

1 Richiami di teoria sui problemi di Cauchy

Il *problema di Cauchy* consiste nel cercare una funzione \mathbf{y} continua e derivabile in un intervallo I_0 di \mathbb{R} contenente un punto x_0 tale che

$$\begin{cases} \mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)) & \forall t \in I_0, \\ \mathbf{y}(t_0) = \mathbf{y}_0, \end{cases} \quad (1)$$

dove \mathbf{f} è una funzione definita e continua su $I_0 \times \mathbb{R}^n$ a valori in \mathbb{R}^n , $\mathbf{y}_0 \in \mathbb{R}^n$ e I_0 non è ridotto al solo punto x_0 . Generalmente l'intervallo I_0 è della forma $[t_0, t_0 + T]$ con $T > 0$.

Il problema di Cauchy (1) è del prim'ordine in quanto interviene sola la derivata prima di \mathbf{y} e richiede una condizione iniziale (data da $\mathbf{y}(t_0) = \mathbf{y}_0$) per evitare di avere certamente più di una soluzione (in effetti, se \mathbf{y} è soluzione di (1) anche $\mathbf{y} + \mathbf{c}$ con \mathbf{c} costante lo sarebbe). Se si considerassero problemi differenziali di ordine p della forma

$$\mathbf{y}^{(p)}(t) = \mathbf{F}(t, \mathbf{y}(t), \mathbf{y}'(t), \dots, \mathbf{y}^{(p-1)}(t)) \quad \forall t \in I_0,$$

servirebbero allora p condizioni iniziali della forma

$$\mathbf{y}(t_0) = \mathbf{y}_{0,1}, \quad \mathbf{y}'(t_0) = \mathbf{y}_{0,2}, \quad \dots, \quad \mathbf{y}^{(p-1)}(t_0) = \mathbf{y}_{0,p},$$

dove $\mathbf{y}_{0,1}, \mathbf{y}_{0,2}, \dots, \mathbf{y}_{0,p}$ sono valori assegnati e \mathbf{y} una funzione di classe $C^p(I_0)$.

Osservazione 1.1 *Le condizioni iniziali non sono l'unico tipo di condizioni che si possono considerare. Ad esempio, si possono considerare condizioni agli estremi dell'intervallo I_0 ed in tal caso si parla di problemi ai limiti. La differenza sostanziale risiede nel fatto che generalmente la soluzione di un problema di Cauchy esiste localmente, vale a dire in un intorno della condizione iniziale, mentre la soluzione di un problema ai limiti se esiste, ha un carattere di globalità.*

Si noti che è sempre possibile ridurre un problema differenziale di ordine p ad un sistema di p equazioni differenziali di ordine 1, ponendo semplicemente

$$\mathbf{z}_1(t) = \mathbf{y}(t), \quad \mathbf{z}_2(t) = \mathbf{y}'(t), \quad \dots, \quad \mathbf{z}_p(t) = \mathbf{y}^{(p-1)}(t).$$

Per questo motivo nel seguito considereremo solo problemi di Cauchy di ordine 1.

1.1 Esistenza della soluzione

Non sempre il problema di Cauchy ammette una soluzione, ad esempio se t_0 è un punto di discontinuità di prima specie per \mathbf{f} la soluzione non esiste. Il teorema più semplice che si può avere a tale riguardo è il seguente:

Teorema 1.1 (di Cauchy-Peano) *Se \mathbf{f} è una funzione continua in un intorno di (t_0, \mathbf{y}_0) , allora esiste un intorno J_0 di t_0 ed una funzione $\mathbf{y} \in C^1(J_0)$ soluzione del problema di Cauchy (1) in J_0 .*

Questo teorema (che si può dimostrare attraverso il teorema delle contrazioni) non garantisce l'unicità, nè fornisce un metodo per la costruzione della soluzione stessa. Ad esempio, il problema di Cauchy

$$\begin{cases} y'(t) = \sqrt[3]{y(t)}, \\ y(0) = 0 \end{cases}$$

ha almeno tre soluzioni (quali?), pur essendo $f(t, y) = \sqrt[3]{y(t)}$ continua in tutto \mathbb{R}^2 .

Prima di passare a teoremi con condizioni più restrittive, introduciamo alcuni concetti. Diciamo che la coppia (I, \mathbf{y}) è una *soluzione locale* di (1) se $I \subseteq I_0$, $\mathbf{y} \in C^1(I)$ che soddisfa (1) $\forall t \in I$. Si dice inoltre che la coppia (J, \mathbf{z}) *prolunga* la soluzione locale (I, \mathbf{y}) se $I \subseteq J$ e $\mathbf{y}(t) = \mathbf{z}(t) \forall t \in I$. La *prolunga strettamente* se $J \neq I$. Una soluzione locale è detta *massimale* se non c'è una soluzione locale che la prolunga strettamente. Infine, diciamo che una soluzione locale (I, \mathbf{y}) è una *soluzione globale* se $I = I_0$.

Ad esempio, il problema di Cauchy

$$\begin{cases} y'(t) = 2xy^2(t), & t \in \mathbb{R}, \\ y(0) = 1, \end{cases}$$

ammette una soluzione massimale $y(t) = 1/(1 - t^2)$ per $t \in (-1, 1)$ e non ammette una soluzione globale.

Si potrebbe in effetti dimostrare che se \mathbf{f} è una funzione continua e definita su $I_0 \times \mathbb{R}^n$ con $I_0 = [t_0, t_0 + T]$ (o $I_0 = [t_0, t_0 + T)$ o $I_0 = [t_0, \infty)$) e (I, \mathbf{y}) è una soluzione massimale non globale di (1), allora I è un intervallo della forma $[t_0, \bar{t})$ e \mathbf{y} è illimitata su I .

L'interesse naturalmente sta nella possibilità di ottenere soluzioni globali. A questo proposito dimostriamo il teorema seguente:

Teorema 1.2 (di esistenza globale) Supponiamo $I_0 = [t_0, t_0 + T]$ e \mathbf{f} continua e definita su $I_0 \times \mathbb{R}^p$. Supponiamo che

$$(\mathbf{f}(t, \mathbf{y}), \mathbf{y}) \leq l(t)(1 + \|\mathbf{y}\|^2) \quad \forall (t, \mathbf{y}) \in I_0 \times \mathbb{R}^p, \quad (2)$$

con $l \in L^1(I_0)$, allora il problema (1) ammette almeno una soluzione globale su I_0 , dove (\cdot, \cdot) denota il prodotto scalare in \mathbb{R}^p e $\|\mathbf{z}\|^2 = (\mathbf{z}, \mathbf{z})$.

Dimostrazione 1.1 Consideriamo una soluzione locale (I, \mathbf{y}) con $I = [t_0, t_0 + \bar{T}]$, allora

$$(\mathbf{y}'(t), \mathbf{y}) = (\mathbf{f}(t, \mathbf{y}(t)), \mathbf{y}(t)), \quad \forall t \in I$$

ovvero

$$\frac{d}{dt} \|\mathbf{y}\|^2 = 2(\mathbf{f}(t, \mathbf{y}(t)), \mathbf{y}(t)), \quad \forall t \in I.$$

Per la (2), si ha

$$\frac{d}{dt} \|\mathbf{y}\|^2 \leq 2l(t)(1 + \|\mathbf{y}\|^2).$$

Dividendo per $1 + \|\mathbf{y}\|^2$ ed integrando su $[t_0, t]$, per $t \in I$ si ha

$$\int_{t_0}^t \frac{1}{1 + \|\mathbf{y}(\tau)\|^2} \frac{d}{d\tau} \|\mathbf{y}(\tau)\|^2 d\tau \leq 2 \int_{t_0}^t l(\tau) d\tau.$$

Grazie alla derivata logaritmica, si ricava allora la seguente disuguaglianza

$$\ln \left(\frac{1 + \|\mathbf{y}(t)\|^2}{1 + \|\mathbf{y}(t_0)\|^2} \right) \leq 2 \int_{t_0}^t l(\tau) d\tau,$$

e quindi

$$1 + \|\mathbf{y}\|^2 \leq (1 + \|\mathbf{y}_0\|^2) \exp \left(2 \int_{t_0}^t l(\tau) d\tau \right) \quad \forall t \in I.$$

Di conseguenza, tutte le soluzioni locali sono limitate su I e quindi anche quelle massimali, ma allora esse sono anche soluzioni globali perché se non lo fossero dovrebbero essere illimitate.

Il teorema precedente può essere esteso ad intervalli qualsiasi pur di modificare leggermente l'ipotesi (2). Inoltre, affinché il teorema sia soddisfatto basta che

$$\|\mathbf{f}(t, \mathbf{y})\| \leq \tilde{l}(t)(1 + \|\mathbf{y}\|) \quad \forall (t, \mathbf{y}) \in I_0 \times \mathbb{R}^p.$$

1.2 Unicità della soluzione

Il problema di Cauchy (1) ammette un'unica soluzione se ammette una soluzione globale e tutte le soluzioni locali sono restrizioni di tale soluzione. L'unicità discende dal seguente teorema:

Teorema 1.3 (di esistenza ed unicità) *Sia $I_0 = [t_0, t_0 + T]$ e \mathbf{f} continua e definita su $I_0 \times \mathbb{R}^p$. Supponiamo che*

$$(\mathbf{f}(t, \mathbf{y}) - \mathbf{f}(t, \mathbf{z}), \mathbf{y} - \mathbf{z}) \leq l(t) \|\mathbf{y} - \mathbf{z}\|^2 \quad \forall (t, \mathbf{y}), (t, \mathbf{z}) \in I_0 \times \mathbb{R}^p, \quad (3)$$

con $l \in L^1(I_0)$, allora il problema (1) ammette un'unica soluzione.

Dimostrazione 1.2 L'esistenza discende immediatamente dal Teorema 1.2 prendendo $\mathbf{z} = \mathbf{0}$.

Consideriamo una soluzione locale (I, \mathbf{z}) ed una soluzione globale (I, \mathbf{y}) e definiamo la funzione

$$\phi(t) = e^{-2L(t)} \|\mathbf{y}(t) - \mathbf{z}(t)\|^2 \quad \text{con } L(t) = \int_{t_0}^t l(\tau) d\tau.$$

Allora

$$\phi'(t) = -2l(t)e^{-2L(t)} \|\mathbf{y}(t) - \mathbf{z}(t)\|^2 + e^{-2L(t)} \frac{d}{dt} \|\mathbf{y}(t) - \mathbf{z}(t)\|^2.$$

D'altra parte

$$\frac{d}{dt} \|\mathbf{y}(t) - \mathbf{z}(t)\|^2 = 2(\mathbf{y}'(t) - \mathbf{z}'(t), \mathbf{y}(t) - \mathbf{z}(t))$$

e quindi

$$\phi'(t) = 2e^{-2L(t)} ((\mathbf{y}'(t) - \mathbf{z}'(t), \mathbf{y}(t) - \mathbf{z}(t)) - l(t) \|\mathbf{y}(t) - \mathbf{z}(t)\|^2) \leq 0.$$

Di conseguenza, $\phi(t) \leq \phi(t_0) = 0$ ed essendo $\phi(t) \geq 0$ per costruzione si deve avere $\phi(t) = 0$ per ogni t e quindi ogni soluzione locale (I, \mathbf{z}) è restrizione di una soluzione globale (I_0, \mathbf{y}) .

Vale inoltre il seguente corollario:

Corollario 1.1 *Supponiamo $I_0 = [t_0, t_0 + T]$ e \mathbf{f} continua e definita su $I_0 \times \mathbb{R}^p$. Supponiamo inoltre che sia Lipschitziana rispetto al secondo argomento di costante $L > 0$ ossia che*

$$\|\mathbf{f}(t, \mathbf{y}) - \mathbf{f}(t, \mathbf{z})\| \leq L \|\mathbf{y} - \mathbf{z}\|, \quad \forall (t, \mathbf{y}), (t, \mathbf{z}) \in I_0 \times \mathbb{R}^p, \quad (4)$$

allora il problema di Cauchy (1) ammette un'unica soluzione.

Dimostrazione 1.3 Se verifichiamo l'ipotesi (3) il corollario è dimostrato. Grazie alla disuguaglianza di Cauchy-Schwarz si ha

$$(\mathbf{f}(t, \mathbf{y}) - \mathbf{f}(t, \mathbf{z}), \mathbf{y} - \mathbf{z}) \leq \|\mathbf{f}(t, \mathbf{y}) - \mathbf{f}(t, \mathbf{z})\| \|\mathbf{y} - \mathbf{z}\| \leq L\|\mathbf{y} - \mathbf{z}\|^2$$

e quindi la (3) con $l(t) = L$.

La condizione (4) è comunque difficile da verificare in pratica. In effetti, se \mathbf{f} è derivabile rispetto a \mathbf{y} allora è anche necessariamente Lipschitziana di costante

$$L = \max_{t \in I_0} \left| \frac{\partial \mathbf{f}}{\partial \mathbf{y}} \right|. \quad (5)$$

In effetti, per il teorema del valor medio esiste $\boldsymbol{\xi}$ tra \mathbf{y} e \mathbf{z} tale che

$$\mathbf{f}(t, \mathbf{y}) - \mathbf{f}(t, \mathbf{z}) = \frac{\partial \mathbf{f}(t, \boldsymbol{\xi})}{\partial \mathbf{y}} ((t, \mathbf{y}) - (t, \mathbf{z})),$$

da cui passando alle norme e massimizzando si trova (5).

Per quanto riguarda la regolarità della soluzione di un problema di Cauchy si può osservare che se \mathbf{f} è di classe C^n relativamente a tutte le variabili allora \mathbf{y} è di classe C^{n+1} .

1.3 Dipendenza della soluzione dai dati

Supponiamo di considerare solo problemi di Cauchy che ammettano un'unica soluzione (e quindi riteniamo valide le ipotesi del Teorema 1.3). Siamo interessati a mostrare come eventuali perturbazioni sul dato iniziale \mathbf{y}_0 o sulla funzione \mathbf{f} possano agire sulla soluzione. L'interesse può essere motivato con l'osservazione che i metodi numerici che introdurremo nel seguito non risolvono il problema di Cauchy esatto (a causa ad esempio dei puri errori di arrotondamento), ma una sua versione perturbata della forma

$$\begin{cases} \bar{\mathbf{y}}'(t) = \mathbf{f}(t, \bar{\mathbf{y}}) + \rho(\bar{\mathbf{y}}) & \forall t \in I_0 = [t_0, t_0 + T], \\ \bar{\mathbf{y}}(t_0) = \mathbf{y}_0 + \rho_0, \end{cases} \quad (6)$$

dove ρ è una funzione continua su I_0 a valori in \mathbb{R}^p e ρ_0 una costante. Supponiamo che tale problema ammetta un'unica soluzione globale e che \mathbf{f} soddisfi l'ipotesi (3).

Calcoliamo

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \|\bar{\mathbf{y}} - \mathbf{y}\|^2 &= (\bar{\mathbf{y}}' - \mathbf{y}', \bar{\mathbf{y}} - \mathbf{y}) = (\mathbf{f}(t, \bar{\mathbf{y}}) - \mathbf{f}(t, \mathbf{y}), \bar{\mathbf{y}} - \mathbf{y}) + \rho(\bar{\mathbf{y}})(\bar{\mathbf{y}} - \mathbf{y}) \\ &\leq l(t) \|\bar{\mathbf{y}} - \mathbf{y}\|^2 + \rho(t)(\bar{\mathbf{y}} - \mathbf{y}). \end{aligned}$$

Definiamo

$$\phi(t) = e^{-2L(t)} \|\bar{\mathbf{y}}(t) - \mathbf{y}(t)\|^2 \quad \text{con } L(t) = \int_{t_0}^t l(\tau) d\tau$$

e, di conseguenza,

$$\|\bar{\mathbf{y}}(t) - \mathbf{y}(t)\|^2 = e^{2L(t)} \phi(t).$$

Allora

$$\frac{1}{2} \frac{d}{dt} \|\bar{\mathbf{y}} - \mathbf{y}\|^2 = l(t) e^{2L(t)} \phi(t) + \frac{1}{2} e^{2L(t)} \phi'(t) = l(t) \|\bar{\mathbf{y}} - \mathbf{y}\|^2 + \frac{1}{2} e^{2L(t)} \phi'(t)$$

e quindi

$$\frac{1}{2} e^{2L(t)} \phi'(t) \leq \rho(t) (\bar{\mathbf{y}}(t) - \mathbf{y}(t)) = \rho(t) e^{L(t)} \sqrt{\phi(t)}.$$

In conclusione

$$\frac{1}{2} \frac{\phi'(t)}{\sqrt{\phi(t)}} \leq \rho(t) e^{-L(t)}$$

o anche ($\forall \mu > 0$)

$$\frac{1}{2} \frac{\phi'(t)}{\sqrt{\phi(t) + \mu}} \leq \rho(t) e^{-L(t)}.$$

Integrando e passando al limite per μ che tende a 0 si ha

$$\sqrt{\phi(t)} - \sqrt{\phi(t_0)} \leq \int_{t_0}^t \rho(\tau) e^{-L(\tau)} d\tau.$$

D'altra parte $\sqrt{\phi(t)} = \|\bar{\mathbf{y}}(t) - \mathbf{y}(t)\| e^{-L(t)}$ e $\sqrt{\phi(t_0)} = \|\mathbf{y}(t_0)\| e^{-L(t_0)}$ e quindi

$$\|\bar{\mathbf{y}}(t) - \mathbf{y}(t)\| \leq \|\mathbf{y}(t_0)\| e^{L(t)-L(t_0)} + e^{L(t)} \int_{t_0}^t \rho(\tau) e^{-L(\tau)} d\tau.$$

2 Risoluzione numerica di equazioni differenziali ordinarie

In questo paragrafo completiamo la panoramica offerta nel corso di Calcolo Numerico A dei metodi numerici per la risoluzione di equazioni differenziali ordinarie. In particolare, presentiamo due classi di metodi: i metodi multistep o multipasso (in breve, MS) ed i metodi di tipo Runge-Kutta.

2.1 I metodi a più passi (o multistep)

Un metodo si dice a q passi ($q \geq 1$) se $\forall n \geq q - 1$, u_{n+1} dipende da u_{n+1-q} , ma non da valori u_k con $k < n + 1 - q$.

In particolare consideriamo nel seguito i soli *metodi multistep (lineari)* a $p + 1$ passi (con $p \geq 0$) definiti dalla seguente relazione

$$u_{n+1} = \sum_{j=0}^p a_j u_{n-j} + h \sum_{j=0}^p b_j f_{n-j} + h b_{-1} f_{n+1}, \quad n = p, p + 1, \dots \quad (7)$$

I coefficienti a_j e b_j , assegnati in \mathbb{R} , individuano lo schema e sono tali che $a_p \neq 0$ o $b_p \neq 0$. Nel caso in cui $b_{-1} \neq 0$ lo schema è implicito, mentre nel caso contrario lo schema è esplicito. Naturalmente, per $p = 0$ si ritrovano gli schemi ad un passo.

Appare evidente che per innescare un metodo multistep servono q condizioni iniziali u_0, \dots, u_{q-1} . Poiché il problema di Cauchy ne fornisce una sola (u_0), una via per assegnare le condizioni mancanti consiste nell'utilizzare metodi espliciti ad un passo di ordine elevato. Degli esempi verranno forniti dai metodi Runge-Kutta.

Riformuliamo il metodo (7) come segue

$$\sum_{s=0}^{p+1} \alpha_s u_{n+s} = h \sum_{s=0}^{p+1} \beta_s f(t_{n+s}, u_{n+s}), \quad n = 0, 1, \dots, N_h - (p + 1). \quad (8)$$

avendo posto $\alpha_{p+1} = 1$, $\alpha_s = -a_{p-s}$ per $s = 0, \dots, p$ e $\beta_s = b_{p-s}$ per $s = 0, \dots, p + 1$. La (8) è una equazione lineare alle differenze.

Definizione 2.1 *L'errore di troncamento locale (LTE) $\tau_{n+1}(h)$ introdotto dal metodo multistep (7) nel punto t_{n+1} (per $n \geq p$) è definito dalla relazione*

$$h\tau_{n+1}(h) = y_{n+1} - \left[\sum_{j=0}^p a_j y_{n-j} + h \sum_{j=-1}^p b_j y'_{n-j} \right], \quad n \geq p \quad (9)$$

dove si è posto $y_{n-j} = y(t_{n-j})$ e $y'_{n-j} = y'(t_{n-j})$ per $j = -1, \dots, p$.

La quantità $h\tau_{n+1}(h)$ è il residuo che si genera nel punto t_{n+1} avendo preteso di “far verificare” alla soluzione esatta lo schema numerico. Indicando con $\tau(h) = \max_n |\tau_n(h)|$ l’errore di troncamento (globale), si ha la seguente definizione:

Definizione 2.2 (Consistenza) *Il metodo multistep è consistente se $\tau(h) \rightarrow 0$ per $h \rightarrow 0$. Se inoltre $\tau(h) = \mathcal{O}(h^q)$, per qualche $q \geq 1$, allora il metodo si dirà di ordine q .*

Si può dare una più precisa caratterizzazione del LTE introducendo il seguente operatore lineare \mathcal{L} associato al metodo MS lineare (7)

$$\mathcal{L}[w(t); h] = w(t+h) - \sum_{j=0}^p a_j w(t-jh) - h \sum_{j=-1}^p b_j w'(t-jh), \quad (10)$$

dove $w \in C^1(I)$ è una funzione arbitraria. Si noti che il LTE è esattamente $\mathcal{L}[y(t_n); h]$. Se supponiamo che w sia sufficientemente regolare e sviluppiamo in serie di Taylor $w(t-jh)$ e $w'(t-jh)$ in $t-ph$, otteniamo

$$\mathcal{L}[w(t); h] = C_0 w(t-ph) + C_1 h w^{(1)}(t-ph) + \dots + C_k h^k w^{(k)}(t-ph) + \dots$$

Di conseguenza, se un metodo MS ha ordine q e $y \in C^{q+1}(I)$, si ha

$$\tau_{n+1}(h) = C_{q+1} h^{q+1} y^{(q+1)}(t_{n-p}) + \mathcal{O}(h^{q+2})$$

Il termine $C_{q+1} h^{q+1} y^{(q+1)}(t_{n-p})$ è detto *errore di troncamento locale principale* (PLTE, dall’inglese *principal local truncation error*) mentre C_{q+1} è la costante dell’errore. Il PLTE è largamente usato nella costruzione di strategie adattive per i metodi MS

Introduciamo ora due famiglie specifiche di metodi multistep.

2.2 I metodi di Adams

Questi metodi vengono derivati dalla forma integrale del problema di Cauchy. Si suppone che i nodi di discretizzazione siano equispaziati, ovvero $t_j = t_0 + jh$, con $h > 0$ e $j \geq 1$; indi, anziché f , si integra il suo polinomio interpolatore su $\tilde{p} + \theta$ nodi distinti, con $\theta = 1$ se i metodi sono espliciti ($\tilde{p} \geq 0$ in tal caso) e $\theta = 2$ se i metodi sono impliciti ($\tilde{p} \geq -1$). Gli schemi risultanti sono dunque *consistenti* per costruzione e hanno la seguente espressione

$$u_{n+1} = u_n + h \sum_{j=-1}^{\tilde{p}+\theta} b_j f_{n-j} \quad (11)$$

I nodi di interpolazione possono essere dati da

1. $t_n, t_{n-1}, \dots, t_{n-\tilde{p}}$, in tal caso $b_{-1} = 0$ e lo schema risultante è esplicito, oppure da
2. $t_{n+1}, t_n, \dots, t_{n-\tilde{p}}$, in questo caso $b_{-1} \neq 0$ e lo schema è implicito.

I metodi *impliciti* sono detti di *Adams-Moulton*, mentre quelli *espliciti* sono detti di *Adams-Bashforth*.

Metodi di Adams-Bashforth (AB) (espliciti)

Per $\tilde{p} = 0$ si ritrova il metodo di Eulero esplicito, essendo il polinomio interpolatore di grado zero nel nodo t_n dato da $\Pi_0 f = f_n$. Per $\tilde{p} = 1$, il polinomio interpolatore lineare nei nodi t_{n-1} e t_n è dato da

$$\Pi_1 f(t) = f_n + (t - t_n) \frac{f_{n-1} - f_n}{t_{n-1} - t_n}.$$

Poiché $\Pi_1 f(t_n) = f_n$, mentre $\Pi_1 f(t_{n+1}) = 2f_n - f_{n-1}$, si ottiene

$$\int_{t_n}^{t_{n+1}} \Pi_1 f(t) dt = \frac{h}{2} [\Pi_1 f(t_n) + \Pi_1 f(t_{n+1})] = \frac{h}{2} [3f_n - f_{n-1}].$$

Si ottiene pertanto lo schema AB a due passi

$$u_{n+1} = u_n + \frac{h}{2} [3f_n - f_{n-1}]. \quad (12)$$

Nel caso in cui $\tilde{p} = 2$, si trova in modo del tutto analogo lo schema AB a tre passi

$$u_{n+1} = u_n + \frac{h}{12} [23f_n - 16f_{n-1} + 5f_{n-2}]$$

mentre per $\tilde{p} = 3$ si trova il metodo AB a quattro passi

$$u_{n+1} = u_n + \frac{h}{24} (55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3})$$

Si osservi che i metodi di Adams-Bashforth usano $\tilde{p} + 1$ nodi e sono a $\tilde{p} + 1$ passi (con $\tilde{p} \geq 0$). In generale, gli schemi di Adams-Bashforth a q passi sono di ordine q . Le costanti dell'errore C_{q+1}^* di questi metodi sono indicate nella Tabella 1.

I metodi di Adams-Moulton (AM) (impliciti)

Per $\tilde{p} = -1$ si ritrova il metodo di Eulero implicito, mentre nel caso $\tilde{p} = 0$, si usa il polinomio interpolatore lineare di f nei nodi t_n e t_{n+1} e si ritrova lo schema di Crank-Nicolson. Nel caso del metodo a due passi ($\tilde{p} = 1$), si costruisce il polinomio di grado 2 interpolante f nei nodi t_{n-1} , t_n , t_{n+1} e si trova un nuovo schema (del terz'ordine) dato da

$$u_{n+1} = u_n + \frac{h}{12} [5f_{n+1} + 8f_n - f_{n-1}] \quad (13)$$

Gli schemi successivi con $\tilde{p} = 2$ e $\tilde{p} = 3$ sono dati rispettivamente da

$$u_{n+1} = u_n + \frac{h}{24} (9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2})$$

$$u_{n+1} = u_n + \frac{h}{720} (251f_{n+1} + 646f_n - 264f_{n-1} + 106f_{n-2} - 19f_{n-3})$$

Si osservi che i metodi di Adams-Moulton usano $\tilde{p}+2$ nodi e sono a $\tilde{p}+1$ passi se $\tilde{p} \geq 0$, con la sola eccezione del metodo di Eulero implicito (corrispondente a $\tilde{p} = -1$) che usa un nodo ed è un metodo ad un passo. I metodi di Adams-Moulton a q passi hanno ordine $q+1$. Le loro costanti dell'errore C_{q+1} sono indicate nella Tabella 1.

q	C_{q+1}^*	C_{q+1}	q	C_{q+1}^*	C_{q+1}
1	$\frac{1}{2}$	$-\frac{1}{2}$	3	$\frac{3}{8}$	$-\frac{1}{24}$
2	$\frac{5}{12}$	$-\frac{1}{12}$	4	$\frac{251}{720}$	$-\frac{19}{720}$

Tabella 1: Costanti dell'errore per i metodi di Adams-Bashforth e di Adams-Moulton di ordine q

2.3 I metodi BDF

I metodi alle differenze all'indietro (*backward differentiation formulae*, brevemente indicati con BDF) sono metodi MS impliciti che si ottengono in maniera complementare ai metodi di Adams. Se infatti in questi si è ricorso all'integrazione numerica per la funzione f , nei metodi BDF si approssima direttamente il valore della derivata prima di y nel nodo t_{n+1} tramite la derivata prima del polinomio interpolatore di y di grado $p+1$ nei $p+2$ nodi

$t_{n+1}, t_n, \dots, t_{n-p}$, con $p \geq 0$. In questo modo si trovano tutti schemi della forma

$$u_{n+1} = \sum_{j=0}^p a_j u_{n-j} + hb_{-1} f_{n+1} \quad (14)$$

con $b_{-1} \neq 0$. Riportiamo nella Tabella 2 i coefficienti dei metodi BDF che sono zero-stabili.

p	a_0	a_1	a_2	a_3	a_4	a_5	b_{-1}
0	1	0	0	0	0	0	1
1	$\frac{4}{3}$	$-\frac{1}{3}$	0	0	0	0	$\frac{2}{3}$
2	$\frac{18}{11}$	$-\frac{9}{11}$	$\frac{2}{11}$	0	0	0	$\frac{6}{11}$
3	$\frac{48}{25}$	$-\frac{36}{25}$	$\frac{16}{25}$	$-\frac{3}{25}$	0	0	$\frac{12}{25}$
4	$\frac{300}{137}$	$-\frac{300}{137}$	$\frac{200}{137}$	$-\frac{75}{137}$	$\frac{12}{137}$	0	$\frac{60}{137}$
5	$\frac{360}{147}$	$-\frac{450}{147}$	$\frac{400}{147}$	$-\frac{225}{147}$	$\frac{72}{147}$	$-\frac{10}{147}$	$\frac{60}{147}$

Tabella 2: I coefficienti dei metodi BDF zero-stabili per $p = 0, 1, \dots, 5$

2.4 Analisi dei metodi multistep

Analizziamo in questo paragrafo le condizioni che assicurano la consistenza e la stabilità dei metodi multistep. L'obiettivo è quello di ricondurre tale verifica al controllo di semplici relazioni algebriche.

Relativamente alla *consistenza* vale il seguente risultato:

Teorema 2.1 *Il metodo multistep (7) è consistente se e solo se sono soddisfatte le seguenti condizioni algebriche sui coefficienti*

$$\sum_{j=0}^p a_j = 1, \quad -\sum_{j=0}^p j a_j + \sum_{j=-1}^p b_j = 1 \quad (15)$$

Se inoltre $y \in C^{q+1}(I)$ per qualche $q \geq 1$, dove y è la soluzione del problema di Cauchy il metodo è di ordine q se e solo se vale la (15) ed inoltre sono soddisfatte le seguenti condizioni

$$\sum_{j=0}^p (-j)^i a_j + i \sum_{j=-1}^p (-j)^{i-1} b_j = 1, \quad i = 2, \dots, q \quad (16)$$

Applichiamo il metodo multistep (7) al problema modello $y' = \lambda y$. La soluzione numerica soddisfa l'equazione lineare alle differenze

$$u_{n+1} = \sum_{j=0}^p a_j u_{n-j} + h\lambda \sum_{j=-1}^p b_j u_{n-j}, \quad (17)$$

per la quale si cercano quindi soluzioni fondamentali della forma $u_k = [r_i(h\lambda)]^k$, $k = 0, 1, \dots$, essendo $r_i(h\lambda)$, per $i = 0, \dots, p$, le radici del polinomio $\Pi \in \mathbb{P}_{p+1}$

$$\Pi(r) = \rho(r) - h\lambda\sigma(r). \quad (18)$$

Si sono indicati rispettivamente con

$$\rho(r) = r^{p+1} - \sum_{j=0}^p a_j r^{p-j}, \quad \sigma(r) = b_{-1} r^{p+1} + \sum_{j=0}^p b_j r^{p-j}$$

il *primo* ed il *secondo polinomio caratteristico* del metodo multistep (7). Il polinomio $\Pi(r)$ si chiama *polinomio caratteristico* associato all'equazione alle differenze (17), e le sue radici $r_j(h\lambda)$ si dicono *radici caratteristiche*.

Evidentemente, le radici di ρ sono date dalle $r_i(0)$, con $i = 0, \dots, p$, e verranno indicate nel seguito semplicemente con r_i . Inoltre, la prima delle condizioni di consistenza (15) comporta che se un metodo multistep è consistente, allora 1 è radice di ρ . Supporremo che tale radice (di consistenza) sia $r_0(0) = r_0$ e chiameremo *principale* la corrispondente radice $r_0(h\lambda)$ del polinomio caratteristico (18).

Diremo che il metodo multistep (7) soddisfa la *condizione delle radici* se tutte le radici r_i sono contenute nel cerchio unitario centrato nell'origine del piano complesso. Nel caso in cui una radice cada sul bordo di tale cerchio, essa deve essere una radice semplice di ρ . Equivalentemente,

$$\begin{cases} |r_j| \leq 1, & j = 0, \dots, p; \\ \text{se inoltre } |r_j| = 1, & \text{allora } \rho'(r_j) \neq 0. \end{cases} \quad (19)$$

Il metodo MS (7) soddisfa la *condizione forte delle radici* se soddisfa la condizione delle radici e se inoltre $r_0 = 1$ è l'unica radice che cade sul bordo del cerchio unitario. Equivalentemente,

$$|r_j| < 1, \quad j = 1, \dots, p. \quad (20)$$

Infine, il metodo MS (7) soddisfa la *condizione assoluta delle radici* se esiste $h_0 > 0$ tale che

$$|r_j(h\lambda)| < 1, \quad j = 0, \dots, p, \quad \forall h < h_0. \quad (21)$$

2.5 Analisi di stabilità e di convergenza per i metodi multistep

Individuiamo ora le relazioni che intercorrono fra le condizioni delle radici e le proprietà di stabilità di un metodo multistep.

Diciamo che il metodo multistep (7) a $p + 1$ passi è zero-stabile se

$$\exists h_0 > 0, \exists C > 0 : \quad \forall h \in (0, h_0], |z_n^{(h)} - u_n^{(h)}| \leq C\varepsilon, \quad 0 \leq n \leq N_h, \quad (22)$$

dove $N_h = \max\{n : t_n \leq t_0 + T\}$ e $z_n^{(h)}, u_n^{(h)}$ sono rispettivamente le soluzioni dei problemi

$$\begin{cases} z_{n+1}^{(h)} = \sum_{j=0}^p a_j z_{n-j}^{(h)} + h \sum_{j=-1}^p b_j f(t_{n-j}, z_{n-j}^{(h)}) + h\delta_{n+1}, \\ z_k^{(h)} = w_k^{(h)} + \delta_k, \quad k = 0, \dots, p \end{cases} \quad (23)$$

$$\begin{cases} u_{n+1}^{(h)} = \sum_{j=0}^p a_j u_{n-j}^{(h)} + h \sum_{j=-1}^p b_j f(t_{n-j}, u_{n-j}^{(h)}), \\ u_k^{(h)} = w_k^{(h)}, \quad k = 0, \dots, p \end{cases} \quad (24)$$

per $p \leq n \leq N_h - 1$, dove $|\delta_k| \leq \varepsilon, 0 \leq k \leq N_h, w_0^{(h)} = y_0$ e $w_k^{(h)}, k = 1, \dots, p$, sono p valori iniziali generati usando un altro schema numerico.

A questo punto si può dimostrare che per un metodo multistep consistente, la condizione delle radici (19) è equivalente alla zero-stabilità.

Il teorema precedente consente di caratterizzare il comportamento, in merito alla stabilità, di diverse famiglie di metodi di discretizzazione.

Nel caso particolare dei metodi ad un passo consistenti, il polinomio ρ ammette la sola radice $r_0 = 1$. Essi dunque *soddisfano automaticamente la condizione delle radici* e sono pertanto zero-stabili.

Per i metodi di Adams (11), il polinomio ρ assume sempre la forma $\rho(r) = r^{p+1} - r^p$. Le sue radici sono pertanto $r_0 = 1$ e $r_1 = 0$ (con molteplicità p) e dunque tutti i metodi di Adams sono zero-stabili.

Infine, i metodi BDF riportati nel paragrafo 2.3 sono zero-stabili purché $p \leq 5$, essendo in tali casi soddisfatta la condizione delle radici.

Possiamo ora dare il seguente risultato di convergenza:

Teorema 2.2 *Un metodo multistep consistente è convergente se e solo se è soddisfatta la condizione delle radici e l'errore sui dati iniziali è un infinitesimo rispetto a h . Inoltre, esso converge con ordine q se sia $\tau(h)$ sia l'errore sui dati iniziali sono infinitesimi di ordine q rispetto ad h .*

Una notevole conseguenza del Teorema 2.2 è il seguente teorema di equivalenza.

Corollario 2.1 (Teorema di equivalenza) *Un metodo multistep consistente è convergente se e solo se è zero-stabile e se l'errore sui dati iniziali tende a zero per h che tende a zero.*

2.6 L'assoluta stabilità nei metodi multistep

Il metodo MS (7) applicato al problema modello genera l'equazione alle differenze (17), la cui soluzione assume la forma

$$u_n = \sum_{j=0}^{k'} \left(\sum_{s=0}^{m_j-1} \gamma_{sj} n^s \right) [r_j(h\lambda)]^n, \quad n = 0, 1, \dots$$

dove le $r_j(h\lambda)$, $j = 0, \dots, k'$ sono le radici distinte del polinomio caratteristico (18). Abbiamo indicato con m_j la molteplicità della radice $r_j(h\lambda)$. È chiaro che la *condizione assoluta delle radici* è necessaria e sufficiente ad assicurare che il metodo MS (7) sia assolutamente stabile se $h \leq h_0$.

Fra i metodi che godono della proprietà di assoluta stabilità sono da preferire quelli per i quali la regione di assoluta stabilità \mathcal{A} è molto estesa o addirittura illimitata. Fra questi, vi sono i metodi *A-stabili* ed i metodi *ϑ -stabili*; per questi ultimi \mathcal{A} contiene la regione angolare definita dagli $z \in \mathbb{C}$ tali che $-\vartheta < \pi - \arg(z) < \vartheta$ con $\theta \in (0, \pi/2)$. In particolare i metodi *A-stabili* sono molto importanti nella risoluzione di problemi *stiff*.

2.7 Metodi predictor-corrector

La risoluzione di un problema di Cauchy non lineare con uno schema implicito, richiede ad ogni passo temporale la risoluzione di un'equazione non lineare. Se si usa il metodo di Newton per risolvere questa equazione, si trova

$$u_{n+1}^{(k+1)} = u_{n+1}^{(k)} - \Phi(u_{n+1}^{(k)})/\Phi'(u_{n+1}^{(k)}),$$

per $k = 0, 1, \dots$, fino a convergenza e richiedendo che il dato iniziale $u_{n+1}^{(0)}$ sia sufficientemente vicino a u_{n+1} .

Alternativamente, si può usare una iterazione di punto fisso della forma

$$u_{n+1}^{(k+1)} = \Psi(u_{n+1}^{(k)}) \quad (25)$$

per $k = 0, 1, \dots$, fino a convergenza. In tal caso, la condizione di convergenza per un metodo di punto fisso comporterà una limitazione sul passo di discretizzazione della forma

$$h < \frac{1}{|b_{-1}|L} \quad (26)$$

dove L è la costante di Lipschitz di f rispetto a y . In pratica, ad eccezione dei problemi *stiff* questa restrizione su h non è troppo penalizzante in quanto considerazioni di accuratezza impongono restrizioni ben maggiori su h . Tuttavia, ogni iterazione di (25) richiede una valutazione della funzione f ed il costo computazionale può essere contenuto solo fornendo un buon dato iniziale $u_{n+1}^{(0)}$. Questo può essere fatto eseguendo un passo di un metodo MS esplicito ed iterando poi per un numero *fissato* m di iterazioni. Così facendo, il metodo MS implicito usato nello schema di punto fisso corregge il valore di u_{n+1} “predetto” dallo schema MS esplicito. Il metodo che si ottiene è detto complessivamente *metodo predictor-corrector*, o metodo PC. Un metodo PC può essere realizzato in vari modi.

Nella sua versione elementare, il valore $u_{n+1}^{(0)}$ viene calcolato tramite un metodo esplicito a $\tilde{p} + 1$ -passi, detto il *predictor* (i cui coefficienti verranno indicati con $\{\tilde{a}_j, \tilde{b}_j\}$)

$$[P] \quad u_{n+1}^{(0)} = \sum_{j=0}^{\tilde{p}} \tilde{a}_j u_{n-j}^{(1)} + h \sum_{j=0}^{\tilde{p}} \tilde{b}_j f_{n-j}^{(0)},$$

dove $f_k^{(0)} = f(t_k, u_k^{(0)})$ e $u_k^{(1)}$ sono le soluzioni calcolate con il metodo PC al passo precedente oppure sono le condizioni iniziali.

A questo punto, si valuta la funzione f nel nuovo punto $(t_{n+1}, u_{n+1}^{(0)})$ (*fase di valutazione*)

$$[E] \quad f_{n+1}^{(0)} = f(t_{n+1}, u_{n+1}^{(0)}),$$

ed infine si esegue una sola iterazione del metodo di punto fisso usando uno schema MS implicito della forma (7), ovvero

$$[C] \quad u_{n+1}^{(1)} = \sum_{j=0}^p a_j u_{n-j}^{(1)} + h b_{-1} f_{n+1}^{(0)} + h \sum_{j=0}^p b_j f_{n-j}^{(0)}.$$

Questo secondo passo della procedura è ora esplicito ed il metodo che si usa è detto *corrector*. La procedura nel suo insieme viene denotata in breve come metodo *PEC* o $P(EC)^1$, in cui P e C indicano una applicazione del metodo predictor e del metodo corrector al tempo t_{n+1} , mentre E indica che è stata effettuata una valutazione di f .

Possiamo generalizzare questa strategia supponendo di eseguire $m > 1$ iterazioni al passo t_{n+1} . Il metodo corrispondente è detto *predictor-multicorrector* e calcola $u_{n+1}^{(0)}$ al tempo t_{n+1} usando il predictor nella forma

$$[P] \quad u_{n+1}^{(0)} = \sum_{j=0}^{\tilde{p}} \tilde{a}_j u_{n-j}^{(m)} + h \sum_{j=0}^{\tilde{p}} \tilde{b}_j f_{n-j}^{(m-1)}. \quad (27)$$

Qui $m \geq 1$ indica il numero (fissato) di iterazioni che vengono eseguite nei passi $[E]$, $[C]$: per $k = 0, 1, \dots, m-1$

$$[E] \quad f_{n+1}^{(k)} = f(t_{n+1}, u_{n+1}^{(k)}),$$

$$[C] \quad u_{n+1}^{(k+1)} = \sum_{j=0}^p a_j u_{n-j}^{(m)} + hb_{-1} f_{n+1}^{(k)} + h \sum_{j=0}^p b_j f_{n-j}^{(m-1)}.$$

Questa implementazione del metodo PC è comunemente indicata come $P(EC)^m$. Una differente implementazione, nota come $P(EC)^m E$, prevede che al termine del processo venga nuovamente valutata la funzione f . Abbiamo quindi

$$[P] \quad u_{n+1}^{(0)} = \sum_{j=0}^{\tilde{p}} \tilde{a}_j u_{n-j}^{(m)} + h \sum_{j=0}^{\tilde{p}} \tilde{b}_j f_{n-j}^{(m)},$$

e per $k = 0, 1, \dots, m-1$,

$$[E] \quad f_{n+1}^{(k)} = f(t_{n+1}, u_{n+1}^{(k)}),$$

$$[C] \quad u_{n+1}^{(k+1)} = \sum_{j=0}^p a_j u_{n-j}^{(m)} + hb_{-1} f_{n+1}^{(k)} + h \sum_{j=0}^p b_j f_{n-j}^{(m)},$$

seguito da

$$[E] \quad f_{n+1}^{(m)} = f(t_{n+1}, u_{n+1}^{(m)}).$$

Introduciamo una semplificazione nelle notazioni. Usualmente, il numero di passi del metodo predictor è maggiore di quello del metodo corrector; di conseguenza definiamo il numero di passi del metodo predictor-corrector come il numero di passi del predictor. Denoteremo questo numero con p .

In ogni metodo predictor-corrector, l'errore di troncamento del *predictor* combinato con quello del *corrector*, genera un nuovo errore di troncamento che ora esaminiamo. Siano \tilde{q} e q , rispettivamente, gli ordini del predictor e del corrector, e supponiamo che $y \in C^{\hat{q}+1}$, dove $\hat{q} = \max(\tilde{q}, q)$. Allora

$$\begin{aligned} y(t_{n+1}) &= \sum_{j=0}^{\tilde{p}} \tilde{a}_j y(t_{n-j}) - h \sum_{j=0}^{\tilde{p}} \tilde{b}_j f(t_{n-j}, y_{n-j}) \\ &= \tilde{C}_{\tilde{q}+1} h^{\tilde{q}+1} y^{(\tilde{q}+1)}(t_{n-p}) + \mathcal{O}(h^{\tilde{q}+2}), \\ y(t_{n+1}) &= \sum_{j=0}^p a_j y(t_{n-j}) - h \sum_{j=-1}^p b_j f(t_{n-j}, y_{n-j}) \\ &= C_{q+1} h^{q+1} y^{(q+1)}(t_{n-p}) + \mathcal{O}(h^{q+2}), \end{aligned}$$

dove $\tilde{C}_{\tilde{q}+1}, C_{q+1}$ sono le costanti dell'errore dei metodi predictor e corrector rispettivamente. Vale il seguente risultato.

Teorema 2.3 *Supponiamo che il metodo predictor abbia ordine \tilde{q} e il metodo corrector abbia ordine q . Allora:*

se $\tilde{q} \geq q$ (o $\tilde{q} < q$ con $m > q - \tilde{q}$), il metodo predictor-corrector ha lo stesso ordine e lo stesso PLTE del corrector;

se $\tilde{q} < q$ e $m = q - \tilde{q}$, allora il metodo predictor-corrector ha lo stesso ordine del corrector, ma diverso PLTE;

se $\tilde{q} < q$ e $m \leq q - \tilde{q} - 1$, allora il metodo predictor-corrector ha ordine pari a $\tilde{q} + m$ (quindi minore di q).

In particolare, si noti che se il predictor ha ordine $q - 1$ ed il corrector ha ordine q , il metodo *PEC* fornisce un metodo di ordine q . Inoltre, i metodi $P(EC)^m E$ e $P(EC)^m$ hanno sempre lo stesso ordine e lo stesso PLTE.

Per quanto riguarda l'assoluta stabilità si noti soltanto che la regione di assoluta stabilità del metodo predictor-corrector è fortemente influenzata dalla regione di assoluta stabilità del metodo predictor e, di conseguenza, non esisteranno in generale metodi predictor-corrector incondizionatamente stabili.