Metodi numerici per equazioni differenziali ordinarie

Mele Giampaolo
appunti del corso della prof Beatrice Meini

10 agosto 2011

Più un modello matematico descrive bene un problema reale, più questo modello è complicato

Indice

Premessa

Questi sono appunti del corso Metodi numerici per equanzioni differenziali ordinarie tenuto presso l'università di Pisa. La maggior parte degli appunti sono stati scritti durante le lezioni e fin ora sono stati revisionati una sola volta, pertanto queste dispense non hanno la pretesa di essere: complete, scritte bene o corrette. Invito chiunque trovi errori o ritenga sia necessario aggiungere/togliere qualcosa a mandarmi una mail all'indirizzo *mele@mail.dm.unipi.it* .
Inoltre, onde sottolineare l'incompletezza di questi appunti, non è trattato il caso dei problemi di tipo Stiff come anche l'aspetto pratico-programmativo, quindi è consigliabile saper usare matlab e sapere a gradi linee come funzionano le varie librerie che risolvono le ODE, ad esempio ODE45, ODE15S. Chiunque volesse aiutarmi e rendere migliori questi appunti può chiedermi il file sorgente.

Mele Giampaolo

Capitolo 1

Problemi ai valori iniziali (IVP)

14/03/2011

1.1 Intruduzione

Spesso dalla modellizzazione di un problema reale nasce la necessità di risolvere un'equazione differenziale della seguente forma

$$x'(t) = f(t, x(t))$$

Dove l'incognita è la funzione x e x' è la derivata di questa funzione rispetto alla variabile temporale t. Assumeremo sempre che x sia abbastanza regolare in modo da porterci applicare i teoremi noti a volte sottointendendo il tutto. E' noto che se abbiamo una condizione iniziale, sotto opportune ipotesi di regolarità della f (ad esempio f lipschitziana o meglio di classe C^1), la soluzione del problema seguente

$$\begin{cases} x'(t) = f(t, x(t)) \\ x(t_0) = x_0 \end{cases}$$

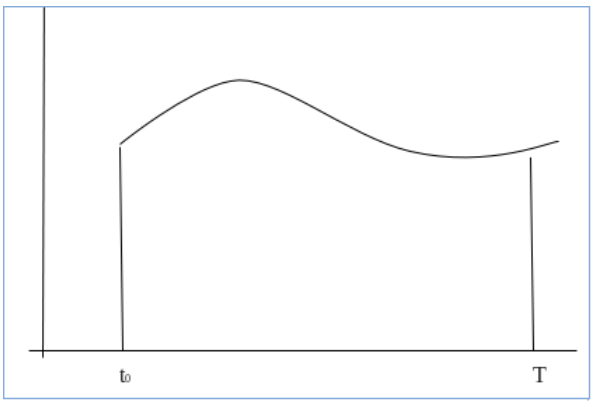
esiste ed è unica. Daltronde quello che in pratica succede è che tanto più il modello che descrive il problema è realistico, tanto più la f è complicata e quindi trovare la funzione x spesso è impossibile, si cercheranno dunque metodi per stimare la soluzione.

Cenni sul metodo di Eulero

Consideriamo al solito il problema

$$\begin{cases} x'(t) = f(t, x(t)) \\ x(t_0) = x_0 \end{cases}$$

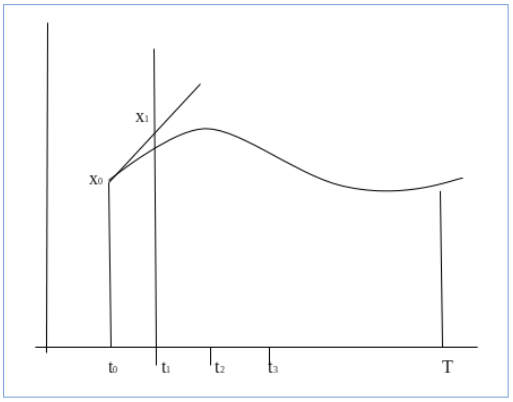
supponiamo di voler trovare la soluzione in un intervallo ristretto $[t_0, T]$



allora possiamo considerare una suddivisione equispaziata:

$$t_0 < t_1 < t_2 < t_3 < ... < t_N = T$$

ovvero chiediamo $t_{k+1} - t_k = \frac{1}{N}$, cerchiamo ora un modo per determinare i valori $x_i = x(t_i)$.
Osserviamo che se conosco $x(t_0)$ allora conosco $x'(t_0)$, quindi conosco la tangente (approssimazione lineare di x in quel punto), quindi a partire da x_0 riesco a stimare x_1, infatti mi basta trovare l'intersezione della tangente in di x in t_0 e intersecarla con la retta parallela all'asse verticale che passa per t_1



riusciamo quindi a costruire una successione di punti x_i con una regola $x_{i+1} = F(x_i)$ dove F è una opportuna funzione. In Realtà il metodo di Eulero fa parte di una famiglia più grande di metodi detti metodi ad un passo, ovvero si genera una successione di punti con la regola

$$x_{i+i} = F(x_i)$$

Dove i punti x_i stimano $x(t_i)$. Si può mostrare che il metodo di Eulero da luogo ad una convergenza puntuale, ovvero x_i tende a stimare sempre meglio il valore assunto realmente da x nel tempo t_i.
La classe ancora più generale di metodi che usano questo approccio è quella dei metodi a k passi, ovvero di usa la regola

$$x_{i+1} = F(x_i, x_{i-1}, ..., x_{i-k+1})$$

Osservazione 1.1. Quando abbiamo considerato la suddivisione con il metodo di Eulero abbiamo scelto la più ingenua, quella equispaziata, daltronde se sapevamo che la soluzione ha un picco in un certo intervallo di tempo è ragionevole inffittire la sudddivisione in quell'intervallo e invece rilassarla su intervalli in cui la soluzione non ha comportamenti insoliti (non ha picchi o non oscilla in modo brusco), daltronde si tratterà più avanti questo aspetto.

Problema a contorno (BVP)

Supponiamo al solito di avere una funzione $f : \Omega \times I \to \mathbb{R}^N$ con le ipotesi di regolarità opportune, dove $\Omega \subset \mathbb{R}^N$ è un aperto e $I \subseteq \mathbb{R}$ è un intervallo, dunque cerchiamo $x : [t_0, T] \to \mathbb{R}^N$ tale che soddisfi

$$\begin{cases} x'(t) = f(t, x(t)) \\ g(x(t_0), x(T)) = 0 \end{cases}$$

dove $g : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^N$.
Questo è noto come problema a contorno o BVP (boundary value problem) e qui non valgono i risultati di esistenza e unicità che si avevano nel caso precedente.

Esempio di BVP

Consideriamo il seguente problema (equazione della molla con $k/g = 1$)

$$\begin{cases} x''(t) = -x(t) \\ x(t_0) = c_1 \\ x'(t_0) = c_2 \end{cases}$$

Questo è un problema con le condizioni iniziali quindi valgono i risultati di esistenza e unicità, la soluzione generale sarà

$$x(t) = \alpha \cos(t) + \beta \sin(t)$$

dalle condizioni a contorno ottengo la soluzione (basta imporre $x(t_0) = c_1$ e $x'(t_0) = c_2$ e risolvere il sistema lineare).

Consideriamo ora il problema BVP (per comodità poniamo $t_0 = 0$)

$$\begin{cases} x''(t) = -x(t) \\ x(0) = c_1 \\ x(T) = c_2 \end{cases}$$

Con un cambio di variabili posso ridurmi al primo ordine

$$y_1(t) = x(t)$$

$$y_2(t) = x'(t)$$

e quindi considerare

$$y(t) = \begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix}$$

$$y'(t) = \begin{pmatrix} y_2(t) \\ -y_1(t) \end{pmatrix} = f(t, y)$$

Le condizioni a contorno diventano

$$g(y(0), y(T)) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y_1(0) \\ y_2(0) \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} y_1(T) \\ y_2(T) \end{pmatrix} - \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = 0$$

Tornando al problema, tenendo conto che la soluzione è sempre della forma

$$x(t) = \alpha \cos(t) + \beta \sin(t)$$

allora dobbiamo imporre

$$\begin{cases} x(0) = \alpha = c_1 \\ x(T) = \alpha \cos(T) + \beta \sin(T) = c_2 \end{cases}$$

Quindi abbiamo il sistema

$$\begin{pmatrix} 1 & 0 \\ \cos(T) & \sin(T) \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$$

Osserviamo subito che se $T \neq k\pi$ allora abbiamo un'unica soluzione (la matrice è invertibile), se invece $T = \pi$ allora $\alpha = c_1$ e $-\alpha = c_2$, quindi se $c_1 \neq -c_2$ non ci sono soluzioni, altrimenti ce ne sono infinite.

Si parlerà più avanti delle DAE (differential algebraic equations), ovvero dei problemi della forma

$$x'(t) = f(t, x(t), z(t))$$

dove z è una funzione che rispetta

$$g(t, x(t), z(t)) = 0$$

1.2 Equazioni alle differenze

16/03/2011 Le equazioni alle differenze sono l'analogo discreto delle equazioni differenziali. Nel caso discreto il tempo non varia in modo continuo ma si avrà una seccussione di tempi $\{t_h\}_{h\in J}$. Tali modelli possono esser usati ad esempio per descrivere la crescita di una popolazione x_0 sarà la popolazione al tempo t_0 e per i tempi successivi

$$x_{n+1} = x_n + \alpha x_n$$

dove α è il tasso di crescita.
E' sensato in questo contesto prendere una successione di tempi discreti piuttosto che continui.
Daltronde le equazioni alle differenze servono anche per approssimare la soluzione del classico problema (continuo)

$$\begin{cases} x' = f(t, x(t)) \\ x(t_0) = x_0 \end{cases}$$

Diamo ora una definizione formale di equazione alle differenze

Sia $J \subseteq \mathbb{N}$ l'insieme dei tempi, $\Omega \subset \mathbb{R}^N$ e $f_i : \Omega \to \Omega$ con $i \in J$, si chiama equazione alle differenze di ordine 1

$$x_{i+1} = f_i(x_i) \quad x_i \in \Omega, \quad i \in J, x_0 \in \Omega$$

Si trova subito che la soluzione è univocamente determinata da x_0.
Per semplicità supporremo sempre $J = \mathbb{N}$.

Si generalizza e si definisce equazione alle differenze di ordine k

$$x_{i+1} = F_i(x_i, x_{i-1}, ..., x_{i-k+1}) \qquad i \geq k-1$$

E' possibile avere anche equazioni alle differenze in forma implicita, ovvero

$$F_i(x_i + 1, x_i, x_{i-1}, ..., x_{i-k+1}) \qquad i \geq k-1$$

Caso lineare

Consideriamo il caso in cui la funzione f_i sia lineare, per comodità consideriamo $N = 1$, quindi avremo $x_i \in \mathbb{R}$.
Dunque un'equazione alle differenze lineare sarà

$$x_{i+1} = \sum_{j=1}^{k} a_{i,j}\ x_{i-j+1} \quad + \quad b_i \qquad a_{i,j} \in \mathbb{R} \quad , \quad b_i \in \mathbb{R}$$

Si osserva subito che fissati i primi $x_0, ..., x_{k-1}$ allora la soluzione esiste ed è unica. E' possibile trasformare questa equazione alle differenze di odine k e renderla di ordine 1, cio comporterà l'aumento della dimensione. Infatti scegliendo

$$X_i = \begin{pmatrix} x_i \\ x_{i+1} \\ \vdots \\ x_{i+k-1} \end{pmatrix}$$

$$A_i = \begin{pmatrix} 0 & 1 & & & & \\ & 0 & 1 & & & \\ & & 0 & 1 & & \\ & & & \ddots & \ddots & \\ & & & & 0 & 1 \\ a_{i+k-1,k} & \cdots & \cdots & \cdots & \cdots & a_{i+k-1,1} \end{pmatrix}$$

$$B_i = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ b_{i+k} \end{pmatrix}$$

Si trova che il problema può esser dunque riscritto come

$$X_{i+i} = A_i \; X_i + B_i$$

Si osserva subito che la matrice A_i è una matrice companion quindi il suo polinomio caratteristico è noto.

Caso lineare con coefficienti costanti

Supponiamo ora che la matrice A_i sia indipendente dai tempi, quindi
$a_{i,j} = a_j$, quindi il problema viene riformulato come

$$x_{i+1} = \sum_{j=1}^{k} a_j x_{i-j+1} + b_i$$

ripetendo lo stesso discorso di prima scriviamo la matrice associata all'equazione

$$A_i = A = \begin{pmatrix} 0 & 1 & & & & \\ & 0 & 1 & & & \\ & & 0 & 1 & & \\ & & & \ddots & \ddots & \\ & & & & 0 & 1 \\ a_k & \dots & \dots & \dots & \dots & a_1 \end{pmatrix}$$

Quindi il problema è

$$X_{i+i} = A \; X_i + B_i$$

Fissiamo le condizioni iniziali

$$x_0 = c_0 \; , \; \dots , \; , x_{k-1} = c_{k-1}$$

Supponiamo che l'equazione sia omogenea, ovvero $B = 0$ e cerchiamo di esprimere le soluzioni.
Per ora tralasciamo le condizioni iniziali, ma consideriamo solo il problema

$$X_{i+i} = A \; X_i$$

Osserviamo che se

$$\left\{ X_i^{(1)} \right\}_{i \geq k-1} \text{ e } \left\{ X_i^{(2)} \right\}_{i \geq k-1}$$

sono soluzioni dell'equazione allora anche le loro combinazioni lineari lo sono

$$\left\{ \alpha X_i^{(1)} + \beta X_i^{(2)} \right\}_{i \geq k-1}$$

Un insieme di soluzioni è linearmente indipendente se: una combinazione lineare è nulla se tutti i coefficienti sono nulli (inteso come per gli spazi vettoriali).

Un sistema fondamentale di soluzioni è un insieme massimale di soluzioni linearmente indipendenti

$$\left\{ X_i^{(j)} \right\}_{i \geq k-1} \qquad j = 1, \dots, k$$

Quindi una generica soluzione sarà della forma

$$x_i = \sum_{j=1}^{k} \alpha_j \; x_i^j$$

dove gli α_j sono dei numeri determinati dalle condizioni iniziali.

Calcolo esplicito del sistema fondamentale di soluzioni Cerchiamo soluzioni della forma $x_i = \lambda^i$, sostituendo si trova

$$\lambda^{i+1} = \sum_{j=1}^{k} a_j \ \lambda^{i-j+1} \qquad i \geq k-1$$

che equivale a

$$\lambda^k = \sum_{j}^{k} a_j \ \lambda^{k-j}$$

definisco dunque il polinomio

$$p(x) = x^k - \sum_{j=1}^{k} a_j \ x^{k-j}$$

quindi le radici di questo polinomio possono esser usate per costruire il sistema di soluzioni fondamentali. Osserviamo che questo altri non è che il polinomio della matrice companion associata all'equazione alle differenze.

Quindi ricapitolando, se $p(\lambda) = 0$ allora $x_i = \lambda^i$ è una soluzione dell'equazione alle differenze. Se $p(x)$ ha tutte le radici distinte $\lambda_1, \lambda_2, \ldots, \lambda_k$ allora, usando che la soluzione ha la forma

$$x_i = \sum_{j=1}^{k} \alpha_j \ \lambda_i^j$$

imponendo le condizioni iniziali (per semplificare la notazione partiamo da x_1 piuttosto che da x_0)

$$x_1 = c_1 \ , \ \ldots, \ , x_k = c_k$$

otteniamo il seguente sistema lineare

$$\begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ \lambda_1 & \lambda_2 & \lambda_3 & \ldots & \lambda_k \\ \lambda_1^2 & \lambda_2^2 & \lambda_3^2 & \ldots & \lambda_k^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda_1^{k-1} & \lambda_2^{k-1} & \lambda_3^{k-1} & \ldots & \lambda_k^{k-1} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_k \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_k \end{pmatrix}$$

La matrice che definisce il sistema è di Vandermonde e sappiamo che è invertibile se $\lambda_i \neq \lambda_j$ per $i \neq j$, che è appunto l'ipotesi sulle radici distinte.

Risolvendo il sistema si trovano i coefficienti α_i e quindi si determina la soluzione.

Se le radici hanno molteplicità maggiore di 1 allora bisogna modificare il ragionamento.

Siano $\lambda_1, \lambda_2, \ldots, \lambda_r$ le radici con molteplicità $m_1, \ldots, m_r$, allora la soluzione generale è

$$x_i = \sum_{j=1}^{r} \sum_{h=1}^{m_j} \alpha_{h,j} \ i^{h-1} \lambda_j^i$$

Diamo solo un'idea del perchè le soluzioni hanno quella forma.
Se una radice λ ha molteplità m allora si avrà $P^{(h)}(\lambda) = 0$ per $h = 0, \ldots, m-1$ (per vederlo basta usare il teorema fondamentale dell'algebra e scrivere il polinomio come prodotto di monomi).
Ad esempio se λ ha molteplicità 2 si avrà che $p(\lambda) = 0$, ma anche $p'(\lambda) = k\lambda^{k-1} - \sum_{j=1}^{k} a_{j,k-1} \ \lambda^{k-j-1} = 0$, quindi λ deve soddisfare anche questa equazione.
Con qualche conto si riesce ad arrivare alla forma generale delle soluzioni scritta prima.
Ma se le radici non sono distinte non si avrà come nel caso precedente un sistema determinato da una matrice di Vandermonde, quindi gli $\alpha_{h,j}$ andranno trovati risolvendo il sistema lineare che di volta in volta si trova.

Stabilità delle equazioni alle differenze

Diremo che una equazione alle differenze è stabile se il polinomio associato $p(x)$ ha radici $|\lambda| \leq 1$, e le radici di modulo 1 hanno molteplicità 1. Diremo che una equazione alle differenze è asintoticamente stabile se $p(x)$ ha radici $|\lambda| < 1$.

Significato Se l'equazione alle differenze è asintoticamente stabile allora tutte le soluzioni sono tali che per il tempo che tende all'infinito $x_i \to 0$, quindi le soluzioni convergono a 0 (indipendente dal fatto che ci siano o meno radici multiple), formalmente

$$\forall x_0, \ldots, x_k \qquad \lim_{i \to \infty} x_i = 0$$

Nel caso in cui l'equazione alle differenze sia stabile non è detto che la soluzione converga, ma sicuramente è limitata, formalmente

$$\forall x_0, \ldots, x_k \qquad |x_i| \leq K$$

Dove K è una costante che dipende dalle condizioni iniziali.

Tutto è conseguenza della forma generale delle soluzioni

$$x_i = \sum_{j=1}^{r} \sum_{h=1}^{m_j} \alpha_{h,j} \; i^{h-1} \lambda_j^i$$

Nel caso asintoticamente stabile tutti i λ_i sono minori di 1 quindi la soluzione va a 0. Nel caso stabile i λ_i che hanno modulo 1 hanno molteplicità 1, quindi il termine i^{h-1} non compare, quindi al limite giro intorno alla circonferenza unitaria (in $\mathbb{C}$), ma comunque la soluzione è limitata.

Osservazione 1.2 (caso non omogeneo)**.** Nel caso non omogeneo i discorsi sulla stabilità e stabilità asintotica si ripetono allo stasso modo dato che una soluzione dell'equazione è $u_i = x_i + y_i$ dove x_i è soluzione dell'omogenea e y_i è una soluzione particolare dell'equazione non omogenea. Non c'è un modo generale per trovare una soluzione particolare dell'equazione non omogenea, ad esempio se B è indipendente dal tempo non è difficile vedere che una soluzione particolare si trova nella forma $x_i = cost$.
Ciò mostra che se l'equazione alle differenze è asintoticamente stabile non è detto che converga a 0, ma convergerà comunque (ad esempio se B è indipendente dal tempo convergerà a alla soluzione particolare). Inoltre la soluzione particolare è indipendente dalle condizioni a contorno $x_0 = c_0, \ldots, x_{k-1} = c_{k-1}$.

Considerazioni Innanzitutto specifichiamo che in genere quando si parla di stabilità si discute la stabilità dei punti, nel nostro caso abbiamo parlato di equazioni stabili sottointendendo che studiavamo la stabilità dell'origine. Inoltre computazionalmente conviene non risolvere l'equazione alle differenze dato che si riesce a trovare con esattezza x_k con le itarazioni. Daltronde avere la soluzione ci permette di studiare cosa succede quando il tempo va all'infinito (quindi per t_h grande), quindi le cose dette sulla stabilità hanno questo scopo. Il concetto di stabilità nasce dalla domanda: cosa succede se perturbo le condizioni iniziali? Chiaramente se l'equazione è asintoticamente stabile dopo molto tempo le soluzioni sono tutte uguali (convergono a 0), mentre nel caso della sola stabilità è necessario uno studio più approfondito.

1.3 Discretizzazione di equazioni differenziali

1.3.1 Metodi di discretizzazione di equazioni differenziali

Come sempre consideriamo il problema

$$\begin{cases} x' = f(t, x(t)) \\ x(t_0) = x_0 \end{cases}$$

E supponiamo di voler stimare la soluzione $x(t)$.
Procediamo come descritto nella prima lezione: discretizziamo l'intervallo

$$t_0 < t_1 < \cdots < t_N = T \quad \text{dove} \quad t_{i+1} = t_i + h_i \qquad h_i > 0$$

(quindi la suddivisione non è necessariamente equispaziata)
approssimiamo l'operatore di derivata con

$$x'(t_i) \simeq \frac{x(t_i + h_i) - x(t_i)}{h_i}$$

Quindi vogliamo stimare i valori che la soluzione assume su i t_i, chiamiamo queste stime $x_i \simeq x(t_i)$ .Inoltre dalle condizioni del problema abbiamo

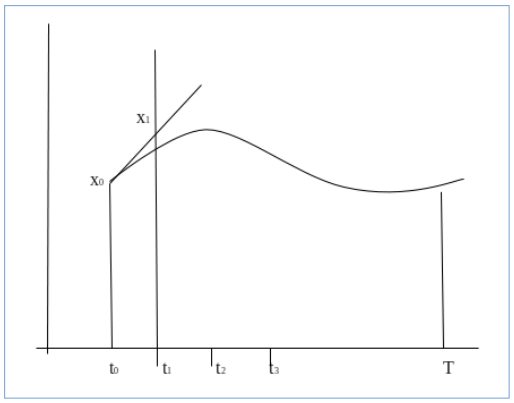
$$x'(t_i) = f(t_i, x(t_i))$$

sostituendo la stima fatta sopra sull'operatore di derivazione otteniamo

$$x_{i+1} = x_i \ + \ h_i \ f(t_i, x_i) \qquad i \geq 0$$

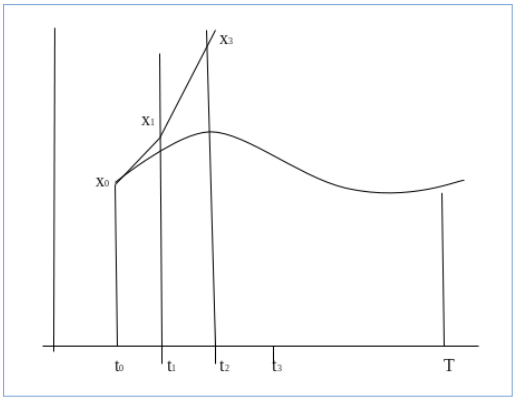
dove è fissato $x_0 = x(t_0)$, quindi abbiamo una equazione alle differenze e per quanto già detto sappiamo che la soluzione esiste ed è unica. Quello appena esposto è il metodo di Eulero.

Osservazione 1.3. Può succedere che andando avanti la stima che si ottiene sia sempre peggiore, infatti supponiamo di aver stimato x_1 a partire da x_0 facendo un passo.



Se facciamo ancora un passo è come risolvere il problema

$$\begin{cases} x' = f(t, x(t)) \\ x(t_1) = x_1 \end{cases}$$



Quindi effettivamente ad ogni passo posso commettere un errore che viene amplificato nei passi successivi, dobbiamo dunque introdurre nuovi tipi di errori per poter analizzare i vari algoritmi di discretizzazione.

1.3.2 Errori di discretizzazione

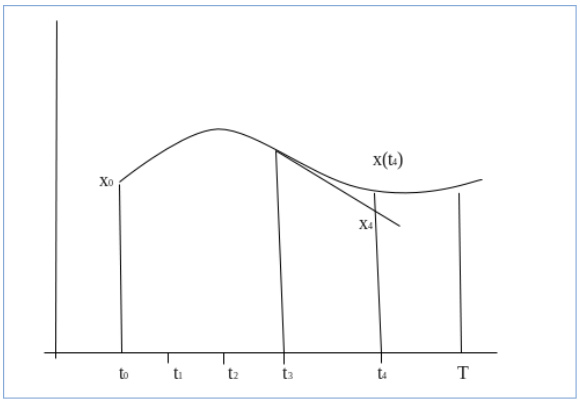
Errore locale di discretizzazione

Introduciamo l'errore locale di discretizzazione per il metodo di Eulero, la generalizzazione sarà ovvia.

$$\delta(x(t+h),h) = \frac{1}{h}\left[x(t+h) - [x(t) + h\ f(t,x(t))]\right]$$

Dove il termine $[x(t) + h\ f(t,x(t))]$ è l'approssimazione ottenuta con un passo a partire dalla soluzione esatta $x(t)$.

Ad esempio, come mostra il disegno, la lunghezza del segmento che unisce $x(t_4)$ ad x_4 è l'errore locale di discretizzazione



Errore globale di discretizzazione

Definiamo l'errore globale di discretizzazione

$$e_i = x(t_i) - x_i$$

La differenza tra l'errore locale e l'errore globale è che nell'errore locale supponiamo di aver fatto soltanto un passo e vediamo quanto abbiamo sbagliato, ad esempio se dobbiamo calcolare l'errore in t_4 supponiamo di aver calcolato bene $x_3 = x(t_3)$ e vediamo l'errore che facciamo nel passo successivo. Ma abbiamo visto prima che gli errori si accumulano, quindi quando calcoliamo x_4 commettiamo sia l'errore che si ottiene con un passo ma dobbiamo tener conto anche degli errori fatti prima per calcolare x_3. L'errore globale tiene conto di ciò, infatti è definito come la differenza tra il valore che assume la funzione al tempo t_i e il valore approssimato x_i.

1.3.3 Metodi consistenti

Diremo che un metodo è consistente se

$$\lim_{h\to 0} \delta(x(t+h),h) \ = \ 0 \qquad\qquad \forall t \in [t_0, T]$$

Sarà di ordine p se

$$\delta(x(t+h),h) = O(h^p)$$

Il metodo di Eulero è consistente infatti usando Taylor

$$x(t+h) = x(t) + hx'(t) + \frac{h^2}{2}x''(\tau)$$

sostituendo $x'(t) = f(t,x(t))$ otteniamo

$$\delta(x(t),h) = \frac{1}{h}\left[h\ f(t,x(t)) + \frac{h^2}{2}\ x''(\tau) - h\ f(t,x(t))\right]$$

sotto buone ipotesi di regolarità che daremo sempre per scontato x'' è limitata, quindi

$$\delta(x(t), h) \leq k\ h$$

Quindi per $h \to 0 \quad \delta(x(t), h) \to 0$ e quindi il metodo è consistente.

Osservazione 1.4. La consistenza del metodo, se la soluzione è abbastanza regolare, non dipende da x ma dal metodo

$$|\delta(x(t), h)| \leq c\ h$$

dove c dipende dal problema.

22/03/2011

Ricapitolazione

L'ultima volta si era considerato il problema

$$\begin{cases} x_{i+1} = \sum_{j=1}^{k} a_j x_{i-j+1} \\ x_0 = c_0\ ,\ \ldots,\ x_{k-1} = c_{k-1} \end{cases}$$

Dalle condizioni a contorno troviamo l'unicità delle soluzioni, infatti se le soluzioni sono della forma

$$x_i = \sum_{r=1}^{k} \alpha_i x_i^{(r)} \qquad i = 0,\ \ldots,\ k-1$$

Allora troviamo il sistema

$$\begin{pmatrix} x_0^{(1)} & x_0^{(2)} & \ldots & x_0^{(k)} \\ x_1^{(1)} & x_1^{(2)} & \ldots & x_1^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k-1}^{(1)} & x_{k-1}^{(2)} & \ldots & x_{k-1}^{(k)} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{pmatrix} = \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_{k-1} \end{pmatrix}$$

La matrice che definisce il sistma è non singolare altrimenti esisterebbe un vettore v nel nucleo che sarebbe combinazione lineare delle colonne della matrice e quindi le soluzioni non sarebbero linearmente indipendenti.

Norme

Definiamo le norme che useremo Se $x \in \mathbb{R}^N$ allora $||x|| = ||x||_2 = \sqrt{\sum_{i=1}^{N} x_i^2}$

Se $x(t) : I \to \mathbb{R}^N$ allora $||x|| = \sup_{t \in I} ||x(t)||_2$

Se $f : I \times \Omega \to \mathbb{R}^N$ allora $||f|| = \sup_{(t,x) \in I \times \Omega} ||x(t)||_2$

Osservazioni sul metodo di Eulero

Torniamo a considerare il problema

$$\begin{cases} x'(t) = f(t, x(t)) \qquad t \in [t_0, T] \\ x(t_0) = x_0 \end{cases}$$

Dove al solito supponiamo f lipschitziana sulla seconda variabile, ovvero

$$||f(t,x) - f(t,y)|| \leq L\ ||x - y|| \qquad \forall t \in I\ ,\ \forall x, y \in \Omega$$

Ricordiamo che il metodo di Eulero consiste nel suddividere l'intervallo $[t_0, T]$

$$t_0 < t_1 < \cdots < t_N = T$$

Poniamo

$$\begin{cases} x_i \simeq x(t_i) \\ x_{i+1} = x_i + h\ f(t_i, x_i) \qquad i = 1, \ldots, N-1 \end{cases}$$

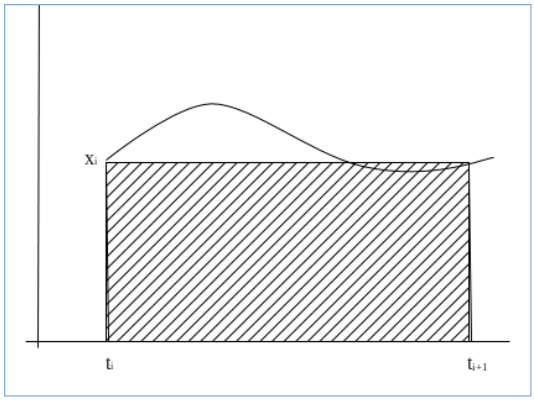
Questo è il metodo di eulero esplicito. Possiamo dare anche un'altra interpretazione del metodo di Eulero, possiamo considerare il problema in forma integrale

$$x(t_{i+1}) = x(t_i) + \int_{t_i}^{t_{i+1}} f(s, x(s))\ ds$$

Quindi l'idea di un metodo per risolvere il problema potrebbe esser quello di stimare l'integrale scitto nella formula sopra ed è quello che effettivamente fa il metodo di Eulero facendo la stima seguente

$$\int_{t_i}^{t_{i+1}} f(s, x(s))\ ds = (t_{i+1} - t_i)\ f(t_i, x_i)$$

Ovvero apprissima l'area sottesa al grafico delle funzione con il rettangolo di base $t_{i+1} - t_i$ e di altezza $f(t_i, x_i)$ come mostrato in figura



Quindi

$$x_{i+1} = x_i + (t_{i+1} - t_i)\ f(t_i, x_i)$$

Si era inoltre visto che

$$||\delta(x(t+h), h)|| \leq M\ h$$

Sotto l'ipotesi $f \in C^2$ abbiamo $M = \sup f''$ e dunque il metodo di Eulero è consistente di ordine 1.

1.3.4 Metodi convergenti

Diremo che un metodo è convergente se l'errore di discretizzazione globale tende a zero, ovvero

$$\lim_{h \to 0} \max_{i=0,\ldots,N-1} ||e_i|| = 0$$

Dove $e_i = x(t_i) - x_i$

Teorema 1.1. Il metodo di Eulero è convergente.

Generalizzazione:
per tutti i metodi ad un passo la consistenza di un metodo implica la convergenza.

Dimostrazione. Per comodità supponiamo la suddivisione equispaziata, ovvero $t_{i+i} = h + t_i$ allora

$$e_{i+1} = x(t_{i+1}) - x_{i+1} = x(t_i + h) - x_{i+1} = x(t_i) + h\ f(t_i, x(t_i)) + h\ \delta(x(t_{i+1}), h) - x_i - h\ f(t_i, x_i)$$

Dove si è sostituito usando la relazione data dal metodo di Eulero, possiamo riaccorpare il tutto in modo da scrivere

$$e_{i+1} = e_i + h\ (f(t_i, x(t_i)) - f(t_i, x_i)) + h\ \delta(x(t_{i+1}, h))$$

A questo punto possiamo passare alle norme e usare la lipschitzianità

$$||e_{i+1}|| \leq ||e_i|| + h\ ||f(t_i, x(t_i)) - f(t_i, x_i)|| + h\ ||\delta(x(t_{i+1}, h))|| \leq (1 + h\ L)||e_i|| + h\ ||\delta(x(t_{i+1}, h))||$$

Posso togliere la dipendenza da t nell'errore locale (lo abbiamo visto nella lezione precedente) e quindi

$$||\delta(x(t + h), h))|| \leq \tau(h) \quad \text{dove} \quad \tau(h) \to 0 \quad \text{per} \quad h \to 0$$

In conclusione abbiamo

$$||e_{i+1}|| \leq (1 + h\ L)\ ||e_i|| + h\ \tau(h)$$

per concludere è necessario il seguente

Lemma 1.1. Sia $\{z_k\}_{k \geq k_0}$ una successione di numeri reali $z_k \in \mathbb{R}$ tale che esistano $\alpha, \beta > 0$ con

$$|z_k| \leq (1 + \alpha)|z_{k-1}| + \beta$$

allora

$$|z_k| \leq e^{\alpha(k-k_0)} \left(|z_{k_0}| + \frac{\beta}{\alpha} \right) - \frac{\beta}{\alpha}$$

(dimostrazione omessa)
Usiamo ora il lemma per concludere, senza perdere di generalità supponiamo $t_0 = 0$, ponendo $\alpha = h\ L$, $\beta = h\ \tau$ e $k_0 = 0$ abbiamo

$$||e_{i+1}|| \leq e^{h\ L\ i} \left(||e_0|| + \frac{\tau}{L} \right) - \frac{\tau}{L} = \frac{\tau}{L} \left(e^{L\ t\ i} - 1 \right) \leq \frac{\tau}{L} \left(e^{L\ T} - 1 \right)$$

Quindi in conclusione

$$||e_{i+1}|| \leq \frac{\tau}{L} \left(e^{L\ T} - 1 \right)$$

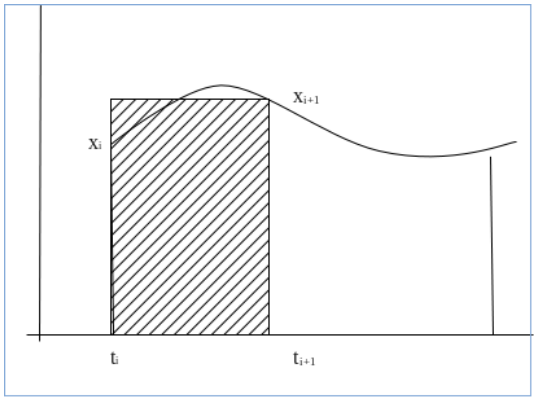
Dato che

$$\frac{\tau}{L} \to 0 \quad \text{per} \quad h \to 0$$

Segue che il metodo converge con lo stesso ordine dell'errore locale, quindi la consistenza implica la convergenza. □

Metodo di Eulero implicito

Riprendendo il discorso fatto prima, ovvero quando abbiamo detto che il metodo di Eulero approssima l'integrale con un rettangolo di altezza $f(t_i)$ e base $t_{i+1} - t_i$, proviamo ora a fare il contrrio, a prendere come altezza $f(t_{i+1})$



Quindi la relazione che abbiamo è

$$x_{i+1} = x_i + h\ f(t_{i+1}, x_{i+1}) \qquad \text{per} \quad i = 0, \dots, N-1$$

Per calcolare x_{i+1} se non so nulla su f devo iterare un metodo di punto fisso

$$x_{i+1}^{(k+1)} = x_i + h\ f(t_{i+1}, x_{i+1}^{(k)}) \quad \text{per} \quad k \geq 0$$

Definendo l'errore locale allo stesso modo praticamente con la stessa dimostrazione si trova che il metodo di Eulero è consistente.

1.4 Stabilità

16/03/2011

1.4.1 Lemma di Gronwall

Il seguente risultato, anche se abbastanza elementare, è molto importante per la teoria che si farà e ci permetterà di fare stime e analisi dell'errore

Lemma 1.2 (Lemma di Gronwall)**.** Sia $x(t) : \mathbb{R} \to \mathbb{R}$ la funzione che soddisfa il problema

$$\begin{cases} x'(t) = a(t)\ x(t) + b(t) & t \geq t_0 \\ x(t_0) = x_0 \end{cases}$$

Dove $a(t), b(t) : \mathbb{R} \to \mathbb{R}$ sono funnzioni continue.
Sia inoltre $y(t) : \mathbb{R} \to \mathbb{R}$ che soddisfa

$$\begin{cases} y'(t) \leq a(t)\ y(t) + b(t) & t \geq t_0 \\ y(t_0) \leq x_0 \end{cases}$$

Allora vale che

$$y(t) \leq x(t) \qquad \forall t \geq t_0$$

Dimostrazione. Consideriamo la differenza $y(t) - x(t)$, vogliamo mostrare che questa differenza è non positiva. Innanzitutto questa differenza soffisfa il seguente problema

$$\begin{cases} y'(t) - x'(t) \leq a(t)\ (y(t) - x(t)) & t \geq t_0 \\ y(t_0) - x(t_0) \leq 0 \end{cases}$$

Definisco

$$\bar{a}(t) = \int_{t_0}^{t} a(t)dt$$

e moltiplico la disuguaglianza per $e^{-\bar{a}(t)}$ ottengo

$$e^{-\bar{a}(t)}(y'(t) - x'(t)) \leq e^{-\bar{a}(t)}\ (y(t) - x(t))\ a(t)$$

chiamo

$$z(t) = e^{-\bar{a}(t)}\ (y(t) - x(t))$$

e osservo che

$$z'(t) = /e^{-\bar{a}(t)}a(t)\ (y(t) - x(t)) + e^{-\bar{a}(t)}(y'(t) - x'(t)) \leq 0$$

quindi

$$z'(t) \leq 0 \qquad \forall t \geq t_0$$

dunque $z(t)$ è una funzione non crescente che in t_0 è non positiva, dunque è non positiva per ogni valore $t \geq t_0$, segue che

$$y(t) - x(t) \leq 0 \qquad \forall t \geq t_0$$

che è appunto la tesi.
Daltronde il problema

$$\begin{cases} x'(t) = a(t)\ x(t) + b(t) & t \geq t_0 \\ x(t_0) = x_0 \end{cases}$$

ha soluzione esplicita

$$x(t) = e^{\bar{a}(t)}\ \left(x_0 + \int_{t_0}^{t} e^{-\bar{a}(s)}\ b(s)\ ds\right)$$

e useremo più avanti questo fatto. □

1.4.2 Problemi perturbati e stima dell'errore

Teorema 1.2 (Teorema di perturbazione)**.**

Consideriamo il problema

$$\begin{cases} x'(t) = f(t, x(t)) & t \in [t_0, T] \\ x(t_0) = x_0 \end{cases}$$

E consideriamo il problema perturbato

$$\begin{cases} y'(t) = f(t, y(t)) + r(t, y(t)) & t \in [t_0, T] \\ y(t_0) = x_0 + z_0 \end{cases}$$

Dove r è una funzione che ,in qualche senso da specificare, è piccola e rappresenta la pertubazione e z_0 rappresenta la perturbazione del dato iniziale. Allora vale che

$$||y(t) - x(t)||e^{L(t-t_0)}\ ||z_0|| + \frac{M}{L}\left(e^{L(t-t_0)} - 1\right)$$

Osservazione 1.5. Il metodo di Eulero opera proprio in questo senso, ad ogni passo perturbiamo il dato iniziale.

Dimostrazione. Considero al solito la funzione differenza

$$z(t) = y(t) - x(t)$$

questa funzione soddisfa

$$\begin{cases} z'(t) = (f(t, y(t)) - f(t, x(t))) + r(t, y(t)) \\ z(t_0) = z_0 \end{cases}$$

Considero questo problema in forma integrale

$$z(t) = z_0 + \int_{t_0}^{t} [f(s, y(s)) - f(s, x(s)) + r(s, y(s))]\, ds$$

passo alla norma

$$||z(t)|| \leq ||z_0|| + \int_{t_0}^{t} [L||y(s) - x(s)|| + ||r(s, y(s))||]\, ds$$

Dove si è usato tutto insieme che la norma dell'integrale è più piccola dell'integrale della norma, poi si è usato il fatto che f è L-lipschitziana e infine la disuguaglianza triangolare.

Poniamo

$$g(t) = ||z_0|| + \int_{t_0}^{t} [\ \ L||z(s)|| + ||r(s, y(s))||\ \ ]\, ds$$

e quindi abbiamo

$$||z(t)|| \leq g(t)$$

calcolando la derivata troviamo

$$g'(t) = L||z(t)|| + ||r(t, y(t))|| \leq Lg(t) + ||r(t, y(t))||$$

quindi abbiamo che $g(t)$ soddisfa il problema

$$\begin{cases} g'(t) \leq L\ g(t) + ||r(t, y(t))|| \\ g(t_0) = ||z_0|| \end{cases}$$

Applico il lemma di Gronwall con

$$g(t) = y(t) \qquad a(t) = L \qquad b(t) = ||r(t, y(t))||$$

e quindi ottengo

$$g(t) \leq e^{L(t-t_0)} \left(||z_0|| + \int_{t_0}^{t} e^{-L(s-t_0)}\ ||r(s, y(s))||\ ds \right)$$

se r è una perturbazione è ragionevole supporre che

$$||r||_{I\times\Omega} \leq M$$

quindi

$$g(t) \leq e^{L(t-t_0)}\ ||z_0|| + \frac{M}{L} \left(e^{L(t-t_0)} - 1 \right)$$

daltronde avevamo che $||z(t)|| \leq g(t)$ quindi

$$||z(t)|| \leq e^{L(t-t_0)}\ ||z_0|| + \frac{M}{L} \left(e^{L(t-t_0)} - 1 \right)$$

Pertanto l'errore si amplifica tanto più quanto t è lontano da t_0, quindi l'idea per non far amplificare l'errore è quella di prendere degli intervalli $[t_0, T]$ molto piccoli. Questo risultato è il migliore che si può avere al caso generale, infatti esistono alcuni problemi per cui la disuguaglianza diventa proprio un'uguaglianza come vedremo più avanti. □

Esempio

$$\begin{cases} x'(t) = a\ x(t) \\ x(0) = 1 \end{cases}$$

Sappiamo già che la soluzione è

$$x(t) = e^{at}$$

Condieriamo il problema perturbato

$$\begin{cases} y'(t) = a\ y(t) \\ y(0) = 1 + \epsilon \end{cases}$$

la soluzione del problema perturbato sarà

$$y(t) = (1 + \epsilon)\ e^{at}$$

la funzione $f(t, x) = a\ x$ è $|a|$-lipschitziana, secondo la stima che abbiamo trovato prima

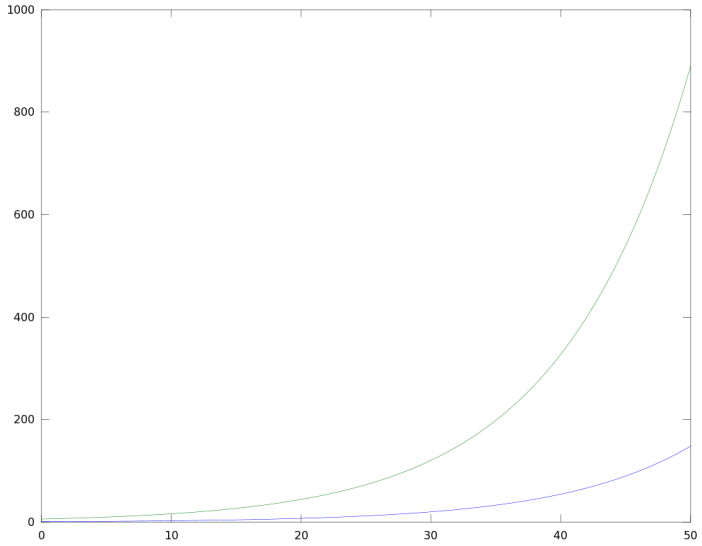
$$|y(t) - x(t)| \leq e^{|a|\ t}\ |\epsilon|$$

daltronde calcolando la differenza

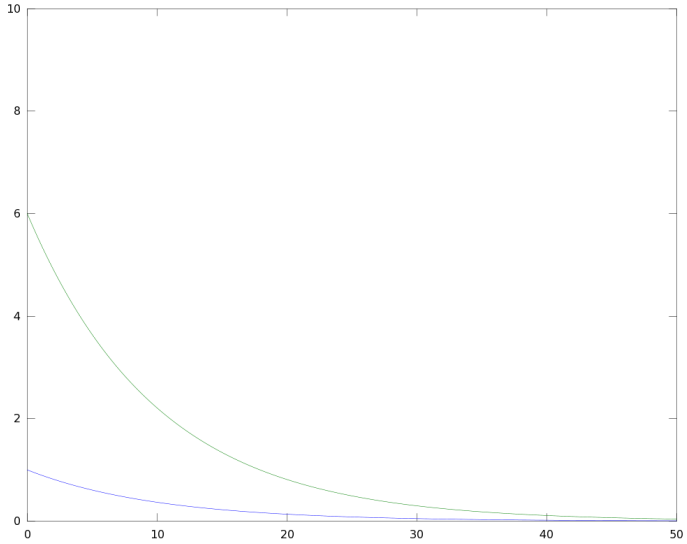
$$|y(t) - x(t)| = |\epsilon| e^{at}$$

Dunque si presentano tre casi

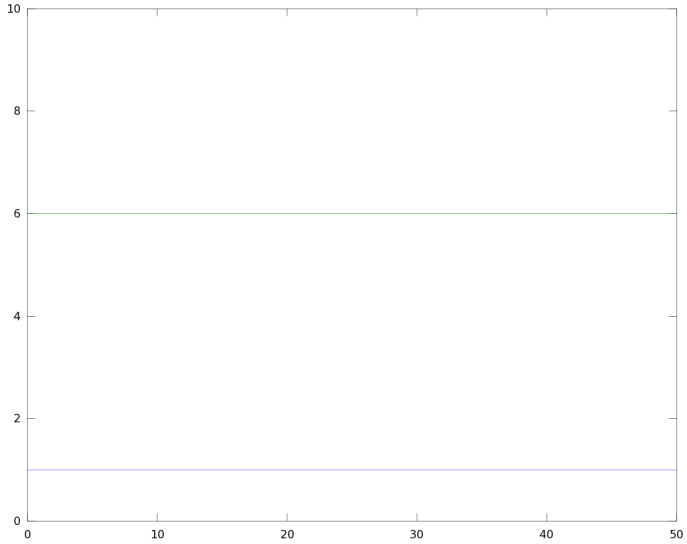
- $a > 0$
  In questo caso la disuguaglianza generale diventa proprio un'uguaglianza (il caso peggiore che ci possa essere) e le soluzioni si allontanano sempre di più, siamo quindi in una condizione di instabilità, ovvero la soluzione vera e la soluzione del problema perturbato si allontanano sempre di più.



- $a < 0$
  In questo caso la disuguaglianza è blanda e non ci da informazioni dato che siamo in una situazione di stabilità asintotica, ovvero la distanza tra le soluzioni dei due problemi diminuisce esponenzialmente.



- $a = 0$
  Anche in questo caso la disuguaglianza non ci da informazioni, siamo in una situazione di stabilità, la distanza tra le soluzioni dei due problemi è costante.



Esempio

Consideriamo ora lo stesso tipo di problema ma nel caso complesso (ciò ci servirà nel caso matriciale)

$$\begin{cases} x'(t) = a\ x(t) & a \in \mathbb{C} \\ x(0) = 1 \end{cases}$$

e il problema perturbato

$$\begin{cases} y'(t) = a\ y(t) & a \in \mathbb{C} \\ y(0) = 1 + \epsilon \end{cases}$$

come prima scriviamo la differenza tra le soluzioni

$$|y(t) - x(t)| = |\epsilon|\ |e^{at}| = |\epsilon||e^{t\ Re(a)+i\ t\ Im(a)}| = |\epsilon|\ e^{t\ Re(a)}$$

Non è difficile capire che siamo nella stessa situazione di prima, ovvero

- $Re(a) < 0$
  stabilità asintotica, la differenza tra le due soluzioni va a zero in tempo esponenziale
- $Re(a) > 0$
  instabilità, la differenza tra le soluzioni va ad infinito in tempo esponenziale
- $Re(a) = 0$
  stabilità, la differenza tra le soluzioni è costante

1.4.3 Sistemi lineari

Partendo dagli esempi precedenti possiamo ora considerare il problema nel caso vettoriale

$$\begin{cases} x'(t) = A\ x(t) \\ x(0) = x_0 \end{cases}$$

Dove ora abbiamo $x(t) : I \to \mathbb{R}^N$ ed $A \in \mathbb{R}^{N\times N}$.
Consideriamo il problema perturbato

$$\begin{cases} y'(t) = A\ y(t) \\ y(0) = y_0 + \epsilon \end{cases}$$

Allora si presentano due casi A *diagonalizzabile* Allora esisterà una matrice di cambiamento di base V tale che

$$A = V\ D\ V^{-1}$$

dove

$$D = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix}$$

Allora il problema lo riscriviamo come $x' = V\ D\ V^{-1}$ Moltiplicando a sinistra per V^{-1}

$$(V^{-1}\ x') = D\ (V^{-1}\ x)$$

Quindi facciamo il cambiamento di coordinate

$$\bar{x} = V^{-1}\ x(t)$$

a questo punto con questo cambiamento ho tante equanzioni lineari come mostrate nei due esempi precedenti,

$$\bar{x}' = D\ \bar{x}$$

ovvero componente per componente avrò

$$\bar{x_i}'(t) = \lambda_i\ x(t)$$

quindi le risolvo tutte e trovo la soluzione

$$\bar{x}(t) = (\bar{x_1}(t), \ldots, \bar{x_n}(t))$$

Daltronde questa è la soluzione nel nuovo sistema di coordinate e non la soluzione del problema originale, quindi ricambio coordinate

$$x(t) = V\ \bar{x}(t)$$

Ad ogni modo se vogliamo studiare la stabilità delle soluzioni, come mostrato, a meno di cambiare base possiamo considerare A già diagonale (il cambiamento di base non altera le proprietà di stabilità di una soluzione), quindi in base agli autovalori λ_i della matrice avremo

- $Re(\lambda_i) \leq 0$ Stabilità, $||y(t) - x(t)|| = cte$
- $Re(\lambda_i) < 0$ Stabilità asintotica, $||y(t) - x(t)|| \to 0$
- $Re(\lambda_i) > 0$ Instabilità, $||y(t) - x(t)|| \to \infty$

Richiamo Possiamo in generale considerare il problema

$$\begin{cases} x'(t) = A\ x(t) \\ x(0) = x_0 \end{cases}$$

un teorema dice che la soluzione è

$$x(t) = e^{tA}\ x_0$$

Dove definiamo l'esponenziale di matrice come

$$e^A = \sum_{i=0}^{\infty} \frac{A^i}{i!}$$

si può dimostrare che questa serie converge e che quella scritta sopra è veramente la soluzione, useremo questo fatto più avanti.

A non diagonalizzabile Se A non è diagonalizzabile allora avrà una forma canonica di Jordan, a meno di cambiamento di coordinate dunque possiamo pensare A già in forma di Jordan, daltronde se siamo in questo caso la matrice A è una matrice a blocchi, dove ogni blocco è un blocco di Jordan, quindi ogni blocco definisce il problema a se stante, quindi dobbiamo risolvere un problema per ogni blocco.
Di conseguenza possiamo considerare A come un singolo blocco di Jordan

$$A = \begin{pmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda & 1 \\ & & & & \lambda \end{pmatrix}$$

Per il richiamo abbiamo che se $Re(\lambda) < 0$ allora $x_0\ e^{tA} \to 0$ quindi ho stabilità, se invece $Re(\lambda) = 0$ e la dimensione del blocco è più grande di 1 allora si ha $x_0\ e^{tA} \to \infty$ quindi instabilità.
Generalizzazione

- Se tutti gli autovalori hanno parte reale negativa allora ho stabilità asintotica
- Se ci sono tutti gli autovalori hanno parte reale non positiva e quelli con parte reale nulla hanno blocchi di Jordan di dimensione 1 allora ho stabilità
- In tutti gli altri casi ho instabilità

Problema perturbato, caso generale

Torniamo ora a considerare il solito problema

$$\begin{cases} x(t) = f(t, x(t)) \quad t \in [t_0, T] \\ x(t_0) = x_0 \end{cases}$$

sia $t_0 < t_1 < T$, dove t_1 è molto vicino a t_0, dunque se faccio un passo stimo $x(t_1)$, al passo successivo dovrò risolvere il problema

$$\begin{cases} y(t) = f(t, y(t)) \quad t \in [t_0, T] \\ y(t_1) = x(t_1) + z_1 \end{cases}$$

dove chiaramente z_1 è un numero piccolo e rappresenta la perturbazione. Voglio fare lo stesso tipo di studio che ho fatto prima ma questa volta voglio vedere come si comporta localmente la differenza dei due problemi. Sia dunque $z(t) = y(t) - x(t)$, allora questa soddisferà

$$\begin{cases} z'(t) = f(t, y(t)) - f(t, x(t)) & t \in [t_0, T] \\ z(t_1) = z_1 \end{cases}$$

Daltronde possiamo usare il teorema di Lagrange in R^n (è uguale a quello in R ma al posto della derivata c'è il gradiente e al posto del prodotto c'è il prodotto scalare)

$$f(t, y(t)) - f(t, x(t)) = J(\xi(t)) \cdot z(t) + g(t, z(t))$$

Dove Lagrange si è usato solo sul secondo argomento e $\xi(t)$ è un opportuno punto tra $x(t)$ e $y(t)$ e la funzione g è il resto di Lagrange (infinitesima per $x(t)$ e $y(t)$ vicini).
Quindi se scelgo t_1 molto vicino a t_0 posso approssimare

$$||f(t, y) - f(t, x)|| \simeq J(t_0) \cdot z(t)$$

Quindi almenlo localmente sarà siddisfatta

$$z'(t) = J(t_0)z(t)$$

E questo è il caso lineare trattato prima.
Daltronde questi ragionamenti non ci portano lontano dato che se un punto è stabile localmente non ci da informazioni dato che non sappiamo quanto t_1 deve esser vicino a t_0.

1.4.4 A-stabilità

Iniziamo subito con il dire che il concetto di A-stabilità non si applica ad un problema ma ad un metodo. Consideriamo il problema test

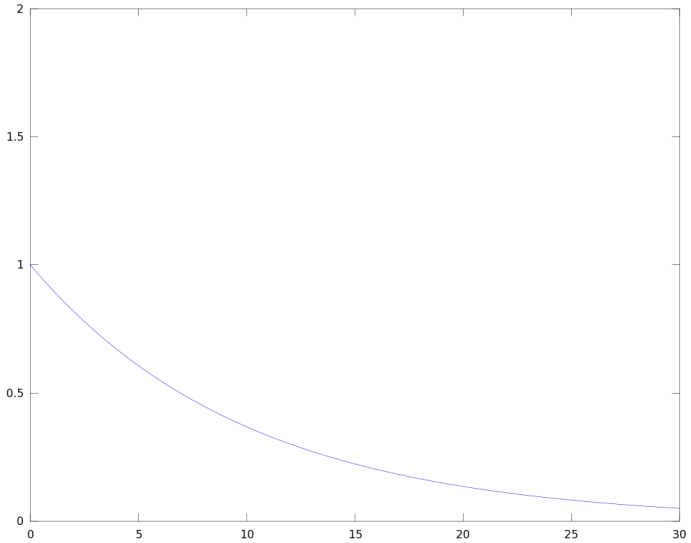
$$\begin{cases} x'(t) = \lambda\ x(t) & \text{con} \qquad Re(\lambda) < 0 \\ x(0) = 1 \end{cases}$$

Sappiamo che il metodo di Eulero genera una successione

$$x_{i+1} = (1 + h\lambda)\ x_i$$

Daltronde sappiamo anche che la soluzione del problema è

$$x(t) = e^{\lambda\ t} \to 0$$



Quindi è sensato chiedere che il metodo rispetti

$$|x_{i+1}| \leq |x_i|$$

se ciò è rispetto allora diremo che il metodo è A-stabile. Quindi nel caso del metodo di Eulero chiederò

$$|1 + h\lambda| \ |x_i| \leq |x_i|$$

e quindi

$$|1 + h\lambda| \leq 1$$

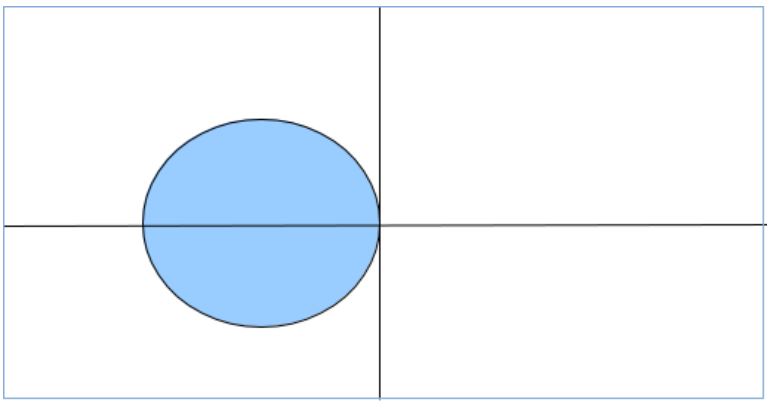
Pultroppo, se il punto è asintoticamente stabile $\lambda < 0$ e tanto più λ è grande in modulo, tanto più la soluzione va a zero velocemente, ma affinchè venga rispettata la proprietà appena vista allora h deve esser piccolissimo e pultroppo le macchine hanno un limite.

1.4.5 Regione di stabilità assoluta

Per il metodo di Eulero chiameremo regione di stabilità assoluta

$$S = \{z = \lambda h \quad \text{tale che} \quad |1 + z| \leq 1\}$$

Ovvero z si trova nel cechio di centro $(-1, 0)$ e raggio 1 nel piano complesso



Generalizzazione Dato un metodo ad un passo definito da

$$x_{i+1} = \phi(h\lambda) \ x_i$$

chiameremo regione di stabilità assoluta

$$L = \{z = h\lambda \quad \text{tale che} \quad |\phi(z)| \leq 1\}$$

Esercizio proposto

Trovare la regione di stabilità del metodo di Eulero implicito.
29/03/2011

1.4.6 A-stabilità del metodo di Eulero implicito

Ricordiamo che la A-stabilità è una proprietà di un metodo, ovvero si considera il problema

$$\begin{cases} x'(t) = \lambda \ x(t) \quad \text{con} \quad Re(\lambda) < 0 \\ x(0) = 1 \end{cases}$$

Quindi sappiamo che la soluzione è $e^{\lambda \ t}$ e questa è decrescente, quindi voglio che anche l'approssimazione conservi questa proprietà, ovvero che

$$|x_{i+1}| \leq |x_i|$$

Abbiamo calcolato per il metodo di Eulero esplicito la regione di stabilità.
Facciamo lo stesso conto per il metodo di Eulero esplicito

$$x_{i+1} = x_i + h\lambda x_{i+1}$$

che diventa

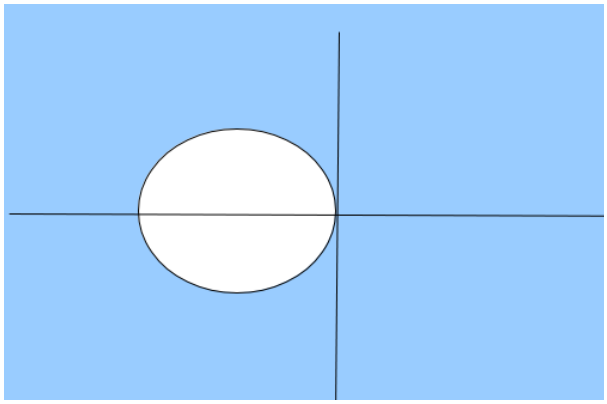
$$x_{i+1} = \frac{x_i}{1 - h\lambda}$$

Se chiediamo che il metodo di Eulero esplicito rispetti la proprietà di decrescenza otteniamo come condizione

$$|1 - h\lambda| \geq 1$$

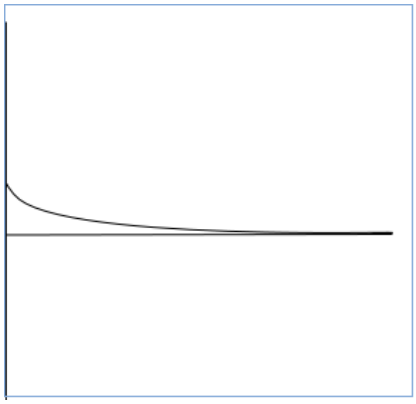
Quindi la regione di stabilità sarà

$$L = \{z \in \mathbb{C} \quad \text{tale che} \quad |1 - z| \geq 1\}$$



Quindi la regione di stabilità del metodo di Eulero implicito è molto più grande, il che ci fa intuire che effettivamente sia preferibile a quello esplicito.

Osservazione 1.6. Se consideriamo il solito problema che prendiamo per la A-stabilità e poniamo $Re(\lambda) < 0$ con $|Re(\lambda)| >> 1$, abbiamo che la soluzione si approssima bene con la funzione nulla.



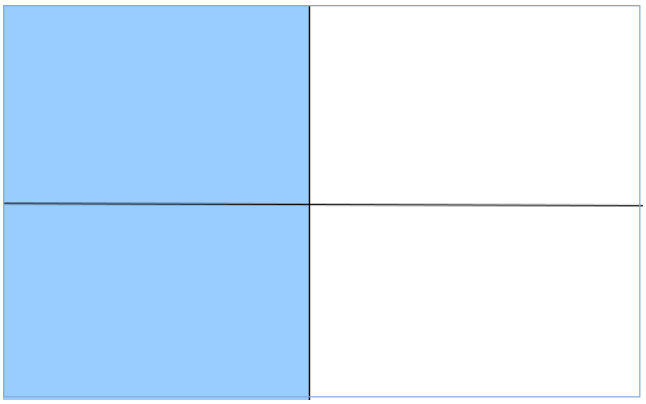
Quindi mi aspetto che sia facile risolvere il problema, daltronde se uso il metodo di Eulero esplicito devo prendere un h molto piccolo per trovarmi nella regione di stabilità quindi il calcolo diventa lungo e se supero la precisione di macchina non è possibile tracciare un'approssimazione coerente della soluzione.
Se invece considero il metodo di Eulero implicito allora senza imporre condizioni su h siamo sempre nella regione di stabilità, ovvero

$$\forall h > 0 \quad \forall \lambda \quad \text{tale che} \quad Re(\lambda) < 0 \quad \text{si ha} \quad z = h\lambda \in L$$

Quindi possiamo scegliere anche un h grande ma ci troviamo sempre nella regione di stabilità. Quindi a volte fa comodo usare metodi impliciti.

1.4.7 Metodi A-stabili

Diremo che un metodo è A-stabile se $\mathbb{C}_< \subseteq L$, dove L è la regione di stabilità e $\mathbb{C}_< = \{z \in \mathbb{C} \text{ tale che } Re(z) \leq 0\}$.



Quindi dire che un metodo è A-stabile equivale a dire che non ci sono condizioni da imporre affinche $h\lambda \in L$, ovvero

$$\forall h > 0 \ \ \forall \lambda \text{ tale che } Re(\lambda) < 0 \text{ si ha } z = h\lambda \in L$$

1.5 Metodi di Runge-Kutta

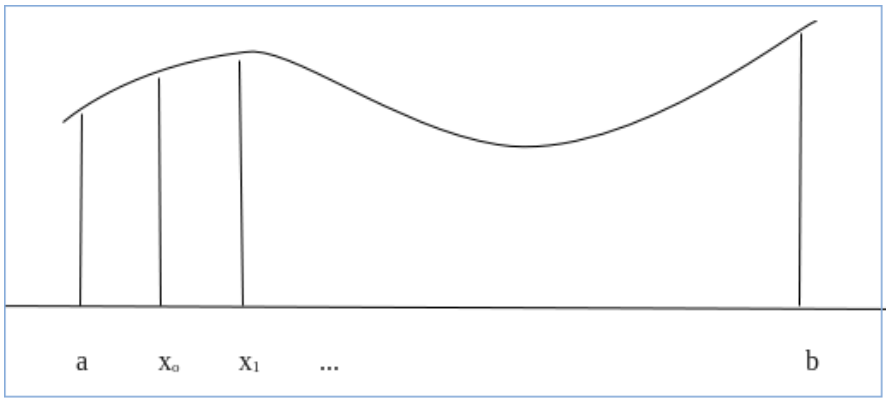
1.5.1 Integrazione approssimata

Sia $f(x) : [a, b] \to \mathbb{R}$ una funzione integrabile, cercheremo di approssimare il suo integrale

$$I(f) = \int_a^b f(x)\ dx$$

Per far ciò al solito considereremo una suddivisione dell'intervallo

$$a = x_0 < x_1 < \cdots < x_N = b$$



I punti x_i sono detti nodi dell'integrazione.

Formula di integrazione approssimata

Definiamo la formula di integrazione approssimata come

$$I(f) \simeq I_{n+1}(f) = \sum_{i=0}^{n} \omega_i\ f(x_i)$$

gli ω_i sono detti coefficienti della formula di integrazione e si definisce l'errore

$$E_{n+1}(f) = I(f) - I_{n+1}(f)$$

Diremo che la formula di integrazione ha grado di precisione k se

$$E_{n+1}(x^j) \begin{cases} = 0 & \text{per } 0 \leq j \leq k \\ \neq 0 & \text{per } j = k+1 \end{cases}$$

Ovvero se la formula di integrazione calcola in modo esatto l'integrale di un polinomio di grado al più k.

1.5.2 Formule di interpolazione

Consideriamo il polinomio $p(x)$ di grado al più n tale che

$$p(x_i) = f(x_i) \quad 0 \leq i \leq n$$

L'idea è di approssimare l'integrale di f con quello di p.

E' ora utile introdurre il polinomio di Lagrange

$$L_i(x) = \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j}$$

Non è difficile verificare che

$$L_i(x_j) \begin{cases} 1 & \text{se } i = j \\ 0 & \text{se } i \neq j \end{cases}$$

Quindi possi scrivere esplicitamente

$$p(x) = \sum_{i=0}^{n} L_i(x)\ f(x_i)$$

Quindi definisco la formula di integrazione per interpolazione come

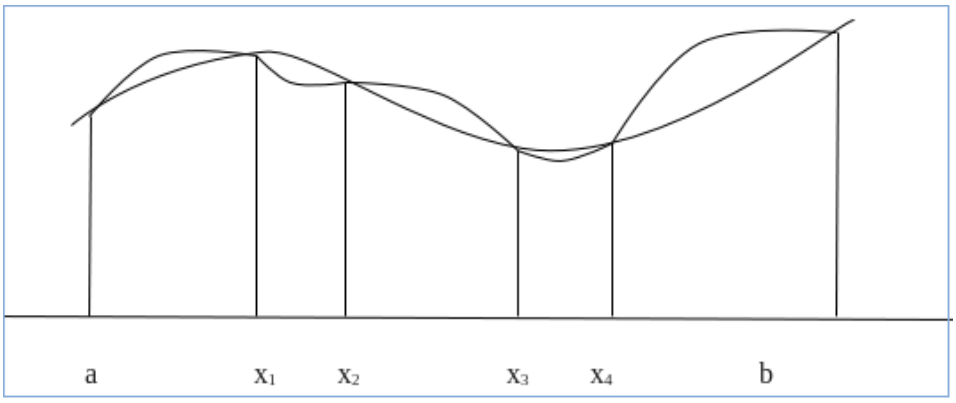
$$I_{n+1}(f) = \int_a^b p(x)\ dx = \sum_{i=0}^{n} f(x_i) \int_a^b L_i(x)\ dx$$

Quindi in questo caso abbiamo che i coefficienti dell'integrazione sono

$$\omega_i = \int_a^b L_i(x)\ dx$$

Quindi i coefficienti non dipendono dalla funzione ma solo dal tipo di suddivisione dell'intervallo che scegliamo.

Nell'immagine c'è la funzione disegnata prima dove fissati cinque punti abbiamo tracciato il polinomio di interpolazione.



Se f è abbastanza regolare sappiamo che il resto dell'interpolazione polinomiale è

$$p(x) - f(x) = \prod_{i=0}^{n} (x - x_i)\ \frac{f^{(i+1)}(\xi)}{(i+1)!}$$

Quindi abbiamo

$$E_{n+1}(f) = \frac{1}{(n+1)!} \int_a^b \prod_{i=0}^{n} f^{(i+1)}(\xi)\ dx$$

Osservazione 1.7. Se f è un polinomio di grado al più n allora $p(x) = f(x)$ quindi l'errore è 0. Pertanto la formula interpolatoria ha grado di precisione almeno n.

Osservazione 1.8. La differenza tra i vari metodi consiste nella scelta dei nodi, fin ora li abbiamo presi equidistanti, ma l'idea può esser quella di sceglierli per rendere massimo il grado di precisione del metodo.

30/03/2011

1.5.3 Metodi di Runge-Kutta

Studieremo ora una classe di metodi per trovare le soluzioni appprossimate del problema generale

$$\begin{cases} x'(t) = f(t, x(t)) \\ x(t_0) = x_0 \end{cases}$$

I metodi di Runge-Kutta sono metodi ad un passo, quindi supponiamo di avere una suddivisione equispaziata

$$t_0 < t_1 < \cdots < t_n$$

Al solito definiamo $x_i \simeq x(t_i)$ l'approssimazione della soluzione nel punto t_i, allora abbiamo già detto che i metodi ad un passo sono quelli della forma

$$x_{i+1} = x_i + h\phi(t_i, h, x_i, x_{i+1})$$

L'idea per costruire metodi ad un passo è la seguente: Data la suddivisione, possiamo suddividere il problema in tanti problemi

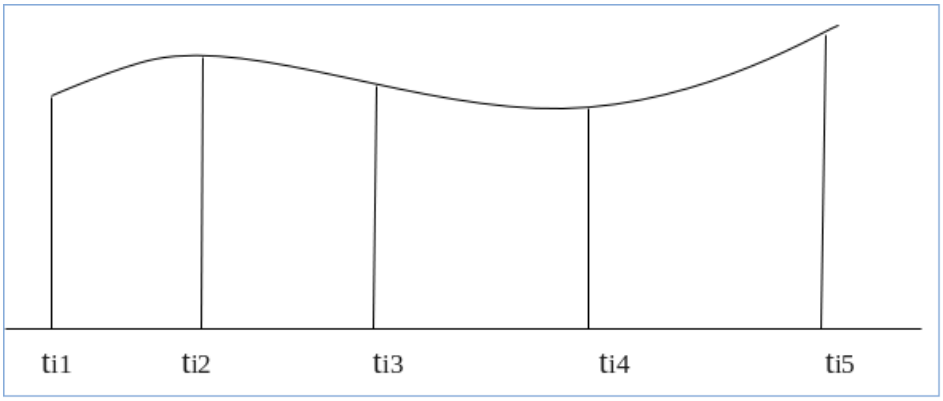
$$\begin{cases} x(t) = f(t, x(t)) \\ x(t_i) = x_i \end{cases}$$

Quindi possiamo trasformare il problema nell'equivalente problema integrale

$$x(t_{i+1}) = x_i + \int_{t_i}^{t_i+1} f(s, x(s))\ ds$$

A questo punto per stimare x_{i+1} possiamo pensare di stimare l'integrale, quindi i vari metodi si differenzieranno in base alla stima che si farà sull'integrale. Abbiamo già osservato che il metodo di Eulero fa questa stima, pensando all'integrale di Riemann, il metodo di Eulero esplicito stima con le somme inferiori e quello esplicito con le somme superiori.

I metodi di Runge-Kutta consistono nello stimare l'integrale della funzione tramite l'integrale di un polinomio che approssima la funzione. Quindi supponendo di avere una suddivisione equispaziata $t_{i+1} = h + t_i$ consideriamo una sottosuddivisione



Ovvero, come da disegno si intuisce, prenderemo come sottosuddivisione

$$t_{i_j} = t_i + \rho_j h \quad 0 \leq \rho_j \leq 1$$

In modo da suddividere l'intervallo $[t_i, t_{i+1}]$ in intervalli più piccoli $[t_{i_1}, t_{i_2}], [t_{i_2}, t_{i_3}], [t_{i_3}, t_{i_4}], \dots$ Quindi avremo

$$\int_{t_i}^{t_{i+1}} f(s, x(s))\ ds \simeq h\ \sum_{j=1}^{m}\ \beta_j f(t_{i_j}, x(t_{i_j}))$$

Dove l'h che esce dalla sommatoria lo abbiamo tirato fuori per normalizzare i β_j, tutto sarà chiaro più avanti. Chiamiamo

$$K_j = f(t_{i_j}, x(t_{i_j}))$$

Daltronde non conosciamo gli $x(t_{i_j})$ quindi dobbiamo fare un'ulteriore approssimazione

$$x(t_{i_j}) = x(t_i + \rho_j) = x(t_i) + \int_{t_i}^{t_i+\rho_j h} f(s, x(s))\ ds$$

Quindi

$$\int_{t_i}^{t_i+\rho_j h} f(s, x(s))\ ds \simeq h\ \sum_{l=1}^{m} \gamma_{j_l}\ f(t_{i_l}, x(t_{i_l}))$$

Pertanto da questi ultimi passaggi troviamo

$$x(t_{i_j}) \simeq x_i + h \sum_{l=1}^{m} \gamma_{j_l} K_l$$

Ora che abbiamo una stima di $x(t_{i_j})$ tornando al conto iniziale possiamo riscrivere K_j come

$$K_j = f\left(t_i + \rho_j h, x_i + h \sum_{l=1}^{m} \gamma_{j_l} K_l\right) \quad 1 \leq j \leq m$$

NB: qui c'è il pericolo di impazzire con gli indici, ricordiamo che i è fissato, l'unica cosa che varia è j.

Ricapitolazione

Un metodo di Runge-Kutta è caratterizzato da queste due equazioni

$$\begin{cases} x_{i+1} = x_i + h\ \sum_{j=1}^{m} \beta_j K_j \\ K_j = f\left(t_i + \rho_j h, x_i + h \sum_{l=1}^{m} \gamma_{j_l} K_l\right) \end{cases} \quad 1 \leq j \leq m$$

Questo è detto metodo di Runge-Kutta a m stadi.

Esempio

Il metodo di Eulero esplicito è un metodo di Runge-Kutta con $m = 1$, $\rho_1 = 0$ e $\gamma_{1_1} = 0$ mentre nel metodo di Eulero impolicito $m = 1$,$\rho_1 = 1$ e $\gamma_{1_1} = 1$

Commenti

Riassumeremo spesso un metodo di Runge-Kutta con la seguente tabella

ρ_1	γ_{1_1}	$\cdots$	γ_{1_m}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
ρ_m	γ_{m_1}	$\cdots$	γ_{m_m}
	β_1	$\dots$	β_m

Definiamo la matrice

$$\Gamma = \begin{pmatrix} \gamma_{1_1} & \cdots & \gamma_{1_m} \\ \vdots & \ddots & \vdots \\ \gamma_{m_1} & \cdots & \gamma_{m_m} \end{pmatrix}$$

Se la matrice Γ è strettamente triangolare inferiore (la diagonale è nulla) allora diremo che il metodo di Runge-Kutta è di tipo esplicito, altrimenti sarà di tipo implicito.

1.5.4 Metodi di Runge-Kutta espliciti

In generale con un metodo di Runge-Kutta esplicito trovare i coefficienti K_j è facile dato che la matrice è strettamente triangolare.
Osserviamo dapprima che

$$\sum_{l=1}^{m} \gamma_{j_l} K_l$$

è uguale alla j-esima riga del prodotto

$$\begin{pmatrix} \gamma_{1_1} & \cdots & \gamma_{1_m} \\ \vdots & \ddots & \vdots \\ \gamma_{m_1} & \cdots & \gamma_{m_m} \end{pmatrix} \begin{pmatrix} K_1 \\ \vdots \\ K_m \end{pmatrix}$$

Quindi i k_j sono facilmente calcolabili, infatti

$$K_1 = f(t_i + h\rho_j, x_i)$$

$$K_2 = f(t_i + h\rho_j, x_i + h\gamma_{2_1} K_1)$$

E così via, in generale

$$K_j = f\left(*, x_i + h \sum_{l=1}^{j-1} \gamma_{j_l} k_l\right) \qquad 1 \leq j \leq m$$

Costo computazione dei metodi di Runge-Kutta

Prima di calcolare il costo computazionale di dei metodi di Runge-Kutta, osserviamo che questo certamente dipenderà dalla funzione f, ad esempio nel metodo di Eulero

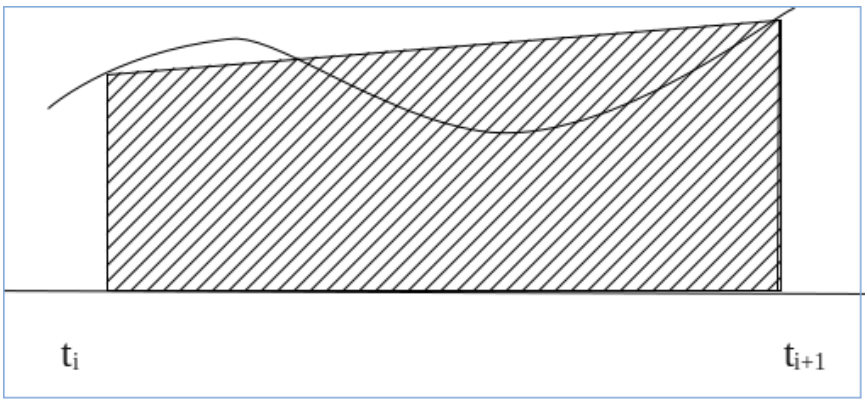
$$x_{i+1} = x_i + h\ f(t_i, x_i)$$

il costo è quello di N valutazioni della funzione $f(t, x)$. Se invece consideriamo un metodo di Runge-Kutta a m stadi succede che ad ogni passo dobbiamo calcolare m volte la funzione $f(t, x)$, il resto lo trascuriamo.
Mostreremo che questi metodi, anche se più costosi, convengono più velocemente dato che l'errore va a zero come h^p con $p > 1$, quindi anche se ad ogni passo devo fare molte valutazioni di f posso prendere un N piccolo.

1.5.5 Esempi

Metodo dei Trapezi

Mostriamo ora alcuni metodi di Runge-Kutta, ad esempio il metodo dei trapezi, l'idea è di approssimare con trapezi piuttosto che con rettangoli l'area sottesa ad una funzione



Il metodo dei trapezi è dato dai coefficienti

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 0 & 1 \\ \hline & 1/2 & 1/2 \end{array}$$

Quindi questo è un metodo di tipo Runge-Kutta a due stadi, scritto esplicitamente

$$x_{i+1} = x_i + \frac{h}{2}\left(f(t_i, x_i) + f(t_{i+1}, x_{i+1})\right)$$

Si osserva che il metodo dei trapezi è una combinazione convessa di Eulero esplicito e implicito. Non è difficile provare che il metodo dei trapezi è consistente di ordine 2 (si usa Taylor e si fanno i soliti conti). Daltronde il metodo dei trapezi è implicito, vogliamo renderlo esplicito, per far ciò introduciamo un altro metodo.

Metodo di Heun

Consideriamo il metodo dei trapezi

$$x_{i+1} = x_i + \frac{h}{2}\left(f(t_i, x_i) + f(t_{i+1}, x_{i+1})\right)$$

Sostituiamo al posto di x_{i+1} nella funzione, un'approssimazione con Eulero

$$x_{i+1} = x_i + h\ f(t_i, x_i)$$

Quindi abbiamo

$$x_{i+1} = x_i + \frac{h}{2}\left(f(t_i, x_i) + f(t_{i+1}, x_i + h\ f(t_i, x_i))\right)$$

che si riassume nella tabella

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 1/2 & 1/2 \end{array}$$

Questo metodo ha consistenza di ordine 2 ma è più difficile provarlo.

Osservazione 1.9. Posso considerare la famiglia di metodi

$$x_{i+1} = x_i + h\ [\alpha f(t_i, x_i) + (1-\alpha) f(t_{i+1}, x_{i+1})]$$

e posso cercare l'α che massimizza l'ordine di consistenza.

Il metodo classico di Runge-Kutta

Consideriamo il classico metodo di Runge-Kutta con 4 stadi

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array}$$

Questo metodo ha ordine di consistenza 4.

Teorema 1.3. In un metodo di Runge-Kutta vale in generale che

$$\sum_{j=1}^{m} \beta_j = 1 \qquad \sum_{l=1}^{m} \gamma_{jl} = \rho_j$$

queste sono condizioni necessarie di consistenza (chiaramente non sufficienti).

Dimostrazione. Dimostriamo la prima

$$\delta(x(t+h),h) = \frac{1}{h}[x(t+h) - x(t) - h \sum_{j=1}^{m} \beta_j \ K_j]$$

per ipotesi

$$\lim_{h\to 0} \delta(x(t+h),h)$$

quindi sviluppiamo

$$\delta(x(t+h),h) = \frac{x(t+h)-x(t)}{h} - \sum_{j=1}^{m} \beta_j \ K_j$$

ma sappiamo che

$$\frac{x(t+h)-x(t)}{h} \to x'(t) = f(t,x(t))$$

Ricordiamo che

$$K_j = f\left(t + h\rho_j, x(t) + h\sum_{l=1}^{m} \gamma_{j_l} K_l\right) \to f(t,x(t))$$

Quindi abbiamo

$$\sum_{j=1}^{m} \beta_j K_j \to \sum_{j=1}^{m} \beta_j f(t,x(t))$$

Quindi necessariamente deve essere

$$\sum_{j=1}^{m} \beta_j = 1$$

Per l'altra relazione il ragionamento è analogo

$$x(t_i + h \ \rho_j) \simeq x(t_i) + h\sum_{l=1}^{m} \gamma_{j_l} K_l$$

Quindi se impongo

$$\lim_{h\to 0} \frac{x(t_i + h\rho_j) - [x(t_i) - h\sum_{l=1}^{m} \gamma_{j_l} K_l]}{h\rho_j} = 0$$

con passaggi analoghi ho la tesi. □

05/04/2011

1.5.6 Condizioni necessarie di consistenza

Supponiamo di avere un metodo di Runge-Kutta di ordine p. *Richiamo*
Ricordiamo che l'ordine di consistenza di un metodo è relativo al metodo e non al problema, quindi se un metodo ha ordine di consistenza p allora per ogni problema l'errore locale di discretizzazione deve andare a zero come h^p dove h è il passo di discretizzazione.

Dunque considero il problema

$$\begin{cases} x'(t) = x(t) + t^{l-1} \quad t \in [0,T] \ \ l \in \mathbb{N} \setminus \{0\} \\ x(0) = 0 \end{cases}$$

Per quando detto nel richiamo, se stiamo considerando un metodo con ordine di consistenza p allora per questo particolare problema bisognerà avere che l'errore locale di discretizzazione deve andare a 0 come h^p. Chiaramente se l'errore locale per questo problema va a zero come h^p non possiamo dedurre che il metodo è consistente di ordine p, quindi imponendo che per questo particolare tipo di problema ci sia consistenza di ordine p troveremo condizione necessarie ma non sufficienti di consistenza.

Si trova subito dal problema posto che

$$x^{(n)}(0) = \begin{cases} 0 & \text{se} \quad 0 \leq n \leq l-1 \\ (l-1)! & \text{se} \quad n \geq l \end{cases}$$

Ricordiamo che un metodo di Runge-Kutta ha la forma

$$\begin{cases} x_{i+1} = x_i + h \ \sum_{j=1}^{m} \beta_j K_j \\ K_j = f\left(t_i + \rho_j h, x_i + h \sum_{l=1}^{m} \gamma_{j_l} K_l\right) \end{cases} \qquad 1 \leq j \leq m$$

Se applichiamo il generico metodo di Runge-Kutta al problema che stiamo esaminado troviamo che

$$K_j = (t_i + h\rho_j)^{l-1} + x_i + h \sum_{l=1}^{m} \gamma_{j_l} K_l$$

Ricordiamo, come si è visto nella lezione precedente, che

$$\sum_{l=1}^{m} \gamma_{j_l} K_l$$

è uguale alla j-esima riga del prodotto

$$\begin{pmatrix} \gamma_{1_1} & \cdots & \gamma_{1_m} \\ \vdots & \ddots & \vdots \\ \gamma_{m_1} & \cdots & \gamma_{m_m} \end{pmatrix} \begin{pmatrix} K_1 \\ \vdots \\ K_m \end{pmatrix}$$

Quindi chiamo $\underline{\mathrm{K}} = (K_1, \ldots, K_m)^T$, allora possiamo riscriverla come

$$(I - h\Gamma)\underline{\mathrm{K}} = \begin{pmatrix} (t_i + h\rho_1)^{l-1} + x_i \\ \vdots \\ (t_i + h\rho_m)^{l-1} + x_i \end{pmatrix}$$

La matrice $(I - h\Gamma)$ è invertibile per h abbastanza piccolo, ma dato che stiamo considerando $h \to 0$ non ci poniamo problemi di questa natura, quindi otteniamo

$$\underline{\mathrm{K}} = (I - h\Gamma)^{-1} \begin{pmatrix} (t_i + h\rho_1)^{l-1} + x_i \\ \vdots \\ (t_i + h\rho_m)^{l-1} + x_i \end{pmatrix}$$

Scriviamo l'errore di discretizzazione al primo passo $t_1 = 0 + h$, quindi

$$\delta(x(h), h) = \frac{1}{h}[x(h) - x(0) - h\underline{\beta}^T \underline{\mathrm{K}}]$$

Definisco per comodità la matrice diagonale

$$D = diag(\rho_1, \ldots, \rho_m)$$

e il vettore

$$\underline{1} = (1, \ldots, 1)^T$$

Quindi abbiamo

$$\begin{pmatrix} \rho_1^{l-1} \\ \vdots \\ \rho_m^{l-1} \end{pmatrix} = D^{l-1}\underline{1}$$

A questo punto scrivo in serie $x(t)$ e in base a quanto osservato prima sulle derivate

$$x(h) = \frac{h^l}{l} + \frac{h^{l+1}}{l(l+1)} + \frac{h^{l+2}}{(l+2)(l+1)l} + \dots$$

Mentre sviluppando in serie l'altro pezzo

$$h^l \underline{\beta}(I - h\Gamma)^{-1} D^{l-1} \underline{1}$$

Per svilupparlo in serie ricordiamo il caso più semplice

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n$$

Quindi se consideriamo le matrici il conto diventa più difficile ma otteniamo sempre lo stesso risultato

$$(I - h\Gamma)^{-1} = I + h\Gamma + ...$$

Quindi andando a sostituire questo pezzo otteniamo

$$h^l \underline{\beta}(I - h\Gamma)^{-1} D^{l-1} \underline{1} = h^l \underline{\beta}^T D^{l-1} \underline{1} + h^{l+1} \underline{\beta}^T \Gamma D^{l-1} \underline{1} + \dots$$

Quindi se suppongo che $\delta(x(h), h) = O(h^p)$ allora ho la condizione

$$\underline{\beta}^T \Gamma^k D^{l-1} \underline{1} = \frac{1}{l(l+1)\dots(l+k)} \qquad 0 \le k \le p - l$$

Quindi ho condizione necessarie che coinvolgono $\underline{\beta}, \Gamma$ e D.
Se supponiamo $l = 1$ allora otteniamo la condizione

$$\underline{\beta}\Gamma^k \underline{1} = \frac{1}{(k+1)!} \qquad 0 \le k \le p - 1$$

Per $k = 0$ troviamo la già nota relazione $\underline{\beta}^T \underline{1} = 1$

Quindi se il metodo che stiamo esaminando è consistente di ordine p allora

$$\delta(x(h), h) = h^p \left(\frac{1}{(p+1)!} - \underline{\beta}\Gamma^p \underline{1} \right) + O(h^{p+1})$$

Cosa possiamo dire sul massimo ordine di convergenza?
Analizziamo i metodi di Runge-Kutta espliciti, ovvero la matrice Γ è strettamente triangolare inferiore, osserviamo ad esempio che se

$$\Gamma = \begin{pmatrix} & & & \\ * & & & \\ * & * & & \\ * & * & * & \\ * & * & * & * \end{pmatrix}$$

allora

$$\Gamma^2 = \begin{pmatrix} & & & \\ & & & \\ * & & & \\ * & * & & \\ * & * & * & \end{pmatrix} \qquad \Gamma^3 = \begin{pmatrix} & & & \\ & & & \\ & & & \\ * & & & \\ * & * & & \end{pmatrix} \qquad \Gamma^4 = \begin{pmatrix} & & & \\ & & & \\ & & & \\ & & & \\ * & & & \end{pmatrix} \qquad \Gamma^5 = \underline{0}$$

Quindi tutte le matrici strettamente triangolari sono nilpotenti, quindi se ho un metodo ad m stadi allora

$$\Gamma^j = \underline{0} \quad j \ge m$$

Quindi un metodo di Runge-Kutta esplicito ha ordine di consistenza al massimo pari al numero di stadi (vedremo che per i metodi impliciti questa cosa non è vera).

Abbiamo mostrato l'altra volta che se ho un metodo a m stadi allora per calcolare x_{i+1} devo fare m valutazioni della funzione $f(t,x)$, quindi il metodo ottimale sarebbe quello per cui l'ordine di convergenza è esattamente uguale al numero di stadi, quindi nasce spontanea la domanda se per ogni m esiste un metodo di Runge-Kutta con m stadi e con consistenza di ordine m, ma la risposta è no e la dimostrazione è difficile.

numero stadi	1	2	3	4	5	6	7	8	...	10
massimo ordine di consistenza	1	2	3	4	4	5	6	6	...	7

06/04/2011

1.5.7 Convergenza e consistenza per metodi ad un passo

Per i metodi ad un passo vale che un metodo è consistente di ordine p se e soltanto se è convergente di ordine p. Mostreremo solo che la consistenza implica la convergenza (il viceversa è senzato e facile da dimostrare) e faremo la dimostrazione solo nel caso esplicito (ma il risultato vale anche se il metodo è implicito). Ricordiamo che un metodo ad un passo è una successione della forma

$$x_{i+1} = x_i + h\ \phi(t_i, x_i, h) \quad 1 \leq i \leq N$$

Supponiamo dunque che il metodo che stiamo considerando sia consistente di ordine p.
Consideriamo il problema con l'intervallo ristretto invece che $[t_0, T]$ prendiamo il sottointervallo $[t_i, t_{i+1}]$ ovvero

$$\begin{cases} x'(t) = f(t, x(t)) & t \in [t_i, t_{i+1}] \\ x(t_i) = x_i \end{cases}$$

Definiamo $x_i(t)$ la soluzione esatta del problema appena scritto.
Allora se applichiamo il metodo ad un passo avremo

$$x_{i+1} = x_i + h\ \phi(t_i, x_i, h) = (x_i - x_i(t_{i+1}) + h\ \phi(t_i, x_i, h)) + x_i(t_{i+1}) = -h\ \delta(x_i(t_{i+1}), h) + x_i(t_{i+1})$$

Osserviamo che per come abbiamo definito il tutto vale che $x_i = x_i(t_i)$ Quindi abbiamo ottenuto

$$x_{i+1} = -\delta(x_i(t_{i+1}), h) + x_i(t_{i+1})$$

Consideriamo ora l'errore globale

$$e_{i+1} = x(t_{i+1}) - x_{i+1} = x(t_{i+1}) - x_i(t_{i+1}) + h\delta(x_i(t_{i+1}), h)$$

Come conseguenza del teorema di perturbazione abbiamo che, se consideriamo i due problemi

$$\begin{cases} x'(t) = f(t,x) & t \in [t_i, t_{i+1}] \\ x(t_i) = x(t_i) \end{cases} \qquad \begin{cases} x'(t) = f(t,x) & t \in [t_i, t_{i+1}] \\ x(t_i) = x_i \end{cases}$$

Abbiamo la seguente minorazione

$$||x(t_{i+1}) - x_i(t_{i+1})|| \leq e^{L(t_{i+1} - t_i)}\ ||x(t_i) - x_i|| = e^{Lh} e_i$$

Quindi andando a sostituire

$$||e_{i+1}|| \leq e^{Lh} ||e_i|| + h \max_{j \leq i} ||\delta(x_j(t_{j+1}), h)||$$

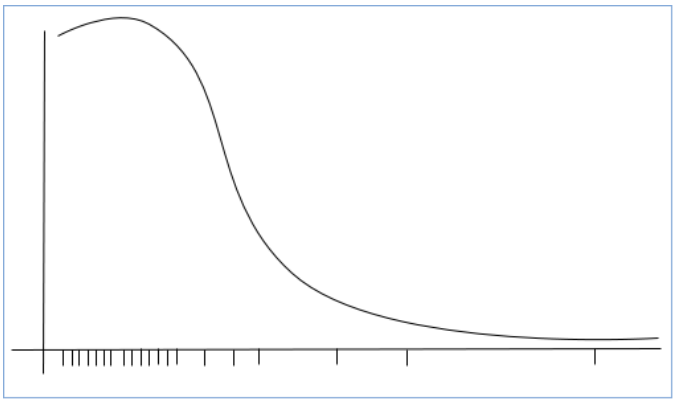
Applicando ricorsivamente questa formula si trova

$$||e_i|| \leq \frac{e^{ihL} - 1}{e^{hL} - 1}\ h\ h \max_{j \leq i} ||\delta(x_j(t_{j+1}), h)|| \leq \frac{e^{T - t_0} - 1}{L} h \max_{j \leq i} ||\delta(x_j(t_{j+1}), h)||$$

Per $h \to 0$ per ipotesi sappiamo che $\delta \to 0$ con l'ordine $O(h^p)$, quindi segue che anche $||e_i||$ va a zero con ordine almeno p.

1.5.8 Stima dell'errore locale

Fin ora abbiamo sempre supposto di avere suddivisioni equispaziate, daltronde se abbiamo una soluzione al classico problema della forma



Sarà necessario infittire la suddivisione quando la soluzione ha picchi o comunque comportamenti poco regolari, mentre quando è circa costante (in questo caso tende a 0) possiamo rilassare la suddivisione. Applichiamo ora il metodo tenendo conto che x_{i+1} dipenderà anche dalla discretizzazione dell'intervallo che si sarà presa $t_{i+1} = t_i + h_i$, quindi denoteremo

$$x_{i+1}^{h_i} = x_i + h_i \ \phi(t_i, x_i, h_i) \quad 0 \leq i \leq N-1$$

NB: la notazione è troppo pesante, quindi spesso ometteremo il fatto che $x_{i+1}^{h_i}$ dipenda da h_i e useremo la classica notazione.

Cercherò gli h_i tali che $h_i \ \delta(x_i(t_{i+1}), h_i) \approx$ tolleranza (in genere 10^{-6})

Definiamo al solito $x_i(t)$ come la soluzione esatta al seguente problema

$$\begin{cases} x'(t) = f(t, x(t)) & t \in [t_i, t_{i+1}] \\ x(t_i) = x_i \end{cases}$$

Per avere la stima richiesta supponiamo di avere due metodi, sul primo vogliamo avere la stima sulla tolleranza che abbiamo scritto sopra, il secondo metodo sarà solo un metodo ausiliario. Denoteremo dunque con x_{i+1} l'approssimazione calcolata con il primo metodo e con $\overline{x_{i+1}}$ la soluzione trovata con il secondo metodo. Vedremo che in qualche senso

$$x_{i+1} - \overline{x_{i+1}} \approx h \ \delta(*)$$

Tecnica di estrapolazione locale

Non abbiamo detto nulla sul metodo ausiliario, infatti i risultati che otterremo dipenderanno dalla scelta di quest'ultimo, mostriamo ora la tecnica di estrapolazione lineare, definiamo il metodo

$$\overline{x_{i+1}} = x_i + \frac{h}{2} \ \phi\left(t_i + \frac{h}{2}, x_i \frac{h}{2} \phi\left(t_i, x_i, \frac{h}{2}\right), \frac{h}{2}\right) + \frac{h}{2} \ \phi\left(t_i, x_i, \frac{h}{2}\right)$$

Suppongo che il primo metodo sia consistente di ordine p, allora per un risultato generale, se un metodo è consistente di ordine p allora

$$\delta(x(t+h), h) = c(x(t)) \ h^p + O(h^{p+1})$$

dove $c(x)$ è una funzione continua (sottolineamo il fatto che questo risultato è generale ma non lo dimostriamo).

Dunque avremo

$$\overline{\delta}(x(t+h), h) = \frac{1}{h}[x(t+h) - x(t) - \dots] = \frac{1}{h}\left[c(x(t)) \ \left(\frac{h}{2}\right)^{p+1} + c\left(x\left(t + \frac{h}{2}\right)\right)\left(\frac{h}{2}\right)^{p+1}\right]$$

$$+O(h^{p+1}) = \left(\frac{h}{2}\right)^p \; c(x(t)) + O(h^{p+1})$$

Osservazione 1.10. I due metodi hanno lo stesso ordine di consistenza ma cambia la costante moltiplicativa (quindi bisogna comunque porre molta attenzione).

In sintesi abbiamo trovato che

$$\overline{\delta}(*) = \frac{1}{2^p}\delta(*) + O(h^{p+1})$$

Quindi

$$x_i(t_i + 1) = x_{i+1} + h\ \delta(x_i(t_{i+1}), h) = \overline{x_{i+1}} + h\ \overline{\delta}(x_i(t_{i+1}), h) + O(h^{p+1})$$

In conclusione ottengo

$$\delta(x_i(t_{i+1}), h) = \frac{1}{h}\left(1 - \frac{1}{2^p}\right)^{-1} (\overline{x_{i+1}} - x_i) + O(h^p)$$

E questa è una stima dell'errore locale di discretizzazione.

Tecniche di Embedding

L'idea delle tecniche di Embedding è di confrontare due metodi con ordine di consistenza diversi, quindi definiamo

- ϕ : $\quad x_{i+1} = x_i + h\ \phi(t_i, x_i, h)$ $\qquad$ consistente di ordine p
- $\overline{\phi}$: $\quad \overline{x_{i+1}} = x_i + h\ \overline{\phi}(t_i, x_i, h)$ $\qquad$ consistente di ordine $q \geq p + 1$

Quindi calcoliamo $x_i(t_{i+1})$ con il primo metodo e poi con il secondo

$$x_i(t_{i+1}) = x_{i+1} + h\ \delta(x_i(t_{i+1}), h) = \overline{x_{i+1}} + h\ \overline{\delta}(x_i(t_{i+i}), h)$$

Quindi ottengo

$$\overline{x_{i+1}} - x_{i+1} = h(\delta(*) - \overline{delta}(*))$$

Usando il fatto che $\delta(*) \simeq O(h^p)$ e $\overline{\delta}(*) \simeq O(h^q)$ otteniamo

$$\overline{x_{i+1}} - x_{i+1} = h\ \delta(*) + O(h^{p+1})$$

in conclusione

$$\delta(x_i(t_{i+1}), h) = \frac{1}{h}(\overline{x_{i+1}} - x_{i+1}) + O(h^p)$$

Quando conviene usare questi metodi e cenni su ODE45

Sembra sia sconveniente cercare una stima dell'errore locale di discretizzazione dato che è necessario usare contemporaneamente due metodi, daltronde ci sono dei casi invece in cui non costa nulla farlo, ad esempio nelle famiglie di metodi di Runge-Kutta espliciti dove il metodo $\overline{\phi}$ è dato dalla matrice $\overline{\Gamma}$ mentre il metodo ϕ è dato dalla matrice Γ che è definita come la matrice $\overline{\Gamma}$ senza l'ultima riga e l'ultima colonna. Abbiamo già detto che

$$x_{i+1} = x_i + h\ \sum_{j=1}^{m} \beta_j K_j$$

$$\overline{x_{i+1}} = x_i + h\ \sum_{j=1}^{m+1} \overline{\beta_j K_j}$$

dove

$$\overline{K_j} = K_j \qquad 1 \leq j \leq m$$

quindi avremo

$$\overline{x_{i+1}} - x_{i+1} = h\ \sum_{j=1}^{m+1} K_j(\beta_j - \overline{\beta_j}) \quad \text{con} \quad \beta_{m+1} = 0$$

Questa è effettivamente la tecnica più usata, ad esempio la implementa matlab con la funzione ODE45 (ordinary equation 4-5) ovvero usa una famiglia di metodi di Runge-Kutta, uno a 5 stadi e l'altro a 6 e con ordine di consistenza l'uno 4 e l'altro 5. A questo punto ci si può chiedere come vengano usate queste stime, supponiamo di avere una stima

$$s_i \approx h\ \delta(x(t_{i+1}), h)$$

Se ad esempio abbiamo fissato un valore tol di tolleranza allora

$$\text{se} \quad \frac{1}{4}\ tol < s_i < 4\ tol \quad \text{allora il passo } h \text{ va bene}$$

Se invece s_i non si trova in quell'intervallo allora se so che

$$s_i \simeq \alpha h^{p+1}$$

allora trovo

$$\alpha = \frac{s_i}{h^{p+1}}$$

e cerco un nuovo $\overline{h}$ tale che $\alpha\overline{h} \simeq 0.9\ tol$, quindi sostitiendo

$$\frac{s_i}{h^{p+1}}\ \overline{h}^{p+1} \simeq 0.9\ tol$$

quindi

$$\overline{h} = \left(\frac{0.9\ tol}{s_i}\right)^{\frac{1}{p+1}}\ h$$

In conclusione ho un metodo per fare una suddivisione che minimizza

$$|h\ \delta(*)| \leq \epsilon$$

Ma l'errore globale rispetta questa minorazione? La risposta è chiaramente no, può succedere che l'errore locale sia maggiorato dalla tolleranza ma quello globale no, infatti compaiono le costanti moltiplicative

$$\delta(x(t+h), h) = c(x(t))\ h^p + O(h^{p+1})$$

$$e_i = O(x(x))\ h^p + O(h^{p+1})$$

può quindi succedere che $d(x(t)) >> c(x(t))$, infatti facciamo in modo che $c(x(t))$ sia molto piccolo ma non facciamo nulla su $d(x(t))$.
Prima abbiamo visto che

$$||e_i|| \leq e^{LT}||\delta(*)||$$

A volte si ha quasi l'uguaglianza. E' comunque possibile imporre che l'errore globale sia più piccolo di una certa tolleranza e si usano tecniche simili a quelle appena viste ma è costosissimo e non ne vale la pena.
NB: $d(x)$ è una funzione che svolge lo stesso ruolo di $c(x)$, l'una descrive l'errore locale, l'altra quello globale.

1.5.9 Ordine di consistenza dei metodi di Runge-Kutta impliciti

I metodi di Runge-Kutta impliciti possono raggiungere ordine di consistenza $2m$ dove m è il numero di stadi, in particolare non vale quanto detto per i metodo espliciti, ovvero per ogni numero m esiste un metodo ad m stadi che ha consistenza massima $2m$, tali metodi li chiameremo metodi di Gauss.

Metodi di Gauss

Il seguente è detto metodo del punto di mezzo, è ad uno stadio e ha consistenza di ordine 2

$1/2$	$1/2$
	1

Un altro metodo di Gauss con 2 stadi e ordine di consistenza 4

*	$1/4$	$\frac{3-2\sqrt{3}}{12}$
*	$\frac{3+2\sqrt{3}}{12}$	$1/4$
	$1/2$	$1/2$

NB: al posto degli * c'è la somma della riga

Metodi di Radau (ordine di consistenza 2m-1)

I metodi di Radau sono così caratterizzati, c'è una condizione sui ρ_j, ovvero esiste un r tale che $\rho_r = 0$ oppure $\rho_r = 1$, il che equivale a prendere t_{i+1} oppure t_i nella sottosuddivisione. Ad esempio un metodo di Radau a 1 stadio e con consistenza 1 è

1	1
	1

Oppure con 2 stadi e con ordine di consistenza 3

*	5/12	-1/12
*	3/4	1/4
	3/4	1/4

Metodi di Lobatto (ordine di consistenza 2m-2)

I metodi di Lobatto sono così caratterizzati, c'è una condizione sui ρ_j, ovvero esistono r e w tali che $\rho_r = 0$ e $\rho_w = 1$, il che equivale a prendere t_{i+1} e t_i nella sottosuddivisione. Ad esempio un metodo di Lobatto con 2 stadi e consistenza 2 è

0	0	0
1	1/2	1/2
	1/2	1/2

Ci sono tantissimi metodi ma mostriamo solo questi come esempio.
12/04/2011

1.5.10 Metodi di Runge-Kutta basati su collocazione

Consideriamo al solito una suddivisione dell'intervallo $[t_0, T]$ data da

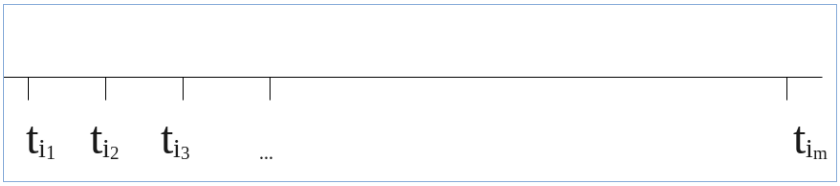
$$t_0 < t_1 < t_2 < \cdots < t_N = T$$

Allora considero come sempre i valori

$$0 \leq \rho_j \leq 1 \qquad j = 1, \ldots, m$$

Quindi considero la sottodivisione

$$t_{i_j} = t_i + h\rho_j$$



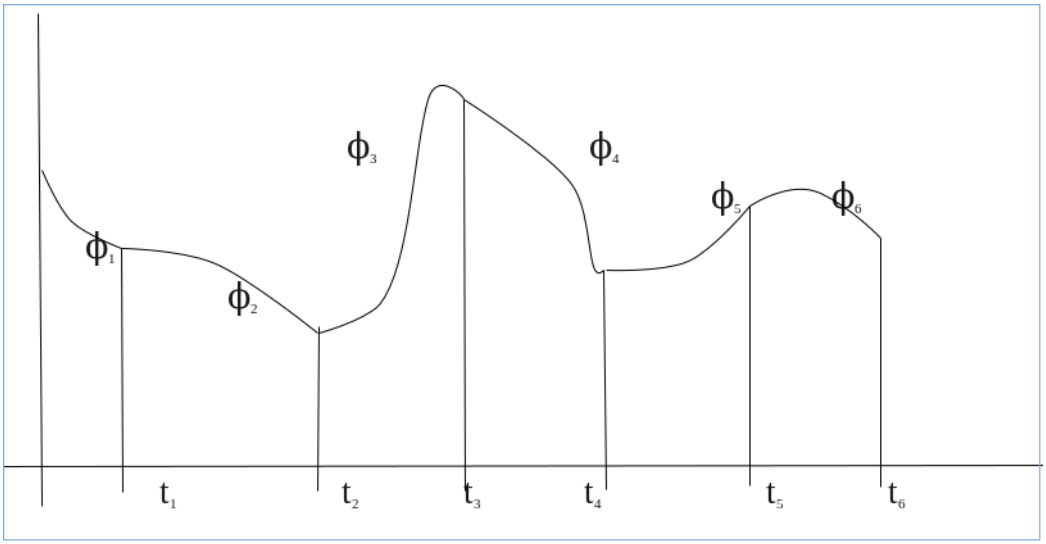
Supponiamo di conoscere $x_i \simeq x(t_i)$, quello che vogliamo è una funzione $\phi_i(t)$ relativa all'intervallo $[t_i, t_{i+1}]$ e voglio che soddisfi il problema

$$\begin{cases} \phi_i'(t_{i_j}) = f(t_{i_j}, \phi_i(t_{i_j})) & 1 \leq j \leq m \\ \phi_i(t_i) = x_i \end{cases}$$

Diremo che la funzione ϕ colloca $x(t)$ nei punti t_{i_j}.
Quindi definisco il punto successivo $x_{i+1} = \phi_i(t_{i+1})$, quindi cercheremo di trovare le funzioni ϕ_i.

Osservazione 1.11. Con i metodi che abbiamo fin ora definito siamo in grado di calcolare una successione $\{x_i\}_{1 \leq i \leq N}$ dove $x_i \simeq x(t_i)$, daltronde non è possibile avere subito una approssimazione di $x(\bar{t})$ se $\bar{t} \neq t_i$, a meno che non interpoli i punti.

Se invece dispongo delle funzioni descritte prima, ovvero $\{\phi_i\}_{1 \leq i \leq N}$ posso farlo



L'approssimazione della soluzione del problema sarà l'incollamento delle funzioni ϕ_i quindi posso valutare la funzione approssimata in ogni punto valutando le funzioni ϕ_i.

Costruiamo ora le funzioni $\phi_i(t)$, queste saranno dei polinomi di grado al più m che verificano

$$\begin{cases} \phi_i'(t_{i_j}) = f(t_{i_j}, \phi_i(t_{i_j})) & 1 \leq j \leq m \\ \phi_i(t_i) = x_i \end{cases}$$

Definisco il polinomio $p(t)$ di grado al più $m-1$ che soddisfa

$$p(t_{i_j}) = f(t_{i_j}, \phi_i(t_{i_j}))$$

La funzione $\phi_i(t)$ sarà la primitiva di $p(t)$ con $\phi_i(t_i) = x_i$, quindi tutto è univocamente determinato. Ricordiamo che i nodi saranno $t_{i_j} = t_i + h\rho_j$, definisco

$$K_j = \phi_i'(t_{i_j})$$

Quindi avremo che

$$\phi_i'(t_i + \tau h) = \sum_{j=1}^{m} L_j(t_i + \tau h)\ K_j \quad \text{con} \quad 0 \leq \tau \leq 1$$

dove L_j sarà il polinomio di Lagrange, esplicitamente

$$L_j(t) = \prod_{r=1, r\neq j}^{m} \frac{t - (t_i + h\ \rho_j)}{t_i + h\ \rho_j - (t_i + \ h\rho_r)}$$

quindi avremo

$$L_j(t_i + \tau h) = \prod_{r=1, r\neq j}^{m} \frac{\tau - \rho_r}{\rho_j - \rho_r}$$

andando a sostituire abbiamo che

$$\phi_i'(t_i + \tau\ h) = \sum_{j=1}^{m} \prod_{r=1, r\neq j}^{m} \frac{\tau - \rho_r}{\rho_j - \rho_r} K_j$$

Posto che $\phi_i(t_i) = x_i$ andando a riscrivere il problema in forma integrale abbiamo

$$x_{i+1} = \phi_i(t_{i+1}) = \phi_i(t_i) + \int_{t_i}^{t_{i+1}} \phi_i'(s)\ ds$$

voglio ora in qualche modo legare la nuova approssimazione con la vecchia, quindi risolvo l'integrale

$$\int_{t_i}^{t_{i+1}} \phi_i'(s)\ ds = h\ \int_0^1 \sum_{j=1}^{m} \prod_{r=1, r\neq j}^{m} \frac{\tau - \rho_r}{\rho_j - \rho_r} K_j d\tau$$

quindi

$$x_{i+1} = x_i + h \sum_{j=1}^{m} \left(\int_0^1 \prod_{r=1, r\neq j}^{m} \frac{\tau - \rho_r}{\rho_j - \rho_r} d\tau \right) K_j$$

posto

$$\beta_j = \int_0^1 \prod_{r=1, r\neq j}^{m} \frac{\tau - \rho_r}{\rho_j - \rho_r} d\tau$$

otteniamo

$$x_{i+1} = x_i + h\ \sum_{j=1}^{m} \beta_j K_j$$

E questo è un metodo di Runge-Kutta, ma dobbiamo ora sciogliere in qualche modo i K_j

$$K_j = \phi_i'(t_{i_j}) = f(t_i + h\ \rho_j, \phi_j(t_i + h\ \rho_j))$$

Osserviamo che

$$\phi_i(t_i + h\rho_j) = \phi_i(t_i) + \int_{t_i}^{t_i + h\ \rho_j} \phi_i'(s)\ ds = \phi_i(t_i) + \int_0^{\rho_j} \phi_i'(s)\ ds =$$

$$= \phi_i(t_i) + \int_0^{\rho_j} \sum_{l=1}^{m} \prod_{r=1, r\neq j}^{m} \frac{\tau - \rho_r}{\rho_l - \rho_r} K_l d\tau = x_i + h \sum_{l=1}^{m} \left[\prod_{r=1, r\neq j}^{m} \frac{\tau - \rho_r}{\rho_l - \rho_r} d\tau \right]\ K_l$$

poniamo

$$\gamma_{j_l} = \int_0^{\rho_j} \prod_{r=1, r\neq j}^{m} \frac{\tau - \rho_r}{\rho_l - \rho_r} d\tau$$

in conclusione abbiamo

$$x_{i+1} = h\ \sum_{l=1}^{m} \gamma_{j_l}\ K_l$$

quindi in conclusione abbiamo il sistema di equazioni non lineari

$$K_j = f\left(t_i + h\ \rho_j, x_i + h\ \sum_{l=1}^{m} \gamma_{j_l}\ K_l \right) \quad j = 1, \ldots, m$$

Per risolvere questo problema si potrebbe pensare ad un metodo di punto fisso, quindi generare una successione

$$\left\{ K_j^{(r)} \right\}_{r\geq 0} \qquad K_j^{(r)} \to K_j$$

quindi la successione sarà

$$K_j^{(r+1)} = f\left(t_i + h\ \rho_j, x_i + h\ \sum_{l=1}^{m} \gamma_{j_l}\ K_l^{(r)} \right) \quad j = 1, \ldots, m$$

daltronde la convergenza è lineare

$$||K_j^{(r+1)} - K_j|| \leq L\ h\ \sum_{l=1}^{m} |\gamma_{j_l}| ||K_j^{(r+1)} - K_j||$$

dove L è la costante di lipschitzianità, quindi il metodo converge (condizione sufficiente) se

$$m\ L\ h \max\{|\gamma_{j_l}|\} < 1$$

Osservazione 1.12. Si potrebbe anche pensare ad un metodo di tipo Gauss Seidel per risolvere il problema sfruttando la matrice Γ, riprenderemo questo aspetto nella prossima lezione.

13/04/2011 Dalla lezione precedente vogliamo risolvere il sistema di equazioni non lineari

$$K_j = f\left(t_i + h\ \rho_j, x_i + h\ \sum_{l=1}^{m} \gamma_{j_l}\ K_l\right) \quad j = 1, \dots, m$$

si è visto che usando i metodi del punto fisso abbiamo convergenza lineare e dobbiamo imporre forti condizioni sul passo di discretizzazione h, ovvero

$$m\ L\ h \max\{|\gamma_{j_l}|\} < 1$$

L'idea è quella di usare invece il metodo di Newton che rispetto al punto fisso ha odine di convergenza locale quadratica.

$$F(\underline{\mathrm{K}}) = \begin{pmatrix} K_1 - f\left(t_i + h\ \rho_1, x_i + h\ \sum_{l=1}^{m} \gamma_{1,l}\ K_l\right)) \\ \vdots \\ K_1 - f\left(t_i + h\ \rho_m, x_i + h\ \sum_{l=1}^{m} \gamma_{m,l}\ K_l\right)) \end{pmatrix}$$

dove abbiamo posto

$$\underline{\mathrm{K}} = (K_1, \dots, K_m)^T$$

Quindi l'obbiettivo è trovare un $\underline{\mathrm{K}}$ tale che $F(K) = 0$ e questa operazione devo farla ad ogni passo del metodo per trovare tutti gli x_i. Il metodo di Newton genera una successione

$$\left\{\underline{\mathrm{K}}^{(r)}\right\}_{r \geq 1} \qquad \underline{\mathrm{K}}^{(r+1)} = \underline{\mathrm{K}}^{(r)} - J(\underline{\mathrm{K}}^{(r)})^{-1} \dot{F}(\underline{\mathrm{K}}^{(r)})$$

dove poniamo

$$J(\underline{\mathrm{K}}) = \frac{\partial F}{\partial \underline{\mathrm{K}}} = \left(\frac{\partial F_1}{\partial K_1}, \dots, \frac{\partial F_m}{\partial K_m}\right)^T$$

quindi la matrice J sarà una matrice a blocchi

$$J = \begin{pmatrix} I - h\ \gamma_{1,1} J_1 & -h\ \gamma_{1,2}\ J_2 & \dots & -h\ \gamma_{1,m}\ J_m \\ -h\ \gamma_{2,1} J_1 & I - h\ \gamma_{2,2}\ J_2 & \dots & -h\ \gamma_{2,m}\ J_m \\ \vdots & \vdots & \ddots & \vdots \\ -h\ \gamma_{m,1} J_1 & -h\ \gamma_{m,2}\ J_2 & \dots & I - h\ \gamma_{m,m}\ J_m \end{pmatrix}$$

J_i è la derivata di f rispetto alla seconda variabile, quindi è una matrice.
La matrice J ha ordine mn quindi il costo computazionale per trovarne l'inversa è $(mn)^3$ ed ovviamente è troppo alto.
Una prima idea per superare questo ostacolo potrebbe essere quella di approssimare le matrici

$$W \simeq J_1 \simeq \dots \simeq J_m$$

Quindi facendo così possi riscrivere J come prodotto tensore

$$J \sim I - \begin{pmatrix} -h\gamma_{1,1} & -h\ \gamma_{1,2} & \dots & -h\ \gamma 1, m \\ \vdots & \vdots & & \vdots \\ -h\gamma_{m,1} & -h\gamma_{m,2} & \dots & -h\gamma_{m,m} \end{pmatrix} \otimes W = I - h\ \Gamma \otimes W$$

e a questo punto si può dimostrare che è possibile trovare l'inversa con un costo nettamente inferiore.
Un'altra strategia è ad esempio quella di calcolare J^{-1} ogni certo numero di passi, quindi supponendo implicitamente che resti pressochè costante durante questi, ad esempio di può calcolare J^{-1} ogni 5 passi.
Nel caso in cui la matrice Γ sia triangolare inferiore allora anche $J \sim I - h\ \Gamma \otimes W$ è triangolare inferiore quindi posso calcolare J^{-1} per sostituzioni (è facile invertire la matrici triangolari).

1.5.11 A-stabilità

Cosideriamo come sempre il problema test

$$\begin{cases} x'(t) = \lambda\ x(t) & t \in [t_0, T] \\ x(0) = 1 \end{cases}$$

dove poniamo $Re(\lambda) < 0$, allora vogliamo che se applichiamo il metodo al problema test deve valere

$$|x_{i+1}| \leq |x_i| \quad \forall i$$

imponiamo questa condizione ad un generico metodo di Runge-Kutta, avremo

$$x_{i+1} = x_i + h \sum_{j=1}^{m} \beta_j K_j$$

dove nel caso del problema test vale

$$K_j = \lambda \left(x_i + h \sum_{l=1}^{m} \gamma_{j,l} K_l \right)$$

osserviamo che

$$\underline{\mathrm{K}} = \begin{pmatrix} K_1 \\ \vdots \\ K_m \end{pmatrix} = \lambda\ x_i \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \lambda\ h\ \Gamma \begin{pmatrix} K_1 \\ \vdots \\ K_m \end{pmatrix}$$

quindi vale che

$$(I - \lambda\ h\ \Gamma)\underline{\mathrm{K}} = \lambda\ x_i\ \underline{1}$$

se il passo di disretizzazione è abbastanza piccolo posso invertire

$$\underline{\mathrm{K}} = \lambda\ x_i (I - \lambda\ h\ \Gamma)^{-1} \underline{1}$$

quindi andando a sostituire ottengo

$$x_{i+1} = x_i + h\underline{\beta}^T \underline{\mathrm{K}} = x_i + \lambda\ h\ x_i\ \underline{\beta}^T (I - \lambda\ h\ \Gamma)^{-1}\ \underline{1} = [1 + h\ \lambda \underline{\beta}^T (I - \lambda\ h\ \Gamma)^{-1} \underline{1}]\ x_i$$

imponiamo ora la condizione di A-stabilità

$$\psi(\lambda h) = 1 + h\lambda \underline{\beta}^T (I - h\lambda\Gamma)^{-1}\ \underline{1}$$

quindi la regione di stabilità dei metodi di Runge-Kutta è

$$L = \{ z \in \mathbb{C} \ \text{ tale che } \ z = h\lambda \ \text{ con } \ |\psi(z)| \leq 1 \}$$

si è già visto che se il metodo è consistente di ordine p, allora abbiamo

$$\underline{\beta}^T \Gamma^j \underline{1} = \frac{1}{(j+1)!} \quad j = 1, \ldots, p$$

usando questa regola, supponendo che il metodo abbia ordine p, possiamo fare lo sviluppo in serie di potenze

$$\psi(z) = 1 + z \sum_{j=0}^{\infty} z^j \underline{\beta}^T \Gamma^j \underline{1} = 1 + z + \frac{z^2}{2} + \cdots + \frac{z^p}{p!} + \sum_{j>p} z^j \underline{\beta}^T \Gamma^{j-1} \underline{1}$$

Quindi se il metodo ha ordine p la prima parte è indipendente dal metodo, l'unico pezzo che dipende dal metodo è

$$\sum_{j>p} z^j \underline{\beta}^T \Gamma^{j-1} \underline{1}$$

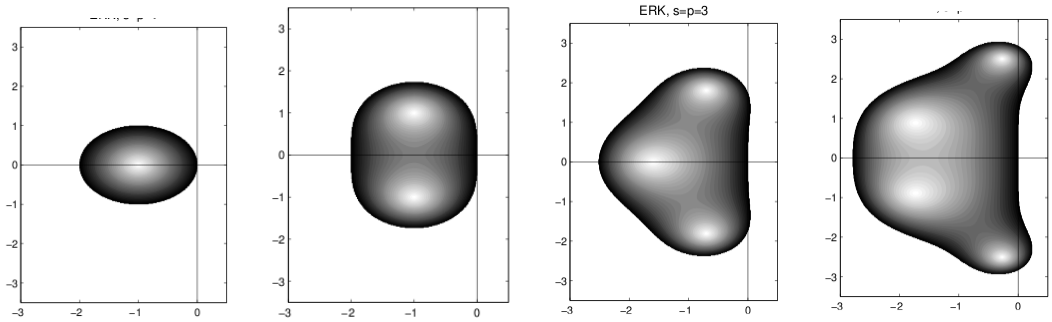
Osserviamo inoltre che se il metodo è esplicito allora Γ è strettamente triangolare, quindi varrà

$$\Gamma^j = 0 \quad j \geq m$$

daltronde sappiamo che $p \leq m$, quindi se l'ordine di consistenza è proprio uguale al numero di stadi, ovvero $m = p$ avremo

$$\psi(z) = 1 + \frac{z^2}{2} + \cdots + \frac{z^p}{p!}$$

Quindi in questi casi la regione di stabilità è indipendente dal metodo.
In realtà si può dimostrare che $p = m \iff m \leq 4$. Di seguito ci sono le regioni di stabilità dei metodi di Runge-Kutta ad 1,2,3,4 stadi con consistenza massima



Nel caso dei metodi di Runge-Kutta impliciti si può dimostrare che

$$\psi(z) = \frac{P(z)}{Q(z)}$$

ovvero è una funzione razionale (rapporto di polinomi).
Tutti i metodi di Runge-Kutta che abbiamo fin ora mostrati sono A-stabili.
NB:Ricordiamo che un metodo è A-stabile se $\mathbb{C}^- \subseteq L$

1.6 Analisi dell'errore

Fin ora abbiamo trascurato gli errori dovuti al fatto di non avere un'artimetica esatta, vogliamo ora fare questo tipo di analisi.
Dunque partiamo dal fatto che non calcoleremo mai esattamente la successione

$$x_{i+1} = x_i + h\ \phi(*)$$

ma calcolaremo una successione $u_i \sim x_i$, ovvero poniamo

$$u_i = x_i + \epsilon_i$$

quindi abbiamo la successione degli errori

$$\epsilon_i = u_i - x_i$$

quindi è ragionevole chiedersi se la successione u_i converge alla soluzione se facciamo tendere a 0 il passo di discretizzazione.
Quindi già a partire dal primo punto (quello noto) abbiamo un errore di troncamento

$$u_0 = x_0 + \rho_0$$

per semplificare i conti faccio l'analisi solo sul metodo di Eulero, quindi avremo

$$u_{i+1} = u_i + h\ f(t_i, u_i) + \rho_{i+1}$$

dove ρ_{i+1} è l'errore locale di arrotondamento. Quindi la successione effettivamente calcolata sarà

$$\begin{cases} u_{i+1} = u_i + h\ f(t_i, u_i) + \rho_{i+1} & 0 \leq i \leq N-1 \\ u_0 = x_0 + \rho_0 \end{cases}$$

Osserviamo che se scelgo un passo di discretizzazione h troppo piccolo allora avrò tanti u_i ma ciò comporta che devo valutare molte volte la funzione f, quindi accumulo errori, quindi devo trovare un h ottimale.

$$E_i = u_i - x(t_i) = u_i - x_i + x_i - x(t_i) = \epsilon_i - e_i$$

dove come sempre e_i è l'errore globale di discretizzazione. Andando avanti con i calcoli

$$\epsilon_{i+1} = u_{i+1} - x_{i+1} = u_i + hf(t_i, u_i) + \rho_{i+1} - x_i - hf(t_i, x_i) =$$
$$= \epsilon_i + h[f(t_i, u_i) - f(t_i, x_i)] + \rho_{i+1}$$

scriviamo ora l'errore globale

$$e_{i+1} = x(t_{i+1}) - x_{i+1} = x(t_i) + hf(t_i, x(t_i)) + h\delta(x(t_{i+1})) - x_i - hf(t_i, x_i) =$$
$$= e_i + h[f(t_i, x(t_i)) - f(t_i, x_i)] + h\delta_{i+1}$$

NB: ogni tanto capita di usare la forma stringata $\delta_{i+1} = \delta(x(t_{i+1}))$
Quindi abbiamo

$$E_{i+1} = \epsilon_{i+1} - e_{i+1} = \epsilon_i + h[\dots] + \rho_{j+1} - e_i - h[\dots] - h\delta_{i+1} =$$
$$= E_i + h[f(t_i, u_i) - f(t_i, x(t_i))] + \rho_{i+1} - h\delta_{i+1}$$

possiamo ora sfruttare la lipschitzianità

$$||E_{i+1}|| \leq ||E_i|| + hL||u_i - x(t_i)|| + ||\rho_{i+1} - h\delta_{i+1}||$$

ovvero, scritto meglio

$$||E_{i+1}|| \leq ||E_i|| + hL||E_i|| + ||\rho_{i+1} - h\delta_{i+1}||$$

supponendo che gli errori siano equilimitati, ovvero che

$$|\rho_i| \leq \rho \quad \forall j$$

e ricordando che gli errori di discretizzazione locale sono tali che

$$|\delta_i| \leq \tau(h) \quad \text{con} \quad \tau(h) \to 0 \quad \text{per} \quad h \to 0$$

ricaviamo che

$$||E_{i+1}|| \leq (1 + hL)||E_i|| + \rho + h\tau(h)$$

Con $||E_0|| = |\rho_0|$ ricordiamo il seguente

Lemma

Data una successione $\{z_k\}_{k \geq k_0}$ con

$$|z_k| \leq (1 + \alpha|z_{k-1}|) + \beta \qquad \alpha, \beta > 0$$

allora vale che

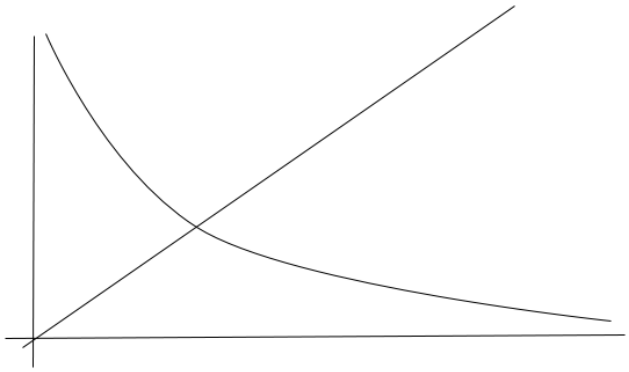
$$|z_k| \leq e^{\alpha(k-k_0)} \left[||z_{k_0}|| + \frac{\beta}{\alpha} \right] - \frac{\beta}{\alpha}$$

posto $\alpha = hL$ e $\beta = \rho + h\tau$, applicando il lemma

$$||E_i|| \leq e^{hLi} \left[||E_0|| + \frac{1}{hL}(\rho + h\tau) \right] - \frac{1}{hL}(\rho + h\tau) \leq$$

$$\leq e^{L(T-t_0)}||E_0|| + \frac{e^{L(T-t_0)} - 1}{L} \left(\frac{\rho}{h} + \tau(h) \right)$$

Quindi ad esempio nel metodo di Eulero che ha consistenza di ordine 1 la funzione $\tau(h)$ è una retta mentre $\frac{\rho}{h}$ è un ramo di iperbole, quindi l'h ottimale è l'intersezione dei due



19/04/2011

1.7 Metodi a più passi (LMM)

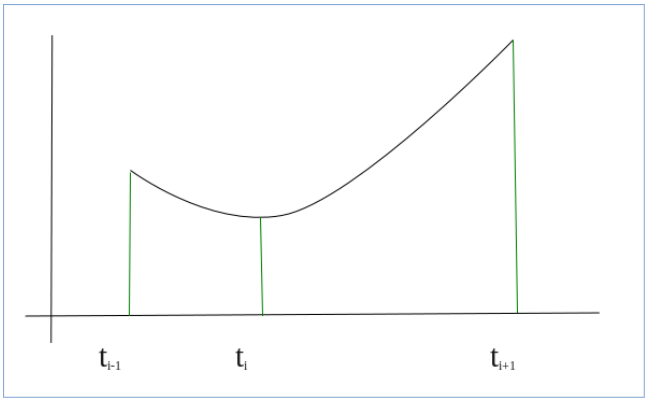
Studiamo ora i metodi per risolvere il solito problema dove l'approssimazione x_{i+1} dipende da $x_i, x_{i-1}, \ldots, x_{i-k+1}$ (metodo a k passi).

Esempio 1.1 (metodo a due passi)**.** Considero come sempre il problema classico e lo trasformo in forma integrale

$$x_{i+1} \simeq x(t_i) + \int_{t_1}^{t_{i+1}} f(t, x(t))\ dt$$

per approssimare l'integrale uso un polinomio $p(t)$ che approssima la funzione $f(t, x(t))$, in particolare chiederò che

$$p(t_j) = f(t_j, x_j) \quad \text{per} \quad i - 1 \leq j \leq i$$



quindi approssimerò

$$\int_{t_1}^{t_{i+1}} f(t, x(t))\ dt = \int_{t_1}^{t_{i+1}} p(t)\ dt$$

con la costruzione che si è fatta nelle lezioni precedenti per costruire il polinomio approssimante devo usare i polinomi di Lagrange quindi facendo i conti

$$\int_{t_1}^{t_{i+1}} p(t)\ dt = \int_{t_1}^{t_{i+1}} \sum_{j=i-1}^{i+1} L_j(t) f(t_j, x_j)\ dt = h \sum_{j=i-1}^{i+1} f(t_j, x_j)\beta_j$$

concludendo, svolgendo gli integrali il metodo che abbiamo appena definito è

$$x_{i+1} = x_i + h\left[\frac{5}{2} f(t_{i+1}, x_{i+1}) + \frac{8}{12} f(t_i, x_i) + \frac{1}{21} f(t_{i-1, x_{i-1}})\right]$$

Osserviamo che in questo caso l'approssimazione di x_{i+1} dipende da lui stesso quindi il metodo è implicito (lo specificheremo meglio dopo), quindi bisogna risolvere l'equazione non lineare per trovare x_{i+1}, cosa che si può far bene se $f(t, x(t))$ è una funzione non troppo complessa.

Metodi a k passi

In generale un metodo a k passi è definito dalla seguente equazione alle differenze

$$\sum_{j=0}^{k} \alpha_j \; x_{i-j+1} = h \; \sum_{j=0}^{k} \beta_j f(t_{i-j+1}, x_{i-j+1}) \quad k \leq i \leq N \quad \alpha_0 \neq 0$$

Dove α_j e β_j sono dei coefficienti asscegnati che dipendono dal metodo. Diremo che il metodo è implicito se $\beta_0 \neq 0$.
Ad ogni metodo assoceremo due polinomi

$$p(\lambda) = \sum_{j=0}^{k} \alpha_j \lambda^{k-j}$$

$$\sigma(\lambda) = \sum_{j=0}^{k} \beta_j \lambda^{k-j}$$

1.7.1 Esempi

Esempio 1.2 (Metodi di Adams-Moulton (impliciti))**.** Riprendiamo come prima il problema scritto in forma integrale

$$x_{i+1} = x(t_i) + \int_{t_1}^{t_{i+1}} f(t, x(t)) \; dt$$

approssimerò l'integrale di $f(t, x(t))$ con l'integrale di un polinomio $p(t)$ tale che

$$p(t_j) = f(t_j, x_j) \qquad i + 1 \leq j \leq i - k + 1$$

dove k sarà il numero di passi del metodo, quindi apporssimo l'integrale della funzione da t_i ad t_{i+1} con l'integrale del polinomio che interpola $f(t, x(t))$ nei punti $(t_j, x(t_j))$ dove $i + i \leq j \leq i - k + 1$, quindi facendo il conto (usando come sempre i polinomi di Lagrange)

$$\int_{t_i}^{t_{i+1}} p(t) \; dt = \sum_{j=0}^{k} \int_{t_i}^{t_{i+1}} L_j(t) f(t_{i-j+1}, x_{i-j+1}) dt$$

ponendo

$$h\beta_j = L_j(t)$$

otteniamo il metodo generale di Adams-Moulton

$$x_{i+1} - x_i = h \sum_{j=0}^{k} \beta_j f(t_{i-j+1}, x_{i-j+1})$$

In realtà volendo usare la terminologia corretta questa è una famiglia di metodi a k passi. I coefficienti sono per i primi due casi

k	β_0	β_1	β_3	$\ldots$
0	1			$\ldots$
1	1/2	1/2		$\ldots$
2	5/12	8/12	$\ldots$	$\ldots$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$

Quindi nel caso $k = 1$ abbiamo il metodo di Eulero.

Osservazione 1.13. Per usare i metodi a k passi è necessario conoscere $x_0, x_1, \ldots, x_{k-1}$, daltronde dal problema è noto solo x_0, quindi l'idea è di usare un metodo ad un passo per calcolare x_1, un metodo a due passi per calcolare x_2 e così via.
Ad esempio se si dispone di una famiglia di metodi come quella di Adams-Moulton è immediato implementare questo algoritmo.

Esempio 1.3 (Metodi di Adams-Bashforth (espliciti))**.** Anche in questo caso i metodi di Adams-Bashforth sono una famiglia di metodi e sono simili alla famiglia di metodi di Adams-Moulton ma pongo $\beta_0 = 0$ per rendere esplicito il metodo.
Quindi approssimo l'integrale di $f(t, x(t))$ tra t_i e t_{i+1} con l'integrale del polinomio che approssima $f(t, x(t))$ interpolando

$$p(t) = f(x_j, t_j) \qquad i + 1 \leq j \leq i - k$$

Anche in questa famiglia i coefficienti si ricavano e sono

k	β_0	β_1	β_2	...
1	1			...
2	3/2	-1/2		...
3	23/12	...	...	...
...	...	...	...	...

Esempio 1.4 (Metodi BBF (impliciti))**.** Consideriamo ora la seguente famiglia di metodi

$$\sum_{j=0}^{k} \alpha_j x_{i-j+1} = h f(t_{i+1}, x_{i+1})$$

L'idea che c'è dietro questi metodi è quella di non approssimare l'integrale (come abbiamo fatto sin ora) ma di approssimare la derivata della soluzione.

Suppongo di avere

$$(t_{i-j+1}, x_{i-j+1}) \qquad 0 \leq j \leq k$$

e considero il polinomio $p(t)$ tale che

$$p(t_{i-j+1}) = x_{i-k+1} \qquad 0 \leq j \leq k$$

so già che vale

$$x'(t_{i+1}) = f(t_{i+1}, x_{i+1}) \qquad 0 \leq i \leq N$$

quindi l'idea è di approssimare $x(t)$ con un polinomio $p(t)$ tale che

$$x'(t_{i+1}) \simeq p'(t_{i+1})$$

usando come sempre i polinomi di Lagrange

$$p(t) = \sum_{j=0}^{k} L_j(t) x_{i-j+1}$$

quindi troviamo

$$\sum_{j=0}^{k} L_j'(t) x_{i-j+1}$$

quindi

$$p'(t_{i+1}) = \frac{1}{h} \sum_{j=0}^{k} \alpha_j x_{i-j-1}$$

quindi trovo il metodo scitto all'inizio.

Anche in questo caso i coefficienti sono noti, abbiamo

$$\beta_j = \begin{cases} 1 & \text{se} \quad j = 0 \\ 0 & \text{altrimenti} \end{cases}$$

mentre

k	α_0	α_1	α_2	...
1	1	-1		...
2	3/2	-2	1/2	...
...	...	...	...	...

20/04/2011

1.7.2 Consistenza

Ricordiamo che i metodi LLM sono definiti dalla seguente equazione alle differenze

$$\sum_{j=0}^{k} \alpha_k x_{i-j+1} = h \sum_{j=0}^{k} \beta_j f(t_{i-j+1}, x_{i-j+1}) \qquad k \leq i \leq N$$

dove $x_0 = x(t_0)$ è noto mentre $x_1, \ldots, x_{k-1}$ supponiamo di averli calcolati in qualche modo.
Abbiamo inoltre associato due polinomi ad un metodo LLM

$$\rho(\lambda) = \sum_{j=0}^{k} \alpha_j \lambda^{k-j}$$

$$\sigma(\lambda) = \sum_{j=0}^{k} \beta_j \lambda^{k-j}$$

osserviamo che $p(\lambda)$ è il polinomio che avremmo associato all'equazione alle defferenze scritta sopra, ma ne parleremo meglio più avanti.

1.7.3 Errore di discretizzazione

Nel caso dei metodi a più passi abbiamo che l'errore di discretizzazione è

$$\delta(x(t+h), t) = \frac{1}{h} \left[\sum_{j=0}^{k} \alpha_j x(t - jh) - h \sum_{j=0}^{k} \beta_j f(t - jh, x(t - jh)) \right]$$

quindi come nel caso dei metodi ad un passo supponiamo che $x_i, \ldots, x_{k-1}$ siano stati calcolati senza errore e siano quindi esattamente $x(t_i), \ldots, x(t_{k-1})$
Diremo che un metodo è consistente se $\delta(x(t), h) \to 0$ per $h \to 0$ per ogni valore $t \in [t_0, T]$

Condizioni necessarie e sufficienti di consistenza

Un metodo LLM è consistente se e soltato se $\rho(1) = 0$ e $\rho'(1) = \sigma(1)$

Per dimostrarlo sviluppiamo in serie

$$x(t - jh) = x(t - kh) + h(k - j)x'(t - kh) + O(h^2) \qquad 0 \leq j \leq k$$

mentre per la derivata prima svliluppiamo

$$x'(t - jh) = x'(t - kh) + O(h)$$

sostituendo queste espressione nell'errore di discretizzazione locale abbiamo

$$\delta(*) = \frac{1}{h} \left[\sum_{j=0}^{k} \alpha_j x(t - kh) + h \left[\sum_{j=0}^{k} (\alpha_j (k - j) - \beta_j) \right] x'(t - kh) + O(h^2) \right]$$

quindi se il metodo è consistente necessariamente deve essere

$$\sum_{j=0}^{k} \alpha_j = 0$$

e anche

$$\sum_{j=0}^{k} (\alpha_j (k - j) - \beta_j) = 0$$

osserviamo che

$$\rho(1) = \sum_{j=0}^{k} \alpha_j$$

mentre

$$\rho'(1) - \sigma(1) = \sum_{j=0}^{k} (\alpha_j(k-j) - \beta_j)$$

Vogliamo ora vedere quando il metodo è consistente di ordine p. Sviluppiamo in un intorno di t

$$x(t - jh) = \sum_{q=0}^{p} \frac{x^{(q)}(t)}{q!} j^q h^q + O(h^{p+1})$$

$$x'(t - jh) = \sum_{q=0}^{p-1} \frac{x^{(q+1)}(t)}{q!} j^q h^q + O(h^p)$$

sostituendo troviamo

$$\delta(*) = \frac{1}{h} \left[\sum_{j=0}^{k} \alpha_j \sum_{q=0}^{p} \frac{x^{(q)}(t)}{q!} j^q h^q - \sum_{j=0}^{k} \beta_j \sum_{q=0}^{p-1} \frac{x^{(q+1)}(t)}{q!} j^q h^{q+1} + O(h^{p+1}) \right]$$

posso riscriverlo come

$$\delta(*) = \frac{1}{h} \left[\sum_{j=0}^{k} \alpha_j \sum_{q=0}^{p} \frac{x^{(q)}(t)}{q!} j^q h^q - \sum_{j=0}^{k} \beta_j \sum_{q=1}^{p} \frac{x^{(q)}(t)}{(q-1)!} j^{q-1} h^q + O(h^{p+1}) \right]$$

in conclusione ho

$$\delta(*) = \frac{1}{h} \left[\sum_{j=0}^{k} \alpha_j x(t) + \sum_{q=1}^{p} h^q x^{(q)}(t) \sum_{j=0}^{k} \left(\alpha_j \frac{j^q}{q!} - \beta_j \frac{j^{q-1}}{(q-1)!} \right) + O(h^{p+1}) \right]$$

si conclude che $\delta(*) = O(h^p)$ quando si verifica che

$$\sum_{j=0}^{k} \alpha_j = 0$$

e

$$\sum_{j=0}^{k} \left(\frac{j^q \alpha_j}{q} - j^{q-1} \ \beta_j \right) = 0 \quad 1 \leq q \leq p$$

quindi è facile verificare quando un metodo è consistente di ordine p.

Quindi è facile generare un metodo consistente di qualsiasi ordine, in generale se voglio un metodo consistente di ordine p e se lo voglio esplicito devo scegliere $2(k+1)$ coefficienti che soddisfano quelle relazioni. Il problema è che in generale per i metodo a più passi la consistenza di un metodo non implica la convergenza. Il problema nasce dall'approssimazione che ho dei primi k punti $x_0, \ldots, x_{k-1}$, un metodo convergente in qualche modo deve dimenticarsene, ma vedremo più un dettaglio questo aspetto.

1.7.4 0-stabilità (zero-stabilità)

Consideriamo il seguente problema

$$\begin{cases} x'(t) = 0 & t \in [0, T] \\ x(0) = 0 \end{cases}$$

Questo ha come soluzione $x(t) = 0$, se usiamo un metodo LLM allora l'equazione alle differenze sarà

$$\sum_{j=0}^{k} \alpha_j x_{i-j+1} = 0$$

sappiamo risolvere questo tipo di equazioni, quindi se $\lambda_1, \dots, \lambda_k$ sono le radici di $\rho(\lambda)$ la generica soluzione la posso scrivere come

$$x_i = \sum_{q=1}^{k} c_q \lambda_q^i$$

dove i coefficienti c_q sono determinati dalle condizioni iniziali $x_0, \dots, x_{k-1}$. Quindi quello che si vuole è che la soluzione non diverga a prescindere dalle condizioni iniziali, quindi in un certo senso voglio eliminare la dipendenza da $x_0, \dots, x_{k-1}$ quindi ad esempio posso chiedere che le radici λ siano di modulo minore o uguale ad 1 e le radici di modulo 1 siano semplici.

Diremo che un LMM è 0-stabile se $\rho(\lambda)$ ha radici di modulo minore o uguale ad 1 e le radici di modulo 1 sono semplici.

Osserviamo che $\rho(1) = 0$ per la condizione di consistenza, quindi esiste sempre una radice di modulo 1, vogliamo dunque che questa radice sia semplice, ovvero $\rho'(1) = \sigma(1) \neq 0$

Dahlst barrier

Un metodo a k passi può avere ordine di consistenza al massimo $k+1$ se k è dispari, $k+2$ se k è pari. Quindi l'ordine di consistenza è limitato dal numero di passi.

Esempio 1.5. Consideriamo il metodo

$$x_{i+1} - x_{i-1} = 2hf(t_i, x_i)$$

i polinomi associati sono

$$\rho(\lambda) = \lambda^2 - 1 \qquad \sigma(\lambda) = \lambda$$

questo metodo è 0-stabile e ha ordine di consistenza 2

Esempio 1.6. Consideriamo il metodo

$$x_{i+1} - x_{i-1} = \frac{h}{3}\left(f(t_{i+1}, x_{i+1}) + 4f(t_i, x_i) + f(t_{i-1}, x_{i-1})\right)$$

quindi

$$\rho(\lambda) = \lambda^2 - 1$$

questo metodo è 0-stabile e ha ordine di consistenza 4 (massimo)

D'ora in avanti considereremo spesso il problema modello

$$\begin{cases} x'(t) = \mu(t)x(t) + g(t) & t \in [0, T] \\ x(t_0) = 0 \end{cases}$$

Le dimostrazioni le faremo su questo problema dato che è possibile ricondurre un problema classico in questa forma o comunque generalizzare le dimostrazioni.
Se applichiamo un metodo a k passi a questo problema otteniamo

$$\sum_{j=0}^{k} \alpha_j x_{i-j+1} = h \sum_{j=0}^{k} \beta_j \left(\mu(t_{i-j+1})\ x_{i-j+1} + g(t_{i-j+1})\right)$$

quindi

$$\sum_{j=0}^{k} (\alpha_j - h\ \beta_j\ \mu(t_{i-j+1})\ x_{i-j+1} = h \sum_{j=0}^{k} \beta_j g(t_{i-j+1})$$

definisco

$$\underline{\mathrm{x}}_i = \begin{pmatrix} x_i \\ \vdots \\ x_{i+k-1} \end{pmatrix} \qquad A_i^h = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ \gamma_i & \gamma_{i+1} & \cdots & \gamma_{i+k-1} \end{pmatrix} \qquad \mathrm{g}_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \dfrac{h - \sum \dots}{\alpha_0 - h\beta_i \dots} \end{pmatrix}$$

dove

$$\gamma_{i-j+k}^h = \frac{h\beta_j \ \mu(t_i - j + k) - \alpha_j}{\alpha_0 - h \ \beta_0 \ \mu(t_{i+k})}$$

con queste notazioni posso riscrivere l'equazione alle differenze come

$$\underline{\mathrm{x}}_{i+1} = A_i \underline{\mathrm{x}}_i + \mathrm{g}_i$$

Se consideriamo il caso $\mathrm{g}_i = 0$ allora

$$\underline{\mathrm{x}}_i = A_i A_{i-1} \dots A_0 \underline{\mathrm{x}}_0$$

affinchè la soluzione non diverga è sensato chiedere che il prodotto delle matrici non diverga, osserviamo che

$$A_i^0 = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ \gamma_i^0 & \cdots & \cdots & \gamma_{i+k-1}^0 \end{pmatrix}$$

dove

$$\gamma_{i-j+k}^0 = -\frac{\alpha_j}{\alpha_0}$$

quindi A_i^0 è la matrice companion associata a $\rho(\lambda)$

Teorema 1.4. Siano i, h tali che $ih \leq T$ e sia

$$h < \widehat{h} = \frac{|\alpha_0|}{|\beta_0| \max_{[0,T]}} |\mu(t)|$$

Assumiamo che il metodo sia 0-stabile e sia

$$||(A^0)^m|| \leq K \quad \forall m$$

Osserviamo che dall'ipotesi di 0-stabilità K esiste perchè gli autovalori di A^0 hanno modulo minore o uguale ad 1 e quelli di modulo 1 sono semplici.
Definiamo

$$M = \max_{[o,T]} \frac{\mu(t)}{|\alpha_0|(|\alpha_0| - h|\mu(t)\beta_0|)} \cdot \left(|\alpha_0| \sum_{j=1}^{k} |\beta_j| + |\beta_0| \sum_{j=1}^{k} |\alpha_j| \right)$$

Allora vale che

$$\forall j \geq 1 \quad \forall i \geq l \qquad || \prod_{j=l}^{i-1} A_j^h || \leq K e^{KMh(i-l)} \leq K e^{KMT}$$

Significato
Se il metodo è 0-stabile e se $h \leq \widehat{h}$ allora il prodotto delle matrici companion scritto prima è limitato. Per una dimostrazione del teorema vedere R. Mattheij and J. Molenaar, Ordinary Differential Equations in Theory and Practice teorema 3.7 (dimostrazione non richiesta).

Teorema 1.5. Supponiamo che il metodo LMM sia 0-stabile e consistente di ordine p. Supponiamo che $|\delta(*)| \leq Ph^p$ dove $P, p > 0$. Supponiamo inoltre $||(A^0)^m|| \leq K \forall m$. Sia

$$\widehat{h} = \frac{|\alpha_0|}{|\beta_0| L}$$

dove L è la costante lipschitziana del problema, sia

$$M = \frac{L}{|\alpha_0|(|\alpha_0| - h|\beta_0|L)} \cdot \left(|\alpha_0| \sum_{j=1}^{k} |\beta_j| + |\beta_0| \sum_{j=1}^{k} |\alpha_j| \right)$$

supponiamo

$$||x_i - x(t_i)|| \leq Qh^q \qquad 1 \leq i \leq k-1$$

allora per ogni scelta di i, h tali che $ih \leq T$ e $h \leq \widehat{h}$ vale che

$$||e_i^h|| \leq e^{KTt_i} \left(\frac{P}{M} h^q + KQh^q \right)$$

Significato
Se un metodo è 0-stabile, consitente di ordine p e i punti iniziali sono approssimati bene, allora ho che per h abbastanza piccolo una maggiorazione dell'errore globale.
Quindi consistenza di ordine p e approssimazione dei punti iniziali di ordine q implica che il metodo è convergente di ordine $O(h^{\min(p,q)})$

idea della dimostrazione. Lo dimostro per il problema test

$$\begin{cases} x'(t) = \mu(t)x(t) + g(t) \\ x(0) = x_0 \end{cases}$$

quindi applicando un metodo a k passi ottengo

$$\sum (\alpha_j - h\beta_j \mu(*))\ x_{i-j+1} = h \sum \beta_j g(*)$$

uso la definizione di $\delta(*)$ e ottengo

$$\sum (\alpha_j - h\beta_j \mu(*))\ x(t_{i-j+1}) = h \sum \beta_j g(*) + h\delta_i$$

faccio la defferenza tra le due espressioni e ottengo

$$\sum (\alpha_i - h\beta_j \mu(*))\ e_{i-j+1} = h\delta_i$$

dove

$$e_i = x(t_i) - x_i$$

quindi ho un'equazione alle differenze. Definisco

$$\underline{\mathrm{e}}_i = \begin{pmatrix} e_i \\ e_{i+1} \\ \vdots \\ e_{i+k-1} \end{pmatrix}$$

Allora ottengo

$$\underline{\mathrm{e}}_{i+1} = A_i^h e_i + \underline{\mathrm{d}}_i$$

dove

$$\underline{\mathrm{d}}_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \delta_i \end{pmatrix}$$

svolgendo i calcoli trovo che

$$\underline{\mathrm{e}}_i = \sum \left(\prod A_*^h \right) \cdot \underline{\mathrm{d}}_* + \prod A^h \underline{\mathrm{e}}_0$$

ora è sufficiente passare alle norme e usare il teorema precedente

$$||e_i|| \leq K \frac{e^{kMhi} - 1}{e^{kMh} - 1} P h^{p+1} + e^{kMhi}\ Q\ h^q$$

□

04/05/2011

Ricapitolazione lemma della lezione precedente ed esempio

L'altra volta abbiamo introdotto le seguenti matrici companion relative ad un LMM

$$\begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ \gamma_i^h & \cdots & \cdots & \gamma_{i+k-1}^h \end{pmatrix}$$

E abbiamo mostrato che se LMM è 0-stabile allora

$$\forall i > l \qquad \left\| \prod_{j=l}^{i-1} A_j^h \right\| \leq K e^{kM(i-l)}$$

Dove M è una costante opportuna e K è tale che

$$\|A_i\| \leq K \quad \text{dove} \quad A := A_i^0$$

avevamo anche ricavato una maggiorazione e ci chiediamo se questa è realistica.
Consideriamo il problema

$$\begin{cases} x'(t) = \lambda x & \lambda < 0 \\ x(0) = 1 \end{cases}$$

Usiamo il seguente metodo (metodo del punto di mezzo esplicito)

$$\frac{1}{2}x_{i+1} - \frac{1}{2}x_{i-1} = hf(t_i, x_i)$$

Questo metodo è consistente e convergente con $\sigma(\lambda) = \lambda^2 - 1$, quindi è anche 0-stabile. Osserviamo che $\sigma(\lambda)$ ha come radici $\{1, -1\}$, mi chiedo quale sia l'effetto della radice -1. Riscriviamo l'equazione alle differenze del problema con il metodo che stiamo usando

$$x_{i+1} = 2\lambda h x_i - x_{i-1} = 0$$

passando alle matrici con la solita notazione

$$\underline{\mathrm{x}}_i = \begin{pmatrix} x_i \\ x_{i+1} \end{pmatrix} \qquad A_i^h = \begin{pmatrix} 0 & 1 \\ 1 & 2\lambda h \end{pmatrix} \qquad \underline{\mathrm{x}}_{i+1} = A_i^h \underline{\mathrm{x}}_i$$

gli autovalori di questa matrice sono

$$\rho_1^h = \sqrt{1 + h^2\lambda^2} + \lambda h \doteq 1 + \lambda h + \frac{1}{2}\lambda^2 h^2$$

$$\rho_2^h = -\sqrt{1 + h^2\lambda^2} + \lambda h \doteq -1 + \lambda h - \frac{1}{2}\lambda^2 h^2$$

quindi

$\rho_1^h < 1$ per h piccolo

$\rho_2^h > 1$ per h piccolo

Quindi la limitazione superiore trovata nel lemma è esponenziale e non si può far nulla dato che abbiamo trovato l'uguaglianza persino per un problema stabile, mostriamolo esplicitamente: le soluzioni sono

$x_{i+1}^h \doteq (1 + \lambda h)x_i^h$ relativa a ρ_1^h

$y_{i+1}^h \doteq (-1 + \lambda h)y_i^h$ relativa a ρ_2^h

quindi abbiamo

$x_i^h \sim e^{\lambda h} x_0 \to 0$

$y_i^h \sim -e^{-\lambda h} y_0 \to \infty$

pertanto non va bene dato che se consideriamo il prodotto

$$\|A_{i-1}^h A_i^h \dots A_l^h\| \leq e^{|\lambda|(1-l)h}$$

e tutto questo è dovuto al fatto che il polinomio $\sigma(\lambda)$ ha un'altra radice di modulo 1. Osserviamo che ho comunque convergenza dato che la produttoria finita di queste matrici companion è limitata da una costante.

1.7.5 Condizioni necessarie e sufficienti di convergenza

Teorema 1.6. Siano $x_1^h, \dots, x_{k-1}^h$ tali che $|x_i^h - x(t_i)| \leq Qh^q$ per $1 \leq i \leq k-1$ (ovvero le prime $k-1$ approssimazioni hanno un errore che va a zero come h^q) dove $Q, q > 0$ e $x_0 = x(t_0)$ non ha errore, allora LMM è convergente se e soltanto se LMM è consistente e 0-stabile.

Dimostrazione. L'implicazione $\Leftarrow$ è stata dimostrata nella lezione precedente, mostriamo dunque l'implicazione $\Rightarrow$. Ricordiamo la definizione di metodo convergente, dato l'errore

$$e_i^h = x(t_i) - x_i^h$$

il metodo è convergente se

$$\lim_{h \to 0} e_i^h = 0 \qquad \forall i \quad \text{tale che} \quad ih < T$$

consideriamo dunque il problema

$$\begin{cases} x'(t) = f(t, x(t)) & t \in [0, T] \\ x(0) = x_0 \end{cases}$$

mostriamo che LMM è 0-stabile. Prendiamo un metodo a più passi lineare

$$\sum_{j=0}^{k} \alpha_j x_{i-j+1} = h \sum_{j=0}^{k} \beta_j f(t_{i-j+1}, x_{i-j+1})$$

supponiamo per assurdo che il metodo non sia 0-stabile, allora possono succedere due cose

caso 1: esiste μ tale che $|\mu| > 1$ e $\rho(\mu) = 0$

caso 2: esiste μ tale che $|\mu| = 1$ e $\rho(\mu) = \rho'(\mu) = 0$ (è radice doppia)

dove

$$\rho(\lambda) = \sum_{j=0}^{k} \alpha_j \lambda^{k-j}$$

allora consideriamo il problema seguente

$$\begin{cases} x'(t) = 0 & t \in [0, T] \\ x(0) = 0 \end{cases}$$

La soluzione sappiamo già che è $x(t) = 0$ quindi applicando il metodo a questo problema abbiamo

$$\sum_{j=0}^{k} \alpha_j x_{i-j+1} = 0$$

definiamo la seguente successione in base a quale caso siamo

$$y_i = \begin{cases} \mu^i y_0 & \text{se siamo nel caso 1} \\ i\mu^i y_0 & \text{se siamo nel caso 2} \end{cases}$$

non è difficile verificare che la successione $\{y_i\}_{i\geq 1}$ risolve l'equzione alle differenze. Definisco la successione

$$\begin{cases} x_i^h = \sqrt{h} y_i & i \geq 1 \\ x_0 = x(0) \end{cases}$$

Verifichiamo le ipotesi del teorema, ovvero che i primi k punti siamo delle buone approssimazioni, infatti

$$|x_i^h - x(t_i)| = \sqrt{h}|y_i| \leq \max_{1\leq i\leq k-1} |y_i| h^{1/2} \quad 1 \leq i \leq k-1$$

Quindi abbiamo che con $Q = \max_{1\leq i\leq k-1} |y_i|$ e $q = 1/2$ verifichiamo le ipotesi del teorema.
Riprendiamo la definizione di convergenza, nel caso 1

$$\lim_{h\to 0} x_i^h = \lim_{h\to 0} \sqrt{h}\mu^i y_0 = \infty \quad ih \leq T$$

Quindi il metodo non converge contro le ipotesi, quindi assurdo. Nel caso 2 si ripetono gli stessi passaggi e compare un fattore i^i che fa divergere il metodo e quindi si giunge nuovamente ad un assurdo, quindi il metodo è 0-stabile. Mostrata la 0-stabilità resta da vedere la consistenza o equivalentemente che $\rho(1) = 1$. Consideriamo il seguente problema

$$\begin{cases} x'(t) = 0 & t \in [0.T] \\ x(0) = 1 \end{cases}$$

Sappiamo che la soluzione di questo problema è $x(t) = 1$. Un qualsiasi metodo applicato a questo problema ci da

$$\sum_{j=0}^{k} \alpha_j x_{i-j+1} = 0 \quad i \geq k$$

Pongo $x_i^h = 1$ per $0 \leq i \leq k-1$ quindi banalmente si verificano le ipotesi del teorema e x_i^h per $i \geq k$ è ottenuto mediante l'equazione alle differenze. Per ipotesi so che il metodo è convergente, quindi

$$\lim_{h\to 0} x_i^h = 1 \quad ih \leq T$$

Segue che

$$0 = \sum_{j=0}^{k} \alpha_j x_{i-j+1} \to \sum_{j=0}^{k} \alpha_j$$

Quindi

$$\rho(1) = \sum_{j=0}^{k} \alpha_j = 0$$

Resta da provare che $\rho'(1) = \sigma(1)$ con $\rho'(1) \neq 0$ (perchè LMM è 0-stabile). Considero allora il seguente problema

$$\begin{cases} x'(t) = 1 & t \in [0, T] \\ x(0) = 0 \end{cases}$$

La soluzione è $x(t) = t$, quindi un qualsiasi metodo diventa

$$\sum_{j=0}^{k} \alpha_j x_{i-j+1} = h \sum_{j=0}^{k} \beta_j = h\sigma(1)$$

Definisco la successione

$$y_i^h = ih\frac{\sigma(1)}{\rho'(1)} \qquad i \geq 0$$

troviamo che (usando $x(t_i) = t_i = ih$)

$$|y_i^h - x(t_i)| = \left| ih\frac{\sigma(1)}{\rho'(1)} - ih \right| = ih \left| \frac{\sigma(1)}{\rho'(1)} - 1 \right| \leq (k-1) \left| \frac{\sigma(1)}{\rho'(1)} - 1 \right| h$$

Dove ponendo

$$Q = (k-1)\left|\frac{\sigma(1)}{\rho'(1)} - 1\right| \quad \text{e} \quad q = 1$$

verifichiamo le ipotesi del teorema sul fatto che i primi punti siano ben approssimati. Mostriamo ora che la successione $\left\{y_i^h\right\}_{i\geq 0}$ soddisfa l'equazione alle differenze, ovvero deve soddisfare

$$\sum_{j=0}^{k} \alpha_j x_{i-j+1} = h\sum_{j=0}^{k} \beta_j = h\sigma(1)$$

sostituendo voglio che sia verificata

$$\sum_{j=0}^{k} \alpha_j (i-j+1) h\frac{\sigma(1)}{\rho'(1)} = h\sigma(1)$$

Devo dunque verificare che

$$\sum_{j=0}^{k} \alpha_j (i-j+1) = \rho'(1)$$

daltronde so che

$$\rho'(1) = \sum_{j=0}^{k} (k-j)\alpha_0$$

quindi ottengo

$$\sum_{j=0}^{k} \alpha_j (i-j+1) = \sum_{j=0}^{k} (k-j)\alpha_j + \sum_{j=0}^{k} \alpha_j + \quad \text{una costante che dipende da } i$$

Daltronde abbiamo già visto prima che

$$\sum_{j=0}^{k} \alpha_j = 0$$

Quindi vale l'identità. Quindi y_i^h converge alla soluzione, ovvero

$$y_i^h \to x(t_i) \quad \forall t = ih$$

Ma sappiamo che

$$y_i^h = ih\frac{\sigma(1)}{\rho'(1)} = x(t_i) = ih$$

pertanto

$$\frac{\sigma(1)}{\rho'(1)} = 1$$

che è appunto la tesi. □

1.7.6 A-stabilità

Consideriamo il problema test

$$\begin{cases} x'(t) = \lambda x(t) & t \in [0,T] \quad \lambda \in \mathbb{C} \quad Re(\lambda) < 0 \\ x(0) = 1 \end{cases}$$

Conosciamo già la soluzione

$$x(t) = e^{\lambda t} \to 0$$

Se applichiamo un generico metodo a più passi otteniamo

$$\sum_{j=0}^{k} \alpha_j x_{i-j+1} = h\lambda\sum_{j=0}^{k} \beta_j x_{i-j+1}$$

ovvero

$$\sum_{j=0}^{k}(\alpha_j - h\lambda\beta_j)x_{i-j+1}$$

La condizione di A-stabilità per i metodi ad un passo era che la successione fosse in modulo non crescente, ovvero $|x_{i+1}| \leq |x_i|$, in questo contesto facciamo una richiesta più debole, chiediamo che $|x_i|$ sia limitato, quindi associamo al metodo un polinomio

$$p(\omega) = \rho(\omega) - h\lambda\sigma(\omega)$$

quindi chiederemo che le radici

$$\omega_1(h\lambda), \ldots, \omega_k(h\lambda)$$

siano tutte di modulo ≤ 1, dunque questa è la condizione di A-stabilità per i metodi a più passi. La regione di stabilità assoluta è definita da

$$L = \{z = h\lambda \quad \text{tale che} \quad |\omega_i(z)| \leq 1 \quad \text{con} \quad 1 \leq i \leq k\}$$

Definizione

$$\psi(z) = \max_{1\leq i\leq k} |\omega_i(z)|$$

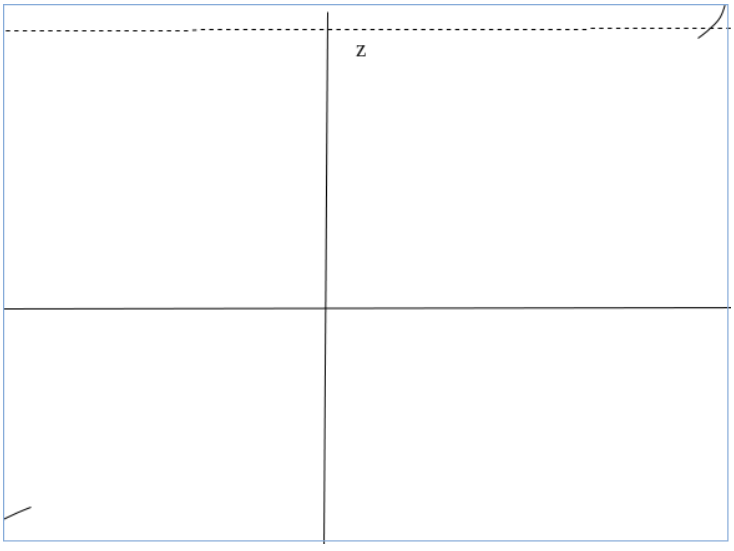
Teorema 1.7. Se $\psi(z) \leq 1 \quad \forall z \in \mathbb{R}$, allora il metodo è implicito.

idea della dimostrazione. Se cerco gli zeri di $p(\omega)$ allora osservo che

$$z = h\lambda = \frac{\rho(\omega)}{\sigma(\omega)}$$

Se $\alpha_0 \neq 0$ il numeratore ha grado k mentre il denominatore ha grado $\leq k$ (precisamente ha grado k se il metodo è implicito, $< k$ se il metodo è esplicito). Se LMM è esplicito allora

$$\frac{\rho(\omega)}{\sigma(\omega)} \simeq \frac{\alpha_0\omega}{\beta_1} \quad \text{per } \omega \text{ grande}$$



Quindi se $|z|$ è abbastanza grande trovo sempre una soluzione ω con $|\omega| > 1$ quindi non posso avere $\mathbb{R}$ come sottoinsieme della regione di stabilità assoluta. □

10/05/2011

A-stabilità

Si consideri al solito il problema test

$$\begin{cases} x'(t) = \lambda x(t) & t \in [0,T] \ , \quad Re(\lambda) < 0 \\ x(0) = 1 \end{cases}$$

Se applichiamo un generico LMM a questo problema otteniamo l'equazione alle differenze

$$\sum_{j=0}^{k} (\alpha_j - h\lambda\beta_j) x_{i-j+1} = 0$$

Consideriamo il polinomio associato a questa equazione alle differenze

$$p(\omega) = \rho(\omega) - h\lambda\sigma(\omega)$$

Se gli zeri di questo polinomio sono in modulo più piccoli di 1 allora abbiamo che la successione $|x_i|$ p non crescente. In generale definiamo

$$\psi(h\lambda) = \max_{1 \le j \le n} \{|\omega_j(h\lambda)| \ \text{ tale che } \ p(\omega_j(h\lambda)) = 0\}$$

La regione di stabilità sarà

$$L = \{z \in \mathbb{C} \ \text{ tale che } \ z = h\lambda \ \ , \ \ \psi(z) \le 1\}$$

Diremo che un metodo è A-stabile se il semipiano sinistro è nella regione di stabilità, ovvero se $\mathbb{C}^- \subseteq L$

Teorema 1.8. Se $\psi(z) \le 1$ per ogni $z \in \mathbb{R}$ allora il metodo è implicito.

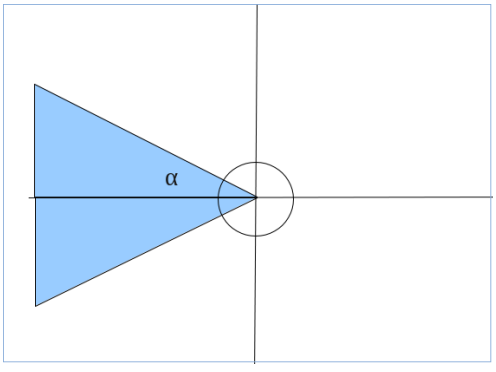
La dimostrazione è stata accennata nella lezione precedente. Come corollario abbiamo subito che se un metodo è esplicito allora non è A-stabile

Teorema 1.9 (Dahlquist barrier)**.** Se un LMM è A-stabile allora ha ordine di convergenza minore o uguale a 2.

A(α)-stabilità

Diremo che un metodo è A(α)-stabile se $V(\alpha) \subseteq L$ dove

$$V(\alpha) = \{z \in \mathbb{C} \ \text{ tale che } \ |\arg(-z)| \le \alpha\}$$



Esempio 1.7 (metodo dei trapezi)**.** Consideriamo il metodo dei trapezi

$$x_{i+1} = x_i + \frac{1}{2} h \left(f(t_i, x_i) + f(t_{i+1}, x_{i+1}) \right)$$

Questo metodo ha convergenza di ordine 2, mostriamo che è A-stabile.

$$\rho(\omega) = \omega - 1 \qquad \sigma(\omega) = \frac{1}{2} + \frac{1}{2}\omega$$

Quindi il polinomio associato all'equazione alle differenze è

$$p(\omega) = \omega - 1 - z(\omega + 1)$$

quindi abbiamo

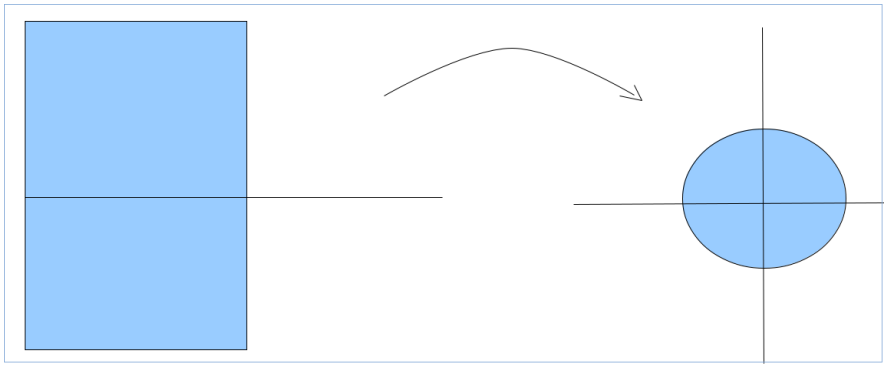
$$\omega = \frac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z} = \frac{2 + z}{2 - z}$$

Trasformazioni di Cayley

Le trasformazioni del tipo

$$C_\gamma(z) = \frac{\gamma + z}{\gamma - z}$$

sono dette trasformazioni di Cayley e hanno la proprietà di mandare $\mathbb{C}^-$ nel cerchio unitario



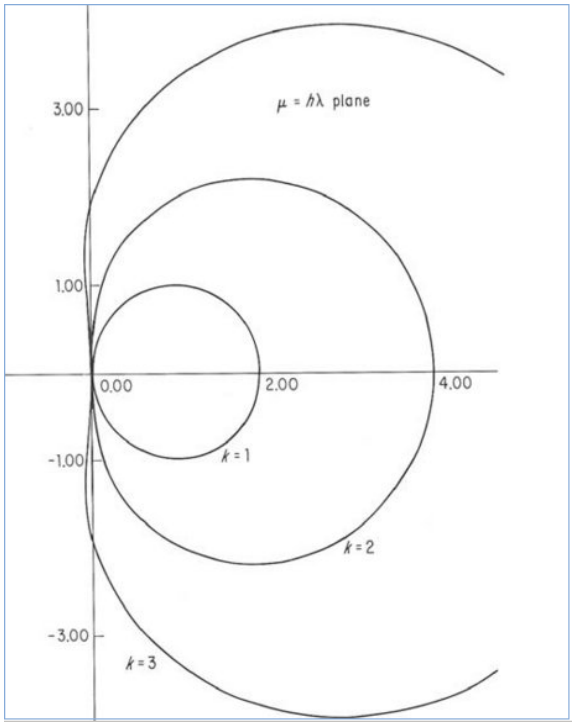
Usando quanto detto per le trasformazioni di Cayley abbiamo che nel metodo dei trapezi

$$z \in \mathbb{C}^- \Rightarrow |\omega| \leq 1$$

Quindi il metodo è A-stabile. Ci sono delle cose che però potrebbero non andare bene, infatti se z è molto grade (ovvero λ è molto grande) allora $|\omega|$ si avvicina al bordo della regione di stabilità assoluta, quindi anche se il metodo è A-stabile non abbiamo troppe garanzie (si pensi ai problemi Stiff).

Metodi BDF

I metodi BDF (ad esempio il risolutore 15s di matlab) sono i metodi preferibili per i problemi Stiff. Per $k = 1$ i metodi BDF coincidono con il metodo di Eulero, per $k = 2$ il metodo è A-stabile, per $k = 3$ non può esserlo per il teorema visto prima.
Quindi per $k = 1$ e $k = 2$ i metodi BDF sono A-stabili, per $k \geq 3$ sono $A(\alpha)$-stabili.



1.7.7 Metodi predictor-corrector

Questa sezione è una parte della lezione 25/05/2011

LMM impliciti

Consideriamo un LMM implicito

$$\sum_{j=0}^{k} \alpha_j x_{i-j+1} = h \sum_{j=0}^{k} \beta_j f(t_{i-j+1}, x_{i-j+1})$$

ovvero chiediamo $\beta_0 \neq 0$, $\alpha_0 = 1$, quindi possiamo riscrivere il tutto come

$$x_{i+1} = h\beta_0 f(t_{i+1}, x_{i+1}) + \left(h \sum_{j=1}^{k} \beta_j f(t_{i-j+1}, x_{i-j+1}) - \sum_{j=1}^{k} \alpha_j x_{i-j+1} \right)$$

chiamando

$$\psi = h \sum_{j=1}^{k} \beta_j f(t_{i-j+1}, x_{i-j+1}) - \sum_{j=1}^{k} \alpha_j x_{i-j+1}$$

possiamo riscriverla come

$$x_{i+1} = h\beta_0 f(t_{i+1}, x_{i+1}) + \psi \qquad \text{per} \qquad k \geq 0$$

dove ψ non dipende da x_{i+1}, quindi per trovare x_{i+1} possiamo usare l'iterazione del punto fisso, quindi generare una successione

$$y^{(k+1)} = h\beta_0 f(t_{i+1}, y^{(k)}) + \psi$$

ci chiediamo se $y^{(k+1)}$ converge ad x_{i+1}, per la teoria nota ci basta dimostrare che tale successione converge, quindi se y è il limite di tale successione allora

$$y^{(k+1)} - y = h\beta_0 \left(f(t_{i+1}, y^{(k)}) - f(t_{i+1}, y) \right)$$

passando alle norme e usando la lipschitzianità

$$\|y^{(k+1)} - y\| \leq h|\beta_0|L\|y^{(k+1)} - y\|$$

quindi la successione converge (a x_{k+1}) se $|\beta_0|L < 1$.

Un'altra idea per usare un LMM implicito può esser quella di usare Newton-Raphson, un approccio completamente diverso è invece quello dei metodi predictor-corrector, questi consistono nell'affiancare all'LMM implicito (corrector) un LMM esplicito (predictor), spiegheremo in dettaglio nei prossimi paragrafi. Indicheremo l'LMM esplicito (predict) come

$$\sum_{j=0}^{k} \widehat{\alpha}_j x_{i-j+1} = h \sum_{j=1}^{k} \widehat{\beta}_j f(t_{i-j+1}, x_{i-j+1})$$

PEC

La strategia PEC (predict, evalutate, correct) consiste in

- P calcola
$$x_{i+1}(P) = -\sum_{j=1}^{k} \widehat{\alpha}_j x_{i-j+1}(P) - h \sum_{j=0}^{k} \widehat{\beta}_j f(t_{i-j+1}, x_{i-j+1}(P))$$
- E calcola
$$f(t_{i+1}, x_{i+1}(P))$$
- C calcola
$$x_{i+1}(C) = -\sum_{j=1}^{k} \alpha_j x_{i-j+1}(C) - h \sum_{j=0}^{k} \beta_j f(t_{i-j+1}, x_{i-j+1}(P))$$

Quindi, detto a parole, la difficoltà di un metodo LMM implicito è il calcolo di $f(t_{i+1}, x_{i+1})$, nei metodi predictor-corrector (ad esempio con la strategia PEC) ci facciamo aiutare da un metodo LMM esplicito ausiliario per questo calcolo.

PECE

La strategia PECE (predict evalute correct evalute) consiste in

- P calcola
$$x_{i+1}(P) = -\sum_{j=1}^{k} \widehat{\alpha}_j x_{i-j+1}(C) - h \sum_{j=0}^{k} \widehat{\beta}_j f(t_{i-j+1}, x_{i-j+1}(C))$$
- E calcola
$$f(t_{i+1}, x_{i+1}(P))$$
- C calcola
$$x_{i+1}(C) = -\sum_{j=1}^{k} \alpha_j x_{i-j+1}(C) - h \sum_{j=0}^{k} \beta_j f(t_{i-j+1}, x_{i-j+1}(P))$$
- E calcola
$$f(t_{i+1}, x_{i+i}(C))$$

Esempio 1.8. Consideriamo il metodo dei trapezi come corrector

$$x_{i+1} = \frac{h}{2}(f(t_i, x_i) + f(t_{i+1}, x_{i+1}))$$

e il metodo di Eulero come predictor

$$x_{i+1} = x_i + hf(t_i, x_i)$$

usando il PECE ebbiamo

$$x_{i+1} = x_i + \frac{h}{2}\left(f(t_i, x_i) + f(t_{i+1}, x_i + hf(t_i, x_i))\right)$$

e questo si riscopre essere il metodo di Heun (che è di tipo Runge-Kutta a 2 stadi con ordine di convergenza 2).

Osservazione 1.14. I metodi predictor-corrector trasformano i metodi impliciti in metodi espliciti, quindi si perde la proprietà sulle regioni di stabilità, quindi sono sconsigliati per i problemi di tipo Stiff.

Consistenza dei metodi PECE

Supponiamo che il metodo predictor sia consistente di ordine q, ovvero

$$\delta^{(\mathrm{P})}(x(t+h), h) = O(h^q)$$

allora

$$\delta^{(\mathrm{PECE})}(x(t+h), h) = \delta^{(\mathrm{C})}(x(t+h), h) - \beta_0 \frac{\partial f}{\partial y}|_{(t+h, x(t+h))} \delta^{(\mathrm{P})}(x(t+h), h) + O(h^{q+2})$$

quindi abbiamo che se

$$\delta^{(\mathrm{C})}(x(t+h), h) = O(h^p) \qquad \text{con} \qquad p \geq q+1$$

allora si conclude che

$$\delta^{(\mathrm{PECE})}(x(t+h), h) = O(h^{q+1})$$

Osserviamo che questo risultato è coerente con l'esempio precedente infatti il metodo di Eulero ha consistenza 1, mentre il metodo dei trapezi 2, quindi abbiamo ottenuto un metodo con ordine di consistenza 2.

In conclusione il corrector deve avere ordine di consistenza superiore al predictor.

Capitolo 2

Problemi al contorno (BVP)

Si consideri il problema

$$\begin{cases} x'(t) = f(t, x(t)) \quad t \in [a, b] \\ g(x(a), x(b)) = 0 \end{cases}$$

Dove $g : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$, questi sono noti come BVP (boundary value problem).

Esempio 2.1. Il problema del determinare le configurazioni di una corda con il vincolo che siano fissati i due estremi si riesce ad esprimere nella forma

$$\begin{cases} x''(t) = h(t, x(t), x'(t)) \\ x(a) = x_a \\ x(b) = x_b \end{cases}$$

Per ricondursi ad un BVP basta porre $y(t) = (x(t), x'(t))$ e riscrivere l'equazione differenziale.

Inizieremo con lo studiare il caso particolare

$$\begin{cases} x'(t) = Ax(t) + r(t) \\ B_a x(a) + B_b x(b) = \beta \end{cases}$$

dove $A \in \mathbb{R}^{n \times n}$, $r : I \to \mathbb{R}^n$, $B_a, B_b \in \mathbb{R}^{n \times n}$ e $\beta \in \mathbb{R}^n$

11/05/2011

Esempio 2.2 (Esempio di BVP)**.**

$$\begin{cases} x'(t) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} x(t) \qquad t \in [0, 1] \\ \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} x(0) + \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} x(1) = \begin{pmatrix} e \\ -1 \end{pmatrix} \end{cases}$$

Al solito avremo $x(t) = (x_1(t), x_2(t))$ quindi il problema diventa della forma descritta nella lezione precedente

$$\begin{cases} x'(t) = Ax(t) \\ B_0 x(0) + B_1 x(1) = \beta \end{cases}$$

dove chiaramente B_0 e B_1 sono matrici mentre β è un vettore.
Sappiamo già che la soluzione generale sarà della forma $x(t) = e^{At} v$ dove $v = x(0)$. Quindi il problema si conduce a determinare $x(0) = (x_1(0), x_2(0))$ mentre in questo caso conosciamo $x_1(1) = e$ e $x_2(0) = -1$.
Daltronde la matrice A è diagonalizzabile

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

Quindi

$$e^{At} = e^{t\begin{pmatrix}2 & 1\\ 1 & 2\end{pmatrix}} = \frac{1}{2}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix} e^{t\begin{pmatrix}3 & 0\\ 0 & 1\end{pmatrix}} \begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}$$

quindi svolgendo il conto

$$e^{At} = e^{t\begin{pmatrix}2 & 1\\ 1 & 2\end{pmatrix}} = \frac{1}{2}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}\begin{pmatrix}e^{3t} & 0\\ 0 & e^{t}\end{pmatrix}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}$$

Sapendo che $e^0 = I$ (intesa come matrice), imponendo le condizioni a contorno abbiamo

$$B_0 I v + B_1 e^A v = \beta$$

$$(B_0 + B_1 e^A)v = \beta$$

Quindi posso risolvere il problema

$$v = (B_0 + B_1 e^A)\beta$$

Nel nostro esempio abbiamo che la soluzione è

$$x(t) = e^{At}\begin{pmatrix}1\\ -1\end{pmatrix}$$

Caso generale

Consideriamo in generale i problemi della forma

$$\begin{cases} x'(t) = Ax(t) + r(t) \\ B_a x(a) + B_b x(b) = \beta \end{cases}$$

Con $t \in [a,b]$, $A, B_a, B_b \in \mathbb{R}^{n\times n}$, $\beta \in \mathbb{R}^n$ ed $r : [a,b] \to \mathbb{R}^n$
Per trovare la soluzione di questo problema iniziamo con il risolvere il problema omogeneo

$$x'(t) = Ax(t)$$

Quindi cerchiamo una $F : \mathbb{R} \to \mathbb{R}^{n\times n}$ tale che $F'(t) = AF(t)$ e che verifichi $det(F(0)) \neq 0$. Sia $p(t)$ una soluzione particolare del sistema non omogeneo, ovvero

$$p'(t) = Ap(t) + r(t)$$

Allora la soluzione generale del problema scritto all'inizio è

$$x(t) = F(t)c + p(t) \qquad c \in \mathbb{R}^n$$

Dove $c \in \mathbb{R}^n$ sarà determinato in modo che valgano le condizioni al contorno. Quindi abbiamo, imponendo le condizioni al contorno

$$B_a(F(a)c + p(a)) + B_b(F(b)c + p(b)) = [B_a F(a) + B_b F(b)]c = \beta - B_a p(a) - B_b p(b)$$

Chiamando $Q = B_a F(a) + B_b F(b)$ abbiamo che la soluzione esiste unica se e soltanto se $det(Q) \neq 0$ (riusciamo a determinare c).

Esempio 2.3. Consideriamo il seguente problema

$$\begin{cases} x''(t) = -x(t) \\ x(0) = 0 \\ x(b) = \beta \end{cases}$$

Questo problema si riconduce a

$$\begin{cases} y'(t) = Ay(t) \\ B_0 y(0) + B_b y(b) = \beta \end{cases}$$

Infatti basta porre

$$y(t) = \begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} = \begin{pmatrix} x(t) \\ x'(t) \end{pmatrix}$$

Quindi abbiamo

$$y'(t) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} y(t)$$

Mentre le condizioni al contorno ci danno

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y_1(0) \\ y_2(0) \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} y_1(b) \\ y_2(b) \end{pmatrix} = \begin{pmatrix} 0 \\ \beta \end{pmatrix}$$

Si trova che

$$F(t) = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}$$

infatti è facile verificare che

$$F'(t) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} F(t)$$

Quindi abbiamo

$$Q = B_0 F(0) + B_b F(b) = \begin{pmatrix} 1 & 0 \\ \cos b & \sin b \end{pmatrix}$$

Quindi la soluzione esiste unica se $det(Q) \neq 0$ ovvero $b \neq k\pi$ con $k \in \mathbb{Z}$.

2.1 Condizionamento

Vogliamo ora confrontare i due seguenti problemi (il problema classico e quello perturbato) per studiare il condizionamento dei BVP al caso lineare

$$\begin{cases} x'(t) = Ax(t) \\ B_a x(a) + B_b x(b) = 0 \end{cases} \qquad \begin{cases} y'(t) = Ay(t) + r(t) \\ B_a y(a) + B_b y(b) = \beta \end{cases}$$

Se $x(t)$ è la soluzione del primo problema e $y(t)$ del secondo, allora $z(t) = y(t) - x(t)$ risolve il problema

$$\begin{cases} z'(t) = Az(t) + r(t) \\ B_a z(a) + B_b z(b) = \beta \end{cases}$$

Ci interessa sapere quanto cresce $\|z(t)\|$ in funzione delle perturbazioni $r(t)$ e β. Suppongo che $Q = B_a F(a) + B_b F(b) = I$, se così non fosse mi basta scegliere $\hat{F} = F(t)Q^{-1}$ come soluzione del sistema omogeneo. (NB: se F è soluzione lo è anche $\hat{F}$). Sia $p(t)$ la soluzione particolare, ovvero

$$\begin{cases} p'(t) = Ap(t) + r(t) \\ B_a p(a) + B_b p(b) = 0 \end{cases}$$

$p(t)$ esiste dato che $det(Q) \neq 0$. La soluzione generale sarà $z(t) = F(t)c + p(t)$, la soluzione del BVP si ottiene determinando c, in questo caso (usando $Q = I$ e $B_a p(a) + B_b p(b) = 0$)

$$c = Q^{-1}(\beta - B_a p(a) - B_b p(b)) = \beta$$

Prima di concludere è necessario introdurre la matrice di Green (serve per definire le soluzioni particolari).

Matrice di Green

In generale vale che la soluzione particolare del problema visto prima è

$$p(t) = \int_a^b G(t,s)r(s)ds$$

Dove G è detta matrice di Green ed è così definita

$$G(t,s) = \begin{cases} F(t)B_aF(a)F^{-1}(s) & \text{se} \quad t \geq s \\ -F(t)B_bF(b)F^{-1}(s) & \text{se} \quad t \leq s \end{cases}$$

Osserviamo subito che G dipende sia da F che dalle condizioni a contorno e che è ben definita per $t = s$.

Tornando al nostro problema abbiamo

$$z(t) = F(t)\beta + p(t) = F(t)\beta + \int_a^b G(t,s)r(s)ds$$

Possiamo maggiorare

$$\|z(t)\| \leq K_1\beta + K_2 \int_a^b \|r(s)\|ds$$

Dove

$$K_1 = \max_{[a,b]} \|F(t)\| \qquad K_2 = \max_{[a,b]} \|G(t,s)\|$$

Definiamo numero di condizionamento del BVP come $K = max\{K_1, K_2\}$

Cenni sul caso non lineare

Dato il BVP non lineare

$$\begin{cases} x'(t) = f(t, x(t)) \\ g(x(a), x(b)) = 0 \end{cases}$$

Possiamo ricondurci (in un certo senso) al caso lineare ponendo

$$A = \frac{\partial f}{\partial y} \quad , \quad B_a = \frac{\partial g}{\partial u} \quad , \quad B_b = \frac{\partial g}{\partial v}$$

Dove $g(u, v)$ è la funzione che da le condizioni al contorno.

2.2 Metodo di shooting

Consideriamo il BVP

$$\begin{cases} x'(t) = f(t, x(t)) \\ g(x(a), x(b)) = 0 \end{cases}$$

Supponiamo di conoscere la soluzione generale dell'equazione differenziale, daltronde non conosco il valore della soluzione nel punto iniziale (ovvero $x(a)$) daltronde sappiamo che $g(x(a), x(b) = 0$ e (sempre supponendo di avere la soluzione generale) conosciamo $x'(a)$. L'idea è quella di approssimare a partire da queste informazioni il valore di $x(a)$.

Perchè metodo di shooting: Prima di spiegare come funziona in pratica il metodo di shooting diamo un'idea informale (che da anche significato alla parola shooting). Quello che faccio è approssimare il valore di $x(a)$, per vedere se l'approssimazione che ho scelto è sensata, dato che conosco la soluzione generale, posso calcolare $g(x(a), x(b))$, se questo numero non è zero allora l'ho sparata quindi partendo da questo errore devo correggere il tiro e cercare una nuova approssimazione di $x(a)$.

Torniamo ora a dare una descrizione pratica del metodo di shooting. Per comodità poniamo $a = 0$, consideriamo i seguenti IVP (initial value problem) al variare di $s \in \mathbb{R}^n$

$$\begin{cases} x_s'(t) = f(t, x_s(t)) \\ x_s(0) = s \end{cases}$$

Diremo che questi sono gli IVP associati al BVP.
Per risolvere il BPV dobbiamo trovare un $s^* \in \mathbb{R}^n$ tale che $g(x_{s^*}(0), x_{s^*}(b)) = 0$. Per trovare s^* useremo un metodo iterativo.

Conclusione: abbiamo ricondotto il BVP ad una serie di IVP, daltronde possiamo subito osservare che abbiamo snaturato il problema, infatti se l'IVP fosse malcondizionato allora una piccola variazione di s^* cambierebbe totalmente la soluzione di IVP e quindi anche quella di BVP. Quindi il malcondizionamento del IVP si propagherebbe in BVP che a priori potrebbe non essere malcondizionato. Vedremo che è possibile rimediare a ciò.

Deflect

Vogliamo ora risolvere il problema della determinazione di s^*. Consideriamo la funzione (di deflect)

$$d(s) := g(s, x_s(b))$$

questa funzione dipenderà chiaramente da s e da $x_s(b)$, è ora chiaro che s^* verifica $d(s^*) = 0$, quindi il problema di riduce a trovare gli zeri di una funzione, possiamo dunque usare il metodo di Newton-Raphson, quindi s^* sarà il limite della successione

$$s_{k+1} = s_k - [J(s_k)]^{-1} d(s_k)$$

dove poniamo

$$J(s_k) = \frac{\partial d}{\partial s}(s_k, x_{s_k}(b))$$

Osserviamo che $d(s_k)$ è calcolabile risolvendo un IVP, nello specifico

$$J = \frac{\partial d}{\partial s} = \frac{\partial g}{\partial u} + \frac{\partial g}{\partial v}\frac{\partial x_s}{\partial s}(b)$$

L'unico pezzo che non siamo in grado di calcolare è $\frac{\partial x_s}{\partial s}$, daltronde se prendiamo IVP (che abbiamo associato al BVP) e deriviamo otteniamo (sotto tutte le ipotesi di regolarità)

$$\frac{\partial}{\partial s} x_s'(t) = \frac{d}{dt}\frac{\partial x_s(t)}{\partial s} = \left[\frac{\partial f}{\partial y}(t, x_s(t))\right]\frac{\partial x_s}{\partial s}(t)$$

poniamo

$$A(t) = \frac{\partial f}{\partial y}(t, x_s(t)) \quad \text{e} \quad y_s(t) = \frac{\partial x_s}{\partial s}(t)$$

Quindi abbiamo un nuovo IVP (a volte lo chiameremo IVP2) associato al BVP

$$y_s'(t) = A(t) y_s(t)$$

Derivando invece la seconda equazione dal IVP associato al BVP (il primo che abbiamo considerato troviamo) anche la condizione iniziale $y_s(0) = I$ (identità come matrice), quindi risolto IVP2 siamo in grado di determinare J e fare le varie iterazioni per trovare uno zero della funzione di deflect.

Ricapitolando brevemente, dato un BVP

$$\begin{cases} x'(t) = f(t, x(t)) \\ g(x(a), x(b)) = 0 \end{cases}$$

per risolverlo con il metodo di shooting dobbiamo associare un primo IVP

$$\begin{cases} x_s'(t) = f(t, x_s(t)) \\ x_s(0) = s \end{cases}$$

e trovare gli zeri della funzione di deflect $d(s) = g(s, x_s(b))$, per far ciò ad ogni iterata di Newton-Raphson dobbiamo risolvere un secondo IVP

$$\begin{cases} y_s'(t) = A(t) y_s(t) \\ y_s(0) = I \end{cases}$$

Commenti

- Dato che i due IVP contengono dei pezzi l'uno dell'altro nella pratica se cerca di usare algoritmi che li risolveno contemporaneamente
- A priori J^{-1} potrebbe esser una matrice malcondizionata
- Il costo computazionale è proibitivo

17/05/2011

Riepilogo

Si è mostrato nella lezione precedente come il risolvere un BVP

$$\begin{cases} x'(t) = f(t, x(t)) \\ g(x(a), x(b)) = 0 \end{cases}$$

si riconduca a risolvere un IVP

$$\begin{cases} x'(t) = f(t, x(t)) \\ x(a) = s \end{cases}$$

dove abbiamo mostrato che il problema è trovare s che soddisfa $g(s, x_s(b)) = 0$.
Ad esempio possiamo trovarlo usando il metodo di Newton-Raphson

$$s^{(k+1)} = s^{(k)} - J(s^{(k)})^{-1} d(s^{(k)})$$

dove d è la funzione di deflect definita come

$$d(s) = g(s, x_s(b))$$

Quindi in sostanza cerchiamo s^* tale che $d(s^*) = 0$.
Il metodo di shooting prevede la risoluzione di due IPV

$$\begin{cases} y'(t) = A(t, x_s(t))y(t) \\ y(a) = I \end{cases}$$

Dove $y : \mathbb{R} \to \mathbb{R}^n$, mentre A è lo jacobiano di g rispetto la seconda variabile.
L'altro IVP da risolvere è

$$\begin{cases} x'(t) = f(t, x(t)) \\ x(a) = s \end{cases}$$

Mostriamo ora il problemi a cui si faceva riferimento nella fine della lezione precedente.

Condizionamento dell'IVP e conseguenze

Supponiamo di voler risolvere un certo BVP della solita forma, abbiamo visto che con il metodo di shooting dobbiamo risolvere ad ogni passo due IVP, daltronde questi potrebbero essere mal condizionati, in tal caso (con il metodo di shooting) il malcondizionamento si scarica sul BVP.
Entrando nel concreto, riprendendo la notazione usata nel paragrafo precedente, una volta avviato il metodo di shooting ad un certo punto avremo $\overline{s}$ che è un'approssimazione di s^*, quindi abbiamo i due IVP

$$\begin{cases} x'(t) = f(t, x(t)) \\ x(a) = \overline{s} \end{cases} \qquad \begin{cases} x'(t) = f(t, x(t)) \\ x(a) = s^* \end{cases}$$

Se $x_{\overline{s}}(t)$ è la soluzione del primo IVP e $x_{s^*}(t)$ è la soluzione del secondo, allora se IVP è mal condizionato $|x_{\overline{s}}(t) - x_{s^*}(t)|$ è un numero grande (lo si può vedere ad esempio usando il Lemma di Gromwall, ma sostanzialmente è stato già dimostrato). Al contrario se abbiamo che l'IVP è stabile allora questo errore è controllato.

Osserviamo che la stabilità di un BVP non implica quella dell'IVP associato, infatti avviamo definito la stabilità di un BVP tramite il numero di condizionamento, pertanto ci sono esempi di BVP stabili i cui IVP associati non lo sono (vedremo poi come superare questo problema).

Esempio 2.4 (fallimento del metodo di Shooting su problemi stabili)**.** Consideriamo il solito problema test

$$\begin{cases} x'(t) = \lambda x(t) & t \in [a,b] \\ x(0) = 1 \end{cases}$$

Questo IVP sappiamo esser stabile se $\lambda \leq 0$ e asintoticamente stabile se $\lambda < 0$ (supponiamo $\lambda \in \mathbb{R}$). Facciamo un cambiamento di variabile finalizzato a trasformare l'IVP in un BVP, ovvero: $t \to b - t$ quindi poniamo $y(t) := x(b-t)$, allora abbiamo

$$y'(t) = -x'(b-t) = -\lambda y(t)$$

Quindi ho un problema ai valori finali (ovvero un BVP)

$$\begin{cases} y'(t) = -\lambda y(t) & t \in [a,b] \\ y(b) = 1 \end{cases}$$

Dato che un cambiamento lineare di coordinate non cambia la stabilità del problema allora il problema è stabile per $\lambda \leq 0$. Consideriamo ora il problema al contorno

$$\begin{cases} \begin{pmatrix} x_1'(t) \\ x_2'(t) \end{pmatrix} = \begin{pmatrix} \lambda & 0 \\ 0 & -\lambda \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} & t \in [a,b] \\ x_1(0) = 1 \\ x_2(1) = 1 \end{cases}$$

In conclusione, per quanto detto sin ora abbiamo che questo problema è stabile per $\lambda \leq 0$ e asintoticamente stabile per $\lambda < 0$. Proviamo ora a risolvere questo problema con il metodo di shooting

$$\begin{cases} x_1'(t) = \lambda x_1(t) \\ x_2'(t) = -\lambda x_2(t) \\ x_1(0) = 1 \\ x_2(0) = s \end{cases}$$

Dove s è da approssimare, quello scritto è l'IVP associato all'ultimo BVP scritto, cosa possiamo dire della stabilità di questo IVP?
Su x_1 abbiamo stabilità se $\lambda \leq 0$ mentre su x_2 abbiamo stabilità se $\lambda > 0$, quindi se $\lambda < 0$ oppure $\lambda > 0$ uno dei due IVP non è stabile.

Esempio 2.5 (BVP stabile con IVP associato instabile)**.** Questo esempio è preso dal testo di U. Ascher, L. Petzold ed è l'esempio 6.4.
Consideriamo il problema

$$\begin{cases} x''(t) = x(t) & t \in [0,T] \\ x(0) = b_1 \\ x(T) = b_2 \end{cases}$$

Per risolverlo definiamo

$$y(t) = \begin{pmatrix} x(t) \\ x'(t) \end{pmatrix} \qquad y'(t) = \begin{pmatrix} y_2(t) \\ y_1(t) \end{pmatrix}$$

Quindi l'IVP diventa

$$\begin{cases} y'(t) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} y_2(t) \\ y_1(t) \end{pmatrix} \\ \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y_1(0) \\ y_2(0) \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} y_1(T) \\ y_2(T) \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \end{cases}$$

Cerco dunque una matrice $\underline{\mathrm{Y}}(t)$ tale che

$$\underline{\mathrm{Y}}\text{'}(t) = A\underline{\mathrm{Y}}(t) \qquad \text{con} \qquad \det \underline{\mathrm{Y}}(0) \neq 0$$

dove

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

(si sta seguendo il solito schema per trovare il sistema fondamentale di soluzione con le matrici di Green). In questo caso si trova subito che

$$\underline{\mathrm{Y}}(t) = \begin{pmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{pmatrix}$$

dove vale $\underline{\mathrm{Y}}(0) = I$, quindi se volessimo andare avanti (cercando le matrici di Green e determinando il valore di c) avremo che la soluzione è della forma $y(t) = \underline{\mathrm{Y}}(t)c + p(t)$, daltronde $\|\underline{\mathrm{Y}}(t)\|$ cresce esponenzialmente quindi IVP non è stabile, mostriamo che però il BVP è stabile.
Definiamo

$$Q = B_0\underline{\mathrm{Y}}(0) + B_T\underline{\mathrm{Y}}(T) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\underline{\mathrm{Y}}(0) + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}\underline{\mathrm{Y}}(T) = \begin{pmatrix} 1 & 0 \\ \cosh T & \sinh T \end{pmatrix}$$

Quindi devo trovare $\widehat{\underline{\mathrm{Y}}}(t) = \underline{\mathrm{Y}}(t)Q^{-1}$ tale che

$$\begin{cases} \widehat{\underline{\mathrm{Y}}'}(t) = A\widehat{\underline{\mathrm{Y}}}(t) \\ B_0\widehat{\underline{\mathrm{Y}}}(0) + B_T\widehat{\underline{\mathrm{Y}}}(T) = I \end{cases}$$

Si trova che

$$\widehat{\underline{\mathrm{Y}}}(t) = \frac{1}{\sinh T}\begin{pmatrix} \sinh(T-t) & \cosh t \\ -\cosh(T-t) & \sinh t \end{pmatrix}$$

Si trova quindi il numero di condizionamento, ovvero

$$K_1 = \max_{t\in[0,1]} \|\widehat{\underline{\mathrm{Y}}}(t)\| \simeq 2 \qquad \text{costante che dipende dalla norma}$$

Quindi K_1 è una costante indipendente da T, stesso discorso su K_2. In conclusione abbiamo trovato un BVP stabile il cui IVP non è stabile, vedremo che questi problemi non ci sono più se utilizziamo il metodo dello shooting multiplo.

18/05/2011

2.3 Metodo di shooting multiplo

Motivazioni per lo shooting multiplo

Si è visto la volta scorsa che il metodo di shooting cerca di trasformare un BVP in un IVP cercando di stimare il valore iniziale, daltronde abbiamo visto che questa trasformazione potrebbe esser malcondizionata, ricordiamo infatti che se

$$\begin{cases} x'(t) = f(t, x(t)) \\ x(t_0) = x_0 \end{cases} \qquad \begin{cases} y'(t) = f(t, y(t)) \\ y(t_0) = x_0 + z_0 \end{cases}$$

Allora abbiamo che

$$\|y(t) - x(t)\| \leq \|z_0\| e^{L(T-t_0)}$$

Quindi ci aspettiamo che se l'itervallo è grande il metodo di shooting fallisca, quindi l'idea è di dividere l'intervallo in sottointervalli e usare il metodo di shooting su ognuno di questi, questo è noto come metodo di shooting multiplo.

Metodo di shooting multiplo

Consideriamo il solito problema

$$\begin{cases} x'(t) = f(t, x(t)) & t \in [a, b] \\ g(x(0), x(b)) = 0 \end{cases}$$

suddividiamo l'intervallo con $t_{i+1} = t_i + h$

$$t_0 = 0 < t_1 < \cdots < t_N = b$$

Considero dunque questi N IVP

$$\begin{cases} y'(t) = f(t, y(t)) & t \in [t_{n-1}, t_n] \qquad 1 \leq n \leq N \\ y(t_{n-1}) = s_{n-1} \end{cases}$$

Spesso chiameremo y_n la soluzione dell'ennesimo IVP ma quando la notazione diventerà troppo pensante lo ometteremo.
Approssimerò la soluzione con $x(t) = y_n(t, s_{n-1})$ per $t \in [t_{n-1}, t_n]$.
C'è da determinare la scelta dei punti iniziali, come prima cosa richiediamo che la soluzione sia continua, quindi le y_n devono raccordarsi in modo continuo, quindi imponiamo

$$y_n(t_n) = s_n \qquad \text{per} \qquad 1 \leq n \leq N - 1$$

L'altra condizione è che valgano le condizioni al contorno

$$g(s_0, y_N(b)) = 0$$

Quindi in definitiva dobbiamo risolvere un sistema non lineare

$$\begin{cases} y_n(t_n) = s_n \\ g(s_0, y_N(b)) = 0 \end{cases}$$

Notazione

Come anticipato prima a volte la notazione si farà pesante, quindi definiamo la seguente notazione

$$y(t, s_{n-1}) := y_n(t)$$

Ovvero $y(t, s_{n-1})$ è la soluzione dell'n-esimo IVP , ovvero parte da s_{n-1}, quindi vale

$$y(t_{n-1}, s_{n-1}) = s_{n-1} \qquad y(t_n, s_{n-1}) = s_n$$

Pertando il sistema non lineare che abbiamo introdotto nel paragrafo precedente possiamo scriverlo come

$$\begin{cases} y(t_n, s_{n-1}) = s_n \\ g(s_0, y(b, s_N)) = 0 \end{cases}$$

Risoluzione del sistema non lineare associato al metodo di shooting

Consideriamo il sistema non lineare

$$\begin{cases} y(t_n, s_{n-1}) = s_n \\ g(s_0, y(b, s_N)) = 0 \end{cases}$$

cerchiamo dunque una soluzione, ovvero un

$$\underline{s} = (s_0, s_1, \ldots, s_{N-1}) \in \mathbb{R}^{N \cdot m}$$

che soddisfa il sistema (ci sono dei casi in cui il sistema ha più soluzioni ma non possiamo controllare questo aspetto).
Riscriviamo il sistema come $h(\underline{s}) = 0$ e risolviamola usando il metodo di Newton-Raphson

$$\begin{cases} \underline{s}^{(k+1)} = \underline{s}^{(k)} - J(s)^{-1} h(\underline{s}^{(k)}) \\ \underline{s}^{(0)} \qquad \text{approssimazione iniziale} \end{cases}$$

Quindi abbiamo

$$J(s^{(k)}) = \frac{\partial h}{\partial s}(\underline{s}^{(k)})$$

Osserviamo che esplicitamente abbiamo

$$h(\underline{s}) = \begin{pmatrix} s_1 - y(t_1, s_0) \\ s_2 - y(t_2, s_1) \\ \vdots \\ s_{N-1} - y(t_{N-1}, s_{N-1}) \end{pmatrix}$$

Quindi lo jabobiano è

$$\frac{\partial h}{\partial s}(\underline{s}^{(k)}) = \begin{pmatrix} -y_{s_0}(t_1) & I & & & & \\ & -y_{s_1}(t_2) & I & & & \\ & & \ddots & & & \\ & & & -y_{s_{N-2}}(t_{N-1}) & I & \\ B_0 & & & & B_b & -y_{s_{N-1}}(t_N) \end{pmatrix}$$

Dove abbiamo usato la notazione

$$y_{s_n}(t_{n+1}) := \frac{\partial y_n}{\partial s_n}(t_{n+1}) \qquad B_0 := \frac{\partial g}{\partial u} \qquad B_b := \frac{\partial g}{\partial v}$$

NB: u, v sono le due variabili di g, ovvero $g = g(u, v)$.
Per trovare i singoli blocchi dello jacobiano devo quindi derivare l'equazione differenziale scritta all'inizio (si ragiona esattamente come nello shooting semplice ma si complica la notazione), ovvero

$$\begin{cases} y_n'(t) = f(t, y(t)) \qquad t \in [t_{n-1}, t_n] \qquad 1 \le n \le N \\ y_n(t_{n-1}) = s_{n-1} \end{cases}$$

ovvero troviamo

$$\begin{cases} \dfrac{\partial y_n'(t)}{\partial s_{n-1}} = \dfrac{\partial f}{\partial y} \dfrac{\partial y_n(t)}{\partial s_{n-1}} \\ \dfrac{\partial y_n(t_{n-1})}{\partial s_{n-1}} = I \end{cases}$$

Per comodità introduciamo la notazione

$$\widehat{y_{n-1}}(t) := \frac{\partial y_n(t)}{\partial s_{n-1}} \qquad A_{n-1} := \frac{\partial f}{\partial y}$$

Quindi possiamo riscrivere il problema come

$$\begin{cases} \widehat{y_{n-1}}'(t) = A_{n-1}(t) \widehat{y_{n-1}}(t) \qquad t \in [t_{n-1}, t_n] \\ \widehat{y_{n-1}}(t_{n-1}) = I \end{cases}$$

Quindi ad ogni passo del metodo di shooting multiplo devo risolvere N IVP del tipo appena mostrato (che a volte chiameremo IVP2) per determinare lo jacobiano, N IVP (di quelli scritti all'inizio che a volte chiameremo IVP1) e poi devo invertire la matrice jacobiana (o equivalentemente risolvere un sistema lineare).

Caso lineare

Consideriamo ora i BVP lineari e vediamo come funziona lo shooting multiplo

$$\begin{cases} x'(t) = A(t)x(t) + q(t) \\ B_0x(0) + B_bx(b) = \beta \end{cases}$$

in questo caso gli IVP1 sono

$$\begin{cases} y_n'(t) = A(t)y_n(t) + g_n(t) \\ y_n(t_{n-1}) = s_{n-1} \end{cases}$$

mentre gli IVP2 sono

$$\begin{cases} \dfrac{\partial y_n'}{\partial s_{n-1}} = A(t) \cdot \dfrac{\partial y_{n-1}(t)}{\partial s_{n-1}} \\ \dfrac{\partial y_n}{\partial s_{n-1}}(t_{n-1}) = I \end{cases}$$

Daltronde in IVP2 dato che siamo nel caso lineare non c'è in realtà la differenza da s quindi IVP1 sono in realtà del tipo

$$\begin{cases} y'(t) = A(t)y(t) & t \in [t_{n-1}, t_n] \\ y(t_{n-1}) = I \end{cases}$$

Quindi lo jacobiano non dipende da s ma solo dai punti di suddivisione degli intervalli

$$J = \begin{pmatrix} -y(t_1) & I & & & \\ & \ddots & \ddots & & \\ & & -y(t_{n-1}) & I & \\ B_0 & & & B_b & y(b) \end{pmatrix}$$

Quindi dato che J è costante il metodo di Newton-Raphson converge in un passo, quindi $h(s) = J \cdot s + v$, quindi se ad esempio scelgo $\underline{s}^0$ allora il metodo converge con una iterazione, quindi abbiamo una soluzione esplicita

$$\underline{s} = -J^{-1} \cdot h(0)$$

In conclusione il sistema non lineare che serve risolvere per trovare una soluzione del BVP ha una soluzione esplicita nel caso in cui il BVP sia lineare, ma comunque bisogna calcolare J^{-1} che a priori potrebbe esser malcondizionata, quindi vogliamo in qualche modo determinare il numero di condizionamento di J, ovvero $\|J\|\|J^{-1}\|$.
Daltronde vale che

$$\|J\| \leq c(\max \|y_n(t_n)\| + 1)$$

dove la presenza di 1 dipende dalla matrice I, le norme di B_0 e B_b le scarico sulla costante c (mi interessa solo come crescono le norme).
Mi aspetto che $\|y_n(t_n)\|$ sia tanto più grande quanto più gli IVP siano instabili.
Nel caso di problemi instabili per ridurre la crescita della norma posso restringere gli intervalli di integrazione e quindi aumentare i punti di discretizzazione.
In conclusione se gli IVP non sono stabili allora bisogna scegliere h piccolo per non avere $\|y_n(t_n)\|$ troppo grande (e quindi un condizionamento dello jacobiano troppo alto).

Vediamo ora cosa succede per $\|J^{-1}\|$

$$J^{-1} = \begin{pmatrix} G(t_0, t_1) & \dots & G(t_0, t_{n-1}) & \phi(t_0) \\ \vdots & \ddots & \vdots & \vdots \\ G(t_{n-1}, t_1) & \dots & G(t_{n-1}, t_{n-1}) & \phi(t_{n-1}) \end{pmatrix}$$

Dove $G(t, s)$ è la matrice di Green definita nei paragrafi precedenti, mentre $\phi(t)$ soddisfa

$$\begin{cases} \phi'(t) = A(t)\phi(t) & t \in [0, b] \\ B_0\phi(0) + B_b\phi(b) = I \end{cases}$$

Quindi la norma di questa matrice dipende dal condizionamento del BVP definito come $K = \max(K_1, K_2)$ dove

$$K_1 = \max_{t\in[0,b]} \|\phi(t)\| \qquad K_2 = \max_{t\in[0,b]} \|G(\cdot,\cdot)\|$$

Allora vale che

$$\|J^{-1}\| \leq NK$$

2.4 Metodo delle differenze finite

Consideriamo un BVP

$$\begin{cases} x'(t) = f(t, x(t)) \\ g(x(0), x(b)) = 0 \end{cases}$$

Dove prendiamo una suddivisione

$$0 = t_0 < t_1 < \cdots < t_N = b \qquad \text{con} \qquad t_{i+1} = t_i + h$$

Supponiamo di avere un metodo ad un passo

$$x_{i+1} = x_i + h\phi(t_i, h, x_i, x_{i+1})$$

Questa è un'equazione alle differenze ma non abbiamo il punto iniziale, possiamo dunque vederla come un sistema non lineare dove $x_0, x_1, \ldots, x_N$ sono incognite, inoltre abbiamo la condizione $g(x_0, x_n) = 0$, quindi esplicitamente dobbiamo risolvere il sistema non lineare

$$\begin{cases} x_1 - x_0 - h\phi(t_0, h, x_0, x_1) = 0 \\ x_2 - x_1 - h\phi(t_1, h, x_1, x_2) = 0 \\ \cdots \\ x_{i+1} - x_i - h\phi(t_i, h, x_i, x_{i+1}) = 0 \\ \cdots \\ x_N - x_{N-1} - h\phi(t_{N-1}, h, x_{N-1}, x_N) = 0 \\ g(x_0, x_N) = 0 \end{cases}$$

che sinteticamente indichiamo con

$$\begin{cases} x_{i+1} - x_i - h\phi(t_i, h, x_i, x_{i+1}) = 0 & 1 \leq i \leq N \\ g(x_0, x_N) = 0 \end{cases}$$

Osserviamo che x_i non è noto, quindi i metodo espliciti non danno vantaggio, pertanto sono preferibili i metodi impliciti dato che sono più stabili. Un metodo alle differenze consiste appunto nel risolvere questo sistema non lineare.

Metodo alle differenze basato sul punto di mezzo (introduzione)

Consideriamo il seguente metodo ad un passo

$$x_{i+1} = x_i + hf\left(\frac{t_i + t_{i+1}}{2}, \frac{x_i + x_{i+1}}{2}\right)$$

questo metodo ha ordine di convergenza 2, il sistema non lineare per impostare il metodo alle differenze è il seguente

$$\begin{cases} x_{i+1} - x_i - hf\left(\dfrac{t_i + t_{i+1}}{2}, \dfrac{x_i + x_{i+1}}{2}\right) = 0 & 1 \leq i \leq N \\ g(x_0, x_N) = 0 \end{cases}$$

Per risolvere questo sistema non lineare usiamo il metodo di Newton-Raphson, indichiamo il sistema non lineare con $h(\underline{x}) = 0$. Quindi abbiamo la successione

$$\underline{x}^{(k+1)} = \underline{x}^{(k)} - J\left(\underline{x}^{(k)}\right)^{-1} h(\underline{x}^{(k)})$$

Dove come sempre J è lo jacobiano, si può mostrare che questa matrice è bidiagonale a blocchi tranne per l'ultima riga.

Consideriamo il caso lineare

$$\begin{cases} x'(t) = A(t)x(t) + q(t) \\ B_0 x(0) + B_b x(b) = \beta \end{cases}$$

Usando il metodo alle differenze basato sul punto di mezzo abbiamo

$$\begin{pmatrix} * & * & & & & \\ & * & * & & & \\ & & \ddots & \ddots & & \\ & & & * & * & \\ * & & & & * & * \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{N-1} \\ x_N \end{pmatrix} = \begin{pmatrix} * \\ * \\ \vdots \\ * \\ * \end{pmatrix}$$

Quindi in realtà ci sarà da risolvere un sistema lineare e successivemente di capire come stimare $\|x(t_i) - x_i\|$ ma entreremo in dettaglio più avanti.

24/05/2011

Errori nel metodo alle differenze basato sul punto di mezzo

Consideriamo il metodo del punto di mezzo

$$x_{i+1} = x_i + hf\left(\frac{t_i + t_{i+1}}{2}, \frac{x_i + x_{i+1}}{2}\right)$$

Si è già detto che questo è un un metodo ad un passo implicito con ordine di consistenza e convergenza 2. Applichiamo il metodo alle differenze associato al punto di mezzo al caso lineare

$$\begin{cases} x'(t) = A(t)x(t) + q(t) \\ B_0 x(0) + B_b x(b) = c \end{cases}$$

Dove $A(t) : \mathbb{R} \to \mathbb{R}^{n \times n}$ mentre $x(t) : \mathbb{R} \to \mathbb{R}^n$, consideriamo la suddivisione

$$0 = t_0 < t_1 < \cdots < t_N = b \qquad , \qquad t_{i+1} = t_i + h$$

Per semplificare la notazione definiamo

$$\widehat{t_i} = \frac{t_i + t_{i+1}}{2}$$

Quindi il sistema non lineare associato al metodo è

$$\begin{cases} x_{i+1} = x_i + h\left(A(\widehat{t_i})\dfrac{x_{i+1} + x_i}{2} + q(\widehat{t_i})\right) & 1 \le i \le N-1 \\ B_0 x_0 + B_b x_N = c \end{cases}$$

Lo scopo è calcolare $(x_0, x_1, \ldots, x_N)$ risolvendo il sistema, quest'ultimo possiamo riscriverlo come

$$h^{-1}\left(I - \frac{h}{2}A(\widehat{t_i})\right)x_{i+1} - h^{-1}\left(I + \frac{h}{2}A(\widehat{t_i})\right)x_i = q(\widehat{t_i}) \qquad 1 \le i \le N-1$$

Per semplificare la notazione poniamo

$$R_{i+1} = h^{-1}\left(I - \frac{h}{2}A(\widehat{t_i})\right) \qquad S_{i+1} = -h^{-1}\left(I + \frac{h}{2}A(\widehat{t_i})\right)$$

Quindi il sistema posso scriverlo in forma matriciale come

$$\begin{pmatrix} S_1 & R_1 & & & \\ & S_2 & R_2 & & \\ & & \ddots & \ddots & \\ & & & S_N & R_N \\ B_0 & & & & B_b \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{N-1} \\ x_N \end{pmatrix} = \begin{pmatrix} q(\widehat{t_0}) \\ q(\widehat{t_1}) \\ \vdots \\ q(\widehat{t_{N-1}}) \\ c \end{pmatrix}$$

Chiamiamo W la matrice scritta sopra. Osserviamo che questa matrice ha la stessa forma di quella del metodo di shooting multiplo applicato al medesimo problema, il costo della risoluzione di questo problema è $O(n^3N)$ dove n è la dimensione del problema, ovvero $x_i \in \mathbb{R}^n$ mentre N è il numero di discretizzazione (in quante parti abbiamo suddiviso l'intervallo $[a, b]$).
A questo punto ci chiediamo da cosa dipende N (in quante parti è sensato suddividere l'intervallo) e quindi quanto vale $\|x(t_i) - x_i\|$.

Sappiamo che il metodo è consistente di ordine 2 quindi

$$\delta(x(t+h), h) = O(h^2)$$

quindi abbiamo che (utilizzando la definizione di errore locale di discretizzazione)

$$x(t_{i+1}) = x(t_i) + h\left(A(\widehat{t}_i)\frac{x(t_{i+1}) + x(t_i)}{2} + q(\widehat{t}_i)\right) + h\delta(x(t_{i+1}), h)$$

La soluzione rispetterà dunque

$$W\begin{pmatrix} x(t_0) \\ x(t_1) \\ \vdots \\ x(t_{N-1}) \\ x(t_N) \end{pmatrix} = \begin{pmatrix} q(\widehat{t}_0) \\ q(\widehat{t}_1) \\ \vdots \\ q(\widehat{t}_{N-1}) \\ c \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_N \\ 0 \end{pmatrix}$$

Facendo la differenza con la soluzione effettivamente calcolata troviamo

$$W\begin{pmatrix} e_0 \\ e_1 \\ \vdots \\ e_{N-1} \\ e_N \end{pmatrix} = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_N \\ 0 \end{pmatrix} \qquad \text{dove} \qquad e_i = x(t_i) - x_i$$

Quindi abbiamo la relazione: $\underline{e} = W^{-1}\underline{\delta}$,dalla quale possiamo calcolare gli errori, in particolare se avessimo $\|W\| \leq K$ allora si avrebbe $\|\underline{e}\| \leq K\|\underline{\delta}\| = O(h^2)$, quindi la consistenza del metodo implicherebbe la convergenza, daltronde ci servono delle stime su W^{-1}. Definiamo la matrice diagonale a blocchi

$$D = \begin{pmatrix} R_1^{-1} & & & & \\ & R_2^{-1} & & & \\ & & \ddots & & \\ & & & R_N^{-1} & \\ & & & & I \end{pmatrix}$$

Osserviamo che le matrici R_i sono invertibili per h piccolo (per vederlo è sufficiente scrivere esplicitamente la matrice D), quindi $\|D\| \leq r$, mentre abbiamo

$$DW = \begin{pmatrix} R_1^{-1}S_1 & I & & & \\ & R_2^{-1}S_2 & I & & \\ & & \ddots & I & \\ & & & R_N^{-1}S_N & I \\ B_0 & & & & B_b \end{pmatrix}$$

Dove abbiamo che ogni singolo blocco possiamo svilupparlo

$$R_i^{-1}S_i = -\left(I - \frac{1}{2}hA(\widehat{t}_{i-1})\right)^{-1}\left(I + \frac{1}{2}hA(\widehat{t}_{i-1})\right) =$$

$$= -\left(I + \frac{1}{2}hA(\widehat{t}_{i-1}) + \frac{1}{2}A(\widehat{t}_{i-1}) + O(h^2)\right) = -(I + hA(\widehat{t}_{i-1})) + O(h^2)$$

Consideriamo ora i problemi

$$\begin{cases} Y_i'(t) = A(t)Y_i(t) & t \in [t_{i+1}] \\ Y_i(t_{i-1}) = I \end{cases}$$

Si può dimostrare che

$$I + hA(\widehat{t}_{i-1}) = Y(t_i) + O(h^2)$$

Quindi

$$DW = \begin{pmatrix} -Y(t_1) & I & & & & \\ & -Y(t_2) & I & & & \\ & & \ddots & I & & \\ & & & -Y(t_N) & I & \\ B_0 & & & & & B_b \end{pmatrix} + E$$

Dove E è una matrice diagonale a blocchi limitata in norma, $\|E\| \leq ch^2$

$$E = \begin{pmatrix} * & & & & & & \\ & * & & & & & \\ & & \ddots & & & & \\ & & & * & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{pmatrix}$$

Pertanto abbiamo (vedere paragrafi precedenti)

$$M^{-1} = \begin{pmatrix} G(t_0, t_1) & \dots & G(t_0, t_{n-1}) & \phi(t_0) \\ \vdots & \ddots & \vdots & \vdots \\ G(t_{n-1}, t_1) & \dots & G(t_{n-1}, t_{n-1}) & \phi(t_{n-1}) \end{pmatrix}$$

dove

$$\begin{cases} \phi'(t) = A(t)\phi(t) \\ B_0\phi(0) + B_b\phi(b) = 0 \end{cases}$$

In conclusione abbiamo che

$$W^{-1} = (M + E)^{-1}D = (M(I + M^{-1}E))D$$

dove $M^{-1}E = O(h)$ e quindi

$$\|W\| \leq cK \qquad \text{dove } K \text{ è il condizionamento del BVP}$$

Conclusione

Il problema è malcondizionato solo se lo il BVP e non interviene per nulla il condizionamento dell'IVP associato, questa è la differenza che si era annunciata tra risolvere un problema con il metodo delle differenze oppure con il metodo di shooting.

25/05/2011

Ricapitolazione della lezione precedente

Riprendiamo l'argomento degli errori nel metodo alle differenze basato sul punto di mezzo utilizzando le stesse notazioni. L'obbiettivo è dare una limitazione a $\|W\|$, abbiamo osservato che possiamo scrivere

$$W = (M + E)^{-1}D = [M(I + M^{-1}E)]^{-1}D = (I + M^{-1}E)^{-1}M^{-1}D$$

dove nella lezione precedente si era anche visto che

$$M = \begin{pmatrix} -Y(t_1) & I & & & \\ & -Y(t_2) & I & & \\ & & \ddots & I & \\ & & & -Y(t_N) & I \\ B_0 & & & & B_b \end{pmatrix}$$

e che è possibile scrivere la sua inversa mediante le matrici di Green

$$M^{-1} = \begin{pmatrix} G(t_0,t_1) & \dots & G(t_0,t_{n-1}) & \phi(t_0) \\ \vdots & \ddots & \vdots & \vdots \\ G(t_{n-1},t_1) & \dots & G(t_{n-1},t_{n-1}) & \phi(t_{n-1}) \end{pmatrix}$$

$$E = \begin{pmatrix} * & & & & & & \\ & * & & & & & \\ & & \ddots & & & & \\ & & & * & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{pmatrix} = O(h^2)$$

ricordiamo la definizione di numero di condizionamento per un BVP

$$K = \max\{K_1, K_2\} \qquad \text{dove} \qquad K_1 = \max_{t\in[0,b]} \|\phi(t)\| \quad K_2 = \max_{t\in[0,T]} \|G(t,s)\|$$

Si è vosto che $\|M^{-1}\| \leq cKN$ dove c è costante, di conseguenza, sapendo che $M^{-1}E = O(h)$ abbiamo

$$(I + M^{-1}E)^{-1}M^{-1}D \doteq M^{-1}D$$

Si era inoltre definito

$$D = \begin{pmatrix} R_1^{-1} & & & & \\ & R_2^{-1} & & & \\ & & \ddots & & \\ & & & R_N^{-1} & \\ & & & & I \end{pmatrix}$$

dunque abbiamo che

$$M^{-1}D = \begin{pmatrix} hG(t_0,t_1) & \dots & hG(t_0,t_{n-1}) & \phi(t_0) \\ \vdots & \ddots & \vdots & \vdots \\ hG(t_{n-1},t_1) & \dots & hG(t_{n-1},t_{n-1}) & \phi(t_{n-1}) \end{pmatrix}$$

quindi si conclude che $\|M^{-1}D\| \leq K$, tornando all'errore, usando la relazione $\underline{e} = W^{-1}\underline{\delta}$ abbiamo

$$\|\underline{e}\| \leq K\|\delta\|$$

dato che il metodo è consistente di ordine 2 abbiamo $\|\delta\| = O(h^2)$, pertanto $\|\underline{e}\| = O(h^2)$ e K è legata al condizionamento del BVP.

Costruzione di metodi alle differenze finite

Per ottenere metodi alle differenze finite con ordine di convergenza alto si usano metodi di Runge-Kutta impliciti, ad esempio quelli di Gauss ad s stadi che hanno ordine di convergenza $2s$, si può dimostrare che vale in generale per i metodi di Gauss

$$\|e\| \leq K\|\delta\| \qquad \text{con} \qquad \|\delta\| = O(h^{2s})$$

Un altro metodo per avere ordine di convergenza alto si usano metodi di estrapolazione, questi consistono nel prendere un metodo convergente con ordine basso, considerare varie discretizzazioni e fare combinazioni lineari, ad esempio consideriamo il metodo del punto di mezzo

$$t_{i+1} - t_i = h \qquad x(t_i) - x_i = ch^2 + O(h^4)$$

Cambio discretizzazione (raddoppio i punti)

$$\widehat{t}_{i+1} - \widehat{t}_i = \frac{h}{2}$$

quindi abbiamo

$$x(t_i) - \widehat{x}_{2i} = c\left(\frac{h}{2}\right)^2 + O(h^2) + O(h^4)$$

ora facendo una combinazione lineare

$$\tilde{x}_i = \frac{4\widehat{x}_{2i} - x_i}{3}$$

quindi si ottiene

$$x(t_i) - \tilde{x}_i = O(h^4)$$

Quindi abbiamo aumentato l'ordine di convergenza (questi metodi sono usati anche negli IVP).

2.5 Metodo di linearizzazione

Consideriamo il solito BVP

$$\begin{cases} x'(t) = f(t, x(t)) \\ g(x(0), x(b)) = 0 \end{cases}$$

L'idea è di generare una successione di BVP lineari la cui soluzione converge a quella del BVP non lineare, quindi avremo

$$\begin{cases} x'(t) = A^{(k)}x(t) + q^{(k)}(t) & t \in [0, T] \\ B_0^{(k)}x(0) + B_b^{(k)}x(b) = \beta^{(k)} & k \geq 0 \end{cases}$$

Chiamiamo $x^{(\tilde{k})}$ la soluzione del $\tilde{k}$-esimo BVP lineare, per $\tilde{k}$ grande vogliamo $x^{(\tilde{k})} \simeq x(t)$, dove $x(t)$ è la soluzione del BVP non lineare.

Mostriamo ora come costruire questa successione, consideriamo $x^{(0)}(t)$ una approssimazione iniziale di $x(t)$ (non ci soffermiamo su come scegliere la prima approssimazione), costruiamo induttivamente $x^{(k+1)}(t)$ a partire da $x^{(k)}(t)$, infatti partendo da

$$x'(t) = f(t, x(t))$$

sostituendo lo sviluppo in serie di Taylor di f abbiamo

$$x^{(k+1)}(t)' \simeq f(t, x^{(k)}(t)) + \frac{\partial f}{\partial y}|_{(t, x^{(k)}(t))}((x^{(k+1)}(t) - x^{(k)}(t)))$$

definiamo

$$A^{(k)}(t) = \frac{\partial f}{\partial y}|_{(t, x^{(k)}(t))}$$

quindi avremo

$$x^{(k+1)}(t)' = A^{(k+1)}(t)x^{(k+1)}(t) + q^{(k+1)}(t)$$

dove abbiamo

$$q^{(k+1)}(t) = f(t, x^{(k)}) - A^{(k+1)}(t)x^{(k)}$$

mentre per le condizioni a contorno (lineari) abbiamo

$$0 = g(x^{(k+1)}(0), x^{(k+1)}(b)) \simeq \frac{\partial g}{\partial u}[x^{(k+1)}(0) - x^{(k)}(0)] + \frac{\partial g}{\partial v}[x^{(k+1)}(b) - x^{(k)}(b)] + g(x^{(k)}(0), x^{(k)}(b))$$

Chiamando

$$B_0^{(k+1)} = \frac{\partial g}{\partial u}[x^{(k+1)}(0) - x^{(k)}(0)] \qquad B_b^{(k+1)} = \frac{\partial g}{\partial v}[x^{(k+1)}(b) - x^{(k)}(b)] \qquad \beta^{(k)} = -g(x^{(k)}(0), x^{(k)}(b))$$

Abbiamo che le soluzioni dei seguenti BVP lineari convergono alla soluzione del BVP non lineare

$$\begin{cases} x^{(k)}(t)' = A^{(k)}(t)x^{(k)} + q^{(k)}(t) \\ B_0^{(k)}x^{(k)}(0) + B_b^{(k)}x^{(k)}(b) = \beta^{(k)} \end{cases}$$

Capitolo 3

Equazioni differenziali algebriche (DAE)

Le DAE (differential algebraic equations) sono problemi della forma

$$g(t, x(t), x'(t)) = 0$$

Quindi non abbiamo un'espressione esplicita della derivata, il caso che analizzeremo più a fondo sarà quello lineare

$$Ex'(t) + Ax(t) = q(t)$$

31/05/2011

Esempio 3.1 (Circuito elettrico)**.** Siamo

$$x_1(t) : \mathbb{R} \to \mathbb{R}^p \qquad x_2(t) : \mathbb{R} \to \mathbb{R}^q$$

un DAE che nasce nello studio dei circuiti elettrici è

$$\begin{cases} x_1'(t) = f(t, x_1(t), x_2(t)) \\ g(t, x_1(t), x_2(t)) = 0 \end{cases}$$

La prima idea per risolvere questi tipi di DAE è quella di trasformarli in IVP, ad esempio se $\|x_2'(t)\| < 1$ allora possiamo approssimare la seconda equazione con

$$\varepsilon x_2'(t) \simeq g(t, x_1(t), x_2(t))$$

dove ε è piccolo, quindi il BVP diventa un IVP

$$\begin{cases} x_1'(t) = f(t, x_1(t), x_2(t)) \\ x_2'(t) = \varepsilon^{-1} g(t, x_1(t), x_2(t)) \end{cases}$$

però chiaramente bisognerà fare stime sull'errore, ovvero stimare la differenza tra la soluzione del BVP e quella dell'IVP che lo approssima e questa è una strada difficile.

Un'altro modo per risolvere questo problema è derivare la seconda equazione

$$\frac{\partial g}{\partial t} + \frac{\partial g}{\partial x_1} x_1' + \frac{\partial g}{\partial x_2} x_2' = 0$$

se vale

$$\det \frac{\partial g}{\partial x_2} \neq 0$$

allora

$$x_2'(t) = -\left(\frac{\partial g}{\partial x_2}\right)^{-1} \left(\frac{\partial g}{\partial t} + \frac{\partial g}{\partial x_1} x_1'\right)$$

Gli inconvenienti sono che la matrice $\frac{\partial g}{\partial x_2}$ potrebbe non essere invertibile, o essere malcondizionata, ma anche il calcolo delle derivate di g potrebbe esser un'operazione malcondizionata.

3.1 Caso lineare

Consideriamo il caso delle DAE lineari, ovvero della forma

$$Ex'(t) + Ax(t) = q(t)$$

dove chiediamo

$$E, A \in \mathbb{R}^{n\times n} \qquad x(t), q(t) : \mathbb{R} \to \mathbb{R}^n$$

osserviamo che se E è invertibile allora ci riconduciamo ad un IVP

$$x'(t) = -E^{-1}Ax(t) - E^{-1}q(t)$$

daltronde E potrebbe non esser invertibile oppure esser malcondizionata, quindi vogliamo evitare di fare l'inversione.

3.1.1 Matrix pencil

Il matrix pencil associato ad una DAE lineare è definito come

$$p(\lambda) = \lambda E + A$$

quindi questo è un polinomio di primo grado con coefficienti matriciali o lo si può vedere come matrice con elementi che sono polinomi di primo grado. Diremo che il matrix pencil è regolare se

$$\exists \bar{\lambda} \qquad \text{tale che} \qquad \det P(\bar{\lambda}) \neq 0$$

In caso contrario diremo che il matrix pencil è singolare.

Esempio 3.2 (matrix pensil singolare)**.**

$$E = \left(\begin{array}{c|ccc} 0 & * & \dots & * \\ 0 & * & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & * & \dots & * \end{array}\right) \qquad A = \left(\begin{array}{c|ccc} 0 & * & \dots & * \\ 0 & * & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & * & \dots & * \end{array}\right)$$

è facile verificare che $\det(\lambda E + A) = 0$ per ogni valore di λ.

Esempio 3.3 (matrix pencil regolare)**.**

$$E = \begin{pmatrix} 0 & * & \dots & * \\ & 0 & \ddots & \vdots \\ & & \ddots & * \\ & & & 0 \end{pmatrix} \qquad A = \begin{pmatrix} a_1 & * & \dots & * \\ 0 & a_2 & \ddots & \vdots \\ & \ddots & \ddots & * \\ & & 0 & a_n \end{pmatrix} \qquad \text{con} \quad a_i \neq 0$$

Allora

$$\det(\lambda E + A) = \prod_{i=1}^{n} a_i \neq 0$$

Se il pencil è regolare e $\det P(\mu) = 0$ allora μ è detto autovalore del pensil

Osservazione 3.1. L'esempio precedente può esser interpretato anche come pensil regolare senza autovalori

Se μ è un autovalore e $P(\mu)v = 0$ con $v \neq 0$, allora v è detto autovettore del pencil.

Osservazione 3.2. Se $\det E \neq 0$ allora

$$\lambda E + A = E(\lambda I + E^{-1}A)$$

quindi gli autovalori del pencil sono, a meno di segno, gli autovalori di $E^{-1}A$

Osservazione 3.3. Se $p(\lambda)$ è regolare allora $\det p(\lambda)$ è un polinomio di grado al più n, quindi gli autovalori del pencil sono al più n e sono esattamente n se E è invertibile.

Consideriamo un DAE con un valore iniziale

$$\begin{cases} Ex'(t) + Ax(t) = q(t) & t \in [0, T] \\ x(0) = x_0 \end{cases}$$

Diremo che la DAE è consistente in x_0 se esiste almeno una soluzione del DAE con in valore iniziale. Diremo che la DAE è risolvibile in x_0 se esiste una sola soluzione del DAE con in valore iniziale.

3.1.2 Esistenza delle soluzioni

Consideriamo un DAE con un valore iniziale

$$\begin{cases} Ex'(t) + Ax(t) = q(t) & t \in [0, T] \\ x(0) = x_0 \end{cases}$$

Il seguente teorema ci dice quando le DAE ai valori iniziali ammettono un'unica soluzione

Teorema 3.1 (Esistenza e unicità della soluzione)**.** Un DAE ai valori iniziali è risolvibile se e soltanto se il matrix pencil è regolare

Dimostrazione. Iniziamo con il dimostrare $\Rightarrow$.

Osserviamo subito che se il probleme è risolvibile, allora dati

$$\begin{cases} Ex'(t) + Ax(t) = q(t) \\ x(0) = x_0 \end{cases} \qquad \begin{cases} Ey'(t) + Ay(t) = q(t) \\ y(0) = x_0 \end{cases}$$

abbiamo che la funzione differenza $x(t) - y(t)$ soddisfa

$$\begin{cases} E(x'(t) - y'(t))A(x(t) - y(t)) = 0 \\ x(0) - y(0) = 0 \end{cases}$$

Quindi possiamo direttamente considerare il problema della forma

$$\begin{cases} Ex'(t) + Ax(t) = 0 & t \in [0, T] \\ x(0) = 0 \end{cases}$$

questo ha un'unica (per ipotesi) soluzione $x(t) = 0$, mostriamo che il pencil è regolare. Supponiamo per assurdo che il pencil $p(\lambda)$ non sia regolare, allora posso trovare

$$\lambda_1, \ldots, \lambda_n, \lambda_{n+1}$$

autovalori con i rispettivi autovettori

$$v_1, v_2, \ldots, v_n, v_{n+1}$$

quindi vale $p(\lambda_i)v_i = 0$, pertanto essendo gli autovettori $n+1$ superiori alla dimensione dello spazio (che è n) allora saranno linearmente dipendenti, quindi

$$\exists \alpha_1, \ldots, \alpha_n, \alpha_{n+1} \qquad \text{tali che} \qquad \sum_{j=1}^{n+1} \alpha_j v_j = 0$$

definiamo la funzione

$$y(t) = \sum_{j=1}^{n+1} \alpha_j e^{\lambda_j t} v_j$$

Osserviamo che questa funzione risolve la DAE, infatti

$$y(t) = \sum_{j=1}^{n+1} \alpha_j \lambda_j e^{\lambda_j t} v_j$$

quindi

$$Ey'(t) + Ay(t) = \sum_{j=1}^{n+1} \alpha_j e^{\lambda_j t}(\lambda_j E + A)v_j = \sum_{j=1}^{n+1} \alpha_j e^{\lambda_j t} p(\lambda_j) v_j = 0$$

dove si è usato $p(\lambda_j)v_j = 0$.
Pertanto $y(t)$ è una dunzione non nulla con

$$y(0) = \sum_{j=1}^{n+1} \alpha_j v_j = 0 \quad \text{(indipendenza lineare)}$$

quindi abbiamo trovato un'altra soluzione, contro l'ipotesi di risolvibilità, quindi assurdo. $\Rightarrow$ diamo solo un'idea della dimostrazione. Il pensil è regolare, quindi siano

$$\lambda_1, \dots, \lambda_k \qquad k \leq n$$

gli autovalori e

$$v_1, \dots, v_n$$

i rispettivi autovettori, allora costruisco

$$x(t) = \sum_{j=1}^{k} \alpha_j e^{\lambda_j t} v_j$$

dove gli $\alpha_1, \dots, \alpha_k$ sono tali che $x(0) = x_0$. Non è difficile completare la dimostrazione mostrando che questa è una soluzione (stessa dimostrazione fatta prima, basta derivare) e che è unica (si supponga per assurdo ne esista un'altra e si mostri che il pencil non è regolare). □

01/06/2011

3.1.3 Forma canonica di Weierstrass-Kronecker

Sia $p(\lambda) = \lambda E + A$ un pencil regolare, allora esistono Q_1, Q_2 matrici non singolari tali che

$$Q_1 E Q_2 = \left(\begin{array}{c|c} I & 0 \\ \hline 0 & N \end{array}\right) \qquad Q_1 A Q_2 = \left(\begin{array}{c|c} C & 0 \\ \hline 0 & I \end{array}\right) \qquad N \in \mathbb{C}^{l\times l},\ C \in \mathbb{C}^{(n-l)\times(n-l)}$$

Dove

$$N = \begin{pmatrix} N_1 & & \\ & \ddots & \\ & & N_r \end{pmatrix} \qquad \text{con} \quad N_i = \begin{pmatrix} 0 & 1 & \\ & \ddots & \ddots \\ & & 0 \quad 1 \end{pmatrix}$$

$$C = \begin{pmatrix} C_1 & & \\ & \ddots & \\ & & C_s \end{pmatrix} \qquad \text{con} \quad C_i = \begin{pmatrix} \lambda_i & 1 & \\ & \ddots & \ddots \\ & & \lambda_i \quad 1 \end{pmatrix}$$

dove λ_i sono gli autovalori del pancil. Quindi vale

$$\det(\lambda E + A) = \gamma \det \left(\begin{array}{c|c} \lambda T + C & 0 \\ \hline 0 & \lambda N + I \end{array}\right)$$

Osservazione 3.4. N è una matrice nilpotente.

3.1.4 Soluzione esplicita

Considerial la soluta DAE

$$Ex'(t) + Ax(t) = q(t)$$

usando la forma canonica di Weierstrass-Kronecker abbiamo

$$Q_1 E Q_2 Q_2^{-1} x'(t) + Q_1 A Q_2 Q_2^{-1} x(t) = Q_1 q(t)$$

chiamiamo

$$\begin{pmatrix} u \\ v \end{pmatrix} = Q_2^{-1} x \qquad \begin{pmatrix} r \\ s \end{pmatrix} = Q_1 q(t)$$

allora riusciamo a spezzzare la DAE nel seguente modo

$$\begin{cases} u'(t) + Cu(t) = r(t) \\ Nv'(t) + v(t) = s(t) \end{cases}$$

Osserviamo subito che la prima equazione è una comune ODE che sappiamo già risolvere. Per la seconda equazione abbiamo osservato prima che N è nilpotente, quindi, derivando

$$Nv''(t) + v'(t) = s'(t)$$

quindi

$$v'(t) = s'(t) - Nv''(t)$$

pertanto

$$v(t) = s(t) - Nv'(t) = s(t) - N(s'(t) - Nv''(t)) = s(t) - Ns'(t) + Nv''(t)$$

andando avanti a derivare si trova

$$v(t) = \sum_{j=0}^{\nu-1} (-1)^j N^j s^{(j)}(t)$$

dove ν è l'ordine di nilpontenza di N.

In conclusione si è ridotta una DAE ad una ODE e una DAE con soluzione esplicita, daltronde volendo calcolare questa soluzione bisogna trovare la forma canonica di Weierstrass-Kronecker, risolvere la ODE, derivare $s(t)$ $\nu - 1$ volte, tutte queste operazioni sono numericamente instabili, quindi consideriamo questo solo come risultato teorico.

Osservazione 3.5. I valori x_0 che rendono consistente la DAE sono quelli per cui $v(0)$ (che è fissato) coincide con le ultime $n - l$ componenti di x_0. A volte ν si chiama indice di nilpontenza.

Indice differenziale

L'indice differenziale di una DAE è il minimo ν tale che le soluzioni della seguente ODE

$$\begin{cases} K(t, x(t), x'(t)) = 0 \\ \frac{d}{dt} K(t, x(t), x'(t)) = 0 \\ \frac{d^2}{dt^2} K(t, x(t), x'(t)) = 0 \\ \qquad \vdots \\ \frac{d^\nu}{dt^\nu} K(t, x(t), x'(t)) = 0 \end{cases}$$

coincidono con le soluzioni della DAE

$$K(t, x(t), x'(t)) = 0$$

3.1.5 Metodi numerici

Consideriamo la DAE

$$Ex'(t) + Ax(t) = f(t)$$

con

$$E = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \qquad A = \begin{pmatrix} -1 & -1 \\ 2 & 0 \end{pmatrix}$$

il metodo di Eulero per questa DAE diventa

$$Ex_{i+1} - Ex_i + hAx_i = hf(t_i)$$

Dato che E è singolare non è possibile calcolare x_{i+1}, daltronde se uso Eulero implicito

$$Ex_{i+1} - Ex_i + hAx_{i+1} = hf(t_{i+1})$$

quindi

$$(E + hA)x_{i+1} = Ex_i + hf(t_{i+1})$$

daltronde $E + hA$ è non singolare per $h \neq 1$, quindi ora posso calcolare x_{i+1}.
Nella pratica non si usa il metodo di Eulero ma i metodi BDF. Un LMM implicito per un DAE è

$$\sum_{j=0}^{k} \alpha_j E x_{i-j+1} - h \sum_{j=0}^{k} \beta_j A x_{i-j+1} = h \sum_{j=0}^{k} \beta_j f(t_{i-j+1})$$

la condizione da imporre per rendere calcolabile x_{i+1} è che $\alpha_0 E - h\beta_0 A$ sia non singolare (è equivalente alla regolarità del pencil).

Teorema 3.2. Un metodo BDF a k passi applicato a una DAE (lineare) con indice di nilpontenza ν è convergente con ordine di convergenza k dopo $(\nu - 1)k + 1$ passi

Esempio 3.4. Mostriamo un caso in cui vale il teorema, prendiamo un metodo BDF con $k = \nu = 1$ (quindi Eulero implicito), supponiamo il problema sia già nella forma canonica di Weierstrass-Kronecker, quindi

$$\left(\begin{array}{c|c} I & 0 \\ \hline 0 & N \end{array}\right) x'(t) + \left(\begin{array}{c|c} C & 0 \\ \hline 0 & I \end{array}\right) x(t) = f(t)$$

poniamo

$$x(t) = \begin{pmatrix} y(t) \\ z(t) \end{pmatrix}$$

allora troviamo

$$\left(\begin{array}{c|c} I + hC & 0 \\ \hline 0 & N + hI \end{array}\right) \begin{pmatrix} y_{i+1} \\ z_{i+1} \end{pmatrix} = \left(\begin{array}{c|c} I & 0 \\ \hline 0 & N \end{array}\right) \begin{pmatrix} y_i \\ z_i \end{pmatrix} + h \begin{pmatrix} g_{i+1} \\ k_{i+1} \end{pmatrix}$$

dato che $\nu = 1$ allora $N = 0$, quindi Eulero implicito diventa

$$\begin{cases} (I + hC)y_{i+1} = y_i + hg_{i+1} \\ hz_{i+1} = hk_{i+1} \end{cases}$$

quindi la seconda equazione ci da già la soluzione esatta (come prevedeva il teorema).