{c}) \left(\max_x \|\nabla f(x_k)\|\right)^{1-\gamma}}{C \max_x \|H_f(x)\|}.$$

■

3.2 Strategie di scelta delle direzioni di ricerca

Le strategie più comuni per scegliere le direzioni d_k sono:

Discesa più ripida (o gradiente) – si sceglie $d_k = -\nabla f(x_k)$. Questo equivale alla direzione in cui la derivata direzionale di f è di modulo maggiore.

Rilassamento – si sceglie $d_k = e_j$ ($j = (k+1) (\mathrm{mod}\, n)$), ovvero ci si muove lungo le direzioni coordinate prese in sequenza. Nel caso di funzioni quadratiche ed in ricerca esatta si riottiene l'algoritmo di Gauss–Seidel.

Direzioni coniugate – Nel caso di una f quadratica definita positiva, con matrice Hessiana A, si scelgono le direzioni d_k in modo tale che

$$(Ad_k, d_j) \begin{cases} > 0 & \text{se } k = j \\ = 0 & \text{se } k \neq j \end{cases} \tag{3.23}$$

(esistono opportune generalizzazioni nel caso di funzioni non quadratiche).

Newton – si sceglie $d_k = -P_k \nabla f(x_k)$ con $P_k = H_f(x_k)^{-1}$ (H_f matrice hessiana di f) nel caso del metodo di Newton propriamente detto, ed una sua opportuna versione approssimata nel caso dei metodi Quasi–Newton.

3.2.1 Discesa più ripida

Inizialmente proposto da Cauchy, questo metodo utilizza la scelta $d_k = -\nabla f(x_k)$. La derivata direzionale di f nella direzione d_k vale

$$\frac{\partial f}{\partial d_k}(x_k) = \frac{(d_k, \nabla f(x_k))}{\|d_k\|} = -\|\nabla f(x_k)\|$$

e d'altra parte per la disuguaglianza di Cauchy-Schwartz si ha anche

$$\frac{|(d_k, \nabla f(x_k))|}{\|d_k\|} \leq \frac{\|d_k\|\|\nabla f(x_k)\|}{\|d_k\|} = \|\nabla f(x_k)\|$$

che mostra come la direzione di ricerca sia quella in cui la derivata direzionale di f è negativa e di modulo massimo (da cui il nome del metodo). Trattandosi di una direzione di discesa, normalmente si impone $\beta_k > 0$.

Complessità Come si è detto a proposito della ricerca esatta, nel caso di funzioni quadratiche la complessità del calcolo di $d_k = -\nabla f(x_k)$ è dell'ordine del numero di elementi non nulli di A. Nel caso di funzioni non quadratiche, è legata alla complessità della dipendenza di una generica componente di ∇f dalle diverse variabili.

Risultati fondamentali

- Convergenza del metodo di discesa più ripida

  Teorema 3.5 *Sia $f(x) \in C^2$, strettamente convessa sull'insieme (che si suppone compatto) Σ_0 definito dalla (3.3), e la successione x_k sia generata tramite l'algoritmo (3.2), con $d_k = -\nabla f(x_k)$.*

  Allora, se i passi β_k sono determinati tramite una delle condizioni:

  i) ricerca esatta

  ii) ricerca parziale di Armijo–Goldstein o Wolfe–Powell

  iii) β_k fissato, sufficientemente piccolo

  la successione x_k converge all'unico punto x^* di minimo per f.

  Dim. Basta osservare che sono soddisfatte le ipotesi dei teoremi 3.1 e 3.2, con $\mathcal{I} \equiv \mathbb{N}$.

  ■

3.2.2 Rilassamento

Per uniformità con la notazione adottata per i metodi iterativi per sistemi lineari, la formulazione generale di un metodo di rilassamento in $\mathbb{R}^n$, per una funzione f di struttura generale, si pone nella forma

$$x_j^{(k+1)} = \underset{t}{\text{argmin}}\ f(x_1^{(k+1)}, \ldots, x_{j-1}^{(k+1)}, t, x_{j+1}^{(k)}, \ldots, x_n^{(k)}) \tag{3.24}$$

(per j *in sequenza* da 1 a n, $k \geq 0$). La minimizzazione rispetto a t in (3.24) viene effettuata in modo esplicito se la funzione è quadratica (nel qual caso si ottiene il metodo di Gauss–Seidel) e tramite una ricerca unidimensionale esatta negli altri casi.

Analogamente a quanto visto per il metodo di Gauss–Seidel, per accelerare la convergenza si può introdurre un *parametro di rilassamento* ω modificando lo schema nella forma SOR:

$$x_j^{(k+1)} = (1-\omega)x_j^{(k)} + \omega x_{j,rel}^{(k+1)} \tag{3.25}$$

in cui $x_{j,rel}^{(k+1)}$ è il secondo membro di (3.24). Come nel metodo di Gauss–Seidel, il valore $\omega = 1$ corrisponde al metodo di rilassamento "puro", e si tenta di utilizzare valori $\omega > 1$ nell'intento di accelerare la convergenza dello schema.

Complessità Nel caso del rilassamento, essendo le direzioni di ricerca determinate a priori, non vi è alcun costo computazionale supplementare per la loro determinazione.

Risultati fondamentali

- Convergenza dei metodi di rilassamento

  Teorema 3.6 *Si supponga $f \in C^1(\mathbb{R}^n)$, f strettamente convessa e coercitiva. Allora la successione $x^{(k)}$ definita da* (2.7) *converge all'unico punto di minimo di f su $\mathbb{R}^n$.*

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

$$f(x_1, x_2) = x_1^2 + x_2^2 - 2(x_1 + x_2) + 2|x_1 - x_2|$$

che è strettamente convessa ed ha minimo (come è 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 $(0,0)$ è di minimo sia rispetto alla variabile x_1 che a x_2, ma non è il minimo assoluto (questo è dovuto al fatto che si tratta di un punto di non differenziabilità).

3.2.3 Direzioni coniugate

Data una matrice simmetrica definita positiva A, definiamo k direzioni linearmente indipendenti $d_0, \ldots, d_{k-1}$ *coniugate* rispetto ad A se soddisfano le condizioni (3.23). Se $d_i \in \mathbb{R}^n$, l'esistenza di n direzioni coniugate linearmente indipendenti è assicurata da un noto teorema di agebra lineare (processo di ortogonalizzazione di Gram–Schmidt).

E' bene chiarire che direzioni che siano coniugate non sono necessariamente direzioni di discesa, ne' tanto meno soddisfano la (3.4), anche a meno del segno. Se supponiamo però che la funzione f da minimizzare abbia la forma (3.7) con A definita positiva, un algoritmo di minimizzazione iterativo nella forma (3.2) ha proprietà di convergenza particolarmente favorevoli se le direzioni di ricerca sono coniugate rispetto alla matrice A, come si vedrà nei risultati fondamentali.

In pratica, la generazione di direzioni coniugate, piuttosto che per ortogonalizzazione, viene usualmente fatta nel corso dell'algoritmo in forma ricorrente, ad esempio secondo lo schema fornito dal *metodo del gradiente coniugato* (CG):

$$\begin{aligned} d_0 &= -\nabla f(x_0), \\ d_k &= -\nabla f(x_k) + \alpha_k d_{k-1} \quad (k \geq 1), \end{aligned} \tag{3.26}$$

in cui gli scalari α_k si ottengono tramite le espressioni (equivalenti, nel caso di una funzione quadratica)

$$\alpha_k = \frac{(\nabla f(x_k), Ad_{k-1})}{(Ad_{k-1}, d_{k-1})} \tag{3.27}$$

$$\alpha_k = \frac{(\nabla f(x_k), \nabla f(x_k) - \nabla f(x_{k-1}))}{\|\nabla f(x_{k-1})\|^2} \quad \text{(Polak–Ribière)} \tag{3.28}$$

$$\alpha_k = \frac{\|\nabla f(x_k)\|^2}{\|\nabla f(x_{k-1})\|^2} \quad \text{(Fletcher–Reeves)} \tag{3.29}$$

e i passi β_k sono calcolati mediante (3.8).

D'altra parte, le direzioni generate mediante (3.28) o (3.29) possono essere definite anche per funzioni non quadratiche, ed in questo caso il passo β_k viene determinato per ricerca unidimensionale esatta o parziale. Per questioni

di robustezza dell'algoritmo, e per ottenere convergenza globale, può essere effettuata ad intervalli periodici (tipicamente, dopo n iterazioni) una reinizializzazione dell'algoritmo consistente nell'applicare un passo di gradiente puro. L'algoritmo viene quindi definito dalle direzioni

$$\begin{aligned} d_k &= -\nabla f(x_k) \qquad\qquad (k = 0, n, 2n, 3n, \ldots), \\ d_k &= -\nabla f(x_k) + \alpha_k d_{k-1} \quad (k \neq 0, n, 2n, 3n, \ldots), \end{aligned} \tag{3.30}$$

Se si lavora in dimensione alta, questo espediente si rende necessario anche nel caso di funzioni quadratiche, quando l'accumulo degli errori di troncamento renda necessario di effettuare più delle n iterazioni teoricamente sufficienti (vedi teorema seguente) per raggiungere il minimo. Si dimostra che la convergenza dell'algoritmo CG per funzioni non quadratiche, in ricerca unidimensionale esatta) è sopralineare, a patto di considerare blocchi di n iterazioni. Più esattamente, sotto opportune ipotesi si ha la maggiorazione

$$\|x_{k+n} - x^*\| \leq C\|x_k - x^*\|^2.$$

Complessità La generazione di direzioni coniugate viene tipicamente effettuata con formule sul tipo della (3.28), (3.29). La complessità prevalente di questo calcolo è legata alla valutazione di $\nabla f(x_k)$ (una per ogni iterazione) ed è quindi, nel caso di funzioni quadratiche, ancora dell'ordine del numero di elementi non nulli di A. Un'altro pregio di questo metodo, che lo rende adatto a problemi di grandi dimensioni, è la bassa occupazione (lineare) di memoria: in particolare, la (3.29) mostra che per calcolare d_k tutto quello che occorre memorizzare dal passo precedente sono la direzione d_{k-1} e la norma $\|\nabla f(x_{k-1})\|$.

Risultati fondamentali

- Convergenza dei metodi CD in n passi

  Teorema 3.7 *Sia data la funzione (3.7), con A definita positiva. Si definisca*

  $$\pi_k = \left\{ x \in \mathbb{R}^n : x = x_0 + \sum_{i=0}^{k-1} a_i d_i \ , \quad a_i \in \mathbb{R} \right\}. \tag{3.31}$$

  Allora, se nello schema (3.2) le direzioni $d_0, \ldots, d_{n-1}$ sono coniugate rispetto ad A, ed i β_k sono definiti da (3.8), 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 è ovviamente garantito, qualunque sia d_0, dalla scelta (3.8) per β. Si tratta di provare poi che, se $\nabla f(x_k)$ è ortogonale alla varietà π_k, allora $\nabla f(x_{k+1})$ è ortogonale alla varietà π_{k+1}, ovvero

$$(\nabla f(x_{k+1}), \bar{x} - x_0) = 0 \tag{3.32}$$

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

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

Poi, scritto $\bar{x}$ come $\bar{x} = x_0 + \bar{a}_0 d_0 + \ldots + \bar{a}_k d_k$, il primo membro di (3.32) diviene

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

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

$$= \bar{a}_k [(\nabla f(x_k), d_k) + \beta_k (A d_k, d_k)] \tag{3.33}$$

dove l'ultimo passaggio è ottenuto applicando la ipotesi di ortogonalità tra $\nabla f(x_k)$ e π_k, e la definizione di direzioni coniugate. E' poi immediato osservare che l'ultimo membro di (3.33) è nullo in virtù della scelta fatta per β_k.

■

- Buona definizione dei metodi CG

Teorema 3.8 *Sia data la funzione (3.7), con A definita positiva. Se nello schema (3.2) le direzioni $d_0, \ldots, d_{n-1}$ sono definite da (3.26), (3.27), ed i β_k sono definiti da (3.8), allora $d_k = 0$ se e solo se $\nabla f(x_k) = 0$.*

Dim. Dalla definizione (3.26), effettuando il prodotto scalare con $\nabla f(x_k)$, si ottiene

$$(\nabla f(x_k), d_k) = -\|\nabla f(x_k)\|^2 + \alpha_k (\nabla f(x_k), d_{k-1}).$$

Poiché dalla (3.6) si ottiene che l'ultimo termine è nullo, si ha

$$(\nabla f(x_k), d_k) = -\|\nabla f(x_k)\|^2. \tag{3.34}$$

Da qui segue immediatamente che, se $d_k = 0$, necessariamente $\nabla f(x_k) = 0$. Se al contrario $\nabla f(x) = 0$, allora in (3.27) si ottiene $\alpha_k = 0$, e di conseguenza, da (3.26), $d_k = 0$.

■

- Convergenza dei metodi CG in n passi

Teorema 3.9 *Sia data la funzione (3.7), con A definita positiva. Se nello schema (3.2) le direzioni $d_0, \ldots, d_{n-1}$ sono definite da (3.26), (3.27), ed i β_k sono definiti da (3.8), allora esiste un indice $m \leq n-1$ tale che, per $i = 1, \ldots, m$ e $j = 0, \ldots, i-1$ si abbia*

$$(\nabla f(x_i), \nabla f(x_j)) = 0, \tag{3.35}$$

$$(Ad_i, d_j) = 0, \tag{3.36}$$

e risulta inoltre $\nabla f(x_{m+1}) = 0$ (cioè x_{m+1} è il punto di minimo per f).

Dim. Notiamo intanto che dalla definizione del metodo una generica direzione d_k risuta sempre coniugata con la precedente, infatti da (3.26), (3.27) si ha

$$(d_k, Ad_{k-1}) = -(\nabla f(x_k), Ad_{k-1}) + \frac{(\nabla f(x_k), Ad_{k-1})}{(Ad_{k-1}, d_{k-1})}(d_{k-1}, Ad_{k-1}) = 0. \tag{3.37}$$

Poiché $\nabla f(x) = Ax - b$, si ha anche

$$\begin{aligned} \nabla f(x_{k+1}) &= Ax_{k+1} - b = \\ &= Ax_k + \beta_k Ad_k - b = \nabla f(x_k) + \beta_k Ad_k \end{aligned} \tag{3.38}$$

ed inoltre, utilizzando la (3.34) in (3.8),

$$\beta_k = \frac{\|\nabla f(x_k)\|^2}{(Ad_k, d_k)}. \tag{3.39}$$

Dimostriamo ora le (3.35), (3.36) per induzione. Se $\nabla f(x_0) = 0$, x_0 è già punto di minimo. Supponiamo quindi che esista un intero $m > 0$ tale che $\nabla f(x_i) \neq 0$ per $i = 0, \ldots, m$.

Iniziamo verificando (3.35), (3.36) per $i = 1$. In questo caso, per la definizione di d_0, si ha

$$(\nabla f(x_1), \nabla f(x_0)) = -(\nabla f(x_1), d_0) = 0$$

dove l'ultima uguaglianza deriva da (3.6). Inoltre, da (3.37) si ha anche

$$(d_1, Ad_0) = 0$$

e quindi sia (3.35) che (3.36) sono soddisfatte.

Dimostriamo ora che se (3.35), (3.36) sono soddisfatte per un generico indice $i < m$, allora sono soddisfatte anche per l'indice $i + 1$, ovvero che per $j = 0, \ldots, i$ si ha

$$(\nabla f(x_{i+1}), \nabla f(x_j)) = 0, \tag{3.40}$$

$$(Ad_{i+1}, d_j) = 0. \tag{3.41}$$

Partiamo dalla (3.40). Utilizzando la (3.38), si ha

$$\begin{aligned} (\nabla f(x_{i+1}), \nabla f(x_j)) &= (\nabla f(x_i) + \beta_i Ad_i, \nabla f(x_j)) = \\ &= (\nabla f(x_i), \nabla f(x_j)) + \beta_i (Ad_i, \nabla f(x_j)). \end{aligned} \tag{3.42}$$

Occorre distinguere i tre casi $j = 0$, $j < i$ e $j = i$. Intanto, se $j < i$, vale (3.35) e quindi

$$(\nabla f(x_{i+1}), \nabla f(x_j)) = \beta_i (Ad_i, \nabla f(x_j)). \tag{3.43}$$

Se ora $j = 0$, si ha $\nabla f(x_0) = -d_0$ e quindi il secondo membro di (3.43) è nullo per l'ipotesi induttiva (3.36). Se $j > 0$, possiamo scrivere per la (3.26) che

$$\nabla f(x_j) = \alpha_j d_{j-1} - d_j, \tag{3.44}$$

che sostituita nella (3.43) dà

$$\begin{aligned} (\nabla f(x_{i+1}), \nabla f(x_j)) &= \beta_i (Ad_i, \alpha_j d_{j-1} - d_j) = \\ &= \beta_i \alpha_j (Ad_i, d_{j-1}) - \beta_i (Ad_i, d_j) = 0, \end{aligned}$$

dove l'ultima uguaglianza si ottiene applicando ancora (3.36). Infine, se $i = j$, utilizzando la (3.44) nella (3.42), si ottiene

$$(\nabla f(x_{i+1}), \nabla f(x_j)) = \|\nabla f(x_i)\|^2 + \beta_i (Ad_i, \alpha_i d_{i-1} - d_i),$$

e quindi, usando anche la (3.36),

$$(\nabla f(x_{i+1}), \nabla f(x_j)) = \|\nabla f(x_i)\|^2 - \beta_i (Ad_i, d_i) = 0,$$

in cui l'ultima uguaglianza è ottenuta applicando (3.39).

Dimostriamo ora la (3.41), distinguendo i due casi $j = i$ e $j < i$. Se $j = i$, d_{i+1} e d_j sono due direzioni consecutive e quindi coniugate per la (3.37). Se $j < i$, si ottiene

$$\begin{aligned}(Ad_{i+1}, d_j) &= (d_{i+1}, Ad_j) = \\ &= -(\nabla f(x_{i+1}), Ad_j) + \alpha_{i+1}(d_i, Ad_j) = \\ &= -(\nabla f(x_{i+1}), Ad_j) = \\ &= -\frac{1}{\beta_j}(\nabla f(x_{i+1}), \nabla f(x_{j+1}) - \nabla f(x_j)) = \\ &= -\frac{1}{\beta_j}(\nabla f(x_{i+1}), \nabla f(x_{j+1})) + \frac{1}{\beta_j}(\nabla f(x_{i+1}), \nabla f(x_j)) = 0.\end{aligned} \tag{3.45}$$

In (3.45), abbiamo usando prima la (3.26), poi la ipotesi induttiva (3.36), infine la (3.38) scritta con $k = i$, ed ancora la (3.40) che abbiamo dimostrato in precedenza. E' così verificata anche la (3.41).

Abbiamo quindi dimostrato che, finché $\nabla f(x_k) \neq 0$, ovvero finché le direzioni di ricerca sono ben definite, esse sono anche coniugate. L'algoritmo converge quindi al più in n passi per il Teorema 3.7.

■

- Equivalenza dei vari metodi CG per funzioni quadratiche

Teorema 3.10 *Sia data la funzione (3.7), con A definita positiva. Si consideri lo schema definito da (3.2), (3.26), (3.8). Allora le espressioni (3.27), (3.28) e (3.29) sono equivalenti.*

Dim. Esprimendo il prodotto Ad_{k-1} tramite la (3.38) e sostituendo nella (3.27), si ottiene

$$\begin{aligned}\alpha_k &= \frac{\frac{1}{\beta_{k-1}}(\nabla f(x_k), \nabla f(x_k) - \nabla f(x_{k-1}))}{\frac{1}{\beta_{k-1}}(\nabla f(x_k) - \nabla f(x_{k-1}), d_{k-1})} = \\ &= \frac{(\nabla f(x_k), \nabla f(x_k) - \nabla f(x_{k-1}))}{(\nabla f(x_k) - \nabla f(x_{k-1}), d_{k-1})} = \\ &= -\frac{(\nabla f(x_k), \nabla f(x_k) - \nabla f(x_{k-1}))}{(\nabla f(x_{k-1}), d_{k-1})},\end{aligned}$$

dove, nell'ultimo passaggio, si è utilizzata la condizione di ricerca esatta (3.6). Da qui la formula di Polak–Ribière (3.28) si ottiene utilizzando la

(3.34), e la formula di Fletcher–Reeves (3.29) utilizzando ulteriormente la (3.35).

■

- Convergenza dei metodi CG con reinizializzazione per funzioni non quadratiche

Teorema 3.11 *Sia $f(x) \in C^2$, strettamente convessa sull'insieme (che si suppone compatto) Σ_0 definito dalla (3.3), e la successione x_k sia generata tramite l'algoritmo (3.2), con d_k definito da (3.30) ed α_k definito (3.28) o (3.29).*

Allora, se i passi β_k sono determinati tramite una delle condizioni:

i) ricerca esatta

ii) ricerca parziale di Armijo–Goldstein o Wolfe–Powell

iii) β_k fissato, sufficientemente piccolo

la successione x_k converge all'unico punto x^* di minimo per f.

Dim. Basta osservare che sono soddisfatte le ipotesi dei teoremi 3.1 e 3.2, con $\mathcal{I} \equiv \{0, n, 2n, 3n, \ldots\}$.

■

3.2.4 Metodo di Newton

In questo caso, la direzione di ricerca viene determinata sulla base di una linearizzazione in x_k del sistema delle condizioni di stazionarietà,

$$\nabla f(x) = 0,$$

cosa che corrisponde a definire x_{k+1} come il minimo della forma quadratica data dallo sviluppo di Taylor di secondo ordine della f in x_k. Questo sviluppo ha la forma

$$T_2(x) = f(x_k) + (\nabla f(x_k), x - x_k) + \frac{1}{2}(H(x_k)(x - x_k), x - x_k)$$

i cui punti stazionari (si tratta di un punto unico se $H(\cdot)$ è definita positiva) soddisfano il sistema di equazioni

$$\nabla T_2(x) = \nabla f(x_k) + (H(x_k), x - x_k) = 0.$$

Imporre la condizione di stazionarietà $\nabla T_2(x_{k+1}) = 0$ porta quindi a definire la direzione di ricerca come

$$d_k = -H(x_k)^{-1}\nabla f(x_k) \tag{3.46}$$

ed a scegliere un passo fisso $\beta \equiv 1$. Per questa strada si ottiene esattamente il metodo di Newton (1.44), (1.45) applicato al sistema $\nabla f(x) = 0$. Più in generale si può lavorare per ricerca esatta o parziale; la condizione di discesa nella direzione d_k è soddisfatta se $H(x_k)$ (e quindi $H(x_k)^{-1}$) è una matrice definita positiva, in questo caso infatti

$$(d_k, \nabla f(x_k)) = -(H(x_k)^{-1}\nabla f(x_k), \nabla f(x_k)) < 0. \tag{3.47}$$

Sotto ipotesi abbastanza naturali, la (3.4) è poi soddisfatta, ed in queste condizioni lo schema di Newton converge quindi globalmente sia in ricerca parziale che esatta (in quest'ultimo caso, ci si attende ovviamente che $\beta_k \to 1$, cioè che asintoticamente la scelta ottima sia proprio quella che corrisponde al metodo di Newton "puro").

Complessità Come si è detto a proposito dei sistemi di equazioni, la direzione di ricerca nel metodo di Newton viene in realtà calcolata come soluzione del sistema lineare

$$H(x_k)d_k = -\nabla f(x_k) \tag{3.48}$$

anziché tramite la (3.46). Supponendo "semplice" il calcolo di H, la complessità è dell'ordine di $O(n^3)$. Diversa è la situazione se, invece di aggiornare H ad ogni iterazione, la si calcola solo nel punto iniziale. In questo caso $H(x_0)$ si può fattorizzare preventivamente e la soluzione del sistema

$$H(x_0)d_k = -\nabla f(x_k) \tag{3.49}$$

diviene di complessità $O(n^2)$. Tuttavia, il metodo di Newton (caratteristica comune anche ai metodi "Quasi-Newton"), non è in grado di trarre vantaggio, né in termini di numero di operazioni, né in termini di occupazione di memoria, dal fatto che la matrice hessiana sia sparsa, tranne che in taluni casi di matrici strutturate. Questo fa sì che il suo utilizzo ottimale sia limitato a problemi di piccola dimensione.

Risultati fondamentali

- Convergenza del metodo di Newton a passo fisso

Teorema 3.12 *Se $f(x) \in C^3$ e la matrice hessiana H è nonsingolare in un punto stazionario x^*, allora esiste un intorno U di x^* tale che, se $x_0 \in U$, la successione x_k generata dall'algoritmo di Newton con passo fisso $\beta_k \equiv 1$ converge con velocità quadratica a x^*.*

- Convergenza del metodo di Newton in ricerca esatta e parziale

Teorema 3.13 *Sia $f(x) \in C^2$, la sua matrice hessiana H sia definita positiva sull'insieme (che si suppone compatto) Σ_0 definito dalla (3.3), e soddisfi la condizione*

$$\inf_{x \in \Sigma_0} \frac{\lambda(x)}{\Lambda(x)} \geq c > 0 \tag{3.50}$$

(dove $\lambda(x)$ e $\Lambda(x)$ sono rispettivamente minimo e massimo autovalore di $H(x)$ al variare di $x \in \mathbb{R}^n$), e la successione x_k sia generata tramite l'algoritmo (3.2), con d_k definito da (3.48) o (3.49).

Allora, se i passi β_k sono determinati tramite una delle condizioni:

i) ricerca esatta

ii) ricerca parziale di Armijo–Goldstein o Wolfe–Powell

la successione x_k converge all'unico punto x^* di minimo per f.

Dim. Si tratta di applicare il Teorema 3.2 verificando quindi che lo schema soddisfa la condizione (3.4). Notiamo intanto che sotto le ipotesi poste la funzione f è strettamente convessa ed ammette quindi un unico minimo globale x^*. Per verificare (3.4) possiamo utilizzare le stime

$$|(d_k, \nabla f(x_k))| \geq \frac{1}{\Lambda(x_k)} \|\nabla f(x_k)\|^2, \tag{3.51}$$

$$\|d_k\| \leq \|H(x_k)^{-1}\| \|\nabla f(x_k)\| \leq \frac{1}{\lambda(x_k)} \|\nabla f(x_k)\| \tag{3.52}$$

dove la norma è quella euclidea, e dove $\lambda(x_k)$ e $\Lambda(x_k)$ sono rispettivamente minimo e massimo autovalore di $H(x_k)$, e quindi i loro reciproci sono rispettivamente massimo e minimo autovalore di $H(x_k)^{-1}$. Utilizzando la positività di $H(x_k)^{-1}$, (3.51) deriva da (3.47), mentre (3.52)

deriva dalla definizione della direzione d_k, ricordando che la norma euclidea di una matrice simmetrica è il suo raggio spettrale. Si arriva quindi alla minorazione

$$\frac{|(d_k, \nabla f(x_k))|}{\|d_k\|\|\nabla f(x_k)\|} \geq \frac{\frac{1}{\Lambda(x_k)}\|\nabla f(x_k)\|^2}{\frac{1}{\lambda(x_k)}\|\nabla f(x_k)\|^2} = \frac{\lambda(x_k)}{\Lambda(x_k)},$$

che, tenendo conto dei segni e della ipotesi (3.50), può essere riscritta

$$\frac{(d_k, \nabla f(x_k))}{\|d_k\|\|\nabla f(x_k)\|} \leq -\frac{\lambda(x_k)}{\Lambda(x_k)} \leq -c.$$

Di conseguenza, la definizione (3.48) di d_k soddisfa (3.4) e si possono applicare i teoremi 3.1 e 3.2 ottenendo la convergenza di tutta la successione x_k. La stessa conclusione si ottiene per la definizione (3.49), poiché chiaramente

$$\frac{(d_k, \nabla f(x_k))}{\|d_k\|\|\nabla f(x_k)\|} \leq -\frac{\lambda(x_0)}{\Lambda(x_0)},$$

che soddisfa ancora la (3.4).

■

3.2.5 Metodi Quasi–Newton

Nei metodi Quasi–Newton si cerca di ovviare alla elevata complessità di una singola iterazione, tipica del metodo di Newton propriamente detto, costruendo direzioni di ricerca della forma

$$d_k = -H_k \nabla f(x_k) \tag{3.53}$$

in cui le matrici H_k sono costruite con l'intento di approssimare (in un senso da specificare) il comportamento di $H(x_k)^{-1}$, ma anche in modo da avere una complessità quadratica, e non cubica, per ogni iterazione. Per motivi di robustezza, è possibile anche nel caso dei metodi Quasi–Newton una reinizializzazione, definendo ad esempio le direzioni di ricerca come

$$\begin{aligned} d_k &= -\nabla f(x_k) \qquad (k = 0, n, 2n, 3n, \ldots), \\ d_k &= -H_k \nabla f(x_k) \qquad (k \neq 0, n, 2n, 3n, \ldots), \end{aligned} \tag{3.54}$$

ovvero riassegnando $H_k = I$ ogni n iterazioni.

Sulle matrici H_k si richiede tipicamente che siano simmetriche e definite positive, in modo che le d_k siano direzioni di discesa:

$$(d_k, \nabla f(x_k)) = -(H_k \nabla f(x_k), \nabla f(x_k)) < 0.$$

Inoltre, poiché sviluppando al primo ordine il gradiente nel punto x_k si ha

$$\nabla f(x_{k-1}) \approx \nabla f(x_k) + H(x_k)(x_{k-1} - x_k),$$

ovvero

$$H(x_k)^{-1}(\nabla f(x_k) - \nabla f(x_{k-1})) \approx x_k - x_{k-1},$$

si richiede che le matrici H_k soddisfino (stavolta in modo esatto) una relazione analoga, detta *relazione Quasi–Newton*:

$$H_k(\nabla f(x_k) - \nabla f(x_{k-1})) = x_k - x_{k-1}. \tag{3.55}$$

La (3.55) ammette in generale infiniti modi di costruire le matrici H_k (meno che nel caso unidimensionale, nel quale si ottiene come unica possibilità il metodo delle secanti). Ad ogni passo, la matrice H_k viene calcolata a partire dall'incremento dei gradienti tra le ultime due iterazioni, $y_k = \nabla f(x_k) - \nabla f(x_{k-1})$, e da quello delle posizioni $s_k = x_k - x_{k-1}$. Indicando per brevità $\bar{H} = H_k$, $y = y_k$, $s = s_k$, si arriva ad esprimere la condizione sul generico aggiornamento H_k come

$$\bar{H}y = s. \tag{3.56}$$

Posto ulteriormente $H = H_{k-1}$, si scrive l'aggiornamento nella forma $\bar{H} = H + \Delta H$. Nelle formule di più note, l'aggiornamento ΔH è una matrice di rango 1 (formula di Broyden) o di rango 2 (formule di Davidon–Fletcher–Powell e Broyden–Fletcher–Goldfarb–Shanno).

Formula di aggiornamento di rango 1 In questa formula, si pone

$$\Delta H = auu^t$$

e si determinano la costante a ed il vettore incognito u in modo che (3.56) sia soddisfatta, ovvero che

$$Hy + auu^t y = s.$$

In pratica, se si sceglie $u = s - Hy$, si ottiene

$$a(s - Hy)(s - Hy)^t y = s - Hy.$$

L'uguaglianza richiede che $a = [(s - Hy)^t y]^{-1}$, da cui

$$\Delta H = \frac{(s - Hy)(s - Hy)^t}{(s - Hy)^t y}$$

Il fatto che la formula possa risultare non ben definita (non ci sono condizioni generali per garantire che $a \neq 0$, né in generale è vero che le matrici H_k siano sempre positive), fa sì che questa formula sia poco utiizzata, e si preferiscano invece gli aggiornamenti di rango 2.

Formula di Davidon–Fletcher–Powell (DFP) Nella formula DFP, l'aggiornamento ΔH è una matrice di rango 2, e analogamente al caso di rango 1 si può scrivere come

$$\Delta H = auu^t + bvv^t.$$

Imponendo la (3.56), si ha

$$auu^t y + bvv^t y = s - Hy.$$

Scegliendo $u = s$ e $v = Hy$, si ottiene ancora

$$ass^t y + bHyy^t Hy = s - Hy$$

che può essere soddisfatta ponendo $a = (s^t y)^{-1}$, $b = -(y^t Hy)^{-1}$. La formula di aggiornamento che si ottiene è quindi

$$\Delta H = \frac{ss^t}{s^t y} + \frac{Hyy^t H}{y^t Hy}. \tag{3.57}$$

Si può dimostrare che se H_0 è simmetrica definita positiva e i passi β_k sono ottenuti per ricerca esatta (o parziale, nella forma di Wolfe–Powell), tutte le matrici H_k sono simmetriche e positive.

Formula di Broyden–Fletcher–Goldfarb–Shanno (BFGS) Questa formula di aggiornamento, pur restando di rango 2, si ottiene da un procedimento di simmetrizzazione concettualmente più complesso ed assume la forma

$$\Delta H = \frac{(s - Hy)s^t + s(s - Hy)^t}{s^t y} - \frac{y^t(s - Hy)}{|s^t y|^2} ss^t. \tag{3.58}$$

Si può dimostrare che questa formula di aggiornamento fornisce matrici H_k simmetriche e definite positive sotto le stesse ipotesi della formula DFP.

Complessità Tutte le operazioni che appaiono in una iterazione dei metodi QN sono di complessità quadratica, sia la costruzione dell'aggiornamento (basata essenzialmente su prodotti colonne–righe di vettori), sia il calcolo della direzione (basato sul prodotto matrice–vettore in (3.53)). Il costo computazionale di una iterazione è quindi quadratico, ma a differenza del metodo di Newton approssimato (3.49), i metodi QN hanno convergenza sopralineare invece che lineare. Resta invece il problema della occupazione di memoria, che di fatto non pemette ai metodi QN di essere applicati in dimensione grande (dalle migliaia di variabili in su).

Risultati fondamentali

- Convergenza dei metodi Quasi–Newton in ricerca esatta

  Teorema 3.14 *Se $f(x) \in C^3$, l'autovalore minimo $\lambda(x)$ della matrice hessiana H soddisfa la limitazione inferiore $\lambda(x) \geq \alpha$ e H_0 è definita positiva, allora la successione x_k definita da (3.2), con d_k definito da (3.53), H_k tramite (3.57) o (3.58) e β_k tramite una ricerca esatta, converge all'unico punto stazionario x^* per ogni $x_0 \in \mathbb{R}^n$ con velocità sopralineare.*

- Convergenza dei metodi Quasi–Newton a passo fisso

  Teorema 3.15 *Se $f(x) \in C^3$ e la matrice hessiana H è strettamente definita positiva in un punto stazionario x^*, allora esistono $\varepsilon, \delta > 0$ tali che, se $\|x_0 - x^*\| < \varepsilon$ e $\|H_0 - H(x^*)^{-1}\| < \delta$, la successione x_k definita da (3.2), con d_k definito da (3.53), H_k tramite (3.57) o (3.58) e $\beta_k \equiv 1$ è ben definita e converge a x^* (almeno) linearmente.*

- Convergenza dei metodi Quasi–Newton con reinizializzazione

  Teorema 3.16 *Sia $f(x) \in C^2$, strettamente convessa sull'insieme (che si suppone compatto) Σ_0 definito dalla (3.3), e la successione x_k sia generata tramite l'algoritmo (3.2), con d_k definito da (3.54) ed H_k tramite (3.57) o (3.58).*

  Allora, se i passi β_k sono determinati tramite una delle condizioni:

  i) ricerca esatta

  ii) ricerca parziale di Armijo–Goldstein o Wolfe–Powell

iii) β_k *fissato, sufficientemente piccolo*

la successione x_k *converge all'unico punto* x^* *di minimo per* f*.*

Dim. Basta osservare che sono soddisfatte le ipotesi dei teoremi 3.1 e 3.2, con $\mathcal{I} \equiv \{0, n, 2n, 3n, \ldots\}$.

■

3.3 Confronto fra i vari schemi

La maggior complessità computazionale e la maggior occupazione di memoria dei metodi di tipo Newton fa si' che il loro uso sia limitato a problemi in dimensione non molto alta (tipicamente n dell'ordine di qualche decina nel metodo di Newton propriamente detto, tra le centinaia e poche migliaia nel caso dei metodi di Newton approssimati), e prevalentemente per problemi pieni. In dimensione più alta, ed in particolare per problemi sparsi, questioni di occupazione di memoria rendono necessario utilizzare metodi di tipo rilassamento, gradiente o gradiente coniugato (con un ovvio vantaggio in quest'ultimo caso). Si noti che la complessità computazionale riportata è quella per singola iterazione.

schema	compless. (pr. pieni)	compless. (pr. sparsi)	occupaz. (pr. pieni)	occupaz. (pr. sparsi)	ordine conv.
rilass.	$O(n^2)$	$O(n)$	$O(n^2)$	$O(n)$	$\gamma = 1$
grad.	$O(n^2)$	$O(n)$	$O(n^2)$	$O(n)$	$\gamma = 1$
CD	$O(n^2)$	$O(n)$	$O(n^2)$	$O(n)$	$1 < \gamma < 2$ (*)
Newt.	$O(n^3)$	$O(n^3)$	$O(n^2)$	$O(n^2)$	$\gamma = 2$
Newt. approx.	$O(n^2)$	$O(n^2)$	$O(n^2)$	$O(n^2)$	$\gamma = 1$
QN	$O(n^2)$	$O(n^2)$	$O(n^2)$	$O(n^2)$	$1 < \gamma < 2$

(*) in caso di funzioni quadratiche, convergenza in n passi.

4 Problemi di minimizzazione vincolata

In questa sezione tratteremo schemi per la soluzione del problema

$$f(x^*) = \min_{x \in S} f(x) \tag{4.1}$$

con $f \in C^1$, o $f \in C^2$ a seconda del metodo utilizzato, ed S usualmente definito tramite vincoli di uguaglianza e/o disuguaglianza. Poiché ogni vincolo di uguaglianza si può vedere come una coppia di vincoli di disuguaglianza, possiamo scrivere senza perdita di generalità S nella forma

$$S = \{x \in \mathbb{R}^n : g_i(x) \leq 0 \quad (i = 1, \ldots, m)\}. \tag{4.2}$$

Ricordiamo che per assicurare la esistenza ed unicità della soluzione, si deve supporre f strettamente convessa ed S non vuoto, convesso e compatto (tipicamente, questo porta a richiedere che anche le funzioni g_i siano convesse). Sotto la ulteriore ipotesi di differenziabilità per f e per i vincoli, il punto di minimo vincolato si può caratterizzare tramite opportune estensioni della condizione di stazionarietà (moltiplicatori di Lagrange nel caso dei vincoli di uguaglianza, moltiplicatori di Kuhn–Tucker o formulazione di punto sella per il caso dei vincoli di disuguaglianza).

Le classi principali di metodi di minimizzazione vincolata sono due:

Metodi primali – In questi metodi (generalmente ottenuti adattando al caso vincolato un metodo iterativo per problemi liberi) si impone il vincolo passo per passo ottenendo quindi una successione di approssimazioni che soddisfano tutte i vincoli. Si tratta di metodi usualmente di costruzione complessa e di convergenza lenta, esclusi alcuni casi più semplici da trattare.

Metodi duali – In questi metodi quello che si risolve non è il problema originale ma un problema ausiliario senza vincoli. Si tratta di schemi di costruzione più semplice anche se non sempre intuitiva, ed in cui non è più garantito che ad ogni passo la approssimazione soddisfi i vincoli (ciò, infatti, accade solo asintoticamente).

4.1 Metodi primali

Di questa classe di schemi verranno dati i due esempi più noti: i metodi del gradiente proiettato e del rilassamento proiettato.

4.1.1 Metodo del gradiente proiettato

Questo metodo è ottenuto dal metodo di massima discesa, lavorando a passo fisso (o meglio, limitato dall'alto) ed effettuando ad ogni passo una proiezione del punto ottenuto dallo schema "libero" all'interno di S. Più formalmente, assegnato x_0 e definito ad ogni passo il punto

$$v_k = x_k - \nabla f(x_k), \tag{4.3}$$

la approssimazione successiva x_{k+1} è data da

$$x_{k+1} = x_k + \beta_k \left[P_S(v_k) - x_k\right] \tag{4.4}$$

dove P_S è la proiezione sull'insieme S, e β_k è ottenuto da una ricerca esatta per $\beta \in [0, 1]$, ovvero

$$f\left(x_k + \beta_k \left[P_S(v_k) - x_k\right]\right) = \min_{\beta \in [0,1]} f\left(x_k + \beta \left[P_S(v_k) - x_k\right]\right). \tag{4.5}$$

L'ipotesi di convessità di S, insieme con le (4.4), (4.5), garantiscono che ad ogni passo $x_k \in S$.

Complessità Il metodo del gradiente proiettato e le sue derivazioni hanno di regola convergenza molto lenta, e sono molto costosi soprattutto nella fase cruciale della proiezione. In generale, anzi, la proiezione di un punto su di un insieme si presenta essa stessa come un problema di minimizzazione vincolata, che può essere risolto in modo semplice solo in un piccolo numero di casi (tipicamente, per insiemi definiti da vincoli lineari). Per tutti questi motivi al momento i metodi primali non vengono reputati competitivi rispetto ai metodi duali più raffinati.

Risultati fondamentali

- Convergenza del metodo del gradiente proiettato

  Teorema 4.1 *Se $f(x) \in C^2(S)$ è una funzione strettamente convessa, S è convesso e chiuso, l'insieme di sottolivello*

  $$S_0 = \{x \in S : f(x) \leq f(x_0)\}$$

  è limitato e la successione x_k è generata tramite l'algoritmo del gradiente proiettato definito da (4.3)–(4.5), allora x_k converge all'unico punto di minimo vincolato x^* di f.

4.1.2 Metodo del rilassamento proiettato

Questo schema è applicabile a problemi con vincoli di tipo n–intervallo, ovvero

$$S = \{x \in \mathbb{R}^n : a_i \leq x \leq b_i \quad (i = 1, \ldots, n)\} \tag{4.6}$$

in cui gli estremi a_i, b_i possono essere finiti o no.

La formulazione del metodo di rilassamento in $\mathbb{R}^n$, viene modificata in questo caso nella forma

$$x_j^{(k+1)} = \underset{a_j \leq t \leq b_j}{\operatorname{argmin}} \quad f(x_1^{(k+1)}, \ldots, x_{j-1}^{(k+1)}, t, x_{j+1}^{(k)}, \ldots, x_n^{(k)}) \tag{4.7}$$

(per j *in sequenza* da 1 a n, $k \geq 1$). La minimizzazione rispetto a t in (4.7) equivale, se la funzione f è convessa, ad una proiezione del minimo unidimensionale libero all'interno dell'intervallo $[a_j, b_j]$ (da cui il nome dell'algoritmo).

Anche nel caso vincolato, si può introdurre un *parametro di rilassamento* ω ponendo lo schema in forma sovrarilassata:

$$x_j^{(k+1)} = P_{[a_j,b_j]} \left[(1-\omega) x_j^{(k)} + \omega \operatorname*{argmin}_t \; f(x_1^{(k+1)}, \ldots, x_{j-1}^{(k+1)}, t, x_{j+1}^{(k)}, \ldots, x_n^{(k)}) \right] \tag{4.8}$$

che corrisponde a proiettare il valore ottenuto dallo schema SOR senza vincoli all'interno dell'intervallo $[a_j, b_j]$.

Complessità Contrariamente a quanto accade per il metodo del gradiente proiettato, vincoli del tipo (4.6) possono essere trattati in modo molto naturale con il metodo del rilassamento. In questo caso la operazione di proiezione non aumenta in modo rilevante la complessità e la perdita di efficienza dello schema risulta legata più che altro al rallentamento della convergenza.

Risultati fondamentali

- Convergenza del metodo di rilassamento proiettato

  Teorema 4.2 *Si supponga $f \in C^1(\mathbb{R}^n)$, strettamente convessa. Allora la successione $x^{(k)}$ definita da (4.7) converge all'unico punto di minimo di f su S.*

4.2 Metodi duali

Anche per questa classe di schemi verranno dati i due esempi più classici: il metodo di penalizzazione e quello di Uzawa. Nel primo caso si utilizza una funzione ausiliaria che rende "poco conveniente" la violazione dei vincoli, nel secondo caso si utilizza la formulazione duale propriamente detta del problema, basata sulla caratterizzazione dei minimi vincolati come punti di sella per la Lagrangiana (vedi Appendice A.5)

4.2.1 Metodo di penalizzazione

Definiamo una funzione (di penalizzazione) $H(x)$ continua (in realtà, per le ragioni che si vedranno, si impone che sia almeno derivabile con continuità) e tale che:

$$H(x) = 0 \quad (x \in S) \quad , \qquad H(x) > 0 \quad (x \notin S). \tag{4.9}$$

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

$$f_\varepsilon(x) = f(x) + \frac{1}{\varepsilon} H(x). \tag{4.10}$$

Si intuisce che per $\varepsilon \to 0$ il comportamento delle soluzioni (non vincolate) del problema penalizzato si avvicini a quello delle soluzioni di (4.1), (4.2). Tale risultato sarà precisato nel paragrafo relativo ai risultati fondamentali.

Una scelta abbastanza naturale per la funzione di penalizzazione H, che ne garantisce la differenziabilità e, nel caso le g_i siano convesse, anche la convessità, è:

$$H(x) = \sum_{i=1}^{m} [g_i(x)^+]^2. \tag{4.11}$$

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

$$H(x) = \sum_{i=1}^{m} g_i(x)^2 \tag{4.12}$$

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

Complessità L'efficienza del metodo di penalizzazione è legata alla scelta del parametro di penalizzazione ε, che se troppo piccolo può introdurre un forte malcondizionamento nel problema. Per vedere ciò, supponiamo di dover risolvere per penalizzazione il problema in $\mathbb{R}^2$:

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

con il vincolo di uguaglianza

$$x_1 = 1.$$

La funzione penalizzata f_ε, con il termine di penalizzazione (4.12), 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 è

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

da cui si vede che valori piccoli di ε causano un aumento incontrollato del numero di condizionamento. D'altra parte, la soluzione del problema penalizzato è il punto $x_\varepsilon = \left(1 - \frac{\varepsilon}{2+\varepsilon}, 0\right)$, che ha quindi distanza $O(h)$ dalla soluzione esatta $x^* = (1, 0)$ (il fatto che per termini di penalizzazione del tipo (4.11) o (4.12) l'errore abbia questo ordine può essere dimostrato in generale).

In queste condizioni, il miglioramento di accuratezza che proviene dalla penalizzazione più precisa può essere completamente annullato dal rallentamento di convergenza nel metodo di minimizzazione.

Risultati fondamentali

- Convergenza del metodo di penalizzazione

  Teorema 4.3 *Indicato con x_ε un punto di minimo in $\mathbb{R}^n$ per (4.10), con $H \in C^0(\mathbb{R}^n)$ che soddisfa (4.9), allora le sottosuccessioni convergenti di x_ε convergono, per $\varepsilon \to 0$, a soluzioni di (4.1), (4.2). In particolare, se f 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

  $$f(x_{\varepsilon_k}) \leq f_{\varepsilon_k}(x_{\varepsilon_k}) \leq \min_S f_{\varepsilon_k}(x) = f(x^*). \tag{4.13}$$

La prima disuguaglianza in (4.13) discende dalla definizione di f_ε, mentre la seconda dal fatto che x_{ε_k} è il minimo di f_{ε_k} su tutto $\mathbb{R}^n$. L'uguaglianza tra gli ultimi due membri di (4.13) 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 è quindi dimostrato, una volta che si verifichi che $\bar{x} \in S$. Da secondo e quarto membro di (4.13) si ha

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

e passando al limite si vede che il primo membro, che è nonnegativo, è anche minore o uguale di $f(x^*) - f(\bar{x})$. Ma perché il primo membro resti limitato per $k \to \infty$ è necessario che $H(\bar{x}) = 0$, ovvero che $\bar{x} \in S$.

■

4.2.2 Metodo di Uzawa

Il metodo di Uzawa è un metodo iterativo in $\mathbb{R}^{n+m}$ per cercare il punto sella per la funzione Lagrangiana di un problema di minimizzazione vincolata (vedi Appendice A.5).

Fissato $\lambda_0 \in \mathbb{R}^m_+$, si aggiornano alternativamente i valori di x_k e λ_{k+1} tramite l'iterazione

$$x_k = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \; L(x, \lambda_k), \tag{4.14}$$

$$\lambda_{k+1} = P_{\mathbb{R}^m_+}\Big[\lambda_k + \beta g(x_k)\Big], \tag{4.15}$$

dove $g(x) = (g_1(x) \cdots g_m(x))^t$ si può interpretare come il gradiente della Lagrangiana rispetto alle variabili λ_i, e $P_{\mathbb{R}^m_+}$ è la proiezione nel cono positivo di $\mathbb{R}^m$. Si tratta quindi di alternare una minimizzazione della Lagrangiana rispetto ad x con una iterazione di gradiente a passo fisso (in questo caso, cercando il massimo) fatto rispetto al vettore λ.

Sono possibili anche varianti in cui entrambi gli aggiornamenti si effettuano allo stesso modo, sia tutti e due tramite ricerca esatta, sia tutti e due a passo fisso o predeterminato.

Complessità Il condizionamento della matrice Hessiana può essere peggiore per la funzione Lagrangiana di quanto non sia per la funzione f, tuttavia non peggiora all'aumentare dell'accuratezza come accade per il metodo di

penalizzazione. In generale le implementazioni attuali del metodo di Uzawa utilizzano metodi a convergenza sopralineare o quadratica sia per la fase di minimizzazione che per quella di massimizzazione. Nelle versioni *inesatte* del metodo non si effettua la minimizzazione esatta rispetto ad x (cosa che richiede due cicli di iterazione nidificati).

Il risultato è una classe di metodi piuttosto efficienti, sicuramente i metodi preferiti al momento in una larga parte dei problemi di ottimizzazione vincolata.

Risultati fondamentali

- Convergenza del metodo di Uzawa

  Teorema 4.4 *Siano $f, g_1, \ldots, g_m \in C^1$ e strettamente convesse. Allora, esiste un valore β_M tale che, per $\beta < \beta_M$ il metodo (4.14)–(4.15) fornisce una successione x_k convergente alla soluzione x^* per ogni scelta di $\lambda_0 \in \mathbb{R}^m_+$. Se inoltre la matrice Jacobiana $J_g(x)$ ha rango pieno, anche la successione λ_k converge ad un λ^* tale che la coppia (x^*, λ^*) è punto di sella per la funzione Lagrangiana.*

4.3 Confronto fra i vari schemi

Allo stato attuale, i metodi primali non vengono reputati competitivi con quelli duali nella soluzione di problemi di minimo vincolato, se non in qualche caso per il metodo di rilassamento proiettato, anche in virtù della sua semplicità. Ancora la semplicità può far scegliere il metodo di penalizzazione, per il resto reputato non troppo accurato e critico dal punto di vista del condizionamento. I metodi (duali) attualmente di uso più diffuso sono quelli basati su generalizzazioni del metodo di Uzawa, in particolare quelle a convergenza sopralineare.

5 Approssimazione di funzioni di una variabile

Data la funzione $f : \mathbb{R} \to \mathbb{R}$, ci si pone il problema di approssimarla in forma di combinazione lineare di funzioni più semplici, cioè nella forma

$$f(x) = c_0\phi_0(x) + c_1\phi_1(x) + \cdots + c_n\phi_n(x) + E_n(x) \tag{5.1}$$

in cui la famiglia $\{\phi_k\}$ si suppone densa in uno spazio opportuno X in cui si approssima f, e si vuole che $\|E_n\|_X \to 0$. Le famiglie tipiche di funzioni utili per l'approssimazione sono:

Polinomiali o polinomiali a tratti – Dal teorema di densità di Weierstass si sa che i polinomi sono densi negli spazi C^k, e di qui si ottiene la densità negli spazi L^p, con p finito.

Polinomi trigonometrici, esponenziali complessi – Hanno proprietà di densità simili a quelle dei polinomi.

5.1 Approssimazioni polinomiali

I criteri di approssimazione più usuali nelle approssimazioni polinomiali sono:

Derivate assegnate in un punto – Corrisponde alla Formula di Taylor.

Interpolazione – Consiste nell'imporre che il valore della combinazione lineare $\sum_i c_i\phi_i(x)$ in (5.1) coincida con quello della funzione in un certo numero di nodi assegnati $x_0, \ldots, x_m$. Perché questo problema abbia soluzione unica si richiede che $m = n$ e che l'insieme dei nodi sia *unisolvente*).

Interpolazione di Hermite – Include come casi particolari entrambe le strategie precedenti: si impone cioè che sia il valore della combinazione lineare (5.1), che quello di un certo numero di sue derivate coincidano con i corrispondenti valori della funzione nei nodi assegnati.

Errore quadratico minimo – Strategia utilizzata in genere per approsimare molti punti con un modello relativamente più semplice (ad esempio, un polinomio di primo grado); la determinazione dei parametri viene effettuata minimizzando un indice (in genere, quadratico) di errore.

Errore di norma minima – Consiste nello scegliere l'approssimazione la cui norma di errore sia minima. In genere non fornisce ricette troppo esplicite, meno che nel caso della norma L^2, in cui dà luogo alle serie di Fourier troncate.

Serie di Fourier troncate – Consistono nell'approssimare f con i primi n termini di uno sviluppo in serie di funzioni ortogonali (in genere, rispetto al prodotto scalare in uno spazio L^2 o L^2_w con peso). Fornisce la migliore approssimazione nella norma associata al prodotto scalare scelto. Permette di operare sia mediante polinomi ortogonali che funzioni trigonometriche.

Esempio E' noto che la formula di Taylor di centro x_0, tra tutti i polinomi di grado assegnato, approssima una funzione regolare con il massimo ordine possibile per $x \to x_0$. Quando però si tratta di approssimare la funzione *in tutto un intervallo*, questa scelta non è in generale la migliore. Dovendo ad esempio approssimare con un polinomio di primo grado la funzione $f(x) = \sin x$ nell'intervallo $[-\pi/2, \pi/2]$, a formula di Taylor di primo grado, centrata nel punto $x_0 = 0$, è data da

$$T_1(x) = x$$

ed il massimo errore di approssimazione è ottenuto negli estremi $\pm\pi/2$ in cui

$$|T_1(\pm\pi/2) - \sin(\pm\pi/2)| = \pi/2 - 1 \approx 0.5708.$$

Un altro modo di approssimare la funzione con un polinomio di primo grado è costruire la retta passante per i due punti estremi $(-\pi/2, -1)$ e $(\pi/2, 1)$. In questo caso, come si verifica facilmente, il polinomio che si ottiene è

$$\Pi_1(x) = \frac{2}{\pi}x$$

ed il massimo errore si ottiene quando

$$\frac{d}{dx}\sin x = \frac{2}{\pi}$$

ovvero per $x = \arccos(2/\pi)$, punti a cui corrisponde l'errore di approssimazione

$$|\Pi_1(\arccos 2/\pi) - \sin(\arccos 2/\pi)| \approx 0.2105.$$

5.1.1 Formula di Taylor

La costruzione e le principali proprietà del polinomio di Taylor,

$$T_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \tag{5.2}$$

sono note dai corsi precedenti. In particolare, la formula di Lagrange dell'errore

$$R_n(x) = \frac{f^{(k+1)}(\xi)}{(k+1)!}(x - x_0)^{k+1} \tag{5.3}$$

mostra che il polinomio di Taylor è una buona approssimazione di f solo in un (piccolo) intorno di x_0. La convergenza di $T_n(x)$ verso $f(x)$ al crescere del grado n avviene solo sotto l'ipotesi di analiticità di f.

5.1.2 Interpolazione

Data una funzione $f(x)$ e $n+1$ punti (detti *nodi* dell'interpolazione) $x_0, \ldots, x_n$, intendiamo trovare un polinomio (interpolatore) $\Pi_n(x)$ di grado al più n tale che:

$$\Pi_n(x_i) = f(x_i). \quad (i = 0, \ldots, n) \tag{5.4}$$

Tale polinomio esiste ed è unico. Se ne possono dare diverse forme, le due principali sono la forma di *Lagrange* e quella di *Newton*.

Polinomio di Lagrange Nella forma di Lagrange, il polinomio $\Pi_n(x)$ si esprime come:

$$\Pi_n(x) = \sum_{i=0}^{n} f(x_i) L_i(x) \tag{5.5}$$

dove

$$L_i(x) = \begin{cases} 1 & \text{se } i = n = 0 \\ \displaystyle\prod_{j \neq i} \frac{x - x_j}{x_i - x_j} & \text{se } n > 0 \end{cases} \tag{5.6}$$

(si ha che $L_i(x_k) = \delta_{ik}$ e che, poiché una costante è sempre interpolata esattamente, $\sum_i L_i(x) \equiv 1$).

Polinomio di Newton Nella forma di Newton, il polinomio interpolatore si scrive:

$$\Pi_n(x) = f[x_0] + f[x_0, x_1](x - x_0) + \cdots + f[x_0, \ldots, x_n](x - x_0) \cdots (x - x_{n-1}). \tag{5.7}$$

dove le costanti $f[\cdots]$ (dette *differenze divise* della funzione f) sono definite ricorsivamente nel modo seguente:

$$f[x_0] = f(x_0), \tag{5.8}$$

$$f[x_0, \ldots, x_k] = \frac{f[x_1, \ldots, x_k] - f[x_0, \ldots, x_{k-1}]}{x_k - x_0}. \tag{5.9}$$

Strategie di infittimento dei nodi Nei problemi di interpolazione, la convergenza del polinomio interpolatore a f dipende dalla regolarità della funzione e dalla legge con cui si infittiscono i nodi. Ad esempio, come si vedrà nei risultati generali, mantenendo i nodi equidistanti si ottiene convergenza solo per una classe di funzioni molto ristretta.

Per ottenere precisioni crescenti nelle approssimazioni per interpolazione, si usano in pratica altre strategie:

Nodi Chebyshev – In questa strategia, si manda effettivamente $n \to \infty$, ma i nodi sono posizionati negli zeri di un polinomio di Chebyshev di grado $n+1$, cioè il polinomio ω_n coincide a meno di costanti moltiplicative con l'$(n+2)$–esimo polinomio di Chebyshev. Nel caso dei nodi di *Chebyshev–Gauss–Lobatto* si ha

$$x_j = a + \frac{b-a}{2}\left(1 - \cos\frac{j\pi}{n}\right),$$

ed in questo caso i nodi estremi sono $x_0 = a$ ed $x_n = b$. Nel caso dei nodi di *Chebyshev–Gauss* si pone

$$x_j = a + \frac{b-a}{2}\left(1 - \cos\frac{(2j+1)\pi}{2n+2}\right),$$

ed i nodi sono tutti interni ad (a, b) (si dimostra che tra tutte le scelte degli $n+1$ nodi di interpolazione, quest'ultima è quella che porta al minimo valore del $\sup_{[a,b]} |\omega_n(x)|$, nel caso in cui gli estremi non siano vincolati ad essere nodi di interpolazione). Entrambe queste famiglie di nodi si addensano alle estremità dell'intervallo.

Interpolazioni composite – Suddividendo l'intervallo di interpolazione $[a, b]$ in sottointervalli $[a_j, b_j]$ (di ampiezza H_j tale che $H = \max_j H_j \to 0$) contenenti ognuno un numero fissato $n+1$ di nodi di interpolazione, si ottiene una interpolazione polinomiale di grado n a tratti. In ogni sottointervallo i nodi sono generalmente (ma non necessariamente) equidistanti.

Se i nodi si scelgono in modo che ce ne siano $n-1$ interni, più due agli estremi in comune con gli intervalli adiacenti, questi nodi estremi garantiscono la continuità della funzione approssimante. Se al contrario i nodi sono tutti interni, si hanno caratteristiche di approssimazione analoghe, ma in generale la funzione approssimante non sarà più continua all'interfaccia tra due sottointervalli.

Stabilità La interpolazione di ordine alto può essere una operazione estremamente sensibile alle perturbazioni. Supponendo che i valori $f(x_i)$ siano affetti da perturbazioni δ_i tali che $|\delta_i| \leq \delta$, ed indicando con Δ la perturbazione risultante sul polinomio interpolatore Π_n, si ha

$$\Pi_n(x) + \Delta = \sum_{i=0}^{n} [f(x_i) + \delta_i] L_i(x) = \Pi_n(x) + \sum_{i=0}^{n} \delta_i L_i(x)$$

da cui

$$|\Delta| \leq \delta \sum_{i=0}^{n} |L_i(x)|. \tag{5.10}$$

La propagazione delle perturbazioni è quindi legata alla cosiddetta *funzione di Lebesgue* $\sum_i |L_i(x)|$ che può assumere valori molto grandi, a meno di utilizzare un grado di interpolazione basso (caso delle interpolazioni composite) o di infittire i nodi in modo opportuno (come nel caso dei nodi di Chebyshev).

Un altro possibile fattore di instabilità, nella forma di Newton del polinomio interpolatore, è la costruzione della tavola delle differenze. Il calcolo delle differenze divise si basa infatti sulla sottrazione di valori molto vicini tra loro (e tanto più vicini con l'infittirsi dei nodi), con conseguente perdita di cifre significative.

Complessità Per calcolare in un punto dato x il polinomio di Lagrange relativo ad un insieme fissato di $n+1$ nodi dalle (5.5)–(5.6), occorre intanto calcolare i valori delle $n+1$ funzioni di base, ognuna delle quali richiede $2n$ sottrazioni, n divisioni ed n prodotti (quindi $4n$ operazioni per ogni funzione L_i) per un totale di $O(4n^2)$ operazioni in virgola mobile. Il calcolo della combinazione lineare (5.5) richiede poi $n+1$ prodotti ed altrettante somme, ed ha quindi complessità trascurabile.

Nel caso del polinomio di Newton, occorre osservare che è possibile calcolarlo con una formula di tipo Horner, e più precisamente:

$$\Pi_n(x) = f[x_0] + (x - x_0)\Big(f[x_0, x_1] + (x - x_1)\Big(f[x_0, x_1, x_2] + \cdots +$$

$$+ (x - x_{n-2})\Big(f[x_0, \ldots, x_{n-1}] + (x - x_{n-1}) f[x_0, \ldots, x_n]\Big) \cdots \Big)\Big)$$

e di conseguenza il calcolo ha complessità lineare. Il numero di operazioni prevalente è quello relativo alla costruzione della tavola delle differenze, che richiede per ogni differenza divisa di effettuare due sottrazioni ed una divisione, che moltiplicate per $(n-1)+(n-2)+\cdots+2+1 = O(n^2/2)$ differenze divise dà una complessità totale di $O(3n^2/2)$ operazioni in virgola mobile.

La situazione è ancora più favorevole alla forma di Newton nel caso in cui, ad un dato insieme di nodi, se ne aggiunga un altro. In questo caso la forma di Lagrange richiede il ricalcolo completo del polinomio, mentre quella di Newton richiede il calcolo dell'ultima riga della tavola delle differenze, e del valore del polinomio (ed ha quindi complessità lineare).

Per quanto riguarda l'efficienza della interpolazione, ovvero la possibilità di ottenere errori più piccoli a parità di numero di valutazioni della funzione f, interpolazioni di ordine alto possono essere molto efficienti con funzioni regolari, mentre situazioni meno regolari si maneggiano meglio con interpolazioni composite, specialmente se la dimensione dei sottointervalli viene variata in base al valore locale della derivata $f^{(n+1)}$. Dal teorema 5.8 si vede che la situazione di massima efficienza è quella in cui l'errore ha lo stesso ordine di grandezza in tutti i sottointervalli, ovvero quando

$$H_j \sim \frac{1}{\left(\sup_{[a_j,b_j]} |f^{(n+1)}(x)|\right)^{\frac{1}{n+1}}}.$$

Esistono naturalmente anche strategie miste in cui si varia sia l'ampiezza degli intervalli che il grado di interpolazione.

Risultati fondamentali

- Convergenza delle interpolazioni in una base generica

  Teorema 5.1 *Si supponga che* $f, \phi_0, \ldots, \phi_n \in C^0([a,b])$*, e che* $\phi_i(x_j) = \delta_{ij}$*. Denotando con* X_n *lo spazio generato dalla base* $\{\phi_0, \ldots, \phi_n\}$*, con*

  $$\varepsilon_n(f) = \inf_{p_n \in X_n} \sup_{x\in[a,b]} |f(x) - p_n(x)| = \min_{p_n \in X_n} \max_{x\in[a,b]} |f(x) - p_n(x)| =$$

  $$= \max_{x\in[a,b]} |f(x) - p_n^*(x)|$$

  e con

  $$\Lambda_n = \max_{x\in[a,b]} \sum_{i=0}^{n} |\phi_i(x)|,$$

  si ha:

  $$\max_{x\in[a,b]} \left| f(x) - \sum_{i=0}^{n} f(x_i)\phi_i(x) \right| \leq (1 + \Lambda_n)\varepsilon_n(f). \tag{5.11}$$

Dim. Notiamo subito che l'ipotesi $\phi_i(x_j) = \delta_{ij}$ non è restrittiva. Infatti, se i nodi $x_0, \ldots, x_n$ costituiscono un insieme unisolvente, anche se questa ipotesi non fosse soddisfatta, sarebbe sempre possibile trovare una base equivalente (nel senso che genera lo stesso spazio X_n) e tale da soddisfarla. Inoltre, sotto questa ipotesi si ha $c_i = f(x_i)$, ed infatti valutando l'interpolata nel nodo x_k, tutti i termini della sommatoria $\sum_i f(x_i)\phi_i(x_k)$ si annullano ad eccezione del termine k–esimo, e quindi

$$\sum_{i=0}^{n} f(x_i)\phi_i(x_k) = f(x_k)\phi_k(x_k) = f(x_k).$$

Scritto ora p_n^* nella forma

$$p_n^*(x) = p_n^*(x_0)\phi_0(x) + \cdots + p_n^*(x_n)\phi_n(x)$$

si ha, per ogni $x \in [a, b]$:

$$\left| f(x) - \sum_i f(x_i)\phi_i(x) \right| \leq |f(x) - p_n^*(x)| + \left| p_n^*(x) - \sum_i f(x_i)\phi_i(x) \right| =$$

$$= \varepsilon_n(f) + \left| \sum_i \Big(p_n^*(x_i) - f(x_i) \Big) \phi_i(x) \right| \leq$$

$$\leq \varepsilon_n(f) + \sum_i \Big| p_n^*(x_i) - f(x_i) \Big| \, |\phi_i(x)| \leq$$

$$\leq \varepsilon_n(f) + \varepsilon_n(f) \sum_i |\phi_i(x)|.$$

Da qui si ottiene immediatamente la (5.11).

■

- Esistenza ed unicità del polinomio interpolatore

Teorema 5.2 *Se i nodi $x_0, \ldots, x_n$ sono distinti, esiste un unico polinomio Π_n di grado minore o uguale ad n che soddisfi le condizioni (5.4). In particolare, se f è essa stessa un polinomio di grado minore o uguale ad n, allora $\Pi_n \equiv f$.*

Dim. La dimostrazione della esistenza verrà data mediante la costruzione dei polinomi di Lagrange e Newton. Per quanto riguarda l'unicità,

supponiamo per assurdo che esistano due polinomi, Π_n^1 e Π_n^2 di grado non superiore ad n e che soddisfino le condizioni (5.4). Possiamo dunque scrivere

$$\Pi_n^1(x_i) = \Pi_n^2(x_i) = f(x_i). \quad (i = 0, \ldots, n)$$

La differenza $\Pi_n^1 - \Pi_n^2$ è ancora un polinomio di grado non superiore ad n che d'altra parte si annulla negli $n+1$ nodi $x_0, \ldots, x_n$. Per il teorema fondamentale dell'algebra si tratta quindi del polinomio identicamente nullo, e di conseguenza $\Pi_n^1 \equiv \Pi_n^2$. Analogamente nel caso in cui f stessa sia un polinomio.

■

- Costruzione del polinomio di Lagrange

Teorema 5.3 *Se i nodi $x_0, \ldots, x_n$ sono distinti, il polinomio interpolatore Π_n si può scrivere nella forma (5.5)–(5.6).*

Dim. La dimostrazione è ovvia se $n = 0$, possiamo quindi supporre che $n > 0$. Notiamo subito che i polinomi L_i sono di grado n, in quanto il loro numeratore è il prodotto di n termini di primo grado; di conseguenza anche Π_n avrà grado non superiore ad n. Dimostriamo ora che $L_i(x_k) = \delta_{ik}$. Infatti, ricordando la formula (5.6):

$$L_i(x_k) = \prod_{j \neq i} \frac{x_k - x_j}{x_i - x_j}$$

si può notare che se $k \neq i$, nella produttoria a secondo membro comparirà un termine con numeratore $x_k - x_k$ (corrispondente a $j = k$), e di conseguenza $L_i(x_k) = 0$. D'altra parte,

$$L_i(x_i) = \prod_{j \neq i} \frac{x_i - x_j}{x_i - x_j}$$

e quindi $L_i(x_i) = 1$ in quanto tutti i termini della produttoria sono unitari. Valutando ora Π_n nel nodo x_k, tutti i termini della sommatoria (5.5) si annullano ad eccezione del termine k–esimo, e si ha in conclusione

$$\Pi_n(x_k) = f(x_k) L_k(x_k) = f(x_k).$$

■

- Costruzione del polinomio di Newton

Teorema 5.4 *Se i nodi $x_0, \ldots, x_n$ sono distinti, il polinomio interpolatore Π_n si può scrivere nella forma (5.7)–(5.8).*

Dim. Come per il polinomio di Lagrange, anche in questo caso la dimostrazione è ovvia se $n = 0$, e possiamo quindi supporre $n > 0$. Cerchiamo un polinomio Π_n che sia nella forma

$$\Pi_n(x) = \Pi_{0\cdots n}(x) = \Pi_{0\cdots n-1}(x) + g(x). \tag{5.12}$$

In (5.12), $\Pi_{i\cdots j}(x)$ indica il polinomio interpolatore costruito sui nodi $x_i, \ldots, x_j$, mentre $g(x)$ è un polinomio di grado n, che si annulla in tutti i nodi $x_0, \ldots, x_{n-1}$. Si può facilmente verificare che (5.12) è un polinomio di grado n, che coincide con $\Pi_{0\cdots n-1}(x)$ (e quindi con $f(x)$) nei nodi $x_0, \ldots, x_{n-1}$. Il polinomio $g(x)$ deve quindi necessariamente avere la struttura

$$g(x) = f[x_0, \ldots, x_n](x - x_0) \cdots (x - x_{n-1}) \tag{5.13}$$

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 $\Pi_n(x_n) = f(x_n)$. Dimostriamo ora che le costanti $f[\cdots]$ soddisfano le relazioni ricorrenti (5.8)–(5.9).

Per prima cosa, se $k = 0$, allora

$$\Pi_0(x) \equiv f(x_0) = f[x_0]$$

e la (5.8) è soddisfatta. Per $k \geq 1$, verifichiamo ora che sussiste la *formula di Neville*

$$\Pi_k(x) = \frac{(x - x_0)\Pi_{1\cdots k}(x) - (x - x_k)\Pi_{0\cdots k-1}(x)}{x_k - x_0}. \tag{5.14}$$

In effetti, calcolando il secondo membro di (5.14) nei nodi $x_1, \ldots, x_{k-1}$ si ottiene

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

$$= \frac{(x_j - x_0)f(x_j) - (x_j - x_k)f(x_j)}{x_k - x_0} = f(x_j)$$

mentre nel nodo x_0 si ha:

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

ed analogamente nel nodo x_k si ottiene:

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

D'altra parte il secondo membro di (5.14) è 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 (5.7) e (5.14), si ha

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

che coincide con la (5.9).

■

- Rappresentazione dell'errore di interpolazione

Teorema 5.5 *Se* $I = [\min(x, x_0, \ldots, x_n), \max(x, x_0, \ldots, x_n)]$ *e se* $f \in C^{n+1}(I)$*, allora l'errore di approssimazione si può rappresentare come:*

$$f(x) - \Pi_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_n(x) \tag{5.15}$$

dove $\omega_n(x) = (x - x_0)(x - x_1)\cdots(x - x_n)$*,* $\xi \in I$*.*

Dim. Supponiamo x fissato e introducendo una variabile ausiliaria z definiamo la funzione

$$G(z) = f(z) - \Pi_n(z) - Q(x)\omega_n(z)$$

in cui $Q(x)$ è una funzione da determinarsi opportunamente. Più precisamente, se imponiamo la condizione $G(x) = 0$, notiamo che si ha:

$$f(x) - \Pi_n(x) = Q(x)\omega_n(x) \tag{5.16}$$

e questo fornisce la struttura dell'errore di interpolazione in x.

Si osservi che per ogni scelta della $Q(x)$, la funzione G si annulla per $z = x_i$, $(i = 0, \ldots, n)$. Imponendo ora che $G(x) = 0$, si ha che per $z \in I$ la funzione G (che è di classe C^{n+1}) si annulla $n+2$ volte. Di conseguenza, per il teorema di Rolle, G' si annulla $n+1$ volte, G'' si annulla n volte e così via fino alla $G^{(n+1)}$ che si annulla almeno una volta. Possiamo quindi scrivere, per un opportuno punto $\xi \in I$:

$$0 = G^{(n+1)}(\xi) = f^{(n+1)}(\xi) - Q(x)(n+1)! \tag{5.17}$$

in cui si è tenuto conto del fatto che $\Pi_n^{(n+1)} \equiv 0$ e che $\omega_n^{(n+1)} \equiv (n+1)!$. Dalla (5.17) si ottiene

$$Q(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}$$

che, utilizzata in (5.16), dimostra la (5.15) (si noti che la dipendenza da x avviene attraverso ξ).

■

- Convergenza uniforme del polinomio interpolatore (scelta generica dei nodi)

Teorema 5.6 *Se è soddisfatta la condizione*

$$\sup_I |f^{(n+1)}(x)| = o\left(\frac{(n+1)!}{(b-a)^{n+1}}\right) \tag{5.18}$$

allora per $n \to \infty$ si ha $\Pi_n(x) \to f(x)$ uniformemente in I, comunque si scelgano $n+1$ nodi distinti $x_0, \ldots, x_n \in I$.

Dim. Ricordiamo che l'ipotesi (5.18) è una condizione sufficiente di analiticità. In effetti, passando all'estremo superiore in (5.15) si ottiene

$$\sup_I |f(x) - \Pi_n(x)| \leq \frac{\sup_I |f^{(n+1)}(x)|}{(n+1)!} \sup_I |\omega_n(x)| \leq$$

$$\leq \frac{\sup_I |f^{(n+1)}(x)|}{(n+1)!}(b-a)^{n+1} \tag{5.19}$$

in cui in $|\omega_n(x)|$ si è maggiorato ogni termine del tipo $|x - x_i|$ con $(b-a)$, indipendentemente dalla posizione dei nodi. E' immediato verificare che se (5.18) è soddisfatta, la maggiorazione (5.19) converge a zero.

■

- Convergenza uniforme del polinomio interpolatore (nodi di Chebyshev)

  Teorema 5.7 *Se $f \in C^k(I)$ $(k \geq 1)$, il polinomio interpolatore $\Pi_n(x)$ costruito sui nodi di Chebyshev converge a $f(x)$ per $n \to \infty$, e vale la stima $|f(x) - \Pi_n(x)| \leq Cn^{-k}$ per ogni $x \in I$.*

- Convergenza uniforme del polinomio interpolatore (approssimazioni composite)

  Teorema 5.8 *Se $f \in C^{n+1}(I)$, l'interpolazione composita $\Pi_{n,H}(x)$, di grado n a tratti su sottointervalli $[a_j, b_j]$ di ampiezza $H_j \leq H$, converge uniformemente a $f(x)$ e, per ogni $x \in I$, vale la stima*

$$|f(x) - \Pi_{n,H}(x)| \leq \frac{1}{(n+1)!} \sup_j \left(H_j^{n+1} \sup_{[a_j,b_j]} |f^{(n+1)}(x)| \right) \tag{5.20}$$

  ed in particolare:

$$|f(x) - \Pi_{n,H}(x)| \leq \frac{\sup_I |f^{(n+1)}(x)|}{(n+1)!} H^{n+1} \tag{5.21}$$

  Dim. Basta osservare che su ogni sottointervallo $[a_j, b_j]$ della approssimazione composita vale la stima (5.19) con $b_j - a_j = H_j \leq H$, e

$$\sup_{[a_j,b_j]} |f^{(n+1)}(x)| \leq \sup_I |f^{(n+1)}(x)|.$$

  ■

5.1.3 Interpolazione di Hermite

Nella interpolazione di Hermite compaiono entrambe le strategie tipiche della formula di Taylor e della interpolazione propriamente detta: si assegnano cioè nei nodi sia il valore del polinomio $H_m(x)$ che quello di un certo numero di derivate, imponendo le condizioni

$$H_m^{(p)}(x_i) = f^{(p)}(x_i) \qquad (i = 0, \ldots, n; p = 0, \ldots, m_i - 1). \tag{5.22}$$

In genere la necessità di disporre delle derivate successive di f limita abbastanza la applicabilità di questo approccio; in qualche caso, si supplisce con derivate stimate (situazione che non considereremo qui).

La forma generale del polinomio di Hermite di grado m relativo a $n+1$ nodi $x_0, \ldots, x_n \in I$ è data da:

$$H_m(x) = \sum_{i=0}^{n} \sum_{k=0}^{m_i-1} f^{(k)}(x_i) L_{ik}(x) \tag{5.23}$$

dove i polinomi (di grado m) $L_{ik}(x)$ sono caratterizzati dalle condizioni

$$L_{ik}^{(p)}(x_j) = \begin{cases} 1 & \text{se } i = j \text{ e } p = k \\ 0 & \text{altrimenti} \end{cases} \tag{5.24}$$

e si suppone $m + 1 = \sum_i m_i$ (pari cioè alla somma di tutte le condizioni imposte)

E' facile verificare, in modo simile a quanto si è fatto a proposito del polinomio di Lagrange, che dalla definizione delle funzioni di base L_{ik} discende che il polinomio (5.23) soddisfa le condizioni (5.22).

Risultati fondamentali

- Costruzione delle funzioni della base di Hermite

  Teorema 5.9 *I polimino L_{ik} soddisfacenti le condizioni (5.24) possono essere calcolati, per $i = 0, \ldots, n$ nella seguente forma ricorrente:*

$$L_{i,m_i-1}(x) = l_{i,m_i-1}(x)$$

$$L_{ik}(x) = l_{ik}(x) - \sum_{p=k+1}^{m_i-1} l_{ik}^{(p)}(x_i) L_{ip}(x) \qquad (k = m_i - 2, \ldots, 0),$$

  dove

$$l_{ik}(x) = \frac{(x - x_i)^k}{k!} \prod_{j \neq i} \left(\frac{x - x_j}{x_i - x_j} \right)^{m_j} \qquad (k = 0, \ldots, m_i - 1).$$

- Rappresentazione dell'errore

  Teorema 5.10 *Se $I = [\min(x, x_0, \ldots, x_n), \max(x, x_0, \ldots, x_n)]$ e se $f \in C^{m+1}(I)$, allora l'errore di approssimazione si può rappresentare come:*

$$f(x) - H_m(x) = \frac{f^{(m+1)}(\xi)}{(m+1)!} \Omega_m(x) \tag{5.25}$$

  dove $\Omega_m(x) = (x - x_0)^{m_0} (x - x_1)^{m_1} \cdots (x - x_n)^{m_n}$, $\xi \in I$.

5.1.4 Errore quadratico minimo

In questa strategia di approssimazione, si cerca di approssimare un numero m usualmente "grande" di punti con un modello (in genere, ma non necessariamente, polinomiale) relativamente semplice, quale ad esempio una retta. Data la scarsezza di parametri a disposizione, si tratterà di approssimare i dati "al meglio", in qualche senso da specificare, ma rinunciando a soddisfare una condizione di passaggio per tutti i punti. Una situazione tipica in cui ciò si rende necessario è quella in cui i dati siano affetti da deviazioni aleatorie, rendendo quindi inutile il tentativo di interpolarli esattamente.

Ponendo la funzione approssimante nella forma di combinazione lineare [2]

$$\pi(x) = a_1\phi_1(x) + \cdots + a_n\phi_n(x), \tag{5.26}$$

il tentativo di imporre la condizione di passaggio per tutti gli m punti (x_i, y_i) porta al sistema (sovradeterminato se $m > n$)

$$\Phi a = y \tag{5.27}$$

in cui si è posto $a = (a_1 \cdots a_n)^t$, $y = (y_1 \cdots y_m)^t$, e

$$\Phi = \begin{pmatrix} \phi_1(x_1) & \cdots & \phi_n(x_1) \\ \vdots & & \vdots \\ \phi_1(x_m) & \cdots & \phi_n(x_m) \end{pmatrix}.$$

La soluzione di questo sistema nel senso del minimo residuo quadratico porta al sistema (vedi §1.3):

$$\Phi^t\Phi a = \Phi^t y \tag{5.28}$$

detto *sistema delle equazioni normali* (in questo caso l'equivalente della condizione di unisolvenza, che garantisce che il sistema delle equazioni normali sia ben posto, è che la matrice Φ abbia rango massimo). Un'altra maniera di vedere lo stesso procedimento si ha notando che il residuo quadratico del sistema (5.27), che viene minimizzato, è la somma degli scarti quadratici $(\pi(x_i) - y_i)^2$, la cui minimizzazione, in statistica bayesiana, corrisponde al criterio della massima verosimiglianza.

La soluzione del sistema delle equazioni normali si effettua tipicamente per fattorizzazione QR della matrice Φ: infatti se $\Phi = QR$, allora il sistema (5.28) si riscrive come

$$\Phi^t\Phi a = R^tQ^tQRa = R^tRa = \Phi^t y$$

e la sua soluzione richiede di risolvere in sequenza i due sistemi triangolari $R^tz = \Phi^t y$ e $Ra = z$, senza bisogno di costruire esplicitamente la matrice $\Phi^t\Phi$ (né, in effetti, di memorizzare la matrice Q).

[2] per semplificare la notazione matriciale, utilizzeremo qui n funzioni di base invece di $n+1$

5.1.5 Approssimazioni in norma

In questa strategia, si cerca un polinomio π_n di grado n per cui si abbia

$$\|f - \pi_n\| = \min_{p \in \mathbb{P}_n} \|f - p\|$$

ovvero per cui l'errore in una certa norma sia minimo tra tutti i polinomi di grado n. I due casi più studiati sono quelli della norma C^0 e della norma L^2. Il primo caso porta ad una teoria molto tecnica (che non viene esposta qui) che dà una caratterizzazione del polinomio di minimo errore, senza peraltro fornire una ricetta esplicita per la sua costruzione (si dimostra comunque che tale polinomio può essere approssimato molto accuratamente, almeno se la funzione f è abbastanza regolare, dalla serie troncata di Fourier–Chebyshev). Il secondo caso può essere risolto in modo relativamente più esplicito e porta alla approssimazione mediante serie di Fourier troncate (vedi §5.2.1).

5.2 Approssimazioni trigonometriche

Benché esista una teoria della interpolazione con polinomi trigonometrici, la strategia usuale di approssimazioni in questo caso è quella delle serie di Fourier troncate (applicabile, peraltro, anche alle famiglie di polinomi ortogonali).

5.2.1 Serie di Fourier troncate

In questa modalità di approssimazione, si utilizza ancora una combinazione lineare nella forma

$$f(x) = c_0\phi_0(x) + c_1\phi_1(x) + \cdots + c_n\phi_n(x) + E_n(x) \tag{5.29}$$

ma si richiede usualmente che le funzioni ϕ_i siano ortogonali rispetto ad un dato prodotto scalare $(\cdot, \cdot)$, ed in questo caso i coefficienti c_i sono definiti da $c_i = (f, \phi_i)/(\phi_i, \phi_i)$. Nella situazione più comune, il prodotto scalare che si utilizza è quello dello spazio $L^2_w([a, b])$ di funzioni di quadrato sommabile rispetto ad un certo peso $w(x)$, ed in conseguenza si avrà

$$\int_a^b \phi_k(x)\phi_j(x)w(x)dx \begin{cases} > 0 & \text{se } k = j \\ = 0 & \text{se } k \neq j \end{cases} \tag{5.30}$$

$$c_k = \frac{\int_a^b \phi_k(x)f(x)w(x)dx}{\int_a^b \phi_k(x)^2 w(x)dx}. \tag{5.31}$$

In pratica gli integrali che compaiono in (5.30), (5.31) vanno valutati a loro volta in modo approssimato. Le scelte più frequenti sono:

Serie di Fourier trigonometriche – Corrispondono a scegliere $w(x) \equiv 1$ ed un sistema trigonometrico per le ϕ_i. Gli integrali (5.31) possono essere calcolati tramite la trasformata di Fourier discreta, o in alcuni casi anche la trasformata veloce (con minore ordine di complessità).

Serie di Fourier–Legendre – Corrispondono a scegliere $w(x) \equiv 1$, e per le ϕ_i il sistema dei polinomi ortogonali di Legendre. In questo caso gli integrali (5.31) possono essere calcolati in modo efficiente tramite quadrature di tipo gaussiano (vedi cap. 6).

Serie di Fourier–Chebyshev – In questo caso, riportandosi convenzionalmente all'intervallo $[-1, 1]$, si sceglie $w(x) = (1 - x^2)^{-1/2}$, e per le ϕ_i il sistema dei polinomi di Chebyshev. E' ancora possibile calcolare i coefficienti c_k tramite quadrature gaussiane (vedi cap. 6).

Complessità Per le principali famiglie di polinomi ortogonali, il calcolo di ogni valore $\phi_i(x)$ si può effettuare in modo ricorrente con complessità lineare; il risultato è un ordine di complessità quadratico a cui si deve aggiungere il calcolo dei coefficienti di Fourier. Questo calcolo richiede di essere effettuato con una accuratezza (e quindi, con un numero di punti) sufficiente a conservare l'ordine della approssimazione. Nel caso dei polinomi ortogonali, si lavora tipicamente con quadrature gaussiane, mentre nel caso della base trigonometrica i coefficienti di Fourier vengono calcolati mediante trasformata discreta, o se possibile trasformata veloce (con ordine di complessità poco più che lineare). Quest'ultima possibilità può essere applicata anche ai polinomi di Chebyshev, se posti in forma trigonometrica.

Risultati fondamentali

- Criterio di miglior approssimazione

  Teorema 5.11 *Si supponga $f \in L^2_w([a, b])$. Allora, tra tutte le combinazioni lineari nella forma (5.29), quella per cui l'errore L^2_w è minimo è data dalla soluzione del sistema lineare*

  $$Mc = F \tag{5.32}$$

  dove $c = (c_0 \cdots c_n)^t$ ed il vettore F e la matrice M sono definiti da

  $$f_k = \int_a^b \phi_k(x) f(x) w(x) dx, \quad m_{ij} = \int_a^b \phi_i(x) \phi_j(x) w(x) dx. \tag{5.33}$$

In particolare, se le funzioni ϕ_i sono mutuamente ortogonali in $L^2_w([a,b])$, la soluzione di minimo errore è data dai coefficienti (5.31).

Dim. E' conveniente minimizzare il quadrato della norma dell'errore, dato da

$$\|E_n\|^2_{L^2_w} = \left\| f - \sum_k c_k \phi_k \right\|^2_{L^2_w} =$$

$$= \left(f - \sum_k c_k \phi_k, f - \sum_j c_j \phi_j \right) =$$

$$= \sum_k \sum_j c_k c_j (\phi_k, \phi_j) - 2 \sum_k c_k (f, \phi_k) + (f, f)$$

e per questa funzione (quadratica definita postiva) delle incognite c_k le condizioni di stazionarietà sono

$$\frac{\partial}{\partial c_k} \|E_n\|^2_{L^2_w} = 2 \sum_j c_j (\phi_k, \phi_j) - 2(f, \phi_k) = 0$$

da cui finalmente si ottiene l'espressione (5.32), (5.33) per i c_k. E' immediato verificare che se le funzioni ϕ_k sono ortogonali tale espressione si riduce alla (5.31), essendo in questo caso M una matrice diagonale.

■

- Convergenza della approssimazione

Teorema 5.12 *Si supponga $f \in L^2_w([a,b])$. Se la famiglia $\{\phi_i\}$ è densa in $L^2_w([a,b])$, allora per $n \to \infty$, $\|E_n\|_{L^2_w} \to 0$. Inoltre, per tutte le famiglie di funzioni viste in precedenza, se $f \in C^k([a,b])$, si ha*

$$\|E_n\|_{L^2_w} \leq C n^{-k}$$

(nel caso del sistema trigonometrico, occorre però che f sia periodica e $C^k(\mathbb{R})$).

Dim. Ci limitiamo a dimostrare la convergenza. Per la densità della base $\{\phi_k\}$, dato un $\varepsilon > 0$ esiste un ordine N_ε e degli scalari $\bar{c}_0, \ldots, \bar{c}_{N_\varepsilon}$ tali che

$$\left\| f - \sum_{k=0}^{N_\varepsilon} \bar{c}_k \phi_k \right\|_{L^2_w} \leq \varepsilon.$$

Ma per il Teorema 5.11 si ha anche, definendo gli scalari c_k tramite le (5.32), (5.33),

$$\left\| f - \sum_{k=0}^{N_\varepsilon} c_k \phi_k \right\|_{L^2_w} \leq \left\| f - \sum_{k=0}^{N_\varepsilon} \bar{c}_k \phi_k \right\|_{L^2_w} \leq \varepsilon,$$

ovvero la convergenza a zero di E_n.

■

5.3 Confronto fra i vari schemi

Nonostante la maggior efficienza delle interpolazioni a nodi non equidistanti (che si vedrà anche nelle formule di quadratura), il loro uso è fortemente limitato dal fatto che la funzione viene valutata in punti, appunto, non equidistanti, ed in generale non razionali. Questo richiede in genere che la funzione sia nota in forma esplicita, mentre nel caso di funzioni (ad esempio provenienti da misure) tabulate ad intervalli costanti, tale strategia non è più applicabile. Lo stesso tipo di ostacolo vi è nella valutazione dei coefficienti di Fourier per approssimazioni in serie di Fourier troncata. Inoltre, lavorando in dimensione maggiore di uno, tutte queste strategie sono facilmente applicabili su rettangoli ma non su geometrie più complesse.

Le approssimazioni di Lagrange composite sono un mezzo robusto e versatile di approssimare funzioni o dati tabulati ad intervalli uniformi, ed in dimensione più alta si prestano ad esempio ad implementazioni su triangoli, che permettono di trattare geometrie molto complicate. Il polinomio di Hermite non viene normalmente utilizzato nella sua forma più generale, ma implementato al terzo grado dà luogo ad approssimazioni composite (splines ed interpolazioni monotone) a derivata prima continua, abbastanza usate in pratica.

Infine, la approssimazione di dati rumorosi, come si è detto, viene effettuata in generale per minimi quadrati, essendo troppo sensibili alle perturbazioni tutte le strategie di interpolazione.

5.4 Esercizi sperimentali

- Al variare del grado n, tracciare il grafico e valutare il massimo modulo di ω_n nei casi di nodi equidistanti in $[-1, 1]$ e di nodi Chebyshev.
- Stesso esercizio per la funzione $\sum_i |L_i(x)|$.

- Verificare le proprietà di convergenza delle approssimazioni di Lagrange globali e composite, sia con nodi uniformi che Chebyshev, a seconda della regolarità della funzione da approssimare.
- Verificare le proprietà di stabilità delle approssimazioni di Lagrange globali e composite, sia con nodi uniformi che Chebyshev, perturbando (leggermente) il valore della funzione in un nodo.
- Verificare che l'interpolazione a nodi equidistanti di una funzione analitica solo su sottoinsiemi di $\mathbb{R}$ può convergere o no alla funzione a seconda dell'intervallo di interpolazione scelto.
- Verificare che il problema esposto sopra non sussiste con nodi di Chebyshev o approssimazioni composite.

6 Integrazione numerica

Data una funzione $f(x)$, di cui sia disponibile una approssimazione come combinazione lineare di funzioni

$$f(x) = c_0\phi_0(x) + c_1\phi_1(x) + \cdots + c_n\phi_n(x) + E_n(x) = \sum_{i=0}^{n} c_i\phi_i(x) + E_n(x) \quad (6.1)$$

(in particolare nella forma polinomiale, o polinomiale a tratti, del paragrafo precedente), il relativo integrale si scrive

$$\int_a^b f(x)dx = \sum_{i=0}^{n} c_i \int_a^b \phi_i(x)dx + \int_a^b E_n(x)dx. \quad (6.2)$$

Nei casi più frequenti, che derivano dalla integrazione di polinomi di Lagrange o Hermite, eventualmente in versione composita, le costanti c_j che appaiono nella (6.2) sono valori del tipo $f^{(k)}(x_i)$ (per il polinomio di Hermite) o $f(x_i)$ (per il polinomio di Lagrange). In quest'ultimo caso si ottiene:

$$\int_a^b f(x)dx = \sum_{i=0}^{n} \alpha_i f(x_i) + \int_a^b E_n(x)dx, \quad (6.3)$$

dove si è posto

$$\alpha_i = \int_a^b L_i(x)dx. \quad (6.4)$$

Il valore aprossimato dell'integrale è dato dall'espressione

$$I_n(f, a, b) = \sum_{i=0}^{n} \alpha_i f(x_i) \quad (6.5)$$

che si indica come *formula di quadratura* (in questo caso, di tipo interpolatorio) associata ad una certa approssimazione di f e all'intervallo $[a, b]$. I punti x_i vengono ancora indicati come *nodi* della quadratura, mentre i coefficienti α_i vengono chiamati *pesi*.

La valutazione dei pesi (6.4) viene fatta di regola su un *intervallo di riferimento* $[\bar{a}, \bar{b}]$ che non coincide necessariamente con $[a, b]$. Se in corrispondenza di $x \in [a, b]$ si ha $t \in [\bar{a}, \bar{b}]$, la trasformazione che lega x a t è

$$x = a + \frac{b-a}{\bar{b}-\bar{a}}(t - \bar{a})$$

e di conseguenza

$$\alpha_i = \int_a^b L_i(x)dx = \frac{b-a}{\bar{b}-\bar{a}} \int_{\bar{a}}^{\bar{b}} L_i(x(t))dt = \frac{b-a}{\bar{b}-\bar{a}} w_i \quad (6.6)$$

in cui i valori w_i sono i pesi associati all'intervallo di riferimento. In (6.6), i polinomi $L_i(x(t))$ altro non sono che i polinomi $\bar{L}_i(t)$ della base di Lagrange associata ai nodi di riferimento t_i. Infatti, poiché $x(t)$ è una trasformazione lineare, si ha $\deg \bar{L}_i = \deg L_i$, ed inoltre

$$\bar{L}_i(t_j) = L_i(x(t_j)) = L_i(x_j) = \delta_{ij}.$$

Una maniera esplicita di caratterizzare l'errore di una formula di quadratura è mediante la (ovvia) maggiorazione

$$\int_a^b E_n(x)dx \leq (b-a)\|E_n\|_{L^\infty([a,b])} = (b-a) \sup_{x\in[a,b]} |E_n(x)|, \tag{6.7}$$

ma generalmente la stima ottenuta per questa strada non è ottimale (come si vedrà a proposito delle formule di Newton–Cotes a nodi dispari). Dimostrare invece maggiorazioni esplicite ed ottimali dell'errore di quadratura è in genere un calcolo piuttosto tecnico.

In modo meno diretto, la accuratezza di una formula di quadratura si caratterizza spesso tramite il suo *grado di precisione*, il massimo intero ν tale che tutti i polinomi di grado non superiore ad ν sono integrati esattamente tramite la (6.3), ovvero

$$\nu = \max\left\{ k : \sum_{i=0}^{n} \alpha_i f(x_i) = \int_a^b f(x)dx, \quad \forall f \in \mathbb{P}_k \right\}.$$

Poiché sia l'integrale che le quadrature nella forma (6.5) sono lineari, il grado di precisione potrebbe essere definito anche come l'esponente ν tale che le potenze x^k sono integrate esattamente per $k = 0, \ldots, \nu$, mentre la potenza $x^{\nu+1}$ non è integrata esattamente.

Si noti che se n è il grado del polinomio interpolatore, si ha $E_k \equiv 0$ (e quindi $\int_a^b E_k = 0$) per $k = 0, \ldots, n$, e di conseguenza $\nu \geq n$. E' però possibile che $\int_a^b E_k = 0$ anche se l'errore di interpolazione non è identicamente nullo, e di qui la possibilità di avere $\nu > n$. Inoltre, poiché ogni polinomio interpolatore interpola esattamente le funzioni costanti, ne risulta che $\nu \geq 0$ e che, come si verifica facilmente (applicando la formula alla funzione $f(x) \equiv 1$),

$$\sum_i \alpha_i = b - a.$$

Anche se le formule di quadratura interpolatorie possono avere le forme più svariate, le due classi principali di formule sono:

Quadrature di Newton–Cotes – Sono basate su polinomi interpolatori a nodi equidistanti, e di regola la convergenza all'integrale esatto si ottiene mediante una strategia di approssimazione composita.

Quadrature gaussiane – Si ottengono con una opportuna scelta dei nodi di quadratura (non equidistanti).

Tali classi di quadrature sono in qualche modo in relazione alle due strategie di infittimento dei nodi che sono state discusse a proposito dell'interpolazione.

Si può osservare che una formula del tipo (6.5) non richiede in linea di principio di essere costruita mediante un polinomio interpolatore, né questo è richiesto dal teorema di convergenza 6.1. Si vedrà però nel teorema 6.2 che, se il grado di precisione della formula è sufficientemente alto, i pesi sono necessariamente gli integrali delle funzioni di Lagrange L_j associate ai nodi x_j scelti.

Esempio Si supponga di voler approssimare l'integrale della funzione f sull'intervallo $[\bar{x}, \bar{x}+h]$ con l'integrale del suo polinomio interpolatore di grado zero $\Pi_0(x) = f(x_0)$ con $x_0 \in [\bar{x}, \bar{x}+h]$. Il valore dell'integrale approssimato è dato da $I_0(f,a,b) = hf(x_0)$: questo è anche il valore esatto dell'integrale di f se f stessa è una funzione costante. E' meno ovvio che si possa ottenere l'integrale esatto di f anche se f non è costante ma è un polinomio di primo grado. In questo caso, l'integrale di f vale

$$\int_{\bar{x}}^{\bar{x}+h} (a+bx)dx = h\left(a + b\left(\bar{x} + \frac{h}{2}\right)\right)$$

e l'integrale approssimato coincide con quello esatto a patto che $x_0 = \bar{x}+h/2$. Come è facile verificare, in questa situazione l'errore di interpolazione ha integrale nullo senza essere identicamente nullo.

D'altra parte, se si costruisce l'integrale approssimato sulla base di una strategia di approssimazione composita di grado zero a tratti (effettuando quindi la stessa operazione su sottointervalli di ampiezza $h \to 0$), per ogni scelta del punto x_0 all'interno del singolo sottointervallo si ottiene convergenza all'integrale esatto: infatti in questo caso l'integrale approssimato altro non è che una somma integrale di f, che converge sotto la sola ipotesi di Riemann–integrabilità.

Queste due strade (da un lato aumentare il grado di precisione con una opportuna scelta dei nodi, dall'altro utilizzare approssimazioni composite) portano alle due principali classi di formule di quadratura.

Risultati fondamentali

- Teorema (di Polya) sulla convergenza delle formule di quadratura interpolatorie

Teorema 6.1 *Sia $I_n(f,a,b)$ definito dalla (6.5), con $x_0,\dots,x_n \in [a,b]$. Condizione necessaria e sufficiente perché si abbia, per ogni $f \in C^0([a,b])$, $I_n(f,a,b) \to \int_a^b f(x)dx$ per $n \to \infty$, è che, al variare di n, esista una costante $M > 0$ tale che*

$$\sum_{i=0}^{n} |\alpha_i| \leq M \tag{6.8}$$

e che, per ogni polinomio di grado fissato $p(x) \in \mathbb{P}_k$, si abbia

$$\lim_{n\to\infty} I_n(p,a,b) = \int_a^b p(x)dx. \tag{6.9}$$

Dim. Ci limitiamo a dimostare la parte relativa alla sufficienza. Per la densità dei polinomi in C^0 (teorema di Stone–Weierstrass), dato un $\varepsilon > 0$, esiste un polinomio $p_\varepsilon \in \mathbb{P}_{k_\varepsilon}$ tale che $|f(x) - p_\varepsilon(x)| < \varepsilon$ per ogni $x \in [a,b]$. D'altra parte,

$$\left|\int_a^b f(x)dx - I_n(f,a,b)\right| \leq \left|\int_a^b f(x)dx - \int_a^b p_\varepsilon(x)dx\right| +$$

$$+ \left|\int_a^b p_\varepsilon(x)dx - I_n(p_\varepsilon,a,b)\right| + |I_n(p_\varepsilon,a,b) - I_n(f,a,b)|.$$

Per il primo dei tre termini a secondo membro si ha:

$$\left|\int_a^b f(x)dx - \int_a^b p_\varepsilon(x)dx\right| \leq \int_a^b |f(x)dx - p_\varepsilon(x)|dx \leq \varepsilon(b-a)$$

mentre per il secondo, dall'ipotesi (6.9),

$$\left|\int_a^b p_\varepsilon(x)dx - I_n(p_\varepsilon,a,b)\right| \to 0.$$

Il terzo termine, per la (6.8), si può maggiorare come:

$$|I_n(p_\varepsilon,a,b) - I_n(f,a,b)| \leq \sum_{i=0}^{n} |\alpha_i[p_\varepsilon(x_i) - f(x_i)]| \leq M\varepsilon.$$

Se ne deduce quindi che tutti e tre i termini possono essere resi arbitrariamente piccoli, da cui la tesi.

■

- Formule di quadratura con pesi generici

Teorema 6.2 *Sia $I_n(f, a, b)$ definito dalla (6.5), e si supponga il grado di precisione $\nu \geq n$. Allora, i pesi α_i sono necessariamente definiti dalla (6.4), in cui L_i indica la base di Lagrange associata all'insieme di nodi x_i scelto.*

Dim. Se il grado di precisione è non minore di n, ognuna delle funzioni L_i (che sono di grado n) sarà integrata esattamente: si ha quindi

$$I_n(L_i, a, b) = \int_a^b L_i(x)dx. \tag{6.10}$$

D'altra parte, poiché $L_i(x_j) = \delta_{ij}$, si ha anche

$$I_n(L_i, a, b) = \sum_j \alpha_j L_i(x_j) = \alpha_i. \tag{6.11}$$

Confrontando (6.10) e (6.11) si ottiene infine la (6.4).

■

6.1 Quadrature di Newton–Cotes

Le formule di quadratura di Newton–Cotes sono costruite a nodi equidistanti. Una formula di Newton–Cotes *semplice* viene costruita con un unico polinomio interpolatore di grado n sull'intervallo di integrazione. In generale, però, le quadrature di Newton–Cotes non vengono utilizzate in questa forma quanto sotto forma *composita*, ovvero dividendo l'intervallo (a, b) in sottointervalli (a_j, b_j) senza aumentare il grado della quadratura utilizzata in ogni sottointervallo. In questa versione, la formula di quadratura è nella forma

$$I_{n,m}(f, a, b) = \sum_{j=0}^{m-1} \sum_{k=0}^{n} \alpha_k^{(j)} f(x_k^{(j)}) \tag{6.12}$$

in cui l'indice j si riferisce al sottointervallo considerato, k al nodo all'interno del j–mo sottointervallo. Porremo inoltre $H_j = b_j - a_j$ e con $H = \max_j H_j$.

Nel caso dell'intervallo singolo, indicando con h il passo tra due nodi, e supponendo che i nodi siano $n + 1$, si possono distinguere due sottoclassi di quadrature: quadrature chiuse e quadrature aperte.

6.1.1 Formule di Newton–Cotes chiuse

In questa classe, i nodi estremi coincidono con gli estremi a e b dell'intervallo di integrazione. Si ha quindi

$$h = \frac{b-a}{n}$$

e, per $i = 0, \ldots, n$:

$$x_i = a + ih.$$

Regola del trapezio La più semplice formula di questo tipo si ottiene per $n = 1$:

$$I_1(f, a, b) = \frac{h}{2}\Big[f(x_0) + f(x_1)\Big]$$

e va sotto il nome di *regola del trapezio*.

Regola di Simpson La formula successiva si ha per $n = 2$:

$$I_2(f, a, b) = \frac{h}{3}\Big[f(x_0) + 4f(x_1) + f(x_2)\Big]$$

ed è nota come *formula di Simpson*.

Formule chiuse di ordine superiore Dall'ordine $n = 7$ nelle formule chiuse compaiono pesi negativi. Per $n = 3, \ldots, 6$ si hanno le quadrature:

$$I_3(f, a, b) = \frac{3h}{8}\Big[f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)\Big],$$

$$I_4(f, a, b) = \frac{4h}{90}\Big[7f(x_0) + 32f(x_1) + 12f(x_2) + 32f(x_3) + 7f(x_4)\Big],$$

$$I_5(f, a, b) = \frac{5h}{288}\Big[19f(x_0)+75f(x_1)+50f(x_2)+50f(x_3)+75f(x_4)+19f(x_5)\Big],$$

$$I_6(f, a, b) = \frac{h}{140}\Big[41f(x_0) + 216f(x_1) + 27f(x_2) + 272f(x_3)+ \\ +27f(x_4) + 216f(x_5) + 41f(x_6)\Big].$$

Formule chiuse composite Se si lavora a nodi equidistanti le formule composite hanno una espressione più maneggevole, ottenuta utilizzando il passo (costante) h tra i nodi ed accorpando i pesi relativi ad uno stesso nodo (se si tratta di nodi comuni a due sottointervalli). Ad esempio, per le quadrature composite dei trapezi e di Simpson si ottiene:

$$I_{1,m}(f,a,b) = \frac{h}{2}\Big[f(x_0) + 2f(x_1) + \cdots + 2f(x_{m-1}) + f(x_m)\Big]$$

$$I_{2,m}(f,a,b) = \frac{h}{3}\Big[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + \cdots + \\ +2f(x_{2m-1}) + 4f(x_{2m}) + f(x_{2m+1})\Big]$$

6.1.2 Formule di Newton–Cotes aperte

Nelle formule aperte i nodi sono tutti interni all'intervallo (a,b), più precisamente si pone:

$$h = \frac{b-a}{n+2}$$

e, per $i = 0, \ldots, n$:

$$x_i = a + (i+1)h.$$

Regola del punto medio La più semplice di queste formule, per $n = 0$, è data da

$$I_0(f,a,b) = 2hf(x_0)$$

e si indica come *formula del punto medio.*

Formula aperta a due punti Per $n = 1$ si ottiene invece

$$I_1(f,a,b) = \frac{3h}{2}\Big[f(x_0) + f(x_1)\Big].$$

Formule aperte di ordine superiore Nelle formule aperte, i pesi negativi compaiono già dall'ordine 2. Comunque, per $n = 2, \ldots, 5$ si ha

$$I_2(f,a,b) = \frac{4h}{3}\Big[2f(x_0) - f(x_1) + 2f(x_2)\Big],$$

$$I_3(f,a,b) = \frac{5h}{24}\Big[11f(x_0) + f(x_1) + f(x_2) + 11f(x_3)\Big],$$

$$I_4(f,a,b) = \frac{6h}{20}\Big[11f(x_0) - 14f(x_1) + 26f(x_2) - 14f(x_3) + 11f(x_4)\Big],$$

$$I_5(f,a,b) = \frac{7h}{1440}\Big[611f(x_0) - 453f(x_1) + 562f(x_2) + \\ +562f(x_3) - 453f(x_4) + 611f(x_5)\Big].$$

Stabilità Supponendo che i valori $f(x_i)$ siano affetti da perturbazioni δ_i tali che $|\delta_i| \leq \delta$, indicando con Δ la perturbazione risultante sul valore dell'integrale approssimato, si ha

$$I_n(f,a,b) + \Delta = \sum_{i=0}^{n} \alpha_i[f(x_i) + \delta_i] = I_n(f,a,b) + \sum_{i=0}^{n} \alpha_i\delta_i$$

che fornisce la maggiorazione

$$|\Delta| \leq \delta \sum_{i=0}^{n} |\alpha_i|.$$

Nonostante sia sempre vero che $\sum_i \alpha_i = b-a$, nelle formule di Newton–Cotes si ha $\sum_i |\alpha_i| \to \infty$ per $n \to \infty$. In presenza di perturbazioni (o anche solo di errori di arrotondamento) non ci si aspetta quindi un buon comportamento per $n \to \infty$, neanche in caso di funzioni estremamente regolari. La situazione è diversa nelle formule di NC composite, in cui, come si vedrà nella dimostrazione del Teorema 6.5, si ha

$$\alpha_k^{(j)} = \frac{H_j}{l} w_k$$

dove $l = n$ se la quadratura è chiusa, $l = n+2$ se la quadratura è aperta. La somma dei moduli dei pesi si può esprimere quindi come

$$\sum_{i=0}^{n} |\alpha_i| = \sum_{j=0}^{m-1} \sum_{k=0}^{n} |\alpha_k^{(j)}| = \sum_{j=0}^{m-1} \sum_{k=0}^{n} \left| \frac{H_j}{l} w_k \right| =$$

$$= \frac{1}{l} \sum_{j=0}^{m-1} H_j \sum_{k=0}^{n} |w_k| = \frac{b-a}{l} \sum_{k=0}^{n} |w_k|,$$

e questa quantità resta costante al variare del numero di nodi (infatti i pesi w_k sono stati calcolati una volta per tutte nell'intervallo di riferimento).

Complessità La complessità di calcolo necessaria ad ottenere un dato errore di quadratura dipende dalla regolarità della funzione e dall'ordine della formula. In generale, la più veloce convergenza delle formule di grado più alto le rende più efficienti a parità di numero di valutazioni di f, a patto che f stessa sia sufficientemente regolare. In caso di funzioni regolari a tratti, è possibile utilizzare strategie adattative a passo non costante, che riducono opportunamente il passo nell'intorno di singolarità di f, in modo da distribuire l'errore di quadratura circa uniformemente tra tutti i sottointervalli (questa è la situazione di massima efficienza). Analogamente a quanto si è detto per l'interpolazione composita, dal Teorema 6.6 si vede che l'errore di quadratura più favorevole si ottiene quando gli errori sugli intervalli elementari sono dello stesso ordine di grandezza, cioè se

$$H_j \sim \frac{1}{\left(\sup_{[a_j,b_j]} |f^{(p)}(x)|\right)^{1/p}}$$

Risultati fondamentali

- Grado di precisione delle formule di Newton–Cotes

  Teorema 6.3 *Il grado di precisione delle formule di quadratura di Newton–Cotes è $\nu = n$ se n è dispari, $\nu = n+1$ se n è pari.*

  Dim. Ci limitiamo a dimostrare che $\nu \geq n$ nel primo caso, $\nu \geq n+1$ nel secondo. Poiché n è il grado del polinomio interpolatore su cui viene costruita la formula di quadratura, il grado di precisione della formula stessa è almeno n. Se poi $f(x)$ è un polinomio di grado $n+1$, allora $f^{(n+1)}(x) \equiv c_{n+1}$ identicamente, e se n è pari, dalla formula di rappresentazione dell'errore (5.15) si ha

  $$\int_a^b E_n(x)dx = \int_a^b \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_n(x)dx = \frac{c_{n+1}}{(n+1)!}\int_a^b \omega_n(x)dx = 0$$

  dove l'ultimo passaggio è giustificato dal fatto che essendo dispari il numero di nodi e di conseguenza dispari rispetto al nodo centrale la funzione ω_n, il suo integrale $\int_a^b \omega_n(x)dx$ è nullo (e ciò implica che il grado di precisione è almeno $n+1$).

  ■

- Espressione esplicita dell'errore di quadratura

Teorema 6.4 *Sia $p = n+1$ se n è dispari, $p = n+2$ se n è pari. Allora, se $f \in C^p([a,b])$, l'errore di quadratura nelle formule di Newton–Cotes semplici si può scrivere come*

$$\int_a^b f(x)dx - I_n(f,a,b) = C_n h^{p+1} f^{(p)}(\xi) \tag{6.13}$$

dove $\xi \in (a,b)$, e la costante C_n vale:

$$C_n = \begin{cases} \frac{1}{(n+1)!}\int_0^n s(s-1)\cdots(s-n)ds < 0 & \textit{chiuse, } n \textit{ dispari} \\ \frac{1}{(n+1)!}\int_{-1}^{n+1} s(s-1)\cdots(s-n)ds > 0 & \textit{aperte, } n \textit{ dispari} \\ \frac{1}{(n+2)!}\int_0^n s^2(s-1)\cdots(s-n)ds < 0 & \textit{chiuse, } n \textit{ pari} \\ \frac{1}{(n+2)!}\int_{-1}^{n+1} s^2(s-1)\cdots(s-n)ds > 0 & \textit{aperte, } n \textit{ pari.} \end{cases}$$

- Convergenza delle formule di NC composite

Teorema 6.5 *Sia f Riemann–integrabile in $[a,b]$. Allora, se la formula di quadratura composita (6.12) è esatta sulle costanti, si ha*

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

Dim. Ricordiamo che se (6.12) è esatta sulle costanti, allora $\sum_k \alpha_k^{(j)} = H_j$. D'altra parte, posto

$$l = \begin{cases} n & \text{se la quadratura è chiusa} \\ n+2 & \text{se la quadratura è aperta} \end{cases}$$

e $h_j = H_j/l$, è 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, $\sum_k w_k = l$). Riscrivendo allora (6.12) tramite queste costanti, si ottiene:

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

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

Nella sommatoria più interna all'ultimo membro, si può riconoscere una somma integrale della f su $[a,b]$. Per $H \to 0$, si ha quindi

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

■

- Ordine di convergenza delle formule di NC composite

Teorema 6.6 *Data la formula di quadratura composita (6.12), se valgono le ipotesi del Teorema 6.4, allora, per qualche costante $C > 0$ si ha*

$$\left| \int_a^b f(x)dx - I_{n,m}(f,a,b) \right| \leq C \sup_j \left(H_j^p \sup_{[a_j,b_j]} |f^{(p)}(x)| \right) \tag{6.14}$$

ed in particolare:

$$\left| \int_a^b f(x)dx - I_{n,m}(f,a,b) \right| \leq C \sup_{[a,b]} |f^{(p)}(x)| \, H^p. \tag{6.15}$$

Dim. E' ovvio che la (6.15) derivi dalla (6.14), una volta maggiorato il sup di un prodotto col prodotto degli estremi superiori. Dimostriamo quindi la (6.14).

Sommando gli errori di quadratura ottenuti dagli intervalli elementari si ha

$$\left| \int_a^b f(x)dx - I_{n,m}(f,a,b) \right| \leq \sum_{j=0}^{m-1} \left| \int_{a_j}^{b_j} f(x)dx - I_n(f,a_j,b_j) \right| =$$

$$= \sum_{j=0}^{m-1} |C_n| \left(\frac{H_j}{l} \right)^{p+1} |f^{(p)}(\xi_j)| =$$

$$= \frac{|C_n|}{l^{p+1}} \sum_{j=0}^{m-1} H_j^{p+1} |f^{(p)}(\xi_j)|$$

dove l'intero l è usato come nel Teorema 6.5 e si è applicata la (6.13) con $\xi_j \in [a_j, b_j]$. Ora, scrivendo $H_j^{p+1} = H_j^p H_j$ e maggiorando il prodotto $H_j^p |f^{(p)}(\xi_j)|$, si ottiene finalmente

$$\left| \int_a^b f(x)dx - I_{n,m}(f,a,b) \right| \leq \frac{|C_n|}{l^{p+1}} \sup_j \left(H_j^p \sup_{[a_j,b_j]} |f^{(p)}(x)| \right) \sum_{j=0}^{m-1} H_j$$

che è nella forma richiesta notando che $\sum_j H_j = b - a$.

■

- Segno dei pesi di quadratura

  Teorema 6.7 *Per $n \to \infty$, nelle formule di Newton–Cotes esistono definitivamente pesi $w_i < 0$ per qualche $i \in [0, n]$, e si ha*

$$\lim_{n\to\infty} \sum_i |w_i| = \infty.$$

6.2 Quadrature gaussiane

Le formule di quadratura gaussiane sono formule a nodi non equidistanti. In modo simile a quanto si è visto per la interpolazione nei nodi di Chebyshev, anche qui gli $n+1$ nodi di quadratura sono collocati negli zeri di un polinomio appartenente ad una famiglia di polinomi ortogonali. Lo spazio in cui si opera è normalmente lo spazio $L^2_w([a, b])$ delle funzioni di quadrato sommabile rispetto ad un dato peso $w(x)$, ed il prodotto scalare è quindi

$$(f, g) = \int_a^b f(x)g(x)w(x)dx. \tag{6.16}$$

Per semplicità di esposizione, ci limiteremo nel seguito al caso $w(x) \equiv 1$. La famiglia di polinomi ortogonali rispetto a questo prodotto scalare è la famiglia $\{P_k\}_{k\geq 0}$ dei *polinomi di Legendre* e viene definita tipicamente sull'intervallo di riferimento $[\bar{a}, \bar{b}] = [-1, 1]$. Nodi e pesi di quadratura sono quindi definiti da

$$P_{n+1}(x_i) = 0 \qquad (i = 0, \ldots, n), \tag{6.17}$$

$$\alpha_i = \int_a^b L_i(x)dx \qquad (i = 0, \ldots, n), \tag{6.18}$$

in cui la base di Lagrange L_i è costruita sui nodi definiti in (6.17) come radici di $P_{n+1} \in \mathbb{P}_{n+1}$. Come si vedrà nei risultati generali, tali radici sono reali, semplici (e quindi in numero di $n + 1$) e tutte interne all'intervallo (a, b).

Il calcolo di nodi e pesi va fatto per via numerica non essendo possibile in generale in forma esplicita; nella tabella 1 si riportano gli insiemi di nodi e pesi per ordini di interpolazione fino a $n = 6$, calcolati sull'intervallo $(-1, 1)$ con otto cifre significative.

Le formule di quadratura gaussiane possono essere implementate in forma composita; a differenza di quanto accade per le formule di Newton–Cotes, però, questo non è necessario per la loro convergenza (in altre parole, le formule gaussiane convergono al valore esatto dell'integrale per $n \to \infty$ sotto la sola ipotesi di continuità di f).

n	t_i	w_i
0	0.0	2.0
1	±0.57735027	1.0
2	±0.77459667 0.0	0.55555556 0.88888889
3	±0.86113631 ±0.33998104	0.34785485 0.65214515
4	±0.90617985 ±0.53846931 0.0	0.23692689 0.47862867 0.56888889
5	±0.93246951 ±0.66120939 ±0.23861918	0.17132449 0.36076157 0.46791393
6	±0.94910791 ±0.74153119 ±0.40584515 0.0	0.12948497 0.27970539 0.38183005 0.41795918

Tabella 1: Nodi e pesi delle formule di Gauss–Legendre per $n = 0, \ldots, 6$

Stabilità Come si vedrà nei risultati generali, nelle formule gaussiane i pesi sono sempre positivi. Questo fatto implica da un lato la convergenza all'integrale esatto nel caso di funzioni continue, dall'altro un buon comportamento in termini di stabilità. Ricordando l'analisi fatta a proposito delle formule di Newton–Cotes, la somma dei moduli dei pesi vale

$$\sum_{i=0}^{n} |\alpha_i| = \sum_{i=0}^{n} \alpha_i = b - a$$

e la sensibilità alle perturbazioni resta quindi costante al variare dell'ordine.

Complessità Dalle stime di errore (di dimostrazione molto tecnica e che non si riportano) si può dedurre una maggiore efficienza delle formule gaussiane rispetto alle formule di Newton–Cotes, nel senso che l'errore di quadratura a parità di numero di valutazioni della funzione f è in generale molto minore, sempre a patto che sia possibile utilizzare formule a nodi non equidistanti. Il calcolo dei nodi e dei pesi nelle formule gaussiane va comunque effettuato per via numerica e ciò porta a una complessità supplementare.

Risultati fondamentali

- Proprietà degli zeri dei polinomi di Legendre

  Teorema 6.8 *Il generico polinomio di Legendre di grado n, $P_n(x)$, ha n radici reali e semplici, tutte interne all'intervallo (a, b).*

  Dim. Supponiamo per assurdo che, nell'intervallo (a, b), $P_n(x)$ abbia $r < n$ radici reali di molteplicità dispari $x_0, \ldots, x_{r-1}$. Definiamo il polinomio

  $$Q(x) = \begin{cases} 1 & \text{se } r = 0 \\ (x - x_0)(x - x_1) \cdots (x - x_{r-1}) & \text{altrimenti.} \end{cases}$$

  Ora, poiché sia questo polinomio che $P_n(x)$ cambiano segno in corrispondenza dei punti $x_0, \ldots, x_{r-1}$, il prodotto $P_n(x)Q(x)$ ha segno costante e poiché non è identicamente nullo si ha anche

  $$\int_a^b P_n(x)Q(x)dx \neq 0$$

  ma ciò è assurdo in quanto il grado di Q è $r < n$, e P_n è ortogonale (rispetto al prodotto scalare (6.16)) a tutti i primi n polinomi

$P_0, \ldots, P_{n-1}$, i quali d'altra parte generano tutto lo spazio dei polinomi di grado non superiore ad $n-1$. Ne segue che $r = n$ e che le radici sono tutte semplici.

■

- Grado di precisione delle formule gaussiane

Teorema 6.9 *La formula di quadratura definita da (6.17), (6.18) ha grado di precisione $\nu = 2n+1$.*

Dim. Data una $f(x)$ nello spazio dei polinomi di grado non superiore a $2n+1$, in base a note proprietà dei polinomi essa può essere scritta come

$$f(x) = p_n(x)\omega(x) + q_n(x)$$

dove come al solito $\omega(x) = (x-x_0)(x-x_1)\cdots(x-x_n)$ e p_n, q_n sono polinomi di grado non superiore ad n. Poiché la formula ha $n+1$ nodi, il suo grado di precisione è almeno n. Si ha inoltre:

$$\int_a^b f(x)dx = \int_a^b p_n(x)\omega(x)dx + \int_a^b q_n(x)dx =$$

$$= \int_a^b p_n(x)\omega(x)dx + I_n(f,a,b) \tag{6.19}$$

dove l'ultimo passaggio è giustificato dal fatto che q_n viene integrato esattamente e che $I_n(f,a,b) = I_n(q_n,a,b)$ (infatti $f(x_i) = q_n(x_i)$ per l'annullarsi di ω nei nodi).

Ora, $\omega(x)$ coincide, a meno di una costante, con l'$(n+2)$–esimo polinomio della famiglia di Legendre ed è ortogonale ai polinomi $P_0, \ldots, P_n$, e quindi (come si è già notato) a tutto lo spazio dei polinomi di grado non superiore ad n. Ne consegue che è nullo il prodotto scalare (p_n, ω), ovvero

$$\int_a^b p_n(x)\omega(x)dx = 0$$

ed in conclusione

$$\int_a^b f(x)dx = I_n(f,a,b).$$

Resta da verificare che il grado di precisione non può essere maggiore di $2n+1$. Infatti, se per assurdo i polinomi di grado $2n+2$ fossero integrati esattamente, lo sarebbe anche il polinomio $\omega(x)^2$, il quale però è nullo in tutti i nodi ed ha integrale positivo.

■

- Convergenza delle formule gaussiane

Teorema 6.10 *Per ogni $f \in C^0([a,b])$, la formula di quadratura definita da (6.17), (6.18) converge al valore esatto dell'integrale, ovvero*

$$\lim_{n\to\infty} I_n(f,a,b) = \int_a^b f(x)dx$$

Dim. Ricordiamo che dal Teorema 6.8, si ha che gli zeri dei polinomi P_n sono reali, distinti ed interni all'intervallo $[a,b]$. La formula di quadratura (6.17), (6.18) è costruita quindi tramite un polinomio interpolatore di grado n a nodi in (a,b) e di conseguenza soddisfa la condizione (6.9). Per applicare il teorema di Polya, verifichiamo ora che si ha $\alpha_i > 0$ per ogni $i = 0,\ldots,n$. Infatti, integrando la funzione positiva L_i^2, si ha ovviamente

$$\int_a^b L_i(x)^2 dx > 0, \tag{6.20}$$

ma d'altra parte, poiché L_i è di grado n, L_i^2 è di grado $2n$ e l'integrale (6.20) può essere valutato esattamente tramite la quadratura (6.17), (6.18). Inoltre, $L_i(x_j)^2 = L_i(x_j) = \delta_{ij}$ e da ciò si ottiene:

$$0 < \int_a^b L_i(x)^2 dx = \sum_j \alpha_j L_i(x_j)^2 = \alpha_i.$$

In conclusione, $\sum_i |\alpha_i| = \sum_i \alpha_i = b - a$ e la convergenza della formula di quadratura si deduce dalla applicazione del teorema di Polya.

■

6.3 Confronto fra i vari schemi

Come si è già detto, esistono situazioni in cui i nodi di quadratura non possono essere che equidistanti, ed in questo caso non ci sono alternative all'uso delle formule di NC. Negli altri casi, le formule gaussiane sono in generale più efficienti, pur presentando il problema del calcolo di nodi e pesi. Se la funzione da integrare è regolare, c'è un forte vantaggio nell'uso di formule di ordine più alto, in caso contrario le formule composite di ordine più basso vengono reputate più robuste. In particolare, l'uso di quadrature composite rende possibili strategie adattative di scelta del passo di integrazione.

Un inconveniente delle formule gaussiane rispetto a quelle di NC composite è poi la scarsa flessibilità nel calcolo di integrali multipli. Come si è visto per l'interpolazione, infatti, con queste formule si possono trattare agevolmente domini rettangolari ma non geometrie complesse. In quest'ultimo caso, si ottengono risultati migliori decomponendo ad esempio l'insieme di integrazione in triangoli o tetraedri, ed applicando una quadratura di ordine relativamente basso su ognuno di essi.

6.4 Esercizi sperimentali

- Verificare che nella integrazione di una funzione analitica (ma non di una funzione con regolarità finita), la accuratezza delle quadrature semplici è ben caratterizzata dal grado di precisione.

- Calcolare, al variare di H, l'ordine di convergenza della quadrature composite a seconda del grado di regolarità della funzione integranda, in particolare se uguale o minore del grado necessario per applicare il Teorema 6.6.

- Calcolare con le varie quadrature l'integrale di una funzione regolare di cui si sia (leggermente) perturbato il valore in un nodo. Confrontare la perdita di accuratezza delle varie quadrature rispetto al caso regolare.

7 Metodi per Equazioni Differenziali Ordinarie

Dato un problema di Cauchy in forma di sistema del primo ordine,

$$\begin{cases} y'(x) = f(x, y(x)) \\ y(x_0) = y_0 \end{cases} \tag{7.1}$$

la filosofia generale dei metodi numerici alle differenze è di fissare un passo di discretizzazione h ed approssimare la soluzione nei punti $x_1 = x_0 + h, x_2 = x_0 + 2h, \ldots, x_k = x_0 + kh, \ldots$ sostituendo alla derivata y' oppurtuni rapporti incrementali. Di regola si richiede che questa approssimazione sia *zero–stabile*, ovvero che dia soluzioni numeriche uniformemente limitate al variare del passo h, e *consistente*, ovvero che i rapporti incrementali usati per approssimare y' convergano (in genere, con ordine $q \geq 1$ rispetto ad h) al valore esatto.

A seconda del modo in cui si costruiscono queste approssimazioni, questi schemi vengono divisi in due classi:

Metodi ad un passo – In questo caso la approssimazione u_{k+1} di $y(x_{k+1})$ si costruisce in base esclusivamente alle informazioni disponibili al passo k–esimo. Presentano il vantaggio della semplicità di programmazione, ma richiedono ad ogni passo un certo numero di valutazioni della funzione f. E' possibile senza eccessive difficoltà realizzarne versioni a passo non costante.

Metodi a più passi – La approssimazione u_{k+1} si costruisce nei metodi a più passi (o multistep) in base alle informazioni calcolate al passo k–esimo e nei p passi precedenti. I primi $p + 1$ passi di "innesco" della soluzione approssimata vengono calcolati per altra via (ad esempio con un metodo ad un passo). Presentano minore complessità computazionale, specialmente con ordini di consistenza alti, ma sono in genere più complessi da implementare ed è più difficile gestire un passo di discretizzazione non costante. Possono presentare fenomeni oscillatori spuri.

D'altra parte, a seconda di come viene calcolata la approssimazione u_{k+1} si può distinguere in entrambi i casi tra

Metodi espliciti – La approssimazione u_{k+1} è fornita esplicitamente a partire dai dati disponibili al passo k, ed eventualmente, ai passi precedenti. Sono più semplici da implementare ma hanno condizioni di stabilità assoluta più restrittive.

Metodi impliciti – In questo caso u_{k+1} va calcolata risolvendo una equazione. Hanno una complessità maggiore ma sono in genere più stabili.

Esempio Consideriamo il problema di Cauchy scalare

$$\begin{cases} y'(x) = y(x) \\ y(0) = 1 \end{cases}$$

e supponiamo di voler approssimare in $x = 1$ la soluzione esatta $y(1) = e$. A questo scopo suddividiamo l'intervallo $[0, 1]$ in N passi di ampiezza $h = 1/N$ e calcoliamo l'approssimazione u_{k+1} di $y(x_{k+1})$, a partire dalla approssimazione u_k di $y(x_k)$ (si procede in avanti a partire dal valore $u_0 = y(0) = 1$ che è noto) sostituendo la derivata y' con il rapporto incrementale

$$\frac{u_{k+1} - u_k}{h}.$$

A seconda che questo rapporto incrementale venga uguagliato al secondo membro della equazione differenziale calcolato in (x_k, u_k) o in (x_{k+1}, u_{k+1}), si ottengono le due relazioni ricorrenti

$$\begin{cases} \dfrac{u_{k+1} - u_k}{h} = u_k \\ u_0 = 1, \end{cases} \tag{7.2}$$

$$\begin{cases} \dfrac{u_{k+1} - u_k}{h} = u_{k+1} \\ u_0 = 1. \end{cases} \tag{7.3}$$

Dalla (7.2) si ottiene immediatamente u_{k+1} come

$$u_{k+1} = (1 + h)u_k = (1 + h)^2 u_{k-1} = \cdots = (1 + h)^{k+1} u_0 = (1 + h)^{k+1}.$$

Tenendo conto del fatto che $h = 1/N$ e che $x_N = 1$, la approssimazione del valore $y(1) = e$ che si ottiene per questa strada vale

$$u_N = \left(1 + \frac{1}{N}\right)^N$$

ed è noto che quando $N \to \infty$ questo valore converge al valore corretto.

In modo analogo, dalla (7.3) si ottiene u_{k+1} come soluzione dell'equazione

$$(1 - h)u_{k+1} = u_k,$$

da cui, allo stesso modo di prima,

$$u_{k+1} = \cdots = \frac{1}{(1 - h)^{k+1}} u_0 = \frac{1}{(1 - h)^{k+1}}.$$

Per questa strada il valore $y(1) = e$ viene approssimato come

$$u_N = \frac{1}{\left(1 - \frac{1}{N}\right)^N}$$

che, per $N \to \infty$, converge ancora al valore corretto $1/e^{-1} = e$. Come si vedrà però a proposito della stabilità, questi due schemi (esplicito il primo, implicito il secondo) hanno proprietà qualitative diverse nell'approssimare soluzioni stabili.

7.1 Metodi ad un passo

La struttura più generale di un metodo ad un passo per Equazioni Differenziali Ordinarie è:

$$\begin{cases} u_{k+1} = u_k + h\Phi(x_k, u_k, u_{k+1}; h) \\ u_0 = y_0 \end{cases} \tag{7.4}$$

in cui quindi le sole informazioni utilizzate per passare da u_k ad u_{k+1} sono quelle disponibili ai passi k e $k+1$. Nella teoria dei metodi ad un passo la condizione di zero–stabilità viene tipicamente sostituita dalla lipschitzianità della funzione Φ rispetto agli argomenti u_k e u_{k+1} (si dimostra che quest'ultima condizione implica la zero–stabilità).

A seconda del fatto che la funzione Φ che compare in (7.4) dipenda o no da u_{k+1}, gli schemi numerici del tipo (7.4) si indicano rispettivamente come impliciti o espliciti.

Risultati fondamentali

- Convergenza degli schemi ad un passo

  Teorema 7.1 *Si supponga che la funzione Φ in (7.4) sia globalmente lipschitziana, ovvero*

  $$\|\Phi(x, u_1, v_1; h) - \Phi(x, u_2, v_2; h)\| \leq L\Big[\|u_1 - u_2\| + \|v_1 - v_2\|\Big]. \tag{7.5}$$

  Allora, se lo schema (7.11) è consistente, è anche convergente, ovvero fissato $\bar{x} > x_0$, per ogni $k = 1, \ldots, (\bar{x} - x_0)/h$ si ha

  $$\|y(x_k) - u_k\| \to 0 \tag{7.6}$$

  per $h \to 0$. Se inoltre lo schema è consistente con ordine q, e la soluzione y è sufficientemente regolare, allora

  $$\|u_k - y(x_k)\| \leq Ch^q. \tag{7.7}$$

Dim. Per semplicità poniamo $h = (\bar{x} - x_0)/N$, con N intero, ed utilizziamo la notazione abbreviata $\Phi(x, u, v; h) \equiv \Phi(u, v)$. Usando la disuguaglianza triangolare, l'errore al passo k–esimo può essere scritto come

$$\begin{aligned}
\varepsilon_k &= \|y(x_k) - u_k\| \leq \\
&\leq \|y(x_k) - y(x_{k-1}) - h\Phi(y(x_{k-1}), y(x_k))\| + \\
&\quad + \|y(x_{k-1}) + h\Phi(y(x_{k-1}), y(x_k)) - u_{k-1} - h\Phi(u_{k-1}, u_k)\| \leq \\
&\leq h\tau(h) + \varepsilon_{k-1} + h\|\Phi(y(x_{k-1}), y(x_k)) - \Phi(u_{k-1}, u_k)\| \leq \\
&\leq h\tau(h) + (1 + hL)\varepsilon_{k-1} + hL\varepsilon_k.
\end{aligned}$$

Nella derivazione di questa maggiorazione, si sono utilizzate la definizione di u_k e quella di errore di consistenza, insieme con la lipschitzianità della funzione Φ. Per $h < h_0 < 1/L$, si può esplicitare ε_k ottenendo

$$\begin{aligned}
\varepsilon_k &\leq \frac{h}{1 - hL}\,\tau(h) + \frac{1 + hL}{1 - hL}\,\varepsilon_{k-1} \leq \\
&\leq \frac{h}{1 - h_0 L}\,\tau(h) + \left(1 + \frac{2hL}{1 - h_0 L}\right)\varepsilon_{k-1} = \\
&= hC_1\tau(h) + (1 + hC_2)\varepsilon_{k-1} \qquad (7.8)
\end{aligned}$$

in cui si è appunto supposto che $h < h_0$, e di conseguenza $1 - hL > 1 - h_0 L$. Iterando (7.8) all'indietro, si ottiene poi

$$\begin{aligned}
\varepsilon_k &\leq hC_1\tau(h) + (1 + hC_2)\varepsilon_{k-1} \leq \\
&\leq hC_1\tau(h) + (1 + hC_2)hC_1\tau(h) + (1 + hC_2)^2\varepsilon_{k-2} \leq \cdots \leq \\
&\leq hC_1\left[1 + (1 + hC_2) + \cdots + (1 + hC_2)^{k-1}\right]\tau(h) + (1 + hC_2)^k\varepsilon_0,
\end{aligned}$$

da cui ricordando che $u_0 = y_0$ (e quindi $\varepsilon_0 = 0$), ed applicando la formula della somma di ordine k di una progressione geometrica di ragione $(1 + hC_2)$, si ha:

$$\begin{aligned}
\varepsilon_k &\leq hC_1\frac{1 - (1 + hC_2)^k}{1 - (1 + hC_2)}\,\tau(h) \leq \\
&\leq hC_1\frac{(1 + hC_2)^N - 1}{hC_2}\,\tau(h). \qquad (7.9)
\end{aligned}$$

Infine, tenendo conto che $h = (\bar{x} - x_0)/N$ e che $e^t \geq (1 + t/N)^N$, si ottiene da (7.9):

$$\varepsilon_k \leq C_1\frac{e^{C_2(\bar{x} - x_0)} - 1}{C_2}\,\tau(h) \qquad (7.10)$$

che dimostra la prima parte del teorema utilizzando la ipotesi di consistenza. Se poi lo schema è consistente con ordine q, e la soluzione è sufficientemente regolare, $\tau(h) < Ch^q$ (si intende che in questo caso $\tau(h)$ viene calcolato *sulla soluzione particolare del problema*) e questo dimostra anche la seconda parte dell'enunciato.

■

7.1.1 Metodi ad un passo espliciti

Nei cosiddetti *metodi espliciti* la funzione Φ che compare in (7.4) non dipende da u_{k+1}, ovvero

$$\begin{cases} u_{k+1} = u_k + h\Phi(x_k, u_k; h) \\ u_0 = y_0, \end{cases} \tag{7.11}$$

ed in questo caso la approssimazione u_{k+1} si può calcolare direttamente usando in (7.11) i valori calcolati al passo k. Diamo di seguito alcuni esempi più noti di metodi espliciti.

Metodo di Eulero Calcola la approssimazione u_{k+1} nella forma

$$u_{k+1} = u_k + hf(x_k, u_k). \tag{7.12}$$

Tale schema è consistente con ordine $q = 1$ e zero–stabile.

Metodo di Heun La approssimazione u_{k+1} viene calcolata nella forma

$$u_{k+1} = u_k + h\left[\frac{1}{2}f(x_k, u_k) + \frac{1}{2}f(x_k + h, u_k + hf(x_k, u_k))\right]. \tag{7.13}$$

Lo schema è consistente con ordine $q = 2$ e zero–stabile.

Metodo di Eulero modificato Si calcola u_{k+1} come

$$u_{k+1} = u_k + hf\left(x_k + \frac{h}{2}, u_k + \frac{h}{2}f(x_k, u_k)\right). \tag{7.14}$$

Anche in questo caso, lo schema è consistente con ordine $q = 2$ e zero–stabile.

Metodi di Runge–Kutta espliciti Rappresentano versioni più generali, e di ordine superiore, dei metodi di Heun o di Eulero modificato. Nei metodi di Runge–Kutta espliciti a r stadi, che sono quelli più comunemente usati, la approssimazione u_{k+1} viene calcolata nella forma (7.11), con

$$\Phi(x_k, u_k; h) = \sum_{i=1}^{r} a_i F_i(x_k, u_k; h) \tag{7.15}$$

$$F_i(x_k, u_k; h) = f\left(x_k + b_i h, u_k + b_i h \sum_{j<i} c_{ij} F_j(x_k, u_k; h)\right) \tag{7.16}$$

(nei metodi più usati, $c_{i,i-1} \neq 0$, $c_{ij} = 0$ altrimenti). Ad esempio il metodo di Eulero esplicito è uno schema di Runge–Kutta ad uno stadio ed i metodi di Heun e di Eulero modificato sono schemi di Runge–Kutta espliciti a due stadi; uno schema a 4 stadi di uso molto comune si ottiene ponendo

$$\Phi(x_k, u_k; h) = \frac{1}{6}(F_1 + 2F_2 + 2F_3 + F_4)$$

in cui:

$$\begin{aligned} F_1(x_k, u_k; h) &= f(x_k, u_k) \\ F_2(x_k, u_k; h) &= f(x_k + h/2, u_k + h/2F_1) \\ F_3(x_k, u_k; h) &= f(x_k + h/2, u_k + h/2F_2) \\ F_4(x_k, u_k; h) &= f(x_k + h, u_k + hF_3) \end{aligned}$$

e si tratta di uno schema zero–stabile di ordine $q = 4$. E' bene osservare, comunque, che dall'ordine 5 in poi non è più possibile ottenere l'ordine di consistenza q mediante uno schema RK esplicito con esattamente q stadi (ad esempio, l'ordine $q = 5$ può essere ottenuto con almeno $r = 6$ stadi).

Complessità L'operazione critica negli schemi espliciti è il calcolo della f, che ad esempio negli schemi di tipo Runge–Kutta va ripetuto tante volte quanti sono gli stadi del metodo. A titolo di esempio, per un sistema differenziale lineare

$$y' = Ay$$

con $y \in \mathbb{R}^n$, ogni valutazione del secondo membro ha una complessità di ordine pari al numero di elementi non nulli della matrice, per esempio $O(2n^2)$ per una matrice piena, da ripetersi per il numero totale di stadi. Ordini di consistenza elevati possono essere realizzati con minore complessità di calcolo mediante metodi a più passi.

Un altro limite degli schemi ad un passo (ma anche a più passi) espliciti, legato questa volta alle limitazioni di stabilità assoluta, riguarda la scarsa efficienza nel risolvere problemi stiff. Infatti, in questo caso, la presenza di autovalori negativi e di modulo grande nella jacobiana di f costringe all'uso di passi h molto piccoli anche solo per ottenere un comportamento qualitativamente corretto della soluzione numerica. In altre parole, il passo h va dimensionato sulla base di considerazioni di stabilità piuttosto che di accuratezza.

Risultati fondamentali

- Covergenza dei metodi nella forma (7.11)

  Teorema 7.2 *Se il metodo (7.11) è esplicito e consistente, allora la (7.5) è soddisfatta (e quindi (7.11) è convergente).*

- Consistenza degli schemi di Eulero e di Runge–Kutta a due stadi

  Teorema 7.3 *Supponendo la funzione $f(\cdot,\cdot)$ sufficientemente regolare, per lo schema di Eulero (7.12) si ha $\tau(h) \leq Ch$. Inoltre, per un generico schema di Runge–Kutta a due stadi nella forma*

$$u_{k+1} = u_k + h\Big[a_1 f(x_k, u_k) + a_2 f(x_k + bh, u_k + bhf(x_k, u_k))\Big] \quad (7.17)$$

  che soddisfi le condizioni

$$\begin{cases} a_1 + a_2 = 1 \\ a_2 b = 1/2 \end{cases} \quad (7.18)$$

  (ed in particolare, per gli schemi (7.13) e (7.14)), si ha $\tau(h) \leq Ch^2$.

  Dim. Ricordiamo che ciò che occorre verificare è che

$$\|\bar{y} + h\Phi(\bar{x}, \bar{y}; h) - y(\bar{x} + h)\| = h\tau(h)$$

  dove $y(\cdot)$ è la soluzione che passa per un dato punto $(\bar{x}, \bar{y})$. Per fare ciò, consideriamo lo sviluppo di Taylor, rispettivamente al primo ed al secondo ordine, della soluzione y nell'intorno del punto $(\bar{x}, \bar{y})$. Si ha, supponendo la f sufficientemente regolare:

$$y(\bar{x} + h) = \bar{y} + hf(\bar{x}, \bar{y}) + O(h^2), \quad (7.19)$$

$$y(\bar{x}+h) = \bar{y}+hf(\bar{x},\bar{y})+\frac{h^2}{2}\left[f_x(\bar{x},\bar{y}) + f_y(\bar{x},\bar{y})f(\bar{x},\bar{y})\right]+O(h^3). \quad (7.20)$$

Per il metodo di Eulero si ha perciò

$$\|\bar{y} + h\Phi(\bar{x},\bar{y};h) - y(\bar{x}+h)\| =$$

$$= \|\bar{y} + hf(\bar{x},\bar{y}) - \bar{y} - hf(\bar{x},\bar{y}) + O(h^2)\| = O(h^2)$$

da cui $\tau(h) = O(h)$. Per i metodi di Runge–Kutta a due stadi si ha intanto, sviluppando opportunamente il secondo membro di (7.17) nell'intorno del punto $(\bar{x},\bar{y})$, per un incremento bh:

$$\bar{y} + h\Big[a_1 f(\bar{x},\bar{y}) + a_2 f(\bar{x}+bh, \bar{y}+bhf(\bar{x},\bar{y}))\Big] =$$

$$= \bar{y} + ha_1 f(\bar{x},\bar{y})+$$

$$+ha_2\left(f(\bar{x},\bar{y}) + hbf_x(\bar{x},\bar{y}) + hbf_y(\bar{x},\bar{y})f(\bar{x},\bar{y}) + O(h^2)\right) =$$

$$= \bar{y} + h(a_1+a_2)f(\bar{x},\bar{y}) + h^2 a_2 b\left(f_x(\bar{x},\bar{y}) + f_y(\bar{x},\bar{y})f(\bar{x},\bar{y})\right) + O(h^3)$$

che confrontato poi con lo sviluppo (7.20) per y, ed applicando le condizioni (7.18), fornisce $\tau(h) = O(h^2)$.

■

- Stabilità assoluta dei metodi di Runge–Kutta espliciti

Teorema 7.4 *Uno schema di Runge–Kutta a q stadi, di ordine q, nella forma (7.15)–(7.16) ha sempre una regione di stabilità assoluta limitata. Posto $z = h\lambda$ ($z \in \mathbb{C}$), la sua regione di stabilità assoluta si ottiene dalla disequazione*

$$|T_q(z)| < 1 \quad (7.21)$$

dove $T_q(z)$ è lo sviluppo di Taylor di ordine q centrato nell'origine di e^z. In particolare la disuguaglianza (7.21) per il metodo di Eulero ha la forma

$$|1+z| < 1 \quad (7.22)$$

(che individua un disco di raggio unitario centrato sul punto $z = -1$), e per i metodi di Runge–Kutta del secondo ordine ha la forma

$$\left|1 + z + \frac{z^2}{2}\right| < 1. \quad (7.23)$$

Dim. Ricordiamo che nella verifica della stabilità assoluta, $f(x, y) = \lambda y$. Dimostriamo intanto per induzione che

$$F_i = \frac{P_i(h\lambda)}{h} u_k \tag{7.24}$$

in cui $P_i(h\lambda)$ è un polinomio di grado non superiore ad i. La (7.24) è chiaramente vera per $F_1 = \lambda u_k$, e d'altra parte dalla definizione (7.16), se (7.24) è vera al passo $l - 1$, si ha:

$$F_l = \lambda \left(u_k + b_l h \sum_{j<l} c_{lj} F_j \right) =$$

$$= \frac{1}{h} \left(\lambda h + b_l \sum_{j<l} c_{lj} h\lambda P_j(h\lambda) \right) u_k \tag{7.25}$$

e quindi (7.24) è vera anche al passo l poiché il termine tra parentesi è un polinomio in $h\lambda$ di grado non superiore ad l. Ancora dalla definizione (7.15) si ha che a sua volta la funzione Φ ha la struttura (7.24) con un polinomio di grado non superiore a q. Finalmente,

$$u_{k+1} = u_k + h\Phi(x_k, u_k; h) = T_q(h\lambda) u_k \tag{7.26}$$

con T_q polinomio di grado q. Poiché per la soluzione esatta si ha

$$y(x_{k+1}) = e^{h\lambda} y(x_k), \tag{7.27}$$

confrontando (7.26) e (7.27) ed applicando la definizione di consistenza con ordine q si ottiene la condizione

$$T_q(h\lambda) = e^{h\lambda} + O(h^{q+1})$$

che è soddisfatta se e solo se T_q è proprio lo sviluppo di Taylor di ordine q della funzione $e^{h\lambda}$. Imponendo poi che $\lim_k u_k = 0$ si ottiene la (7.21). In particolare, nel caso (7.22), il valore $|1 + z|$ equivale alla distanza di z dal punto (posizionato sull'asse reale) $-1 + i \cdot 0$.

■

7.1.2 Metodi ad un passo impliciti

Nei *metodi impliciti* la funzione Φ dipende realmente da u_{k+1}, ed il calcolo della nuova approssimazione va fatto risolvendo (7.4) rispetto a u_{k+1}, in generale mediante schemi iterativi se la f è nonlineare. Anche in questo caso diamo un paio di esempi importanti di metodi impliciti ad un passo, entrambi A–stabili.

Metodo di Eulero implicito Calcola u_{k+1} tramite l'equazione

$$u_{k+1} = u_k + hf(x_k, u_{k+1}). \tag{7.28}$$

L'ordine di consistenza è $q = 1$.

Metodo di Crank–Nicolson u_{k+1} si calcola risolvendo

$$u_{k+1} = u_k + h\left[\frac{1}{2}f(x_k, u_k) + \frac{1}{2}f(x_k + h, u_{k+1})\right]. \tag{7.29}$$

Lo schema ha ordine di consistenza $q = 2$.

Complessità Nei metodi impliciti la complessità computazionale associata ad ogni passo è legata alla soluzione della equazione (o del sistema di equazioni) (7.4) rispetto a u_{k+1}. A prima vista, essendo (7.4) nella forma di una equazione di punto fisso, si potrebbe pensare di risolverla per sostituzioni successive, e si dimostra (vedi teor. 7.7) che questo metodo può essere reso convergente, a patto però che h soddisfi delle limitazioni che riducono fortemente il vantaggio di usare uno schema con grandi margini di stabilità. In pratica il sistema (7.4) viene quindi tipicamente risolto mediante metodi a convergenza sopralineare (Newton, secanti), con una opportuna condizione di innesco (vedi §7.2.1).

Se l'equazione è lineare, nella forma

$$y' = Ay,$$

allora anche (7.4) è un sistema lineare. Ad esempio nel caso del metodo di Eulero implicito,

$$u_{k+1} = u_k + hAu_{k+1}$$

il sistema da risolvere ad ogni passo è

$$(I - hA)u_{k+1} = u_k \tag{7.30}$$

ed in questo caso la complessità della sua soluzione, *a patto di fattorizzare preventivamente la matrice* $I - hA$, è $O(2n^2)$, e quindi confrontabile con la complessità del metodo esplicito.

Risultati fondamentali

- Convergenza dei metodi di Eulero implicito e di Crank–Nicolson

Teorema 7.5 *I metodi di Eulero impicito (7.28) e di Crank–Nicolson (7.29) sono consistenti con ordine rispettivamente $q = 1$ e $q = 2$, e soddisfano la condizione di lipschitzianità (7.5).*

Dim. Iniziamo con la consistenza. Dobbiamo verificare che

$$\|\bar{y} + h\Phi(\bar{x}, \bar{y}, y(\bar{x}+h); h) - y(\bar{x}+h)\| = h\tau(h)$$

con $y(\cdot)$ soluzione regolare passante per $(\bar{x}, \bar{y})$. Ricordiamo che gli sviluppi di Taylor, rispettivamente di primo e secondo ordine, di y nell'intorno del punto $(\bar{x}, \bar{y})$ sono (7.19) e (7.20). Per il metodo di Eulero implicito si ha, scrivendo $y(\bar{x}+h) = \bar{y} + O(h)$:

$$\|\bar{y} + h\Phi(\bar{x}, \bar{y}, y(\bar{x}+h); h) - y(\bar{x}+h)\| =$$

$$= \|\bar{y} + hf(\bar{x}, \bar{y} + O(h)) - \bar{y} - hf(\bar{x}, \bar{y}) + O(h^2)\| =$$

$$= \|\bar{y} + hf(\bar{x}, \bar{y}) + O(h^2) - \bar{y} + hf(\bar{x}, \bar{y}) + O(h^2)\| = O(h^2)$$

da cui $\tau(h) = O(h)$. Nel caso del metodo di Crank–Nicolson, ponendo $y(\bar{x}+h) = \bar{y} + hf(\bar{x}, \bar{y}) + O(h^2)$, si ottiene:

$$\bar{y} + h\Phi(\bar{x}, \bar{y}, y(\bar{x}+h); h) =$$

$$= \bar{y} + h\left[\frac{1}{2}f(\bar{x}, \bar{y}) + \frac{1}{2}f(\bar{x}+h, \bar{y} + hf(\bar{x}, \bar{y}) + O(h^2))\right] =$$

$$= \bar{y} + \frac{h}{2}f(\bar{x}, \bar{y}) + \frac{h}{2}\left(f(\bar{x}, \bar{y}) + hf_x(\bar{x}, \bar{y}) + hf_y(\bar{x}, \bar{y})f(\bar{x}, \bar{y}) + O(h^2)\right) =$$

$$= \bar{y} + hf(\bar{x}, \bar{y}) + \frac{h^2}{2}\left(f_x(\bar{x}, \bar{y}) + f_y(\bar{x}, \bar{y})f(\bar{x}, \bar{y})\right) + O(h^3)$$

che confrontato poi con lo sviluppo (7.20) per y, dà $\tau(h) = O(h^2)$.

Per quanto riguarda la proprietà di lipschitzianità (7.5), indicata con L_f la costante di Lipschitz della funzione $f(x, \cdot)$, si ha per il metodo di Eulero implicito

$$\|\Phi(x, u_1, v_1; h) - \Phi(x, u_2, v_2; h)\| = \|f(x, v_1) - f(x, v_2)\| \leq L_f\|v_1 - v_2\|$$

mentre per il metodo di Crank–Nicolson si ha

$$\|\Phi(x, u_1, v_1; h) - \Phi(x, u_2, v_2; h)\| =$$

$$= \frac{1}{2}\|f(x, u_1) + f(x, v_1) - f(x, u_2) - f(x, v_2)\| \leq$$

$$\leq \frac{1}{2}\Big[\|f(x,u_1)-f(x,u_2)\|+\|f(x,v_1)-f(x,v_2)\|\Big] \leq$$
$$\leq \frac{L_f}{2}\Big[\|u_1-u_2\|+\|v_1-v_2\|\Big]$$

■

- Stabilità assoluta dei metodi di Eulero implicito e Crank–Nicolson

Teorema 7.6 *I metodi di Eulero impicito (7.28) e di Crank–Nicolson (7.29) sono A–stabili. In particolare, la regione di stabilità assoluta del metodo di Eulero implicito è l'esterno del disco di raggio unitario centrato sul punto $z = 1$, mentre la regione di stabilità assoluta del metodo di Crank–Nicolson è il semipiano dei complessi a parte reale negativa.*

Dim. Ponendo $f(x,y) = \lambda y$, con $\lambda \in \mathbb{C}$, si ha per il metodo di Eulero implicito:

$$u_{k+1} = u_k + h\lambda u_{k+1}$$

da cui si ottiene

$$u_{k+1} = \frac{1}{1-h\lambda}u_k. \tag{7.31}$$

Poiché la disequazione

$$\left|\frac{1}{1-h\lambda}\right| < 1$$

equivale a

$$|1-h\lambda| > 1,$$

la regione di stabilità assoluta è costituita dai punti $z = h\lambda$ che hanno distanza più che unitaria dal punto $1 + i \cdot 0$.

Per quanto riguarda il metodo di Crank–Nicolson, si ottiene:

$$u_{k+1} = u_k + h\left[\frac{1}{2}\lambda u_k + \frac{1}{2}\lambda u_{k+1}\right]$$

da cui si ha

$$u_{k+1} = \frac{1+\frac{h\lambda}{2}}{1-\frac{h\lambda}{2}}u_k. \tag{7.32}$$

A sua volta, la condizione

$$\left|\frac{1+\frac{h\lambda}{2}}{1-\frac{h\lambda}{2}}\right| < 1$$

equivale a

$$|2 + h\lambda| < |2 - h\lambda| \tag{7.33}$$

e questa disuguaglianza è soddisfatta da tutti i punti $z = h\lambda \in \mathbb{C}$ per i quali la distanza dal punto $-2 + i \cdot 0$ è minore della distanza dal punto $2 + i \cdot 0$ (ovvero il semipiano a parte reale negativa di $\mathbb{C}$).

■

- Risolubilità della equazione (7.4)

Teorema 7.7 *Se la funzione Φ in (7.4) soddisfa (7.5), e se $h < 1/L$, ad ogni passo l'equazione (7.4) si può risolvere rispetto a u_{k+1} per sostituzioni successive, nella forma*

$$u_{k+1}^{(n+1)} = u_k + h\Phi(x_k, u_k, u_{k+1}^{(n)}; h).$$

Dim. Scritta in modo più esplicito, l'equazione da risolvere ad ogni passo è

$$v = u_k + h\Phi(x_k, u_k, v; h) \tag{7.34}$$

in cui si è indicata con v l'incognita u_{k+1}. Scritta la (7.34) nella forma $v = T(v)$, è immediato verificare che il secondo membro è una contrazione, appunto, se $h < 1/L$.

■

7.1.3 Metodi a passo variabile

Abbiamo supposto fin qui che il passo h resti fisso durante l'algoritmo, ma è chiaro che nessuno dei metodi descritti lo richiede. Nei *metodi a passo variabile*, impliciti o espliciti, il passo può essere variato, normalmente con lo scopo di rendere circa costante l'errore locale di troncamento. Questa è infatti la situazione di migliore efficienza dello schema.

Per stimare l'errore di consistenza, supponiamo che esso abbia l'espressione

$$\tau(x, y, h) = C(x, y)h^p,$$

e, a partire dal punto $(x_k, u_u) = (x_k, y(x_k))$ definiamo in corrispondenza all'ascissa x_{k+2} le due approssimazioni $u_{k+2}^{(h)}$ e $u_{k+2}^{(2h)}$, ottenute rispettivamente con due avanzamenti dello schema a passo h e con un solo avanzamento a passo $2h$. Considerando che la variazione di $C(x, y)$ tra il passo k ed il passo

$k+1$ è un infinitesimo in h, si ottiene che, a meno di infinitesimi di ordine superiore,

$$u_{k+2}^{(h)} \approx y(x_{k+2}) + 2C(x_k, u_k)h^{p+1},$$
$$u_{k+2}^{(2h)} \approx y(x_{k+2}) + C(x_k, u_k)(2h)^{p+1}.$$

Sottraendo, si ha quindi

$$\begin{aligned} u_{k+2}^{(h)} - u_{k+2}^{(2h)} &\approx 2C(x_k, u_k)h^{p+1} - C(x_k, u_k)(2h)^{p+1} = \\ &= 2(1-2^{p+1})h \cdot C(x_k, u_k)h^p, \end{aligned}$$

da cui l'errore di consistenza si approssima come

$$\tau(x_k, u_k, h) \approx \frac{u_{k+2}^{(h)} - u_{k+2}^{(2h)}}{2(1-2^p)h}. \tag{7.35}$$

In un metodo a passo adattativo, l'avanzamento avviene con un passo h_k dipendente dall'indice k secondo lo schema

$$u_{k+1} = u_k + h_k \Phi(x_k, u_k, u_{k+1}; h_k),$$

con l'intento di tenere in tutto l'intervallo $[x_0, \bar{x}]$ l'errore di consistenza (stimato mediante (7.35)) tra due soglie opportunamente fissate σ_1 e σ_2, usando un fattore di adattamento $\rho > 1$. Si pone $k = 0$, $h_0 = h$, e si procede secondo l'algoritmo che segue.

1. Calcola $u_{k+2}^{(h_k)}$ e $u_{k+2}^{(2h_k)}$. Se $x_k + 2h_k \geq \bar{x}$, STOP. Altrimenti, stima $\tau(x_k, u_k, h_k)$ secondo la (7.35).

2. Se $\sigma_1 \leq \|\tau(x_k, u_k, h_k)\| \leq \sigma_2$, poni $x_{k+2} = x_k + 2h_k$, $u_{k+2} = u_{k+2}^{(h_k)}$, $h_{k+2} = h_k$. Incrementa $k \leftarrow k+2$ e vai a 1.

3. Se $\|\tau(x_k, u_k, h_k)\| < \sigma_1$, poni $x_{k+2} = x_k + 2h_k$, $u_{k+2} = u_{k+2}^{(h_k)}$, $h_{k+2} = \rho h_k$. Incrementa $k \leftarrow k+2$ e vai a 1.

4. Se $\|\tau(x_k, u_k, h_k)\| > \sigma_2$, poni $h_k \leftarrow h_k/\rho$ e vai a 1.

7.2 Metodi a più passi

La approssimazione u_{k+1} di $y(x_{k+1})$ si costruisce nei metodi a più passi lineari (o Linear Multistep Methods) in base alle informazioni calcolate negli ultimi p passi, secondo lo schema generale:

$$\begin{cases} u_{k+1} = \sum_{j=0}^{p} a_j u_{k-j} + h \sum_{j=-1}^{p} b_j f(x_{k-j}, u_{k-j}) \\ u_0 = y_0 \\ u_1, \ldots, u_p \ \text{ dati} \end{cases} \tag{7.36}$$

in cui il calcolo dei primi p punti della soluzione si effettua tipicamente con un metodo ad un passo di ordine sufficientemente alto, o mediante sviluppo di Taylor. Lo schema (7.36) è detto schema a $p+1$ passi; se $b_{-1} = 0$ si tratta di uno schema esplicito, mentre se $b_{-1} \neq 0$ lo schema è implicito e u_{k+1} va calcolato risolvendo (7.36), vista come equazione nonlineare in u_{k+1}. Il quadro di riferimento più naturale per studiare lo schema (7.36) è quello delle equazioni alle differenze (vedi §A.6).

Due esempi di metodi multistep, rispettivamente espliciti ed impliciti, sono i seguenti (che rientrano entrambi nella classe dei cosiddetti *metodi di Nyström*):

Metodo del punto medio Approssimando l'integrale che appare nell'equazione di Volterra con la quadratura del punto medio sull'intervallo $[x_{k-1}, x_{k+1}]$, si ottiene lo schema esplicito

$$u_{k+1} = u_{k-1} + 2hf(x_k, u_k)$$

che ha il secondo ordine di consistenza.

Metodo di Simpson Approssimando invece l'integrale dell'equazione di Volterra con la formula di Simpson, ancora sull'intervallo $[x_{k-1}, x_{k+1}]$, si ottiene lo schema implicito

$$u_{k+1} = u_{k-1} + \frac{h}{3}[f(x_{k-1}, u_{k-1}) + 4f(x_k, u_k) + f(x_{k+1}, u_{k+1})]$$

che è di quarto ordine.

In questi due esempi (come anche nei metodi di Adams, che si descriveranno in seguito), si ha $a_{\bar{j}} = 1$, $a_j = 0$ per $j \neq \bar{j}$, e la struttura dello schema diviene

$$u_{k+1} = u_{k-\bar{j}} + h \sum_{j=-1}^{p} b_j f(x_{k-j}, u_{k-j})$$

che si può interpretare come una versione discreta della equazione di Volterra

$$y(x_{k+1}) = y(x_{k-\bar{j}}) + \int_{x_{k-\bar{j}}}^{x_{k+1}} f(x, y(x))dx \tag{7.37}$$

in cui il secondo membro sia stato approssimato con una formula di quadratura. Utilizzando formule di quadratura aperte su x_{k+1} si ottiene uno schema multistep esplicito, mentre con formule chiuse si ottiene uno schema implicito (ad esempio, se $\bar{j} = p = 0$ e si approssima l'integrale su $[x_k, x_{k+1}]$ con una formula dei rettangoli che utilizzi il valore a sinistra, si ritrova il

metodo di Eulero in avanti, con la formula del trapezio si ritrova il metodo di Crank–Nicolson).

Un'altra possibilità è invece quella di avere (come accade nei metodi BDF) più termini nella prima sommatoria, ma uno solo nella seconda, ed in questo caso lo schema si può interpretare come un tentativo di approssimare la derivata, che compare nella formulazione differenziale del problema di Cauchy, in modo più accurato (ovvero con un ordine di consistenza più alto). Lasciando solo il termine $(k+1)$–esimo nella seconda sommatoria si ottiene la forma

$$u_{k+1} = \sum_{j=0}^{p} a_j u_{k-j} + hb_{-1} f(x_{k+1}, u_{k+1})$$

che possiamo riscrivere come

$$\frac{u_{k+1} - \sum_{j=0}^{p} a_j u_{k-j}}{hb_{-1}} = f(x_{k+1}, u_{k+1}) \tag{7.38}$$

in cui il primo membro si può rileggere appunto come una approssimazione della derivata $y'(x_{k+1})$ (nel caso più semplice, quello del rapporto incrementale sinistro, questa strategia fornisce il metodo di Eulero all'indietro).

Nell'analisi dei metodi multistep lineari, un ruolo chiave è giocato dai polinomi

$$\rho(\zeta) = \zeta^{p+1} - \sum_{j=0}^{p} a_j \zeta^{p-j} \tag{7.39}$$

$$\sigma(\zeta) = \sum_{j=-1}^{p} b_j \zeta^{p-j} \tag{7.40}$$

$$P(\zeta) = \rho(\zeta) - h\lambda\sigma(\zeta). \tag{7.41}$$

Si vedrà in particolare che le radici del polinomio ρ sono legate alla zero–stabilità dello schema, mentre quelle del polinomio P alla stabilità assoluta (in corrispondenza ad una certa coppia di valori h e λ).

Complessità I metodi multistep permettono di ottenere ordini di consistenza molto elevati con un basso costo computazionale. In particolare, dalla (7.36) si può notare che ogni passo del metodo richiede solo il calcolo di $f(x_k, u_k)$ (o di $f(x_{k+1}, u_{k+1})$ se il metodo è implicito), essendo già stati calcolati in precedenza gli altri valori $f(x_{k-j}, u_{k-j})$, laddove ad esempio nei metodi di Runge–Kutta tutti gli stadi del metodo vanno ricalcolati ad ogni passo.

La soluzione del sistema associato ai metodi impliciti può essere ancora effettuata per sostituzioni successive, con limitazioni (vedi Teorema 7.14) simili a quelle che si presentano negli schemi ad un passo. Utilizzando invece il metodo di Newton, tecniche di tipo Predictor–Corrector (vedi §7.2.1) rappresenta comunque una ulteriore possibilità di ottenere una buona approssimazione iniziale.

Nel caso di sistemi differenziali lineari del tipo $y' = Ay$, il vettore u_{k+1} si ottiene nei metodi impliciti come soluzione del sistema

$$(I - hb_{-1}A)u_{k+1} = \sum_{j=0}^{p} a_j u_{k-j} + h \sum_{j=0}^{p} b_j A u_{k-j}$$

la cui implementazione efficiente prevede di tenere memorizzati i vettori Au_{k-j} già calcolati in precedenza, come anche di fattorizzare preventivamente la matrice $(I - hb_{-1}A)$.

Risultati fondamentali

- Convergenza dei metodi multistep lineari

  Teorema 7.8 *Se lo schema (7.36) è consistente e zero–stabile e se* $u_k \to y(x_k)$ *per* $k = 1, \ldots, p$ *e* $h \to 0$*, allora, fissato* $\bar{x} > x_0$*, per ogni* $k \in [0, (\bar{x} - x_0)/h]$ *si ha*

  $$\|u_k - y(x_k)\| \to 0$$

  per $h \to 0$*. Se inoltre lo schema è consistente con ordine* q*, i valori di innesco* $u_1, \ldots, u_p$ *sono calcolati con un errore* $O(h^q)$*, e la soluzione* y *è sufficientemente regolare, allora*

  $$\|u_k - y(x_k)\| \leq Ch^q.$$

- Consistenza dei metodi multistep lineari

  Teorema 7.9 *Lo schema (7.36) è consistente con ordine* $q \geq 1$ *se e solo se*

  $$\sum_{j=0}^{p} a_j = 1 \ , \quad \sum_{j=0}^{p} (-j)^i a_j + i \sum_{j=-1}^{p} (-j)^{i-1} b_j = 1 \qquad (7.42)$$

  per $i = 1, \ldots, q$ *(dove si conviene di porre* $(-j)^{i-1} = 1$ *se* $j = 0$ *ed* $i = 1$*).*

Dim. Si tratta di verificare che

$$y(x_{k+1}) = \sum_{j=0}^{p} a_j y(x_{k-j}) + h \sum_{j=-1}^{p} b_j f(x_{k-j}, y(x_{k-j})) + O(h^{q+1}) =$$

$$= \sum_{j=0}^{p} a_j y(x_{k-j}) + h \sum_{j=-1}^{p} b_j y'(x_{k-j}) + O(h^{q+1}). \qquad (7.43)$$

D'altra parte, per $j \neq 0$, i valori $y(x_{k-j})$ e $y'(x_{k-j})$ hanno gli sviluppi di Taylor

$$y(x_{k-j}) = \sum_{i=0}^{q} \frac{(-jh)^i}{i!} y^{(i)}(x_k) + O(h^{q+1}) =$$

$$= \sum_{i=0}^{q} (-j)^i \frac{h^i}{i!} y^{(i)}(x_k) + O(h^{q+1}),$$

$$y'(x_{k-j}) = \sum_{i=1}^{q} \frac{(-jh)^{i-1}}{(i-1)!} y^{(i)}(x_k) + O(h^q) =$$

$$= \frac{1}{h} \sum_{i=1}^{q} i(-j)^{i-1} \frac{h^i}{i!} y^{(i)}(x_k) + O(h^q),$$

che possono includere formalmente anche il caso $j = 0$ (caso in cui non occorre effettuare alcuno sviluppo) qualora si ponga convenzionalmente $(-j)^m = 1$ se $j = m = 0$. Sostituendo questi sviluppi nel secondo membro della (7.43) si ottiene:

$$\sum_{j=0}^{p} a_j \sum_{i=0}^{q} (-j)^i \frac{h^i}{i!} y^{(i)}(x_k) + h \sum_{j=-1}^{p} b_j \frac{1}{h} \sum_{i=1}^{q} i(-j)^{i-1} \frac{h^i}{i!} y^{(i)}(x_k) + O(h^{q+1}) =$$

$$= y(x_k) \sum_{j=0}^{p} a_j + \sum_{i=1}^{q} \frac{h^i}{i!} y^{(i)}(x_k) \left[\sum_{j=0}^{p} (-j)^i a_j + \sum_{j=-1}^{p} i(-j)^{i-1} b_j \right] + O(h^{q+1}) \qquad (7.44)$$

in cui l'ultimo passaggio è ottenuto scambiando le sommatorie e raccogliendo i termini rispetto alle derivate $y^{(i)}(x_k)$. E' immediato verificare che l'ultimo membro di (7.44) è lo sviluppo di Taylor di $y(x_{k+1})$ se e solo se le condizioni (7.42) sono soddisfatte per $i = 1, \ldots, p$.

■

- Massimo ordine di consistenza dei metodi multistep lineari (*prima barriera di Dahlquist*)

Teorema 7.10 *Non esistono metodi multistep a $p+1$ passi, zero–stabili, nella forma (7.36), con ordine di consistenza maggiore di $p+2$ se p è pari, di $p+3$ se p è dispari.*

- Zero–stabilità dei metodi multistep lineari (*condizione delle radici*)

Teorema 7.11 *Un metodo multistep nella forma (7.36) è zero–stabile se e solo se, indicando con ζ_i $(i = 1, \ldots, p+1)$ le radici del polinomio $\rho(\zeta)$ definito dalla (7.39), si ha $|\zeta_i| \leq 1$ per ogni i, ed inoltre tutte le radici tali che $|\zeta_i| = 1$ sono radici semplici.*

- Stabilità assoluta dei metodi multistep lineari

Teorema 7.12 *Un metodo multistep nella forma (7.36) è assolutamente stabile in corrispondenza ad un determinato valore $z = h\lambda \in \mathbb{C}$ se e solo se tutte le radici $\zeta_i(h\lambda)$ $(i = 1, \ldots, p+1)$ del polinomio $P(\zeta)$ definito da (7.41) soddisfano la condizione*

$$|\zeta_i(h\lambda)| < 1. \tag{7.45}$$

Dim. Come sempre, si utilizza il problema modello $f(x, y) = \lambda y$, con $\lambda \in \mathbb{C}$. Dato (si veda l'appendice A.6) che le soluzioni fornite dallo schema sono combinazioni lineari di soluzioni elementari nella forma $u_k = \zeta^k$ (o $u_k = k^m \zeta^k$ per le radici multiple), con $\zeta \in \mathbb{C}$. Utilizzando questa forma nella (7.36), si ha

$$\zeta^{k+1} = \sum_{j=0}^{p} a_j \zeta^{k-j} + h \sum_{j=-1}^{p} b_j \lambda \zeta^{k-j}$$

da cui, raccogliendo il termine ζ^{k-p}, si ottiene la condizione

$$\zeta^{p+1} - \sum_{j=0}^{p} a_j \zeta^{p-j} = h\lambda \sum_{j=-1}^{p} b_j \zeta^{p-j}$$

che si può scrivere come $P(\zeta) = 0$, con $P(\zeta)$ definito in (7.41). Si ha quindi $\lim_k u_k = \lim_k k^m \zeta^k = 0$ se e solo se tutte le radici di $P(\zeta)$ soddisfano $|\zeta_i| < 1$.

■

- A–stabilità dei metodi multistep lineari (*seconda barriera di Dahlquist*)

Teorema 7.13 *Non esistono metodi multistep lineari di ordine maggiore di 2 nella forma (7.36) che siano A–stabili. Se espliciti, i metodi multistep lineari non possono essere A–stabili per alcun ordine.*

- Risolubilità della equazione (7.36)

Teorema 7.14 *Se la funzione $f(x,\cdot)$ è lipschitziana con costante L_f, e se $h < 1/(b_{-1}L_f)$, ad ogni passo l'equazione (7.36) si può risolvere rispetto a u_{k+1} per sostituzioni successive.*

Dim. Del tutto analoga a quella del Teorema 7.7.

■

7.2.1 Metodi di Adams

La prima grande classe di metodi multistep è quella dei cosiddetti *metodi di Adams* per i quali $a_0 = 1$, $a_j = 0$ $(j = 1, \ldots, p)$, e che sono costruiti sulla base della formulazione (7.37) integrando su $[x_k, x_{k+1}]$ il polinomio interpolatore relativo ai valori $f(x_{k-j}, u_{k-j})$. A seconda della natura esplicita o implicita del metodo, si possono dividere in due classi: metodi di Adams–Bashforth e metodi di Adams–Moulton.

Metodi di Adams–Bashforth Per questi metodi $b_{-1} = 0$ e quindi si tratta di metodi espliciti. Il polinomio interpolatore è costruito sui nodi $x_{k-p}, \ldots, x_k$ ed è quindi di grado p. Ad esempio, per $p = 0$ (schema di AB ad un passo) il polinomio interpolatore è costante e identicamente uguale a $f(x_k, u_k)$. Il suo integrale su $[x_k, x_{k+1}]$ vale $hf(x_k, u_k)$ e lo schema risultante è

$$u_{k+1} = u_k + hf(x_k, u_k)$$

ovvero lo schema di Eulero esplicito. Nella tabella 2 sono riportati i coefficienti b_j per schemi fino a quattro passi.

Metodi di Adams–Moulton In questi metodi $b_{-1} \neq 0$; si tratta perciò di metodi impliciti. Il polinomio interpolatore è costruito su $x_{k-p}, \ldots, x_{k+1}$ ed è di grado $p + 1$. Ad esempio, per $p = 0$ (schema di AM ad un passo) il polinomio interpolatore è il polinomio di primo grado passante per i punti

p	b_0	b_1	b_2	b_3
0	1			
1	$\frac{3}{2}$	$-\frac{1}{2}$		
2	$\frac{23}{12}$	$-\frac{16}{12}$	$\frac{5}{12}$	
3	$\frac{55}{24}$	$-\frac{59}{24}$	$\frac{37}{24}$	$-\frac{9}{24}$

Tabella 2: Coefficienti dei metodi di Adams–Bashforth per $p = 0, \ldots, 3$

$\Big(x_k, f(x_k, u_k)\Big)$ e $\Big(x_{k+1}, f(x_{k+1}, u_{k+1})\Big)$. Il suo integrale su $[x_k, x_{k+1}]$ vale $h/2\Big[f(x_k, u_k) + f(x_{k+1}, u_{k+1})\Big]$ e lo schema risultante è

$$u_{k+1} = u_k + \frac{h}{2}\Big[f(x_k, u_k) + f(x_{k+1}, u_{k+1})\Big]$$

ovvero lo schema di Crank–Nicolson. Nella tabella 3 sono riportati i coefficienti b_{-1} e a_j per schemi fino a quattro passi.

Risultati fondamentali

- Consistenza dei metodi di Adams

  Teorema 7.15 *I metodi di Adams a $p+1$ passi sono consistenti con ordine $p+1$ se espliciti, con ordine $p+2$ se impliciti.*

  Dim. La definizione di consistenza, scritta per un metodo di Adams e applicata confrontando (7.37) e (7.36), fornisce la condizione

$$\left\| y(x_k) + \int_{x_k}^{x_{k+1}} f(x, y(x))dx - \left[y(x_k) + h\sum_j b_j f(x_j, y(x_j)) \right] \right\| = h\tau(h)$$

p	b_{-1}	b_0	b_1	b_2	b_3
0	$\frac{1}{2}$	$\frac{1}{2}$			
1	$\frac{5}{12}$	$\frac{8}{12}$	$-\frac{1}{12}$		
2	$\frac{9}{24}$	$\frac{19}{24}$	$-\frac{5}{24}$	$\frac{1}{24}$	
3	$\frac{251}{720}$	$\frac{646}{720}$	$-\frac{264}{720}$	$\frac{106}{720}$	$-\frac{19}{720}$

Tabella 3: Coefficienti dei metodi di Adams–Moulton per $p = 0, \ldots, 3$

che equivale a

$$\left\| \int_{x_k}^{x_{k+1}} f(x, y(x))dx - h \sum_j b_j f(x_j, y(x_j)) \right\| = h\tau(h)$$

in cui il secondo termine a primo membro è l'integrale del polinomio interpolatore associato alla funzione $f(x, y(x))$. Nel caso di schemi di Adams espliciti, questo polinomio è di grado p ed approssima quindi la funzione a meno di infinitesimi $O(h^{p+1})$. L'integrale viene quindi approssimato (ricordando la (6.7)) a meno di infinitesimi $O(h^{p+2})$, e ne risulta che $\tau(h) = O(h^{p+1})$. Per gli schemi impliciti, il polinomio ha grado $p + 1$, e seguendo lo stesso ragionamento si ha $\tau(h) = O(h^{p+2})$.

■

- Zero–stabilità dei metodi di Adams

Teorema 7.16 *I metodi di Adams sono zero–stabili.*

Dim. Si tratta di verificare che la condizione delle radici è soddisfatta. In effetti, il polinomio $\rho(\zeta)$ definito dalla (7.39) ha per tutti i metodi di Adams la forma

$$\rho(\zeta) = \zeta^{p+1} - \zeta^p$$

p	b_{-1}	a_0	a_1	a_2	a_3
0	1	1			
1	$\frac{2}{3}$	$\frac{4}{3}$	$-\frac{1}{3}$		
2	$\frac{6}{11}$	$\frac{18}{11}$	$-\frac{9}{11}$	$\frac{2}{11}$	
3	$\frac{12}{25}$	$\frac{48}{25}$	$-\frac{36}{25}$	$\frac{16}{25}$	$-\frac{3}{25}$

Tabella 4: Coefficienti dei metodi BDF per $p = 0, \ldots, 3$

ed ha quindi una radice di molteplicità p nell'origine, ed una radice semplice in $\zeta = 1$. La condizione delle radici è quindi rispettata.

■

7.2.2 Metodi BDF

Una seconda classe di metodi LMM, utilizzata in particolare nei problemi stiff, è quella dei metodi BDF (acronimo di *Backward Difference Formulae*, formule alle differenze all'indietro). Si tratta di metodi impliciti ($b_{-1} \neq 0$) costruiti uguagliando $f(x_{k+1}, u_{k+1})$ con una approssimazione di $y'(x_{k+1})$ basata sugli ultimi $p+1$ passi. Un modo di costruire questa approssimazione è quello di interpolare i valori $u_{k-p}, \ldots, u_{k+1}$ e derivare il polinomio interpolatore nel punto x_{k+1} (tale approssimazione della derivata ha ordine di consistenza $p + 1$). Ad esempio, se $p = 0$, il polinomio interpolatore in questione è la retta passante per i punti (x_k, u_k) e (x_{k+1}, u_{k+1}), la cui derivata è il rapporto incrementale

$$\frac{u_{k+1} - u_k}{h}$$

e uguagliando questa quantità a $f(x_{k+1}, u_{k+1})$ si ottiene il metodo di Eulero all'indietro. Nella tabella 4 sono riportati i coefficienti a_j e b_{-1} per schemi fino a quattro passi.

Risultati fondamentali

- Consistenza dei metodi BDF

Teorema 7.17 *I metodi BDF sono consistenti, con ordine pari al numero dei passi.*

Dim. Ci limitiamo a dimostrare la consistenza, tenendo conto del fatto che in questo caso il primo membro di (7.38) è la derivata in x_{k+1} del polinomio interpolatore $\Pi_{k+1,\dots,k-p}$ associato ai valori $u_{k+1},\dots,u_{k-p}$. Una volta sostituito $y(x_j)$ a u_j, si tratta quindi di dimostrare che la derivata del polinomio interpolatore in x_{k+1} converge al valore $y'(x_{k+1})$ per $h \to 0$.
Utilizziamo la forma di Newton del polinomio, considerando i nodi nell'ordine inverso da x_{k+1} a x_{k-p}. Ricordiamo intanto che la derivata di un prodotto di funzioni si ottiene sommando un numero corrispondente di termini in cui "si deriva una funzione per volta", ad esempio:

$$D(fgh) = f'gh + fg'h + fgh'.$$

Nel caso dei polinomi $(x-x_{k+1})(x-x_k)\cdots(x-x_{k-m})$, si ha $D(x-x_j) = 1$ e quindi

$$\begin{gathered} D[(x-x_{k+1})(x-x_k)\cdots(x-x_{k-m})] = \\ = (x-x_k)(x-x_{k-1})\cdots(x-x_{k-m}) + \\ +(x-x_{k+1})(x-x_{k-1})\cdots(x-x_{k-m}) + \cdots + \\ +(x-x_{k+1})(x-x_k)\cdots(x-x_{k-m+1}) \end{gathered}$$

da cui si ha, tenendo conto che il passo tra i nodi è h:

$$D\Big[(x-x_{k+1})(x-x_k)\cdots(x-x_{k-m})\Big]_{x=x_{k+1}} =$$

$$= (x_{k+1}-x_k)(x_{k+1}-x_{k-1})\cdots(x_{k+1}-x_{k-m}) = O(h^{m+1}).$$

Derivando in x_{k+1} il polinomio di Newton si ottiene perciò

$$D\Big[\Pi_{k+1,\dots,k-p}(x)\Big]_{x=x_{k+1}} = y[x_{k+1},x_k] +$$

$$+y[x_{k+1},x_k,x_{k-1}]\,O(h) + \cdots + y[x_{k+1},\dots,x_{k-p}]\,O(h^{p+1})$$

ed infine, passando al limite per $h \to 0$ e tenendo conto che $y[x_{k+1},x_k]$ è il rapporto incrementale:

$$\lim_{h\to 0} D\Big[\Pi_{k+1,\dots,k-p}(x)\Big]_{x=x_{k+1}} = y'(x_{k+1})$$

(in cui si è anche tenuto conto del fatto che le differenze divise successive coincidono a meno di costanti con i rapporti incrementali di ordine superiore al primo, e restano quindi limitate per $h \to 0$ se y è regolare).

■

- Stabilità dei metodi BDF

  Teorema 7.18 *I metodi BDF sono zero–stabili per $p \leq 5$. La loro regione di stabilità assoluta è il complementare di un insieme limitato, ed include sempre il semiasse reale negativo.*

7.2.3 Metodi Predictor–Corrector

Nei metodi Predictor–Corrector viene utilizzata l'idea di una soluzione iterativa del metodo implicito (Teorema 7.14), utilizzando però un metodo esplicito per fornire una buona approssimazione iniziale del vettore u_{k+1}. Date le limitazioni che la soluzione iterativa impone sul passo h, applicare questa strategia ai metodi BDF implicherebbe di non utilizzare appieno le loro proprietà di stabilità. Questo problema non si pone invece per i metodi di Adams: in questo caso il calcolo viene effettuato tramite un metodo di Adams–Moulton a p passi (corrector) in cui la approssimazione iniziale per la soluzione iterativa del metodo implicito viene fornita da un metodo di Adams–Bashforth, sempre a p passi (predictor). Se si effettuano N iterazioni nel metodo di sostituzioni successive, lo schema ha la forma:

$$\begin{cases} u_{k+1}^{(0)} = \sum_{j=0}^{p} \alpha_j u_{k-j} + h \sum_{j=0}^{p} \beta_j f(x_{k-j}, u_{k-j}) & \text{(P)}; \\ u_{k+1}^{(n+1)} = \sum_{j=0}^{p} a_j u_{k-j} + h \sum_{j=0}^{p} b_j f(x_{k-j}, u_{k-j}) + h b_{-1} f(x_{k+1}, u_{k+1}^{(n)}) & \text{(C)}; \\ u_{k+1} = u_{k+1}^{(N)}; \end{cases} \tag{7.46}$$

ad esempio accoppiando gli schemi di Eulero esplicito (predictor) e Crank–Nicolson (corrector) con questa modalità si ottiene

$$\begin{cases} u_{k+1}^{(0)} = u_k + h f(x_k, u_k); \\ u_{k+1}^{(n+1)} = u_k + \frac{h}{2}[f(x_k, u_k) + f(x_{k+1}, u_{k+1}^{(n)})]; \\ u_{k+1} = u_{k+1}^{(N)}. \end{cases}$$

Un modo consueto di impiegare questa tecnica consiste nell'effettuare *una sola* iterazione del corrector. In questo caso lo schema risultante è esplicito; ad esempio utilizzando la coppia Eulero/Crank–Nicolson in questo modo si ottiene la forma

$$\begin{cases} u_{k+1}^{(0)} = u_k + h f(x_k, u_k); \\ u_{k+1} = u_k + \frac{h}{2}[f(x_k, u_k) + f(x_{k+1}, u_{k+1}^{(0)})], \end{cases}$$

che corrisponde allo schema (esplicito) di Heun, che è di secondo ordine come il metodo di Crank–Nicolson.

Risultati fondamentali

- Consistenza dei metodi Predictor–Corrector

  Teorema 7.19 *Se nel metodo (7.46) lo schema Predictor è di ordine $q-1$ e lo schema Corrector è di ordine q, allora, per ogni $N \geq 1$, l'ordine di consistenza globale del metodo è q.*

  Dim. Iniziamo verificando l'enunciato per $N = 1$. In questo caso, ponendo $u_{k-j} = y(x_{k-j})$, dalla definizione di consistenza applicata allo schema predictor si ha

  $$u_{k+1}^{(0)} = \sum_{j=0}^{p} \alpha_j y(x_{k-j}) + h \sum_{j=0}^{p} \beta_j f(x_{k-j}, y(x_{k-j})) = y(x_{k+1}) + O(h^q)$$

  che d'altra parte, sostituito nello schema corrector, dà

  $$u_{k+1}^{(1)} = \sum_{j=0}^{p} a_j y(x_{k-j}) + h \sum_{j=0}^{p} b_j f(x_{k-j}, y(x_{k-j})) + h b_{-1} f(x_{k+1}, u_{k+1}^{(0)}) =$$

  $$= \sum_{j=0}^{p} a_j y(x_{k-j}) + h \sum_{j=0}^{p} b_j f(x_{k-j}, y(x_{k-j})) +$$

  $$+ h b_{-1} f(x_{k+1}, y(x_{k+1}) + O(h^q)) =$$

  $$= \sum_{j=0}^{p} a_j y(x_{k-j}) + h \sum_{j=0}^{p} b_j f(x_{k-j}, y(x_{k-j})) +$$

  $$+ h b_{-1} f(x_{k+1}, y(x_{k+1})) + O(h^{q+1}) = y(x_{k+1}) + O(h^{q+1})$$

  in cui nell'ultimo passaggio si è utilizzata la definizione di consistenza di ordine q per lo schema corrector. Raccogliendo i due termini $O(h^{q+1})$, si verifica quindi immediatamente che lo schema risultante è anch'esso consistente con ordine q. Infine, sostituendo di nuovo la approssimazione $u_{k+1}^{(N)}$ all'interno del corrector si verifica facilmente che, se l'enunciato del teorema vale alla iterazione N, allora vale anche per l'iterazione $N+1$ (in questo caso, il termine di resto dominante proviene dall'errore di consistenza del solo corrector).

  ■

7.3 Confronto fra i vari schemi

Poiché in presenza di sufficiente regolarità, gli schemi di ordine più alto permettono errori più piccoli a parità di numero di valutazioni di f, e che tra i metodi di ordine alto quelli a più passi sono computazionalmente meno costosi, sembrerebbe che gli schemi a più passi siano in generale da preferire. Tuttavia, i metodi ad un passo, in particolare i metodi di Runge–Kutta, si prestano meglio ad implementazioni a passo variabile, utili in condizioni di forte disuniformità del campo di velocità della equazione differenziale. Inoltre, i metodi ad un passo non presentano la necessità di essere "innescati", cosa che rende relativamente più complicati i codici di tipo multistep.

7.4 Esercizi sperimentali

- Verificare sperimentalmente la diversità di ordine di convergenza tra i vari metodi alle differenze per EDO.

- Verificare la mancanza di stabilità assoluta dei metodi espliciti quando siano applicati a problemi con autovalori negativi di modulo grande. Verificare che questo problema non si pone nei metodi impliciti A–stabili. Cosa accade in un problema con autovalori negativi in cui solo un autovalore abbia modulo grande?

- Dare un esempio di sistema differenziale con traiettorie chiuse e verificare il comportamento dei vari schemi rispetto alla chiusura delle traiettorie. Darne una spiegazione, almeno geometrica, per gli schemi di primo ordine.

A Alcuni risultati utili

In questa appendice si raccolgono alcuni teoremi che, pur non rientrando strettamente nel campo dell'Analisi Numerica, sono tuttavia di una certa utilità, e d'altra parte non vengono usualmente trattati negli altri corsi di base.

A.1 Matrici trasformanti

Data una operazione di scambio o combinazione lineare di righe o colonne in una matrice A, si può rappresentare questa operazione tramite prodotto per una matrice detta trasformante. Più precisamente, una operazione sulle righe è rappresentata dal prodotto

$$\tilde{A} = TA \tag{A.1}$$

in cui la matrice trasformante T è ottenuta effettuando la stessa operazione sulle righe della matrice identità. In modo analogo, una operazione sulle colonne è rappresentata dal prodotto

$$\tilde{A} = AT \tag{A.2}$$

dove la matrice T è ora ottenuta effettuando la stessa operazione sulle colonne della matrice identità. Dato per buono che per trasposizione una delle due forme è riconducibile all'altra, dimostriamo la (A.1). Indicando con a_i, $\tilde{a}_i$ rispettivamente le colonne di A e di $\tilde{A}$, una trasformazione sulle righe di A viene effettuata indipendentemente su ognuna delle colonne e quindi può essere rappresentata dal prodotto a sinistra, colonna per colonna, per una matrice T:

$$\tilde{a}_i = Ta_i.$$

Considerando la matrice completa, si ottiene quindi (A.1). Infine, poiché $T = TI$, la matrice di trasformazione si ottiene effettuando la trasformazione stessa sulla matrice identità.

A.2 Perturbazione di sistemi lineari

Consideriamo innanzitutto la situazione in cui in un sistema

$$Ax = b \tag{A.3}$$

venga perturbato della quantità δb il vettore b dei termini noti. La soluzione del sistema perturbato soddisferà

$$A(x + \delta x) = b + \delta b \tag{A.4}$$

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

Teorema A.1 *Dato il sistema (A.3), in cui A, x e b soddisfino (A.4), l'errore relativo sulla soluzione è maggiorato da*

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

Dim. Da (A.3), (A.4) si ottiene

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

da cui, utilizzando le proprietà delle norme matriciali,

$$\|\delta x\| \leq \|A^{-1}\| \|\delta b\|. \tag{A.5}$$

Moltiplicando ambo i membri per $\|b\|/\|x\|$ ed usando la maggiorazione $\|b\| \leq \|A\| \, \|x\|$, si ha infine da (A.5):

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

■

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

$$(A + \delta A)(x + \delta x) = b + \delta b \tag{A.6}$$

ed la corrispondente stima dell'errore è data da:

Teorema A.2 *Dato il sistema (A.6), in cui A, x e b soddisfino (A.3) , e supponendo che $\|\delta A\| < 1/\|A^{-1}\|$, l'errore relativo sulla soluzione è maggiorato da*

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

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

A.3 Stime di Gershgorin

Le stime di Gershgorin rappresentano il risultato di localizzazione degli autovalori più semplice e di uso più frequente.

Teorema A.3 *Data una matrice quadrata $A = (a_{ij}) \in \mathbb{C}^{n\times n}$, e definiti i dischi del piano complesso*

$$C_i = \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j\neq i} |a_{ij}| \right\},$$

tutti gli autovalori di A appartengono alla loro unione $\bigcup_i C_i$. Inoltre se l'unione dei dischi C_i ha più componenti connesse disgiunte, ogni componente connessa contiene esattamente tanti autovalori quanti sono i dischi da cui è formata.

Dim. Consideriamo un autovalore λ ed un autovettore u associato tale che $\|u\|_\infty = 1$. Indichiamo con k l'indice della componente di modulo massimo di u, cioè

$$1 = \|u\|_\infty = \max_j |u_j| = |u_k|.$$

Scrivendo ora la k–ma riga della uguaglianza $Au = \lambda u$ ed isolando il termine sulla diagonale si ha

$$(\lambda - a_{kk})u_k = \sum_{j\neq k} a_{kj}u_j$$

da cui passando ai moduli e ricordando che u_k è la componente di massimo modulo, si ottiene finalmente

$$|\lambda - a_{kk}| \leq \sum_{j\neq k} |a_{kj}|\,|u_j| \leq \sum_{j\neq k} |a_{kj}|.$$

Per quanto riguarda la seconda parte dell'enunciato, consideriamo la matrice $D = \mathrm{diag}(a_{jj})$ e, al variare di $t \in [0,1]$, la famiglia di matrici

$$B(t) = D + t(A - D)$$

per la quale si ha $B(0) = D$ (che ha gli autovalori a_{jj}), $B(1) = A$. Poiché gli autovalori di $B(t)$ dipendono in modo continuo dal parametro t, e poiché se due dischi non hanno intersezione per $t = 1$ non ne hanno per alcun valore di $t \in [0,1]$, allora applicando la prima parte dell'enunciato a tutte le matrici della famiglia si ottiene che ogni componente connessa dell'insieme $\bigcup_i C_i$ per

$t = 1$ contiene gli stessi autovalori che contenevano a $t = 0$ i dischi che la compongono.

■

Una stima alternativa si può ottenere sommando i moduli per colonne, come nel teorema seguente.

Teorema A.4 *Data una matrice quadrata $A = (a_{ij}) \in \mathbb{C}^{n\times n}$, e definiti i dischi del piano complesso*

$$C_i = \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j\neq i} |a_{ji}| \right\}.$$

tutti gli autovalori di A appartengono alla loro unione $\bigcup_i C_i$. Inoltre se l'unione dei dischi C_i ha più componenti connesse disgiunte, ogni componente connessa contiene esattamente tanti autovalori quanti sono i dischi da cui è formata.

Dim. Ci si riporta al teorema precedente, una volta notato che A e A^t hanno gli stessi autovalori.

■

A.4 Polinomi di Bernstein

Il teorema di densità dei polinomi in C^0, dimostrato originariamente da Weierstrass in modo non costruttivo, fu più tardi (1912) ridimostrato da Bernstein esibendo una successione costruita esplicitamente di polinomi uniformemente convergenti alla funzione f sull'intervallo di riferimento $[0, 1]$. Il polinomio (detto di Bernstein) di grado n associato alla funzione $f(x)$ sull'intervallo $[0, 1]$ è dato da:

$$B_n[f](x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) \binom{n}{k} x^k (1-x)^{n-k}. \tag{A.7}$$

Vale quindi il teorema:

Teorema A.5 *Data una funzione $f \in C^0([0, 1])$ la famiglia di polinomi (A.7) converge uniformemente ad f per $n \to \infty$.*

I polinomi di Bernstein *non sono polinomi interpolatori*, ed in generale non coincidono in alcun punto con la f; inoltre se f stessa è un polinomio di grado non superiore ad n, *non è vero in generale che $B_n \equiv f$*. Il loro ordine

di convergenza è molto basso, in particolare l'ordine massimo si ottiene se $f \in C^2([0,1])$, nel qual caso si ha

$$\sup_{[0,1]} |f(x) - B_n[f](x)| \leq \frac{C}{n}.$$

A.5 Sistema di Kuhn–Tucker e punti sella

E' noto che le condizioni di stazionarietà vincolata nel caso di vincoli di uguaglianza portano al metodo dei moltiplicatori di Lagrange. L'estensione di questa tecnica al caso dei vincoli di disuguaglianza porta al cosiddetto metodo dei moltiplicatori di Kuhn–Tucker. Anche in questo caso si utilizza la funzione Lagrangiana $L : \mathbb{R}^{n+m} \to \mathbb{R}$, che viene definita come

$$L(x, \lambda) = f(x) + \sum_i \lambda_i g_i(x) \tag{A.8}$$

Si consideri il problema (4.1), (4.2), e per semplicità si supponga che sia la funzione f che le g_i siano derivabili e strettamente convesse. Allora valgono le seguenti caratterizzazioni del punto di minimo di f vincolato all'insieme S:

Teorema A.6 *Dato il problema di minimo vincolato (4.1), (4.2), con le funzioni $f, g_1, \ldots, g_m \in C^1(\mathbb{R}^n)$ e strettamente convesse, un punto x^* è soluzione se e solo se esiste un vettore λ^* tale che sia soddisfatto il seguente sistema:*

$$\begin{cases} \nabla_x L(x^*, \lambda^*) = \nabla f(x^*) + \sum_i \lambda_i^* \nabla g_i(x^*) = 0 \\ g_i(x^*) \leq 0 \\ \lambda_i^* \geq 0 \qquad\qquad (i = 1, \ldots, m) \\ \lambda_i^* g_i(x^*) = 0. \end{cases} \tag{A.9}$$

Il sistema (A.9) è detto sistema di Kuhn–Tucker. L'ultima condizione del sistema richiede che almeno uno tra λ_i e g_i si annulli nella soluzione ottimale. Quello che ci si aspetta è che se si annulla il moltiplicatore λ_i^*, il punto di minimo soddisfa l'i–simo vincolo con la disuguaglianza stretta (cioè è un punto stazionario interno), mentre se si annulla $g_i(x^*)$, allora λ_i^* diviene il moltiplicatore di Lagrange associato a questo vincolo di uguaglianza. Se si scrive un vincolo di uguaglianza come doppio vincolo di disuguaglianza, allora il sistema (A.9) si riduce al metodo dei moltiplicatori di Lagrange.

La seconda caratterizzazione viene effettuata attraverso la funzione Lagrangiana:

Teorema A.7 *Dato il problema di minimo vincolato (4.1), (4.2), con le funzioni $f, g_1, \ldots, g_m \in C^1(\mathbb{R}^n)$ e strettamente convesse, una coppia (x^*, λ^*) è soluzione del sistema (A.9) se e solo se è un punto di sella per la funzione Lagrangiana L, ovvero se:*

$$\begin{cases} L(x^*, \lambda^*) = \min\limits_{x \in \mathbb{R}^n} L(x, \lambda^*) \\ L(x^*, \lambda^*) = \max\limits_{\lambda \in \mathbb{R}^m_+} L(x^*, \lambda) \end{cases} \tag{A.10}$$

dove $\mathbb{R}^m_+$ è il cono positivo dello spazio $\mathbb{R}^m$.

Si noti che nelle ipotesi di convessità e derivabilità fatte, queste due caratterizzazioni forniscono una unica soluzione, che è il minimo globale vincolato. Le condizioni di Kuhn–Tucker e di punto sella valgono in realtà sotto ipotesi più larghe, pur non fornendo in generale, in altri casi, una condizione necessaria e sufficiente. Si noti anche che tecnicamente la condizione di punto sella non richiede la differenziabilità.

A.6 Equazioni alle differenze lineari a coefficienti costanti

Si indica con il termine *equazione alle differenze lineare a coefficienti costanti di ordine $p + 1 \geq 1$* la relazione (che per semplicità supporremo scalare)

$$v_{k+1} + \alpha_0 v_k + \alpha_1 v_{k-1} + \cdots + \alpha_p v_{k-p} = g(k), \tag{A.11}$$

con $\alpha_0, \ldots, \alpha_p$ costanti reali. Se $g(k) \equiv 0$, l'equazione alle differenze si dirà *omogenea*.

Esplicitando v_{k+1} in (A.11), la soluzione (ovvero la successione v_k) può essere costruita in avanti una volta assegnati $p + 1$ valori iniziali $v_0, \ldots, v_p$.

Valgono per (A.11) risultati analoghi a quelli dimostrati per le equazioni differenziali ordinarie lineari:

- La soluzione generale di (A.11) si ottiene sommando alla soluzione generale del problema omogeneo una soluzione particolare del problema non-omogeneo;
- Le soluzioni del problema omogeneo costituiscono uno spazio vettoriale di dimensione $p + 1$;
- E' possibile (a patto di aumentare la dimensione del problema) riscrivere ogni equazione alle differenze di ordine $p + 1$ in forma di sistema alle differenze di ordine uno.

Nel caso di (A.11), quast'ultima costruzione si effettua nel modo seguente. Si definiscano

$$U_k = \begin{pmatrix} v_{k-p} \\ \vdots \\ v_k \end{pmatrix}, \quad G_k = \begin{pmatrix} g(k-p) \\ \vdots \\ g(k) \end{pmatrix}.$$

L'equazione (A.11) si può riscrivere come un sistema del primo ordine nella forma

$$U_{k+1} = BU_k + G_k$$

dove

$$B = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & 1 \\ -\alpha_p & \alpha_{p-1} & \cdots & \cdots & -\alpha_0 \end{pmatrix}.$$

In questa formulazione, la soluzione del problema omogeneo si può scrivere in modo compatto come

$$U_k = B^k U_0$$

e la struttura generale delle soluzioni si ottiene esaminando gli elementi delle matrici B^k. Poiché, in caso di cambi di base

$$B = T^{-1}\widetilde{B}T,$$

le potenze successive hanno la forma

$$B^k = T^{-1}\widetilde{B}^k T,$$

non è restrittivo considerare una matrice diagonale a blocchi in forma di Jordan, ed in particolare (avendo indicato con ζ un autovalore di B) un singolo blocco di dimensione $q \times q$

$$\widetilde{B} = \begin{pmatrix} \zeta & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \zeta & 1 \\ 0 & \cdots & 0 & \zeta \end{pmatrix},$$

le cui potenze successive, per $k \geq q$, si possono facilmente calcolare come

$$\widetilde{B}^k = \begin{pmatrix} \zeta^k & k\zeta^{k-1} & \cdots & \binom{k}{q-1}\zeta^{k-q+1} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \zeta^k & k\zeta^{k-1} \\ 0 & \cdots & 0 & \zeta^k \end{pmatrix}.$$

Si può quindi riconoscere che gli elementi di B^k sono generati da combinazioni lineari di soluzioni elementari linearmente indipendenti di tipo

$$v_k = \zeta^k$$

nel caso di blocchi di Jordan di dimensione uno, e di tipo

$$v_k = k^m \zeta^k \qquad (m = 0, \ldots, q-1)$$

per blocchi di Jordan di dimensione q.

Si noti che, per $k \to \infty$, entrambi i tipi di soluzione sono asintoticamente stabili per $|\zeta| < 1$. Inoltre, le soluzioni del primo tipo restano limitate per $|\zeta| \leq 1$, mentre quelle del secondo tipo solo per $|\zeta| < 1$.

B Definizioni

Analitica – una funzione $f : \mathbb{R} \to \mathbb{R}$ si dice *analitica* in $[a, b]$ se è sviluppabile in una serie di Taylor convergente su $[a, b]$.

Coercitiva – una funzione $f : \mathbb{R}^n \to \mathbb{R}$ si dice *coercitiva* se

$$\lim_{\|x\|\to\infty} f(x) = +\infty.$$

In questo caso un corollario del teorema di Weierstrass assicura l'esistenza di minimi per f in insiemi chiusi (*anche non limitati*) a patto che f sia continua (o anche semicontinua inferiormente).

Consistente – un metodo ad un passo per Equazioni Differenziali Ordinarie si dice *consistente* (rispettivamente *consistente con ordine* q) se, definito $\bar{u}$ tramite l'equazione

$$\bar{u} = \bar{y} + h\Phi(h, \bar{x}, \bar{y}, \bar{u}),$$

e $y(x)$ come la soluzione di

$$\begin{cases} y'(x) = f(x, y(x)) \\ y(\bar{x}) = \bar{y} \end{cases}$$

(con $f(\cdot, \cdot)$ sufficientemente regolare) si ha $|\bar{u} - y(\bar{x} + h)| = h\tau(\bar{x}, \bar{y}, h)$ con $\tau(\bar{x}, \bar{y}, h) = o(1)$ per $h \to 0$ (rispettivamente, $\tau(\bar{x}, \bar{y}, h) \leq Ch^q$). Nel caso di metodi a più passi, l'analoga proprietà si esprime come

$$\left| y(\bar{x} + h) - \sum_{j=0}^{p} a_j y(\bar{x} - jh) - h \sum_{j=-1}^{p} b_j f(\bar{x} - jh, y(\bar{x} - jh)) \right| = h\tau(\bar{x}, \bar{y}, h)$$

sempre con la condizione $\tau(\bar{x}, \bar{y}, h) = o(1)$ (rispettivamente, $\tau(\bar{x}, \bar{y}, h) \leq Ch^q$). La funzione $\tau(x, y, h)$ si indica come *errore di consistenza* (e può essere comodo denotarla con $\tau(h)$ sottintendendo la uniformità della stima rispetto ad x e y).

Convessa – una funzione $f : \mathbb{R}^n \to \mathbb{R}$ si dice *convessa* se per ogni coppia di punti $x, y \in \mathbb{R}^n$ e per ogni $\theta \in (0, 1)$ si ha

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta) f(y)$$

e *strettamente convessa* se la disuguaglianza precedente è stretta. Una funzione C^2 è convessa se e solo se la sua matrice hessiana è sempre semidefinita positiva; se poi la hessiana è sempre definita positiva allora la funzione è strettamente convessa (in questo caso non vale il viceversa).

Definita positiva (negativa) – una matrice simmetrica A si dice *definita positiva (negativa)* se

$$(Ax, x) > 0 \quad ((Ax, x) < 0)$$

per ogni $x \in \mathbb{R}^n$. Se la disuguaglianza precedente è soddisfatta eventualmente con il segno di uguaglianza la matrice si dice *semidefinita positiva (negativa)*. Se nessuno di questi casi è soddisfatto, allora la matrice si dice *indefinita*. Se A è definita positiva (negativa), allora ha autovalori tutti reali positivi (negativi); se è semidefinita, allora ha anche autovalori nulli; se è indefinita può avere autovalori (reali) positivi, negativi o nulli.

Densa – una famiglia numerabile di funzioni $\{\phi_1, \ldots, \phi_k, \ldots\}$ si dice *densa* in uno spazio funzionale X se per ogni $f \in X$ ed ogni $\varepsilon > 0$ esiste un intero N tale che si abbia

$$\left\| \sum_{k=1}^{N} c_k \phi_k - f \right\|_X < \varepsilon.$$

Dominante diagonale – una matrice A si dice *dominante diagonale per righe (per colonne)* se

$$|a_{ii}| \geq \sum_j |a_{ij}| \qquad \left(|a_{ii}| \geq \sum_k |a_{ki}| \right).$$

Se la disuguaglianza è soddisfatta con il segno stretto, la matrice si dice *strettamente dominante diagonale* (rispettivamente per righe o per colonne).

Integrabile – una funzione $f : [a, b] \to \mathbb{R}$ si dice *Riemann–integrabile* se, posto $x_0 = a$, $x_N = b$ e costruita una generica decomposizione $\{x_1, \ldots, x_{N-1}\}$ di $[a, b]$, al variare di N e degli $N-1$ punti $x_1, \ldots, x_{N-1} \in [a, b]$, le sue somme integrali soddisfano la condizione

$$\sup \sum_{k=1}^{N} m_k (x_k - x_{k-1}) = \inf \sum_{k=1}^{N} M_k (x_k - x_{k-1})$$

dove

$$m_k := \inf_{(x_{k-1}, x_k)} f(x) \ , \quad M_k := \sup_{(x_{k-1}, x_k)} f(x)$$

Norma – una *norma* $\|\cdot\|$ è una applicazione da uno spazio vettoriale X in $\mathbb{R}$ che soddisfa le seguenti proprietà:

$$\|x\| \geq 0\ , \quad \|x\| = 0 \text{ se e solo se } x = 0;$$

$$\|cx\| = |c|\ \|x\| \quad (c \in \mathbb{R});$$

$$\|x + y\| \leq \|x\| + \|y\|.$$

Le tre norme di uso più comune in Analisi Numerica sono la norma euclidea, la norma $\|\cdot\|_\infty$ e la norma $\|\cdot\|_1$, definite da:

$$\|x\|_2 = \left(\sum_i x_i^2\right)^{1/2}\ , \quad \|x\|_\infty = \max_i |x_i|\ , \quad \|x\|_1 = \sum_i |x_i|$$

Se X è lo spazio degli operatori lineari limitati su uno spazio vettoriale Y, allora si richiede di regola che la norma soddisfi anche le ulteriori proprietà

$$\|AB\| \leq \|A\|\ \|B\|;$$

$$\|Ax\|_Y \leq \|A\|\ \|x\|_Y$$

e si indica come *norma naturale* (associata ad una certa norma su Y) la seguente norma su X:

$$\|A\| := \sup_{x \neq 0} \frac{\|Ax\|_Y}{\|x\|_Y}.$$

In particolare, le tre norme matriciali associate rispettivamente alla norma euclidea, alla norma $\|\cdot\|_\infty$ e alla norma $\|\cdot\|_1$ sui vettori sono:

$$\|A\|_2 = \rho(A^t A)^{1/2}\ , \quad \|A\|_\infty = \max_i \sum_j |a_{ij}|\ , \quad \|A\|_1 = \max_j \sum_i |a_{ij}|$$

dove si è indicato con $\rho(\cdot)$ il raggio spettrale (vedi) di una matrice.

Una norma matriciale compatibile con la norma euclidea sui vettori, non coincidente con la norma naturale ma di calcolo più semplice, è la norma di Frobenius:

$$\|A\|_F = \left(\sum_{i,j} |a_{ij}|^2\right)^{1/2}\ .$$

Numero di condizionamento – data una matrice nonsingolare A ed una norma matriciale $\|\cdot\|$, si indica come *numero di condizionamento* di A il numero reale positivo

$$K(A) := \|A\| \, \|A^{-1}\|.$$

Il numero di condizionamento dipende dalla norma usata ed in qualche caso si esplicita questo fatto indicandolo con $K_*(A)$ se riferito ad una data norma $\|\cdot\|_*$. Se la norma è una norma naturale, allora $\|I\| = 1$ e per la submoltiplicatività della norma matriciale si ha

$$1 = \|I\| \leq \|A\| \, \|A^{-1}\| = K(A)$$

e quindi $K(A) \geq 1$. Se $K(A)$ è molto grande la matrice A ed il sistema lineare associato si dicono *malcondizionati*. Nel caso di sistemi nonlineari tipicamente si indicano come malcondizionati sistemi in cui sia tale la matrice Jacobiana del sistema.

Ortogonale (matrice) – una matrice Q si dice *ortogonale* se le sue colonne sono vettori mutuamente ortogonali rispetto al prodotto scalare euclideo. Se le colonne sono anche di norma euclidea unitaria, la sua inversa vale

$$Q^{-1} = Q^t.$$

Ortogonali (funzioni) – le funzioni di una famiglia $\{\phi_k\}$ si dicono *mutuamente ortogonali* rispetto ad un prodotto scalare $(\cdot, \cdot)$ se

$$(\phi_k, \phi_j) \begin{cases} > 0 & \text{se } k = j \\ = 0 & \text{se } k \neq j \end{cases}$$

Pieno – si indica come *piena* una matrice con un numero di elementi non nulli dello stesso ordine del numero totale di elementi della matrice, ovvero $O(n^2)$ al variare della dimensione n. Il corrispondente sistema lineare si indicherà come sistema *pieno*.

Polinomi di Chebyshev – si indica con questo nome la famiglia di polinomi definiti in modo ricorsivo da

$$\begin{cases} T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x) & k \geq 1 \\ T_0 = 1 \\ T_1 = x. \end{cases}$$

Tali polinomi sono mutuamente ortogonali in $L^2_w([-1,1])$ rispetto al peso $w(x) = (1-x^2)^{-1/2}$, più precisamente

$$\int_{-1}^{1} (1-x^2)^{-1/2} T_j(x) T_k(x) dx = \begin{cases} \pi & \text{se } k = j = 0 \\ \pi/2 & \text{se } k = j \neq 0 \\ 0 & \text{se } k \neq j \end{cases}.$$

Si dimostra anche che i polinomi di Chebyshev possono essere posti, per $x \in [-1,1]$, nella forma trigonometrica equivalente

$$T_k(x) = \cos(k \arccos x).$$

Polinomi di Legendre – si indica con questo nome la famiglia di polinomi definiti in modo ricorsivo da

$$\begin{cases} P_{k+1}(x) = \frac{2k+1}{k+1} x P_k(x) - \frac{k}{k+1} P_{k-1}(x) & k \geq 1 \\ P_0 = 1 \\ P_1 = x. \end{cases}$$

Tali polinomi sono mutuamente ortogonali in $L^2([-1,1])$, e più precisamente

$$\int_{-1}^{1} P_j(x) P_k(x) dx = \begin{cases} \frac{1}{k+1/2} & \text{se } k = j \\ 0 & \text{se } k \neq j \end{cases}.$$

Prodotto scalare – si dice *prodotto scalare* su uno spazio vettoriale X una applicazione $(\cdot,\cdot) : X \times X \to \mathbb{R}$ tale che

$$(x,y) = (y,x)$$

$$(x,x) \geq 0 \ , \quad (x,x) = 0 \text{ se e solo se } x = 0;$$

$$(cx,y) = c(x,y) \quad (c \in \mathbb{R});$$

$$(x+y,z) = (x,z) + (y,z).$$

Si dice *norma associata* al prodotto scalare $(\cdot,\cdot)$ la norma definita da

$$\|x\| := \sqrt{(x,x)}$$

Proiezione su un insieme chiuso – si dice *proiezione* di un punto $x \in \mathbb{R}^n$ su di un insieme chiuso $A \subseteq \mathbb{R}^n$ la applicazione $P_A : \mathbb{R}^n \to \mathbb{R}^n$ definita da

$$y = P_A(x) \iff |y - x| = \min_{z \in A} |z - x|.$$

Tale applicazione è in generale multivoca, ma diviene univoca se l'insieme A è convesso.

Raggio spettrale – si chiama *raggio spettrale* ρ di una matrice quadrata A, il massimo modulo dei suoi autovalori:

$$\rho(A) = \max_i |\lambda_i(A)|.$$

Si dimostra che il raggio spettrale $\rho(A)$ è maggiorato da ogni norma naturale di A, e coincide con la norma euclidea nel caso di matrici simmetriche.

Simile – si indica come matrice *simile* ad una data matrice quadrata A, la matrice ottenuta tramite una trasformazione del tipo

$$\tilde{A} = TAT^{-1}.$$

Si dimostra che le due matrici A e $\tilde{A}$ hanno gli stessi autovalori.

Sistema trigonometrico – si dice *sistema trigonometrico* relativo ad un intervallo $[0, T]$ il sistema di funzioni

$$\left\{1, \sin\frac{2\pi x}{T}, \cos\frac{2\pi x}{T}, \ldots, \sin\frac{2\pi kx}{T}, \cos\frac{2\pi kx}{T}, \ldots\right\}.$$

Tale famiglia è densa nello spazio $L^2([0, T])$.

Sparso – si indica come *sparsa* una matrice con un numero di elementi non nulli piccolo rispetto al numero totale di elementi della matrice, ovvero $o(n^2)$ al variare della dimensione n. Il corrispondente sistema lineare si indicherà come sistema *sparso*. Un caso notevole è quello delle matrici a banda per cui si ha che il numero di elementi non nulli è $O(n)$.

Stabilità assoluta – uno schema numerico per Equazioni Differenziali Ordinarie si dice *assolutamente stabile* in corrispondenza ad un numero complesso dato λ e ad un passo fissato h se, applicato al problema modello

$$\begin{cases} y'(x) = \lambda y(x) \\ y(0) = y_0 \end{cases}$$

produce una successione u_k che soddisfa la condizione

$$\lim_{k\to\infty} u_k = 0$$

ovvero se produce soluzioni discrete asintoticamente stabili. L'insieme dei valori $z = h\lambda \in \mathbb{C}$ tali che lo schema è assolutamente stabile in corrispondenza al prodotto dei due valori h e λ è detto *insieme di stabilità assoluta*

dello schema. Se tale insieme include tutto il semipiano dei complessi a parte reale negativa, lo schema si dice *A-stabile*, mentre se include un cono di semiapertura θ intorno al semiasse reale negativo, lo schema si dice θ-stabile. Si dimostra che la condizione di A-stabilità implica quella di zero–stabilità.

Stiff – una Equazione Differenziale Ordinaria $y' = f(x, y)$ si dice *stiff* se la matrice jacobiana del secondo membro ha autovalori negativi di modulo molto grande. Ciò corrisponde alla situazione in cui alcune componenti della soluzione convergono molto velocemente ad un punto di equilibrio.

Unimodale – una funzione $f : \mathbb{R} \to \mathbb{R}$ si dice *unimodale* nell'intervallo $[a_0, b_0]$, se esiste 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 è soddisfatta se f ha minimo in $[a_0, b_0]$ ed è convessa.

Unisolvente – L'insieme di nodi $\{x_0, \ldots, x_n\}$ si dice *unisolvente* per la base di funzioni $\{\phi_0, \ldots, \phi_n\}$ se il problema di imporre che la combinazione lineare $\sum_i c_i \phi_i(x)$ abbia valori assegnati nei nodi $x_0, \ldots, x_n$ ha soluzione unica. Scritto cioè il sistema lineare

$$\sum_{i=0}^{n} c_i \phi_i(x_j) = f_j \ , \quad (j = 0, \ldots, n)$$

la condizione di unisolvenza equivale a richiedere che la matrice

$$\Phi = \begin{pmatrix} \phi_0(x_0) & \cdots & \phi_n(x_0) \\ \vdots & & \vdots \\ \phi_0(x_n) & \cdots & \phi_n(x_n) \end{pmatrix}$$

sia nonsingolare. Nel caso della interpolazione polinomiale la condizione di unisolvenza (in una dimensione) equivale alla condizione di avere nodi distinti. Se inoltre si utilizza la base naturale $\phi_i = x^i$, allora la matrice Φ è una matrice di Vandermonde.

Zero–stabilità – uno schema numerico per Equazioni Differenziali Ordinarie si dice *stabile* (o *zero–stabile*) se perturbazioni piccole dei dati iniziali producono perturbazioni piccole della soluzione. Più formalmente, fissati $x > x_0$, $\varepsilon > 0$ e due insiemi di dati inziali u_i e v_i tali che $\|u_i - v_i\| < \varepsilon$ (per $i = 0, \ldots, p$, se il metodo è a $p+1$ passi), si ha per le corrispondenti soluzioni:

$$\|u_k - v_k\| < C\varepsilon$$

per ogni $k \in 0, \ldots, (x - x_0)/h$, con C dipendente da x ma non da h. Nel caso lineare ciò corrisponde alla equilimitatezza delle soluzioni discrete al variare di h su tutto l'intervallo $[x_0, x]$.

Riferimenti bibliografici

Referenze generali

[1] E. Isaacson, H. B. Keller, *Analysis of numerical methods*, Wiley

[2] V. Comincioli, *Analisi Numerica: metodi, modelli, applicazioni*, Mc Graw–Hill

[3] A. Quarteroni, R. Sacco e F. Saleri, *Matematica Numerica*, Springer

Capitoli 1, 2

[4] D. Bini, M. Capovani, O. Menchi, *Metodi numerici per l'algebra lineare*, Zanichelli

[5] G. Golub, C. F. Van Loan, *Matrix computation*, John Hopkins Press

[6] J. H. Wilkinson, *The algebraic eigenvalue problem*, Oxford University Press

[7] J. Ortega, W. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, Academis Press

Capitoli 3, 4

[8] R. Fletcher, *Practical methods of optimization*, J. Wiley

[9] P. E. Gill, W. Murray, *Practical optimization*, Academic Press

[10] M. Minoux, *Mathematical Programming, theory and algorithms*, J. Wiley

Capitolo 7

[11] C. W. Gear, *Numerical Initial Value Problems in ordinary differential equations*, Prentice–Hall

[12] M. Crouzeix, A. L. Mignot, *Analyse Numèrique des equations differentielles*, Masson