Soluzione numerica di equazioni differenziali

con codici in Matlab/Octave

Stefano De Marchi
Dipartimento di Matematica "Tullio Levi-Civita",
Università di Padova

March 21, 2019

Introduzione

Queste pagine sono gli appunti del corso di Metodi Numerici per Equazioni Differenziali che gli autori hanno tenuto dall'AA. 2007-08, per il corso di laurea triennale in Matematica Applicata della Facoltà di Scienze dell'Università degli Studi di Verona. Al lettore è richiesta la familiarità con `Matlab`, `MATrix LABoratory`, o la sua versione freeware `Octave`, di cui si è spesso fatto uso nel testo per scrivere pezzi di codici che implementano alcuni degli esempi e algoritmi numerici. Chi desiderasse conoscere Matlab, la sua sintassi e il suo utilizzo, rimandiamo alla lettura del libro [16] oppure alla miriade di manuali disponibili in rete.

Per quanto riguarda `Octave`, il manuale è incluso nel download del package che si trova al link

`http://www.gnu.org/software/octave/`.

Gli appunti sono organizzati in capitoli, corrispondenti agli argomenti trattati in un corso di base di metodi numerici per equazioni differenziali.

In ogni capitolo c'è una sessione di *Esercizi proposti*: si tratta di una raccolta di esercizi proposti nel corso degli ultimi tre anni nei vari appelli, compiti e compitini da parte dell'autore.

Il testo non ha la pretesa di essere sostitutivo di libri molto più completi e dettagliati disponibili in letteratura, come ad esempio i libri [1, 3, 4, 11, 15, 16, 17, 20], ma come traccia di riferimento per un corso introduttivo alla soluzione numerica di equazioni differenziali. Pertanto l'invito è di consultare anche i testi citati in bibliografia, sia per cultura personale, ma soprattutto per un completamento della preparazione.

Ringrazio fin d'ora tutti coloro che ci segnaleranno sviste ed errori e ci daranno dei consigli per eventuali miglioramenti.

Stefano De Marchi
Dipartimento di Matematica "Tullio Levi-Civita"
Università di Padova.

Indice

Elenco delle Figure

Elenco delle Tabelle

Capitolo 1

Generalità sulle equazioni differenziali ordinarie

Le equazioni differenziali sono il linguaggio che la matematica, la fisica, l'ingegneria e molte altre scienze applicate, usano per modellizzare i fenomeni reali.

Un'equazione differenziale è un'equazione che coinvolge una o più derivate della funzione incognita. Ad esempio

$$y'(x) = f(x, y(x)), \tag{1.1}$$
$$y''(x) = f(x, y(x), y'(x)), \tag{1.2}$$

sono due equazioni differenziali di `primo` e `secondo` ordine, rispettivamente. Diremo che un'equazione differenziale ha `ordine` p, se p è l'ordine massimo di derivazione che appare nella sua formula. L'equazione (1.1) è del `primo` ordine, mentre (1.2) è del secondo ordine.

Vediamo un paio di esempi significativi.

Esempio 1. **Legge di Stefan-Bolzmann**. Consideriamo un corpo di massa m a temperatura interna T in un ambiente a temperatura T_e. La velocità di trasferimento di calore tra il corpo e l'ambiente è regolato dalla legge di `Stefan-Bolzmann`

$$v(t) = \epsilon\gamma S(T^4(t) - T_e^4) ,$$

dove $\epsilon = 5.6 \cdot 10^{-8} J/(m^2\, K^4\, s)$ (è detta *costante di Bolzmann*), γ, che è costante, viene detta `emissività` del corpo, S è la superficie del corpo e v la velocità di trasferimento. Sapendo che l'energia termica vale $E(t) = m\, C\, T(t)$, con C che indica il `calore specifico` e

$$\left| \frac{dE(t)}{dt} \right| = |v(t)| ,$$

allora varrà l'equazione

$$\frac{dT(t)}{dt} = -\frac{v(t)}{m\, C} \tag{1.3}$$

che rappresenta un'equazione differenziale del primo ordine non lineare, che dovremo risolvere per comprendere il fenomeno descritto.

Esempio 2. **Crescita di una popolazione**. Sia $y(t)$ una popolazione di batteri (ma questo esempio si può generalizzare al caso di una popolazione di persone) posta in un ambiente limitato, ovvero dove non possono vivere più di B batteri. Assumiamo $y_0 \ll B$. Sia $C > 0$ il fattore di crescita, allora la velocità di cambiamento dei batteri al tempo t sarà proporzionale al numero dei batteri presenti al tempo t, ma limiata dal fatto che non possono vivere più di B batteri. L'equazione differenziale corrispondente, nota come *equazione logistica*, è

$$\frac{d\,y(t)}{dt} = Cy(t)\left(1 - \frac{y(t)}{B}\right), \tag{1.4}$$

che è ancora un'equazione del primo ordine la cui soluzione mi da il numero dei batteri presenti al tempo t.

Se avessimo due popolazioni in competizione, il problema si modellizzarà mediante il *sistema di ordinarie*

$$\begin{cases} \dfrac{d\,y_1(t)}{dt} = C_1 y_1(t)\,(1 - b_1 y_1(t) - d_2 y_2(t)) \\ \dfrac{d\,y_2(t)}{dt} = -C_2 y_2(t)\,(1 - b_2 y_1(t) - d_1 y_2(t)) \end{cases} \tag{1.5}$$

con $C_1 > 0$ e $C_2 > 0$ che sono i fattori di crescita delle rispettive popolazioni; d_1, d_2 sono i coefficienti che governano il tipo d'interazione tra le due popolazioni e b_1, b_2 sono legati alla disponibilità di nutrienti. Le equazioni (1.5) sono le famose equazioni di `Lotka-Volterra` che sono un sistema di equazioni differenziali ordinarie del primo ordine non-lineari.

$\diamond\diamond$

In generale un' equazione differenziale ordinaria ammette sempre infinite soluzioni. Per ottenere una particolare soluzione dobbiamo fissare alcune condizioni che possono essere `condizioni iniziali`, `condizioni al bordo` o ai limiti oppure entrambe. Il numero delle condizioni in genere dipende dall'ordine dell'equazione differenziale.

Ad esempio, l'equazione (1.4) ammette la `soluzione generale` o `famiglia` di soluzioni, posto $\varphi(t;k) = \mathrm{e}^{Ct+k}$, $k \in \mathbb{R}$,

$$y(t) = \frac{B\varphi(t;k)}{1+\varphi(t;k)}, \quad t \geq 0\,.$$

Se chiediamo che $y(0) = 1$, allora $\varphi(0;k) = e^k$. Imponendo $y(0) = 1$ si otterrà il valore $k^* = -\log(B-1)$ e la particolare soluzione

$$\boxed{y(t) = \frac{B\varphi(t;k^*)}{1+\varphi(t;k^*)} = \frac{B\mathrm{e}^{Ct+k^*}}{1+\mathrm{e}^{Ct+k^*}}\,.}$$

Problema di Cauchy. Sia I un intervallo della retta reale. Il problema di Cauchy, consiste nel determinare una funzione $y : I \longrightarrow \mathbb{R}$ tale che

$$\begin{cases} y'(t) = f(t, y(t)) \\ y(t_0) = y_0\,, \end{cases} \tag{1.6}$$

con $f : I \times \mathbb{R} \to \mathbb{R}$, $t_0 \in I$. y_0 viene chiamato *dato o valore iniziale*. Se f non dipende da t, cioè $f = f(y)$, allora l'equazione differenziale si dice `equazione autonoma`. Vale il seguente teorema

Teorema 1. *Supponiamo che $f(t, y)$ soddisfi le condizioni*

- *sia continua rispetto a t e y;*
- *sia lipschitziana rispetto a y, ovvero $\exists\, L > 0$ t.c.*

$$|f(t, y_1) - f(t, y_2)| \leq L|y_1 - y_2|,\ \forall t \in I,\ y_1, y_2 \in \mathbb{R}.$$

Allora la soluzione del problema (1.6) esiste ed è unica.

Osservazione. Solo un numero limitato di equazioni differenziali ammette una soluzione in forma esplicita. Nella quasi maggioranza dei casi però la soluzione è nota in forma implicita. Quasi sempre invece possiamo trovare la soluzione numericamente.

Un paio di esempi

- $y'(t) = \dfrac{y - t}{y + t}$, ha soluzioni che verificano la forma implicita $y(t) = \dfrac{1}{2}\log(y^2 + t^2) + \arctan(y/t) + C$.
- $y'(t) = \mathrm{e}^{-t^2}$, ha soluzione esprimibile solo con uno sviluppo in serie.

Capitolo 2

Metodi numerici per problemi ai valori iniziali

2.1 Metodi numerici per problemi ai valori iniziali

2.1.1 Metodi di Eulero

Consideriamo il problema di Cauchy (1.6). Dato l'intervallo $I = [t_0, T]$, $T < \infty$, prendiamo un passo $h = (T - t_0)/N$, con $N \geq 1$ che indica il numero dei sottointervalli in cui suddivideremo I, e i punti equispaziati t_n, $0 \leq n \leq N$. Sia poi u_n il valore approssimato della soluzione $y(t_n)$, ovvero $u_n \approx y_n := y(t_n)$, ottenuto con il metodo discreto che costruiamo approssimando $y'(t)$ con il rapporto incrementale

$$y'(t_n) \approx \frac{y_{n+1} - y_n}{h}\,, \tag{2.1}$$

dove $y_{n+1} = y(t_{n+1})$ e $y_n = y(t_n)$. Sostituendo in (1.6), ricordando che approssimiamo con u_n la soluzione vera $y_n := y(t_n)$, otteniamo la formula del *metodo di Eulero esplicito (EE)*

$$u_{n+1} = u_n + h\, f_n\,, n = 0, 1, \ldots, N-1 \tag{2.2}$$

ove abbiamo usato la notazione $f_n = f(t_n, u_n)$.

Se invece dell'approssimazione (2.1) usassimo

$$y'(t_{n+1}) = \frac{y_{n+1} - y_n}{h} \tag{2.3}$$

oppure

$$y'(t_n) = \frac{y_n - y_{n-1}}{h} \tag{2.4}$$

otterremo il *metodo di Eulero implicito (EI)* (o all'indietro)

$$u_{n+1} = u_n + h\, f_{n+1}\,, n = 0, 1, \ldots, N-1 \tag{2.5}$$

dove $f_{n+1} = f(t_{n+1}, u_{n+1})$.

Pertanto, poiché $u_0 = y_0$, l'insieme dei valori $u_1, \ldots, u_N$ rappresentano la *soluzione numerica* del nostro problema.

ESEMPIO 3. Consideriamo ancora l'`equazione logistica` (1.4),

$$y'(t) = Cy(t)\left(1 - \frac{y(t)}{B}\right).$$

Se usiamo l'approssimazione con Eulero esplicito (2.2) essa diventa

$$u_{n+1} = u_n + Ch\, u_n(1 - u_n/B)\; n \geq 0\,.$$

Con Eulero implicito (2.5) essa diventa

$$u_{n+1} = u_n + Ch\, u_{n+1}(1 - u_{n+1}/B)\; n \geq 0\,.$$

In quest'ultimo caso appare evidente che usando un metodo implicito per il calcolo della soluzione al passo t_{n+1}, si dovrà risolvere un'equazione non lineare. Vedremo più oltre, che nonostante i metodi impliciti siano più costosi essi sono più stabili.

2.1.2 Analisi di convergenza del metodo EE

Iniziamo con una definizione.

Definizione 1. *Un metodo numerico si dice* `convergente` *di ordine p se*

$$|u_n - y_n| \leq C(h)\,, \forall\, n = 0, 1, ..., N\,, \tag{2.6}$$

con $C(h) = \mathcal{O}(h^p)$.

La definizione ci dice che per la convergenza è necessario che l'errore assoluto sia un infinitesimo di ordine p rispetto ad h. Nel caso del metodo EE, per l'errore assoluto possiamo scrivere

$$e_n = u_n - y_n = (u_n - u_n^*) + (u_n^* - y_n)\,, \tag{2.7}$$

dove u_n^* rappresenta la soluzione numerica calcolata in t_n a partire dalla soluzione esatta y_{n-1} al tempo t_{n-1} (vedi Fig. 2.1). Vediamo di analizzare le due parti che compongono l'errore assoluto:

- $u_n^* - y_n$: errore prodotto da un passo del metodo EE (errore algoritmico);

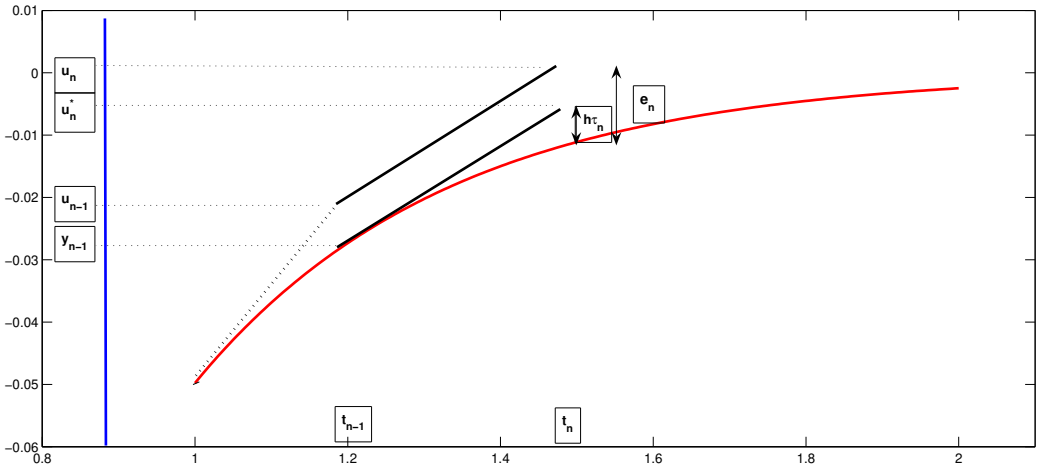

Figura 2.1: Metodo di Eulero esplicito: analisi dell'errore

- $u_n - u_n^*$: propagazione dell'errore da t_{n-1} a t_n.

Il metodo EE sarà convergente se entrambi i termini in (2.7) risulteranno essere infinitesimi al tendere di h a zero.

Ora, sapendo che $u_{n-1}^* = y_{n-1}$, per il primo termine di (2.7), possiamo scrivere:

$$u_n^* - y_n = y_{n-1} - y_n + hf_{n-1} = \frac{h^2}{2} y''(\xi_n), \quad \xi_n \in (t_{n-1}, t_n)\,. \tag{2.8}$$

Posto $\tau_n(h) = \dfrac{u_n^* - y_n}{h}$, che chiameremo `errore locale di troncamento` (esso rappresenta l'errore che si sarebbe commesso se si forzasse la soluzione esatta a soddisfare il metodo numerico), possiamo allora definire l'`errore globale di troncamento` come

$$\tau(h) := \max_{0\le n\le N} |\tau_n(h)|\,. \tag{2.9}$$

Sia $M = \max_{t\in[t_0,T]} |f'(t, y(t))|$, dalla (2.8) l'errore globale di troncamento assume la forma

$$\tau(h) = \frac{Mh}{2}\,. \tag{2.10}$$

Da cui $\lim_{h\to 0} \tau(h) = 0$.

Definizione 2. *Un metodo numerico si dice* `consistente` *quando l'errore globale di troncamento è infinitesimo per $h \to 0$. Inoltre se $\tau(h) = \mathcal{O}(h^p)$, $p \ge 1$ allora il metodo si dice consistente di ordine p.*

Da questa definizione e dalla relazione (2.10) deduciamo che il *metodo EE è consistente.*

Analizziamo il secondo termine della (2.7) che possiamo scrivere come

$$u_n^* - u_n = e_{n-1} + h[f(t_{n-1}, y_{n-1}) - f(t_{n-1}, u_{n-1})]\,.$$

Essendo f, per ipotesi, lipschitziana rispetto al secondo argomento, per cui $|f(t_{n-1}, y_{n-1}) - f(t_{n-1}, u_{n-1})| \le L|y_{n-1} - u_{n-1}|$, allora

$$|u_n^* - u_n| \le (1 + hL)|e_{n-1}|\,.$$

Complessivamente l'errore assoluto in (2.7), assumendo $e_0 = 0$, si può maggiorare come segue

$$\begin{aligned} |e_n| &\le |u_n - u_n^*| + |u_n^* - y_n| \le (1+hL)|e_{n-1}| + h|\tau_n(h)| \quad \text{iterando} \\ &\le (1+hL)^n|e_0| + (1+hL)^{n-1}h\tau(h) + \cdots + h\tau(h) \\ &\le \left((1+hL)^{n-1} + \cdots + (1+hL) + 1\right) h\tau(h) \\ &= \frac{(1+hL)^n - 1}{hL} h\tau(h) \le \frac{e^{L(t_n - t_0)} - 1}{L}\tau(h)\,, \end{aligned}$$

avendo ricordato che $t_n - t_0 = nh$.

Infine

$$\boxed{|e_n| \le \frac{e^{L(t_n - t_0)} - 1}{L}\,\frac{Mh}{2}\,, \quad \forall\, n = 0, 1, \ldots, N.} \tag{2.11}$$

Da cui deduciamo che il *metodo di EE converge* con ordine 1. In tal caso, l'ordine di convergenza è quello dell'ordine di troncamento locale.

Vediamo ora alcune utili osservazioni.

1. Se f soddisfa l'ulteriore condizione

$$\frac{\partial f(t,y)}{\partial y} \le 0, \;\; \forall t \in [t_0, T]$$

facciamo vedere che invece di (2.11) otterremo la stima

$$\boxed{|e_n| \le \frac{Mh}{2}(t_n - t_0)\,.} \tag{2.12}$$

Infatti,

$$\begin{aligned} u_n^* - u_n &= e_{n-1} + h[f(t_{n-1}, y_{n-1}) - f(t_{n-1}, u_{n-1})] \\ &= e_{n-1} + h\,\frac{\partial f(t_{n-1}, \eta_n)}{\partial y}(y_{n-1} - u_{n-1}) \\ &= (1 + h\frac{\partial f(t_{n-1}, \eta_n)}{\partial y})e_{n-1}\,. \end{aligned}$$

Pertanto $|u_n^* - u_n| \leq |e_{n-1}|$ se vale la `condizione di stabilità`

$$h < \frac{2}{\max_t \left| \frac{\partial f(t,y)}{\partial y} \right|} . \tag{2.13}$$

In definitiva, per l'errore avremo

$$|e_n| \leq |u_n^* - u_n| + |e_{n-1}| \leq \underbrace{|e_0|}_{=0} + nh\tau(h) = nh\frac{Mh}{2} = \frac{Mh}{2}(t_n - t_0) ,$$

2. La consistenza è necessaria per la convergenza, altrimenti ad ogni passo il metodo accumulerebbe errori non infinitesimi e quindi quando $h \to 0$ si avrebbe $e_n \nrightarrow 0$.

3. La stima (2.11) non tiene conto del fatto che la soluzione numerica u_n è calcolata in aritmetica finita. Se teniamo conto anche degli errori di arrotondamento, e_n per $h \to 0$ esploderebbe come $\mathcal{O}(1/h)$ (cf. [1]). Pertanto, una scelta ragionevole non è far tendere h a zero, ma $h > h^*$ (con h^* in genere piccolissimo).

2.1.3 θ metodo, Crank-Nicolson e Heun

Dato il problem $y'(t) = f(t, y(t))$, consideriamo la seguente approssimazione

$$u_{n+1} = u_n + h[\theta f_{n+1} + (1-\theta) f_n] \,;\; \theta \in [0,1] , \tag{2.14}$$

con, al solito, $f_n = f(t_n, u_n)$ e analogamente per f_{n+1}. Questo metodo rappresenta una famiglia di metodi. Infatti per $\theta = 0$ si ottiene il metodo di Eulero esplicito; per $\theta = 1$ si riottiene il metodo di Eulero implicito. Inoltre per $\theta \neq 0$ il metodo è implicito.

Metodo di Crank-Nicolson
Un caso che merita particolare attenzione è per $\theta = \frac{1}{2}$ che corrisponde al *metodo di Crank-Nicolson* (in breve CN) detto anche *metodo del trapezio*. Il metodo di CN si chiama anche del trapezio, poiché

$$y(t) - y(t_0) = \int_{t_0}^{t} y'(\tau)\, d\,\tau = \int_{t_0}^{t} f(\tau, y(\tau))\, d\,\tau$$

si conclude applicando la formula dei trapezi all'ultimo integrale. Quindi, per ottenere la fomula (2.14) con $\theta = 1/2$ su ogni sottointervallo $[t_{n-1}, t_n]$ si applica la formula dei trapezi.

Il metodo di CN si può anche ottenere sommando membro a membro le espressioni dei metodi di EE e di EI. Infatti,

$$\begin{aligned} u_{n+1} &= u_n + h\, f_n \;\; (EE) \\ u_{n+1} &= u_n + h\, f_{n+1} \;\; (EI) \\ - - - - - &- - - - - - - - \\ u_{n+1} &= u_n + \frac{h}{2}[f_n + f_{n+1}] \;\; (CN) . \end{aligned}$$

Il metodo di CN è un metodo `implicito ad 1 passo`.

Analizziamo ora l'errore di troncamento locale.

$$h\tau_n(h) = y_n - y_{n-1} - \frac{h}{2}[f_n + f_{n-1}] = \int_{t_{n-1}}^{t_n} f(t, y(t))\, dt - \frac{h}{2}[f_n + f_{n-1}]\,,$$

con $f_n = f(t_n, y(t_n))$ e $f_{n-1} = f(t_{n-1}, y(t_{n-1}))$. Il termine

$$-\frac{h}{2}[f_n + f_{n-1}]$$

esprime l'errore commesso usando la formula dei trapezi (semplice) che appunto vale $-\frac{(t_n - t_{n-1})^3}{12} f''(\xi_n)$. Pertanto, se $y \in \mathcal{C}^3([t_{n-1}, t_n])$ (meglio se lo è su tutto $[0, T]$), allora

$$\tau_n(h) = -\frac{h^3}{12} y'''(\xi_n), \quad \xi_n \in (t_{n-1}, t_n)\,. \tag{2.15}$$

Possiamo allora dire che il metodo di CN è `consistente di ordine 2`. Più oltre faremo vedere che il metodo di CN è anche convergente.

Metodo di Heun
Si ottiene come per il CN con la differenza che $f(t_{n+1}, u_{n+1})$ viene sostituita con $f(t_{n+1}, u_n + hf_n)$, ovvero usando EE per calcolare u_{n+1} rendendolo un *metodo esplicito*:

$$u_{n+1} = u_n + \frac{h}{2}[f_n + f(t_{n+1}, u_n + hf_n)]\,. \tag{2.16}$$

2.1.4 Zero-stabilità

Si valuta su intervalli limitati e fissati.

Definizione 3. *Un metodo numerico per risolvere un IVP su $I = [t_0, T]$ si dice* `zero-stabile`*, se esiste una costante $C > 0$ t.c. per ogni $\delta > 0$ per ogni h*

$$|z_n - u_n| \le C\, h, \quad 0 \le n \le N_h\,,$$

con C che dipende dalla lunghezza di I e z_n che rappresenta la soluzione del problema perturbato con massima perturbazione pari a δ.

Infatti, nello studio di stabilità, si vede come una perturbazione sul dato iniziale si ripercuote sul risultato.

Osservazioni.

1. Per un *metodo consistente ad un passo*, la zero-stabilità deriva dal fatto che f è continua e Lipschitziana. In tal caso la costante C dipende da $\mathrm{e}^{L(T-t_0)}$ con L che è la costante di Lipschitz.

2. Nel caso del metodo di EE, la soluzione z_n di cui alla definizione sarebbe quella che si ottiene risolvendo il problema

$$\begin{cases} z_{n+1} = z_n + h[f(t_n, z_n) + \eta_n]\,, \quad n = 0, ..., N_h - 1\,, \\ z_0 = y_0 + \eta_0\,. \end{cases}$$

supponendo che $\max_n |\eta_n| \leq \delta$.

Ma c'è un altro modo di valutare la zero-stabilità: analizzando il `polinomio caratteristico` associato al metodo numerico. Vedremo più oltre, quando parleremo dei metodi multipasso o `metodi multistep` come costruire il polinomio caratterisco associato ad un metodo numerico per IVP. Di tale polinomio ci interessa la `condizione delle radici`.

Definizione 4. *Un metodo numerico si dice* `zero stabile` *se il polinomio caratteristico* $p(r)$ *ha le radici che soddisfano le condizioni*

$$\begin{aligned} |r_j| &\leq 1, \quad \forall\, j = 0, \ldots, p\,, \\ p'(r_j) &\neq 0, \quad \textit{quando } |r_j| = 1\,. \end{aligned}$$

Il polinomio caratteristico associato al θ-metodo (che include EE, EI e CN) è $p(r) = r - 1$. Pertanto esso ha una sola radice, $r = 1$, che soddisfa la condizione delle radici e quindi `il` θ`-metodo,` $\theta \in [0, 1]$`, è zero-stabile.`

Vale il seguente importantissimo **teorema di equivalenza di Lax** (noto anche come teorema di Lax-Richtmeyer) la cui dimostrazione si trova ad esempio in [13, 18].

Teorema 2. *Ogni metodo consistente è convergente se e solo se è zero-stabile.*

In altre parole

$$\boxed{\text{Consistenza + (zero) Stabilità} \iff \text{Convergenza}}$$

2.1.5 Stabilità assoluta

La stabilità assoluta si ricerca su intervalli illimitati, ovvero per $t \to \infty$. Lo studio si fa ricorrendo al *problema modello*

$$\begin{cases} y'(t) = \lambda y(t)\,, \quad t \in (0,\infty), \quad \lambda \in \mathbb{R}_- \\ y(0) = 1 \end{cases} \tag{2.17}$$

del quale si conosce la soluzione esplcita $y(t) = \mathrm{e}^{\lambda t}$ che descresce verso 0 al tendere di $t \to +\infty$ (essendo $\lambda < 0$).

Analizziamo il comportamento (al limite) dei metodi EE, EI e CN per comprendere meglio il significato della stabilità assoluta.

- La discretizzazione di (2.17) con il *metodo di EE*, ricordando che $u_0 = 1$, porta alla formula
$$u_{n+1} = (1 + \lambda\, h)^{n+1}\,, \quad n \geq 0\,.$$
Pertanto $\lim_{n\to\infty} u_n = 0$ se e solo se $|1 + \lambda\, h| < 1$ ovvero se e solo se
$$h < 2/|\lambda|\,. \tag{2.18}$$
La (2.18) rappresenta la `condizione di assoluta stabilità` del metodo di Eulero esplicito.

- La discretizzazione di (2.17) con il *metodo di EI*, sempre ricordando che $u_0 = 1$, porta alla formula
$$u_{n+1} = \left(\frac{1}{1 - \lambda\, h}\right)^{n+1}\,, \quad n \geq 0\,.$$
Pertanto $\lim_{n\to\infty} u_n = 0$ per ogni valore di $h > 0$.

- Infine, la discretizzazione di (2.17) con il *metodo di CN*, sapendo che $u_0 = 1$, porta alla formula
$$u_{n+1} = \left[\frac{\left(1 + \frac{\lambda h}{2}\right)}{\left(1 - \frac{\lambda h}{2}\right)}\right]^{n+1}\,, \quad n \geq 0\,,$$
che, come per il metodo di EI, tende a zero per $n \to \infty$.

Le formule discrete di EE, EI e CN per il problema test ci danno le seguenti informazioni: il metodo di Eulero esplicito è `condizionatamente stabilite` mentre i metodi di Eulero implicito e di Crank-Nicolson sono `incondizionatamente stabili` (essendo la loro assoluta stabilità indipendente dalla scelta del passo h).

Questi calcoli ci permettono di fare un'osservazione di carattere generale: i metodi impliciti, pur essendo computazionalmente più costosi, sono in genere più stabili.

Vale la seguente

Definizione 5. *I metodi numerici incondizionatamente stabili per il problema modello (2.17) con* $\lambda \in \mathbb{C}$, $\mathtt{Re}\lambda < 0$ *sono detti* `A-stabili`.

Una variate del problema modello, consiste nel prendere λ invece che costante, una funzione negativa $\lambda(t)$. L'esame di stabilità assoluta per EE porta alla condizione $\max_{t\in[0,\infty)} |\lambda(t)|$.

Alternativamente, potremmo variare il passo h in modo adattivo tenendo conto dell'andamento locale $|\lambda(t)|$. Questa idea porta ad una variate del metodo EE nota come `metodo di Eulero esplicito adattivo`. Da un punto di vista implementativo facciamo vedere un pseudocodice per il metodo di Eulero esplicito adattivo applicato al problema modello.

$u_0 = y_0$;
$h_0 = 2\eta/|\lambda(t_0)|$ ($\eta < 1$ mi permette di soddisfare la (2.18))
for $n = 0, 1, \ldots$ (oppure while($\mathrm{t_n} \le \mathrm{t_{fin}}$))
$t_{n+1} = t_n + h_n$
$u_{n+1} = u_n + h_n\lambda(t_n)u_n$
$h_{n+1} = 2\eta/|\lambda(t_{n+1})|$
end for

Esempio 4. Consideriamo il problema

$$\begin{aligned} y'(t) &= -(10\mathtt{e}^{-t} + 1)y(t), \quad t \in (0, \infty) \\ y(0) &= 1 \end{aligned}$$

la cui soluzione analitica è $y(t) = \mathtt{e}^{t+10(\mathtt{e}^{-1}-1)}$. Ora essendo $|\lambda(t)| = (10\mathtt{e}^{-t} + 1)$ funzione decrescente, la stabilità assoluta ci verrà garantita prendendo il passo $h < h_0 = 2/|\lambda(0)| = 2/11 \approx 0.18$.

Regione di stabilità assoluta

Consideriamo ancora il problema modello (2.17) con $\lambda \in \mathbb{C}$, $\mathrm{Re}\lambda < 0$ la cui soluzione analitica è $y(t) = \mathtt{e}^{\lambda t}$ che tende a zero per $t \to +\infty$. Infatti $y(t) = \mathtt{e}^{\mathtt{Re}\lambda}\mathtt{e}^{\mathtt{Im}\lambda} = \mathtt{e}^{\mathtt{Re}\lambda t}(\sin(\mathtt{Im}\lambda t) + i\cos(\mathtt{Im}\lambda t))$.

Definizione 6. *La* `regione di stabilità assoluta` *di un metodo numerico è la regione del piano complesso*

$$\mathcal{R} := \{z : \ z = \lambda h\},$$

dove il metodo è A-stabile.

Vediamo quali sono le regioni di assoluta stabilià dei tre metodi finora studiati. Posto $z = \lambda h$, le regioni sono come segue.

EE. È l'insieme dei numeri complessi tali che

$$|1+z|<1\,, \tag{2.19}$$

ovvero il cerchio unitario centrato in $(-1,0)$.

EI. In questo caso, la regione è data dall'insieme dei numeri complessi che soddisfano la disuguaglianza

$$\left|\frac{1}{1-z}\right|<1\,, \tag{2.20}$$

ovvero $|1-z|>1$ che è il complementare del cerchio centrato in $(1,0)$.

CN. La regione si ottiene risolvendo la disequazione

$$\left|\frac{1+z/2}{1-z/2}\right|<1\,, \tag{2.21}$$

che si riduce solo alla disequazione $z<0$ cioè ai numeri complessi negativi, $\mathbb{C}_-$.

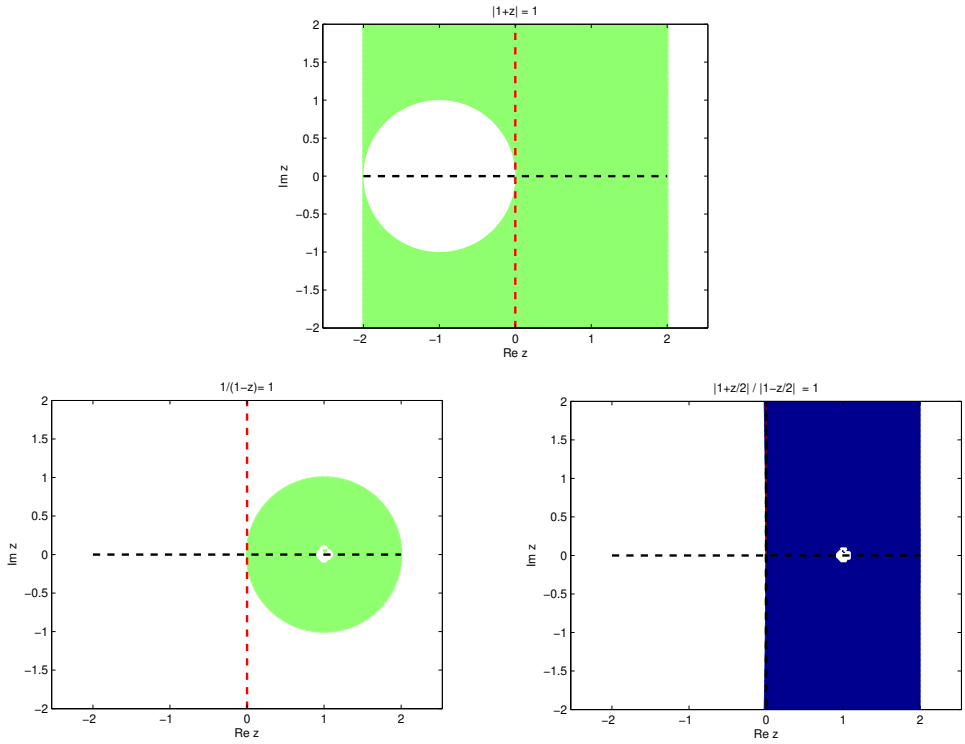

Figura 2.2: Regioni di stabilità dei metodi (dall'alto al basso, da sx a dx) EE, EI e CN disegnate come $f(z)=1$, con f una delle tre funzioni indicate nel membro di sx delle (2.19),(2.20) e (2.21). L'area bianca individua la regione di assoluta stabilità. Il "bollino" bianco indica il "centro" dell'area

Da quest'analisi segue un'altra proprietà dei metodi A-stabili: sono i metodi la cui regione di assoluta stabilità deve includere $\mathbb{C}_-$. Ecco un altro modo di vedere che EI e CN sono A-stabili.

2.1.6 Stabilità del θ-metodo

Applicando il metodo al problema $y' = f(x, y)$ si perviene alla formula discreta

$$u_{n+1} = u_n + h[\theta f_{n+1} + (1-\theta) f_n],\ n \geq 0,\ \theta \in [0,1]$$

il cui `errore di troncamento locale` nel punto x è

$$\tau_h(x) = \left(\frac{1}{2} - \theta\right) h y''(x) + \mathcal{O}(h^2)\,. \tag{2.22}$$

Infatti,

$$\begin{aligned} u_n^* - y_n &= y_{n-1} - y_n + h[\theta f_n + (1-\theta) f_{n-1}] \\ &= [y_{n-1} + h(1-\theta) f_{n-1}] - [y_n - h\theta f_n] \\ &= \frac{h^2}{2}(1-\theta) y''(\xi_1) - \frac{h^2}{2}\theta y''(\xi_2) + \mathcal{O}(h^2) \end{aligned}$$

dividendo per h e valutando in un punto x le derivate, si ottiene la relazione (2.22).

Se applicato al problema modello (2.17), porta all'espressione

$$u_{n+1} = \frac{1-(1-\theta)\lambda h}{1+\theta\lambda h} u_n \quad n \geq 0\,.$$

Posto

$$r(x) = \frac{1-(1-\theta)x}{1+\theta x},\quad x = \lambda h, \tag{2.23}$$

la soluzione discreta verificherá la relazione

$$u_n = r^n(x) = r^n(\lambda h)\,.$$

Ricordando che la soluzione di (2.17)è $y(x) = \mathsf{e}^{\lambda x}$ che è limitata per ogni $x \in [0, \infty)$, la soluzione discreta mimerà la soluzione analitica se $|r(\lambda h)| < 1$. In pratica si tratta di studiare la funzione $x \to r(x)$. Sappiamo che $r(0) = 1$, $\lim_{x\to+\infty} r(x) = 1 - 1/\theta$ e $r'(x) < 0$. Vediamo alcuni casi

- Quando $0 < \theta < 1/2$ il limite $1 - 1/\theta < 0$ per cui esiste un punto $\bar{x} = 2/(1-2\theta)$ (intersezione tra la retta $y = -1$ e $r(x)$) oltre il quale la funzione è in modulo maggiore di 1. Pertanto in questo caso, se $\lambda h \leq \bar{x}$ il metodo è stabile.
- Se $1/2 < \theta < 1$, il limite $1 - 1/\theta \geq -1$ e quindi lo schema risulta incondizionatamente stabile.
- $\theta = \{0, 1\}$ sono EE e EI, rispettivamente, di cui abbiamo giá parlato.

Osservazione. Per $\theta = 1/2$ si può avere instabilità per grandi valori $h\lambda$ che corrisponde alle equazioni dette `stiff` (ovvero `dure da risolvere`). In questi casi si suggerisce un compromesso tra $\theta = 1/2$ (maggiore ordine di precisione) e $\theta = 1$ (maggiore stabilità), prendendo ad esempio $\theta = 2/3$. Il θ-metodo che ne corrisponde, sarà del primo ordine ma con una costante d'errore inferiore e quindi da preferire allo stesso metodo CN.

$\diamond\diamond$

Concludiamo con un paio di interessanti esempi.

Esempio 5. Studiare la stabilità del seguente problema del secondo ordine

$$y'' - 4y' - 5y = 0, \;\; y(0) = 1, \; y'(0) = -1\,.$$

È facile provare che la soluzione analitica è $y(x) = \mathtt{e}^{-x}$. Supponiamo ora di perturbare il valore iniziale $y(0) = 1$ con $y(0) = 1 + \epsilon$. Rifacendo i calcoli, si trova la soluzione $y_\epsilon(x) = \left(1 + \frac{5}{6}\epsilon\right)\mathtt{e}^{-x} + \frac{\epsilon}{6}\mathtt{e}^{5x}$. Plottando la soluzione per diversi valori ϵ si osserva quanto sia malcondizonato il problema perchè molto sensibile anche a piccole perturbazioni (vedi Fig. 2.3).

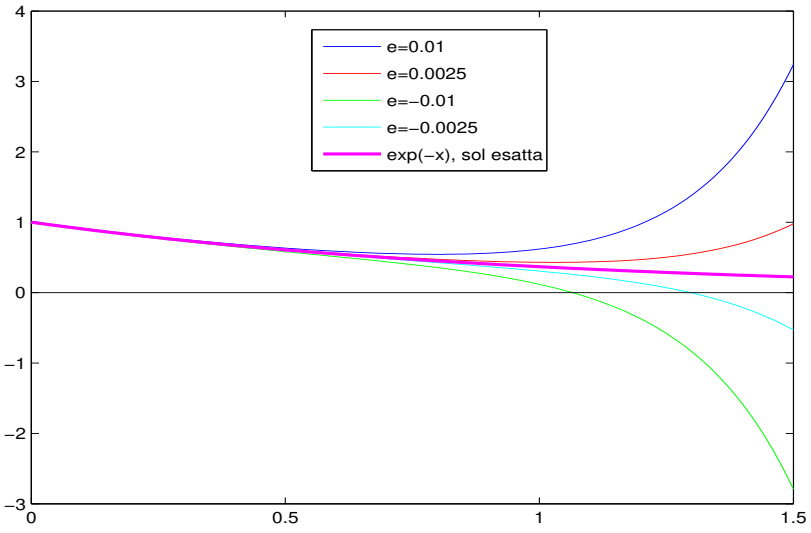

Figura 2.3: Instabilità della soluzione del problema dell' Esempio 5

Naturalmente un metodo numerico, se non opportunamente scelto, risulterà essere instabile.

Esempio 6. Consideriamo il problema

$$\begin{cases} y' = -2y + 1 \\ y(0) = 1 \end{cases} \tag{2.24}$$

la cui soluzione esatta è $y(x) = \frac{1}{2}\left(\mathtt{e}^{-2x} + 1\right)$. Discretizzando la derivata con `differenze finite centrali` otteniamo la successione

$$u_{n+1} + 4hu_n - u_{n-1} = 2h\,, \quad n \geq 1\,. \tag{2.25}$$

Si tratta di una equazione alle differenze la cui soluzione generale è

$$u_n = c_1\lambda_1^n + c_2\lambda_2^n + \frac{1}{2}\,,$$

con 1/2 soluzione particolare. Ora, l'equazione caratteristica associata alla (2.25) è $\lambda^2 + 4h\lambda - 1 = 0$ le cui soluzioni sono

$$\lambda_{1,2} = -2h \pm \sqrt{1+4h^2}\,.$$

Sviluppando in serie di Maclaurin le radici, arrestandoci al primo ordine, otteniamo le approssimazioni

$$\lambda_1 = 1 - 2h + \mathcal{O}(h^2)\,, \quad \lambda_2 = -(1+2h) + \mathcal{O}(h^2)\,,$$

cosicché

$$u_n = \frac{1}{2} + c_1\left(1 - 2h + \mathcal{O}(h^2)\right)^n + c_2(-1)^n\left(1 + 2h + \mathcal{O}(h^2)\right)^n\,.$$

Sia $x_n = nh$. Osservando che

$$\lim_{h\to 0}(1+2h)^n = \lim_{h\to 0}(1+2h)^{(1/2h)2x_n} = \mathtt{e}^{2x_n}\,,$$

e analogamente $\lim_{h\to 0}(1-2h)^n = \mathtt{e}^{-2x_n}$, la soluzione discreta per $h \to 0$ si comporterà come

$$u_n = \left(c_1\mathtt{e}^{-2x_n} + \frac{1}{2}\right) + c_2(-1)^n\mathtt{e}^{2x_n}\,. \tag{2.26}$$

Essendo $u_0 = 1$ e u_1 valore della soluzione esatta in $x_1 = h$ avremo $c_1 = 1/2$, che ci dice che il primo termine della (2.26) tende alla soluzione esatta. Il secondo termine è, per così dire `estraneo` e dovuto al fatto che abbiamo sostituito un'equazione del primo ordine con un'equazione alle differenze del secondo ordine. In altri termini, $(-1)^n\mathtt{e}^{2x_n}$ è un termine dovuto all'`errore di discretizzazione`.

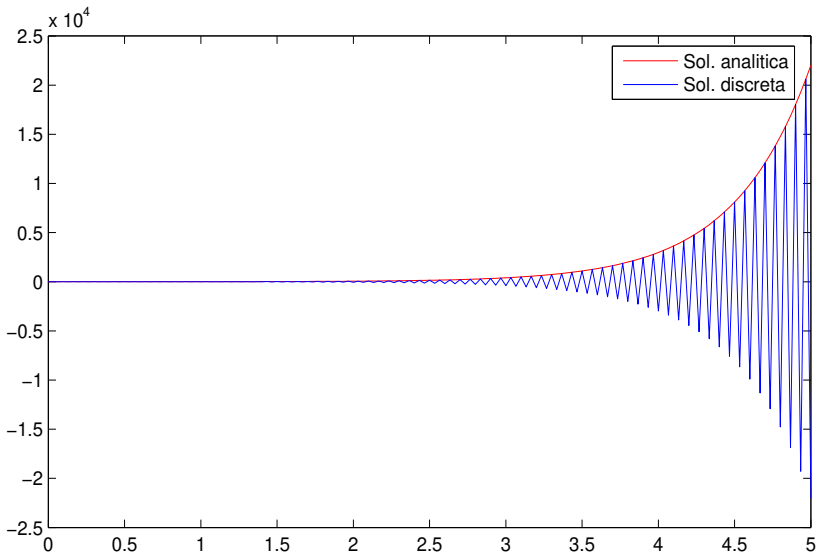

Figura 2.4: Instabilità della soluzione del problema dell' Esempio 6 con $c_2 = 1,\ h = 1/30$ e intervallo $[0, 5]$.

2.1.7 Metodi di Runge-Kutta

Si tratta di metodi per IVP che consentono di raggiungere un ordine di accuratezza maggiore.

Consideriamo ancora una volta il problema $y'(t) = f(t, y(t))$ con condizione iniziale $y(t_0) = y0$.

Nella sua formulazione generale, un `metodo di Runge-Kutta` si scrive con l'iterazione

$$u_{n+1} = u_n + h\, F(t_n, u_n, h; f)\,, \quad n \geq 0\,, \tag{2.27}$$

con (la funzione) incremento F tale che

$$F(t_n, u_n, h; f) = \sum_{i=1}^{s} b_i\, K_i\,,$$

$$K_i = f(t_n + c_i h, u_n + h \sum_{j=1}^{s} a_{i,j} K_j)\,; \quad i = 1, \ldots, s.$$

s viene detto il numero di `stadi` del metodo di R-K. I coefficienti $a_{i,j}$, b_i e c_i caratterizzano completamente il metodo. Di solito si raccolgono nella cosidetta **matrice di Butcher** (vedi Tabella 2.1).

c_1	$a_{1,1}$	$a_{1,2}$	$\cdots$	$a_{1,s}$
c_2	$a_{2,1}$	$a_{2,2}$	$\cdots$	$a_{2,s}$
$\vdots$			$\ddots$	
c_s	$a_{s,1}$	$a_{s,2}$	$\cdots$	$a_{s,s}$
	b_1	b_2	$\cdots$	b_s

$\mathbf{c}$	A
	$\mathbf{b}^T$

Tabella 2.1: Matrice di Butcher per un metodo di Runge-Kutta ad s stadi

con ovvio significato dei termini nella formulazione (compatta) matriciale. Assumeremo che

$$c_i = \sum_{j=1}^{s} a_{i,j}, \quad i = 1, ..., s\,. \tag{2.28}$$

- Se $a_{i,j} = 0,\ j \geq i,\ i = 1, \ldots, s$, allora i coefficienti K_i si determinano esplicitamente in funzione degli stadi predecenti $K_1, \ldots, K_{i-1}$. Son questi i metodi di RK `espliciti`.
- In caso contrario, lo schema è di tipo `implicito`. Pertanto per il calcolo di K_i richiederà la soluzione di un sistema non-lineare di dimensione s.
- Alternativamente, per ridurre il costo computazionale nel caso implicito, si usano schemi `semi-impliciti`

$$K_i = f(t_n + c_i h, u_n + h a_{i,i} K_i + h \sum_{j=1}^{i-1} a_{i,j} K_j)$$

che ad ogni passo richiede solo la soluzione di s equazioni non-lineari indipendenti.

Circa la `consistenza`, osservando che

$$h\tau_{n+1}(h) = y_{n+1} - y_n - hF(t_n, y_n, h; f)\,,$$

con $y(t)$, soluzione del problema di Cauchy continuo. Si dimostra che affinchè il metodo di R-K sia consistente (ovvero affinché $\lim_{h\to 0} \tau(h) = \lim_{h\to 0} \max_n |\tau_n(h)| = 0$) deve essere verificata la condizione

$$\sum_{i=1}^{s} b_i = 1\,. \tag{2.29}$$

Inoltre, il metodo si dirà consistente di ordine $p \geq 1$ rispetto ad h se l'errore di troncamento globale $\tau(h) = \mathcal{O}(h^p), \quad h \to 0$.

Infine, per quando riguarda la `convergenza` dei metodi di R-K, basta osservare che trattandosi di metodi ad un passo, la consistenza implica la stabilità. Infatti, avendo il polinomio caratteristico $p(r) = r - 1$ la cui unica radice soddisfa la condizione della definizione

ordine	1	2	3	4	4	5	6	6	7
s	1	2	3	4	5	6	7	8	9

Tabella 2.2: Numero di stadi e ordine dei metodi di Runge-Kutta fino a $s = 9$

4 di zero stabilità e quindi grazie al Teorema 2 possiamo concludere che `i metodi ad un passo consistenti sono convergenti`.

Un'ultima osservazione, che ritroveremo nella prossima sezione quando parleremo dei metodi multi-step. Se un metodo di R-K ha un errore di troncamento locale $\tau_n(h) = \mathcal{O}(h^p), \ \forall\, n$ allora anche l'ordine di convergenza sarà p. Per i metodi di R-K espliciti vale la seguente importante proprietà

Proposizione 1. *Un metodo di R-K esplicito a s stadi non puó avere ordine maggiore di s. Inoltre non esistono metodi di R-K espliciti a s stadi di ordine maggiore o uguale a* 5 *(vedi Tabella 2.2).*

Si noti come in Tabella 2.2 il numero massimo di stadi in corrispondenza al quale l'ordine non è inferiore al numero massimo di stadi stesso si ottiene per $s \leq 4$.

Come si deriva un metodo di Runge-Kutta?

Un metodo di R-K si costruisce chiedendo che nello sviluppo in serie di Taylor della soluzione esatta $y_{n+1} = y(t_{n+1})$, essa coincida col maggior numero di termini dello stesso sviluppo della soluzione approssimata u_{n+1}, supponendo di eseguire un **unico** passo del metodo di R-K a partire dalla soluzione esatta y_n.

Come esempio, cerchiamo di dedurre un `metodo esplicito di R-K a` $s = 2$ `stadi`

$$u_{n+1} = y_n + hF(t_n, y_n, h; f) = y_n + h(b_1K_1 + b_2K_2) \tag{2.30}$$

$$\texttt{con } K_1 = f(t_n, y_n) \quad \texttt{e } K_2 = f(t_n + h\,c_2,\ y_n + hc_2K_1) \quad (c_2 = a_{2,1})\,. \tag{2.31}$$

Sviluppiamo K_2 con Taylor in un intorno di t_n

$$K_2 = f_n + hc_2(f_{n,t} + K_1 f_{n,y}) + \mathcal{O}(h^2)\,, \tag{2.32}$$

dove $f_{n,z}$ indica la derivata parziale di f rispetto a z valutata in (t_n, y_n) (con $z = t \vee z = y$).

Sostituendo (2.32) in (2.30), si ottiene

$$\begin{aligned} u_{n+1} &= y_n + hb_1f_n + hb_2(f_n + hc_2(f_{n,t} + K_1f_{n,y})) + \mathcal{O}(h^3) \\ u_{n+1} &= y_n + h(b_1 + b_2)f_n + h^2c_2b_2(f_{n,t} + f_nf_{n,y}) + \mathcal{O}(h^3)\,. \end{aligned}$$

Sviluppando allo stesso modo la soluzione esatta

$$y_{n+1} = y_n + hy'_n + \frac{h^2}{2}y''_n + \mathcal{O}(h^3)\,.$$

Pertanto uguagliando gli sviluppi di y_{n+1} e u_{n+1} otterremo le condizioni

$$\begin{cases} b_1 + b_2 = 1 \\ c_2 b_2 = \frac{1}{2}\,. \end{cases} \tag{2.33}$$

L'equazione (2.33) ha *infinite soluzioni*. Vediamo due soluzioni particolari.

- Quando $b_1 = b_2 = \frac{1}{2}$ otteremo $c_2 = 1$. Da cui $K_1 = f_n$ e $K_2 = f(\underbrace{t_n + h}_{t_{n+1}}, u_n + hf_n)$. Il metodo a 2 stadi che si ottiene è noto col nome di **metodo di Heun**

$$u_{n+1} = u_n + \frac{h}{2}(f_n + f(t_{n+1}, u_n + hf_n))\ n \geq 0\,. \tag{2.34}$$

  Il metodo di Heun si ottiene anche a partire dal Crank-Nicolson sostituendo f_{n+1} con $f(t_{n+1}, u_n + hf_n)$, ovvero usando Eulero esplicito per u_{n+1} in f_{n+1}. Il metodo di Heun, che è esplicito, è interessante perchè ha trasformato un metodo implicito (il CN) in uno esplicito pur mantendo lo stesso ordine di convergenza. La matrice di Butcher del metodo di Heun è

0	0	0
1	1	0
	$\frac{1}{2}$	$\frac{1}{2}$

Tabella 2.3: Matrice di Butcher del metodo di Heun

- Quando $b_1 = 0,\ b_2 = 1$ otteremo $c_2 = \frac{1}{2}$. Il metodo corrispondente viene detto **metodo di Eulero modificato**:

$$u_{n+1} = u_n + hf(t_n + \frac{h}{2},\, u_n + \frac{h}{2}f_n),\ n \geq 0\,. \tag{2.35}$$

  La matrice di Butcher del metodo di Eulero modificato è

0	0	0
$\frac{1}{2}$	$\frac{1}{2}$	0
	0	1

Tabella 2.4: Matrice di Butcher del metodo di Eulero modificato

$\mathbf{c}$	A
	$\mathbf{b}^T$
	$\hat{\mathbf{b}}^T$
	$\mathbf{E}^T$

Tabella 2.6: Matrice di Butcher per due metodi di Runge-Kutta ad s stadi ed il relativo errore

Non possiamo concludere questa sezione, senza ricordare il più famoso dei metodi di R-K, ovvero il **metodo di Runge-Kutta di ordine 4**

$$\begin{aligned} K_1 &= f(t_n, u_n) \\ K_2 &= f(t_n + \tfrac{h}{2}, u_n + \tfrac{h}{2}K_1) \\ K_3 &= f(t_n + \tfrac{h}{2}, u_n + \tfrac{h}{2}K_2) \\ K_4 &= f(t_{n+1}, u_n + hK_3) \\ u_{n+1} &= u_n + \tfrac{h}{6}(K_1 + 2K_2 + 2K_3 + K_4)\,, \end{aligned} \tag{2.36}$$

con matrice di Butcher

0	0			
$\frac{1}{2}$	$\frac{1}{2}$			
$\frac{1}{2}$		$\frac{1}{2}$		
1	0	0	1	
	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$

Tabella 2.5: Matrice di Butcher del metodo di Runge-Kutta di ordine 4

Adattività del passo per metodi di R-K

Essendo metodi ad un passo, i metodi di R-K si adattano bene al cambio di passo d'integrazione per adattare la soluzione numerica all'andamento dell'errore. Ció richiede la conoscenza di un buon stimatore d'errore "a posteriori". Stimatori a posteriori, si realizzano in 2 modi:

1. usando lo stesso metodo di R-K con due passi differenti (es. h e $h/2$);
2. usando 2 metodi di R-K di ordini diversi ma con lo stesso numero s si stadi.

Il primo metodo implica in genere un costo computazionale elevato dovuto all'aumento delle valutazioni funzionali. Il secondo metodo richiede un po' d'attenzione. Supponiamo di avere 2 metodi a s stadi di ordini p e $p+1$ rispettivamente. Usando le matrici di Butcher possiamo scrivere l'errore come segue: dove gli arrays $\mathbf{c}$, A e $\mathbf{b}$ sono relativi al metodo di ordine p mentre $\mathbf{c}$, A e $\hat{\mathbf{b}}$ sono relativi al metodo di ordine $p+1$. Inoltre $\mathbf{E} = \hat{\mathbf{b}} - \mathbf{b}$ oppure

$\mathbf{E} = \mathbf{b} - \hat{\mathbf{b}}$. Infatti,

$$u_{n+1} = u_n + h \sum_{i=1}^{s} b_i K_i$$
$$\hat{u}_{n+1} = \hat{u}_n + h \sum_{i=1}^{s} \hat{b}_i K_i$$

Da cui, sottraendo membro a membro, otteniamo

$$u_{n+1} - \hat{u}_{n+1} = u_n - \hat{u}_n + h \sum_{i=1}^{s} E_i K_i$$

come indicato in Tabella 2.6. Quindi, usando questa tecnica, essendo i K_i sempre gli stessi, non servono ulteriori valutazioni di funzione, cosicché se l'errore si mantiene entro una certa soglia il passo viene mantenuto altrimenti si dimezza (calcolandone un nuovo valore). L'unico inconveniente di questo metodo è che ha la tendenza a sottostimare l'errore e non è pertanto del tutto affidabile quando h è "grande".

Il più usato di questi schemi di `Runge-Kutta-Fehlberg` è quello del quarto ordine formato da uno schema esplicito di R-K di ordine 4 accoppiato con uno di ordine 5. Tale schema è implementato nella funzione `ode45` di Matlab/Octave che si trova nel toolbox `funfun`. Altre funzioni che stimano l'errore in modo *adattivo*, presenti nello stesso toolbox, sono: `ode23` e `ode23tb`. Quest'ultima calcola K_1 usando la formula del trapezio mentre calcola K_2 usando una formula `backward` (che descriveremo nella prossima sezione) di ordine 2.

Regione di stabilità per metodi di R-K

Consideriamo il problema modello (2.17) al quale applichiamo uno schema ad s stadi. Si ottiene

$$K_i = \lambda \left(u_n + h \sum_{j=1}^{s} a_{i,j} K_j \right) \tag{2.37}$$
$$u_{n+1} = u_n + h \sum_{i=1}^{s} b_i K_i \,. \tag{2.38}$$

Indichiamo con $\mathbf{K} = (K_i, \ldots, K_s)^T$, $\mathbf{1} = (1, \ldots, 1)^T$. Allora le relazioni precedenti si riscrivono compattamente come

$$\mathbf{K} = \lambda(\mathbf{1} u_n + h\, A\, \mathbf{K})$$
$$u_{n+1} = u_n + h\, \mathbf{b}^T\, \mathbf{K}$$

da cui $\mathbf{K} = (I - h\lambda A)^{-1}\mathbf{1}\lambda\, u_n$. Pertanto

$$u_{n+1} = \left[1 + h\lambda\, \mathbf{b}^T(I - h\lambda A)^{-1}\mathbf{1}\right] u_n = R(h\lambda)u_n\,, \tag{2.39}$$

con A e $\mathbf{b}$ descritti nello schema di Butcher. Si pone allora la seguente definizione

Definizione 7. *Un metodo di R-K è* `assolutamente stabile` *quando la regione di stabilità* $|R(h\lambda)| < 1$. *In tal caso la successione* $\{u_n\}$ *è infinitesima.*

La regione di stabilità assoluta sarà

$$\mathcal{A} = \{z :\ z = h\lambda \in \mathbb{C},\ \texttt{t.c.}\ |R(h\lambda)| < 1\}\ .$$

Vediamo ora come sono fatte alcune regioni di stabilità assoluta per metodi di R-K espliciti. Se un metodo di R-K è esplicito, allora la matrice A è triangolare inferiore in senso stretto, pertanto è facile vedere che

$$R(h\lambda) = \frac{\texttt{det}(I - h\lambda A + h\lambda\mathbf{1}\mathbf{b}^T)}{\texttt{det}(I - h\lambda A)}\,. \tag{2.40}$$

Ma il denominatore vale 1, quindi la funzione $R(h\lambda)$ è un polinomio in $h\lambda$. Nel caso $s = 1, 2, 3, 4$ la funzione è

$$R(h\lambda) = \sum_{i=0}^{s} \frac{(h\lambda)^i}{i!}\,.$$

Il grafico delle regioni di assoluta stabilità dei metodi di R-K espliciti as $s = 1, 2, 3, 4$ stadi si trova in Figura 2.5

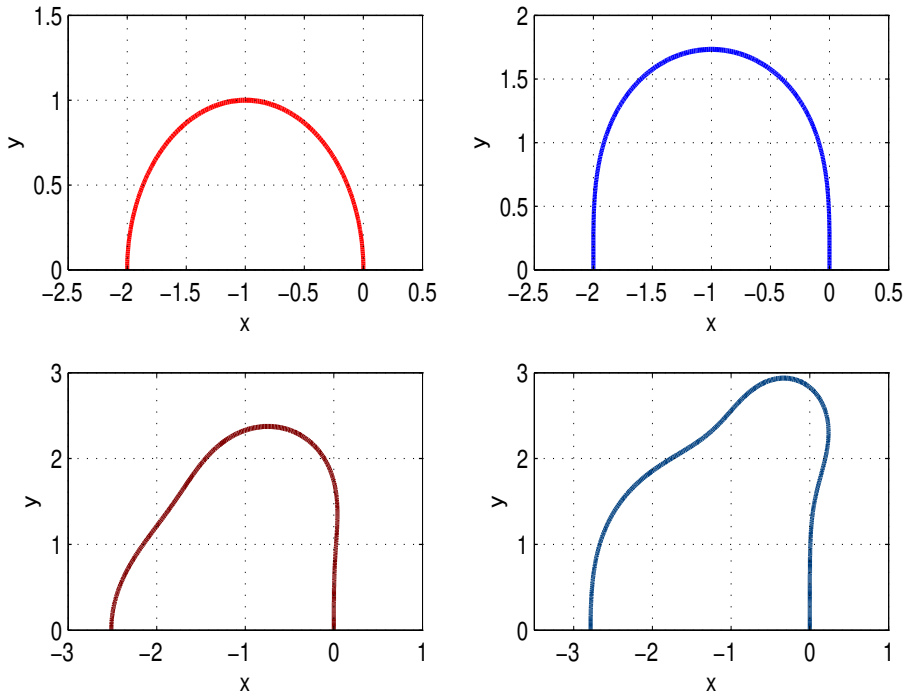

Figura 2.5: Regioni di stabilità assoluta dei metodi di R-K espliciti per $s = 1$ (alto sx), $s = 2$ (alto dx), $s = 3$ (basso sx) e $s = 4$ (basso dx).

Un semplice programma Matlab per il plot, nel semipiano $y > 0$ delle regioni di stabilità dei metodi di Runge-Kutta espliciti a $s \leq 4$ stadi, è il seguente

```
clear;
x=[-3:4/100:1];
y=[0:3/100:3];

f1=inline('abs(1+x+i*y)-1');
f2=inline('abs(1+x+i*y+(x+i*y).^2/2)-1');
f3=inline('abs(1+x+i*y+(x+i*y).^2/2+(x+i*y).^3/6)-1');
f4=inline('abs(1+x+i*y+(x+i*y).^2/2+(x+i*y).^3/6+(x+i*y).^4/24)-1');

subplot(2,2,1) ezplot(f1,[-3,1,0,3]) title('');

subplot(2,2,2) ezplot(f2,[-3,1,0,3]) title('');

subplot(2,2,3) ezplot(f3,[-3,1,0,3]) title('');

subplot(2,2,4) ezplot(f4,[-3,1,0,3]) title('');
```

2.2 Sistemi di equazioni differenziali

Se indichiamo con $Y(t) = (y_1(t), \ldots, y_m(t))^t$, con $Y'(t) = (y_1'(t), \ldots, y_m'(t))^t$ e con $F(t, Y(t))$ il vettore le cui componenti sono $f_i(t, y_1(t), \ldots, y_m(t)),\ i = 1, \ldots, m,\ t \in [t_0, T]$, allora un IVP per sistemi si scriverà compattamente come

$$\begin{cases} Y'(t) = F(t, Y(t)), \ \ t \in (t_0, T] \\ Y(t_0) = Y_0 \end{cases} \tag{2.41}$$

con ovvio significato dei simboli usati.

Perché è importante lo studio dei sistemi di equazioni differenziali? Una delle risposte è perchè consentono di risolvere equazioni differenziali di ordine > 1

$$y^{(m)}(t) = f(t, y(t), y'(t), \ldots, y^{(m-1)}(t))\ ,\ t \in (t_0, T]\,, \tag{2.42}$$

la cui soluzione, se esiste, è una famiglia di funzioni definite a meno di m costanti, che sono i cosidetti *gradi di libertà* del sistema. Se pertanto definiamo le costanti

$$y(t_0) = y_0,\ y'(t_0) = y_1,\ \ldots,\ y^{(m-1)}(t_0) = y_{m-1}\ ,$$

e poniamo

$$w_1(t) = y(t),\ w_2(t) = y'(t), \ldots,\ w_m(t) = y^{(m-1)}(t)$$

allora (2.42) si riscrive come il sistema seguente

$$\begin{cases} w_1' = w_2 \\ w_2' = w_3 \\ \vdots \\ w_m' = f(t, w_1, w_2, \ldots, w_m) \end{cases} \tag{2.43}$$

con le condizioni iniziali

$$\begin{cases} w_1(t_0) = y_0 \\ w_2(t_0) = y_1 \\ \vdots \\ w_m(t_0) = y_{m-1} \end{cases} \tag{2.44}$$

Vediamo alcuni dei metodi che abbiamo visto finora per IVP.

- θ `metodo`.

$$\begin{cases} U_{n+1} = \theta F(t_{n+1}, U_{n+1}) + (1-\theta) F(t_n, U_n), \ \ \theta \in [0,1],\ n \geq 0 \\ U_0 = Y_0 \end{cases} \tag{2.45}$$

- Metodo di R-K di ordine 4. Per semplicità ci poniamo nel caso $m = 2$ (ovvero un sistema 2×2).

$$\begin{cases} y' = F(t, y, v) \\ v' = G(t, y, v) \\ y(t_0) = \alpha, \;\; v(t_0) = \beta \, . \end{cases}$$

Il metodo di R-K diventa

$$\begin{array}{ll} K_1 = hF(t_n, u_n, w_n) & M_1 = hG(t_n, u_n, w_n) \\ K_2 = hF(t_n + \frac{h}{2}, u_n + \frac{K_1}{2}, w_n + \frac{M_1}{2}) & M_2 = hG(t_n + \frac{h}{2}, u_n + \frac{K_1}{2}, w_n + \frac{M_1}{2}) \\ K_3 = hF(t_n + \frac{h}{2}, u_n + \frac{K_2}{2}, w_n + \frac{M_2}{2}) & M_3 = hG(t_n + \frac{h}{2}, u_n + \frac{K_2}{2}, w_n + \frac{M_2}{2}) \\ K_4 = hF(t_{n+1}, u_n + K_3, w_n + M_3) & M_4 = hG(t_{n+1}, u_n + K_3, w_n + M_3) \\ u_{n+1} = u_n + \frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4) & w_{n+1} = w_n + \frac{1}{6}(M_1 + 2M_2 + 2M_3 + M_4) \, . \end{array} \tag{2.46}$$

2.2.1 Analisi di stabilità

In generale, dato il sistema $Y' = F(t, Y)$, $Y \in \mathbb{R}^n$, $F : \mathbb{R} \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$, non riusciamo a determinare esplicitamente la soluzione.

Nel caso in cui il sistema sia lineare, ovvero $Y' = A \cdot Y$ con $A \in \mathbb{R}^{n \times n}$, allora detti λ_j, $j = 1, \ldots, n$ gli n autovalori di A, che assumiamo distinti, sappiamo che possiamo diagonalizzare A e riscrivere il sistema come

$$\begin{aligned} Y' &= T \Lambda T^{-1} \, Y \\ T^{-1} Y' &= \Lambda \, T^{-1} Y \\ W' &= \Lambda \, W \end{aligned}$$

Dall'ultima equazione, si evince, che il sistema è ora equivalente a un sistema di n equazioni differenziali disaccoppiate, per cui alla fine

$$Y(t) = \sum_{i=1}^{n} c_j \mathrm{e}^{\lambda_j \, t} V_j$$

con c_j costanti e V_j che formano una base di autovettori di A. Posto, $\Lambda = -\max_t \rho(A)$ (il segno $-$ perchè per la convergenza alla soluzione analitica gli autovalori devono aveer parte reale negativa!), allora potremo applicare ad ogni equazione l'analisi di stabilità del caso scalare. Osserviamo che il sistema $Y' = A\,Y$ lo possiamo anche scrivere come $Y' = \Phi(t, Y(t))$ da cui $\dfrac{\partial \Phi(t)}{\partial Y} = A$. Pertanto il valore Λ è il naturale sostituto di λ usato nel caso scalare.

Circa l'**assoluta stabilità**, che valuta l'andamento di $Y(t)$, $t \to +\infty$, quello che chiederemo è che $\|Y(t)\| \to 0$. Questa condizione sarà di certo verificata se $\mathtt{Re}\lambda_j < 0$, $\forall\, j$. Infatti, essendo

$$\mathtt{e}^{\lambda_j t} = \mathtt{e}^{\mathtt{Re}\,\lambda_j\, t}(\cos(\mathtt{Im}\lambda_j) + i\sin(\mathtt{Im}\lambda_j)$$

avremo $\|Y(t)\| \to 0$.

2.3 Equazioni stiff

Se consideriamo il problema

$$\begin{cases} y'(t) = -100y(t), & t > 0 \\ y(0) = 1 \end{cases}$$

la condizione (**??**) per il metodo di Eulero impone $k < 1/50 = 0.02$. D'altra parte, la soluzione analitica del problema per $t^* = 0.4$ è minore di 10^{-17} (e dunque, trascurabile, nel senso che $y(0) - y(t^*) = y(0)$, in precisione doppia). Dunque, con poco più di 20 passi il metodo di Eulero arriva a calcolare adeguatamente la soluzione sino a t^*.

`stiff.m`

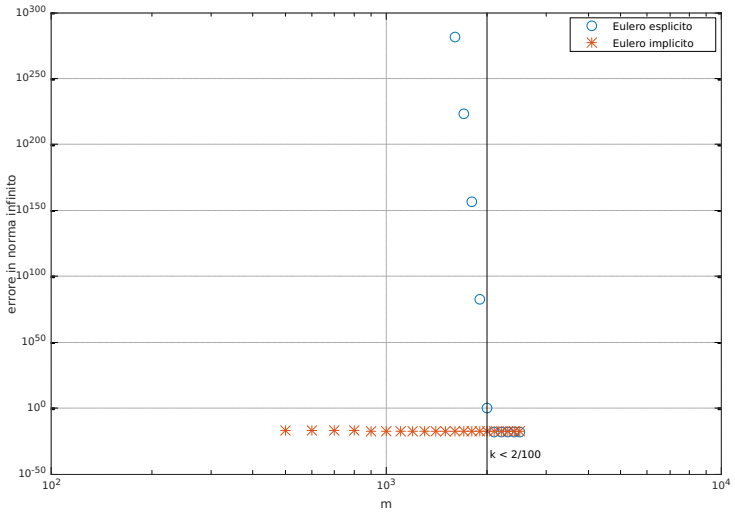

Figura 2.6: Eulero esplicito e Eulero implicito per la soluzione di (2.47) fino al tempo $t^* = 40$.

Qual è dunque il problema? Eccolo:

$$\begin{cases} \boldsymbol{y}'(t) = \begin{bmatrix} -100 & 0 \\ 0 & -1 \end{bmatrix} \boldsymbol{y}(t), & t > 0 \\ \boldsymbol{y}(0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \end{cases} \tag{2.47}$$

La soluzione analitica è

$$\boldsymbol{y}(t) = \begin{bmatrix} \mathrm{e}^{-100t} \\ \mathrm{e}^{-t} \end{bmatrix}$$

e la sua norma infinito è minore di 10^{-17} per $t^* = 40$. Poiché però per poter calcolare la prima componente serve un passo temporale $k < 0.02$, sono necessari più di 2000 passi (vedi Figura 2.6), anche se la prima componente diventa trascurabile dopo pochi passi e la seconda non richiederebbe un così elevato numero di passi. Dunque, anche se il metodo è convergente e il passo, per esempio, $k = 0.1$ garantisce un errore locale proporzionale a $k^2 = 0.01$, il metodo di Eulero non può essere usato con tale passo. Usando il metodo di Eulero implicito sarebbe possibile invece usare un passo piccolo all'inizio e poi, quando ormai la prima componente è trascurabile, si potrebbe incrementare il passo, senza pericolo di esplosione della soluzione. Per questo semplice problema, sarebbe possibile calcolare le due componenti separatamente. Nel caso generale, però, il sistema non è disaccoppiato. Per l'analisi, ci si può ricondurre, eventualmente in maniera approssimata, ad uno disaccoppiato e ragionare per componenti. Infatti, se A è una matrice diagonalizzabile,

$$\boldsymbol{y}'(t) = A\boldsymbol{y}(t) \Leftrightarrow \boldsymbol{z}'(t) = D\boldsymbol{z}(t) \Leftrightarrow \boldsymbol{z}(t) = \exp(tD)\boldsymbol{z}_0$$

ove $AV = VD$, $D = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_d\}$, e $\boldsymbol{y}(t) = V\boldsymbol{z}(t)$. Poi

$$\boldsymbol{y}'(t) = A\boldsymbol{y}(t) + \boldsymbol{b} \Leftrightarrow \boldsymbol{z}'(t) = D\boldsymbol{z}(t) + V^{-1}\boldsymbol{b} \Leftrightarrow$$
$$\Leftrightarrow \boldsymbol{z}(t) = \boldsymbol{z}_0 + t\varphi_1(tD)(D\boldsymbol{z}_0 + V^{-1}\boldsymbol{b})$$

ove

$$\varphi_1(\lambda) = \begin{cases} \dfrac{\mathrm{e}^{\lambda} - 1}{\lambda} & \text{se } \lambda \neq 0 \\ 1 & \text{se } \lambda = 0 \end{cases}$$

Infine (considerando un problema autonomo per semplicità e sviluppando in serie di Taylor)

$$\boldsymbol{y}'(t) = \boldsymbol{f}(\boldsymbol{y}(t)) \Leftrightarrow \boldsymbol{y}'(t) \approx \boldsymbol{f}(\boldsymbol{y}_n) + J(\boldsymbol{y}_n)(\boldsymbol{y}(t) - \boldsymbol{y}_n)$$

ove J è la matrice jacobiana

$$J(\boldsymbol{y}_n) = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{y}}(\boldsymbol{y}_n)$$

e, se J è diagonalizzabile, ci si riconduce al caso precedente. Dunque, si ha sempre a che fare con gli autovalori di J (nel caso J non sia diagonalizzabile, si ragiona in maniera equivalente con blocchi di Jordan) e il più piccolo di questi è quello che determina la restrizione massima sul passo temporale.

Definizione 1. *Un sistema di ODEs si dice* stiff *in un intorno di t_n se esiste almeno una coppia di autovalori λ_1, λ_2 della matrice jacobiana J_n tali che*

- $Re(\lambda_1) < 0$, $Re(\lambda_2) < 0$
- $Re(\lambda_1) \ll Re(\lambda_2)$

In pratica, può essere molto difficile capire se un sistema non lineare presenta regioni di *stiffness* o meno. Altrettanto difficile è rispondere alla domanda: per un problema stiff, conviene usare un metodo esplicito con passo piccolo o un metodo implicito? È chiaro che il metodo esplicito è di facile implementazione e applicazione, ma richiede molti passi temporali (e vedi **??**). Il metodo implicito richiede la soluzione ad ognuno dei "pochi" passi di un sistema, in generale, non lineare.

2.3.1 Risoluzione di un metodo implicito per problemi stiff

Consideriamo, per semplicità, il problema

$$\boldsymbol{y}'(t) = A\boldsymbol{y}(t)$$

con A stiff e *simmetrica*. La restrizione sul passo per il metodo di Eulero esplicito è

$$h < \frac{2}{\rho_{\max}}$$

ove $\rho_{\max}$ è il raggio spettrale di A. Applicando il metodo di Eulero implicito e le iterazioni di punto fisso per risolvere l'equazione (per assurdo, poiché l'equazione da risolvere è lineare), siccome

$$\|A\boldsymbol{x} - A\boldsymbol{y}\|_2 \leq \|A\|_2\|\boldsymbol{x} - \boldsymbol{y}\|_2 = \rho_{\max}\|\boldsymbol{x} - \boldsymbol{y}\|_2$$

si avrebbe la restrizione

$$h < \frac{1}{\rho_{\max}}$$

dunque una restrizione ancora più severa.

Da questo esempio si deduce che i metodi impliciti per problemi stiff vanno risolti con il metodo di Newton (eventualmente modificato).

Consideriamo ora il sistema

$$Y' = A\,Y + \Phi(t)\,,$$

con $A \in \mathbb{R}^{n\times n}$, $\Phi(t) \in \mathbb{R}^n$. In questo caso, la soluzione si scrive come

$$Y(t) = \sum_{i=1}^{n} c_j \mathtt{e}^{\lambda_j\, t} V_j + \Psi(t)$$

dove $\Psi(t)$ indica la soluzione particolare. Assumendo ancora che $\mathtt{Re}\lambda_j < 0,\ \forall\, j$, allora $Y(t) \to \Psi(t),\ t \to +\infty$. Possiamo allora interpretare

- $\Psi(t)$ come `soluzione allo stato stazionario` (ottenibile per tempi grandi);
- $Y(t)$ come `soluzione allo stato transitorio` (ovvero per tempi finiti).

Analizziamo cosa succede nello stato stazionario usando uno schema numerico con regione di assoluta stabilità limitata. Per un tale schema, il passo h sarà limitato da quantità che dipendono da $M = \max_{1\leq j\leq n}\{|\lambda_j|\}$. Tanto più grande è M tanto minore sarà l'intervallo di tempo in cui la corrispondente componente della soluzione darà un contributo significativo alla soluzione del problema. Pertanto, dovremo usare un passo piccolo per descrivere una componente del problema che si annulla per grandi valori di t: assurdo!!

Pertanto, per caratterizzare la difficoltà risolutiva di un sistema differenziale, se supponiamo

$$\sigma \leq \mathtt{Re}\lambda_j \leq \tau < 0,\ \forall\, j$$

e definiamo il *quoziente di stiffness* la quantità

$$r_s = \frac{\sigma}{\tau}\,, \tag{2.48}$$

potremo dare la seguente definizione.

Definizione 8. *Un sistema di ODE lineare a coefficienti costanti si dice* **stiff** *se* $\mathtt{Re}\lambda_j < 0$ *e* $r_s \gg 1$.

Considerare lo spettro di A per caratterizzare lo *stiffness* porta a due osservazioni:

- se $\tau \approx 0$ il problema è certamente *stiff* solo quando $|\sigma|$ è molto grande;
- la scelta di opportune condizioni iniziali può influire sullo *stiffness* del problema.

Dalla seconda osservazione qui sopra, segue che una definizione esatta e rigorosa di *stiffness* non è possibile. Ad esempio, in [12, p. 220] si definisce *stiff* un sistema di ODE quando approssimato con un metodo numerico con regione di assoluta stabilità limitata, per ogni condizione iniziale per cui il problema di Cauchy ha soluzione, il metodo usa un passo di discretizzazione troppo piccolo rispetto a quello necessario per descrivere la soluzione esatta.

Esempio 7. Un interessante esempio di sistema differenziale *stiff* 2×2 è il sistema di Van der Pol

$$\begin{cases} y_1' = y_2 \\ \\ y_2' = 1000(1-y_1^2)y_2 - y_1 \end{cases} \tag{2.49}$$

con l'aggiunta delle condizioni iniziali $y_1(0) = 2$ e $y_2(0) = 0$. In Matlab/Octave per implementare il sistema di Van der Pol, possiamo scrivere la funzione

```
function yp=vanderp(t,y,a)
yp=[y(2); a*(1-y(1)^2)*y(2)-y(1)];
return
```

dove a gioca il ruolo di parametro di *stiffness*, maggiore sarà il suo valore e più difficile da risolvere sarà il sistema (nella formulazione originale di Van der Pol, $a = 1000$, cfr. (2.49)). Inoltre, usando la funzione Matlab/Octave `ode45` sull'intervallo $(0, 3000)$, il metodo richiede un passo h troppo piccolo (dell'ordine di $1.e-6$) rendendone poi l'esecuzione praticamente infinita. Invece l'uso delle funzioni Matlab appropriate per sistemi *stiff*, `ode15s, ode23s, ode23tb` o in Ocatave `ode5r` consentirà di risolvere il sistema in tempi accettabbili mediante l'uso di passi ragionevolmente piccoli (relativamente alla scelta del parametro a). Il programma seguente per Matlab consente di fare questo raffronto

```
clear
% provo la soluzione del sistema di Van der Pol
% con varie tecniche

t0=clock;
[t,u]=ode15s(@vanderp,[0 3000],[2 0],[],1000);
%ODE15S Solve stiff differential equations and DAEs
disp('Tempo richiesto per risolvere il sistema con ode15s')
etime(clock,t0)
disp('Min passo usato= ')
min(diff(t))

plot(t,u(:,1));

pause
close all
%----------------------------------------
t0=clock;
[t,u]=ode23s(@vanderp,[0 3000],[2 0],[],1000);
%ODE23S  Solve stiff differential equations, low order method.
disp('Tempo richiesto per risolvere il sistema con ode23s')
etime(clock,t0)
disp('Min passo usato= ')

min(diff(t))
plot(t,u(:,1));

pause
close all
%----------------------------------------
t0=clock;
[t,u]=ode23tb(@vanderp,[0 3000],[2 0],[],1000);
%ODE23TB Solve stiff differential equations, low order method.
disp('Tempo richiesto per risolvere il sistema con ode23tb')
etime(clock,t0)
```

```
disp('Min passo usato= ')
min(diff(t))
plot(t,u(:,1));
```

Sistemi non-lineari

Consideriamo

$$Y'(t) = F(t, Y(t))\,, \tag{2.50}$$

dove $F : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ è derivabile. Per studiare la stabilità, in questo caso, una possibilità è di linearizzare in un intorno del punto $(\tau, Y(\tau))$, ottenendo

$$Y'(t) = F(\tau, Y(\tau)) + J_F(\tau, Y(\tau))[Y(t) - Y(\tau)]\,.$$

Facciamo osservare, su un esempio, che il problema qui è che gli autovalori della matrice jacobiana J_F non sono sufficienti a descrivere il comportamento della soluzione esatta.

Esempio 8. Consideriamo il sistema

$$Y'(t) = \begin{bmatrix} -\frac{1}{2t} & \frac{2}{t^3} \\ -\frac{t}{2} & -\frac{1}{2t} \end{bmatrix} Y(t) = A(t)Y(t)\,.$$

In questo caso,

$$Y(t) = c_1 \begin{bmatrix} t^{-3/2} \\ -\frac{1}{t}\sqrt{t} \end{bmatrix} + c_2 \begin{bmatrix} 2t^{-3/2}\log t \\ (1-\log t)\sqrt{t} \end{bmatrix}$$

che per $t > \sqrt[4]{12} \approx 1.86$ ha norma euclidea che diverge quando $c_1 = 1,\ c_2 = 0$. Gli autovalori di $A(t)$ sono $\frac{-1\pm 2i}{2t}$, ovvero hanno parte reale negativa.

Come osservazione conclusiva, facciamo notare come il caso non-lineare va studiato con metodi *ad-hoc* riformulando l'idea stessa di stabilità (cfr. [12]).

Capitolo 3

Metodi multi-step

3.1 Metodi multi-step

I metodi a più passi, che in inglese si dicono `multi-step`, consentono di ottenere un ordine di accuratezza elevato nella determinazione di u_{n+1} in due modi fondamentali:

1. coinvolgendo valori precedenti della soluzione discreta, $u_{n-k},\ k = 0, ..., p,\ p \geq 1$: sono questi i metodi espliciti di `Adams-Bashforth` e impliciti `Adams-Moulton`.

2. approssimando l'equazione differenziale al tempo t_{n+1}, $y'(t_{n+1})$, con differenze finite all'indietro: sono i cosidetti metodi `Backward` o `BFD`.

In pratica sono metodi *con memoria*, nel senso che usano informazioni giá acquisite per approssimare la soluzione numerica dell' equazione differenziale.

Vediamo in dettaglio i due approcci.

3.1.1 Metodi di Adams-Bashforth e Adams-Moulton

Consideriamo

$$y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} f(t, y(t))dt\,. \tag{3.1}$$

Se approssimiamo l'integrale con una formula di quadratura su un opportuno insieme di nodi, otteniamo sia metodi espliciti che impliciti.

Metodi espliciti di Adams-Bashforth

Come appena detto, consideriamo l'uguaglianza (3.1) e approssiamo l'integrale con una formula di quadratura di tipo interpolatorio su nodi equispaziati $t_{n-k}, \ldots, t_n$, cosicché

$$f(t, y(t)) = y'(t) \approx \sum_{i=n-k}^{k} \ell_i^{(k)}(t) f_i = \sum_{i=1}^{k+1} \ell_i^{(k)}(t) f_i \,, \tag{3.2}$$

dove $\ell_i^{(k)}$ indica il polinomio elementare di Lagrange di grado k, che nel caso di nodi equispaziati (di passo h) possiamo scrivere come

$$\ell_i^{(k)}(s) = \prod_{j=1, j\neq i}^{k+1} \frac{(s-j)}{i-j}$$

dove $s(l) = t_1 + (l-1)h,\ l = 1, \ldots, k+1$. Pertanto in $[t_n, t_{n+1}]$ avremo

$$\int_{t_n}^{t_{n+1}} \ell_i^{(k)} dt = h \int_0^1 \prod_{j=1, j\neq i}^{k+1} \frac{s-j}{i-j} ds \,,$$

da cui si ottiene la formula generale

$$u_{n+1} = u_n + h \sum_{i=1}^{k+1} \beta_i \, f_{n-i+1} \,, \tag{3.3}$$

con i β_i che assumo i valori riportati in Tabella 3.1. Una delle formule più usate è la `AB3`,

k	β_1	β_2	β_3	β_4	β_5	
0	1					Eulero Esplicito
1	$\frac{3}{2}$	$\frac{-1}{2}$				
2	$\frac{23}{12}$	$-\frac{16}{12}$	$\frac{5}{12}$			AB3
3	$\frac{55}{24}$	$-\frac{59}{24}$	$\frac{37}{24}$	$-\frac{9}{24}$		
4	$\frac{1901}{720}$	$-\frac{2774}{720}$	$\frac{2616}{720}$	$-\frac{1274}{720}$	$\frac{251}{720}$	

Tabella 3.1: Coefficienti delle formule di Adams-Bashforth fino all'ordine 4

ovvero la formula esplicita di Adams-Bashforth di ordine 3

$$u_{n+1} = u_n + \frac{h}{12}(23 f_n - 16 f_{n-1} + 5 f_{n-2}) \tag{3.4}$$

che si ottiene approssimando l'integrale con un polinomio di grado 2 sui nodi t_{n-2}, t_{n-1} e t_n.

Metodi impliciti di Adams-Moulton

Un esempio a tre passi del quarto ordine è l'`AM4`

$$u_{n+1} = u_n + \frac{h}{24}(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2})\,, \tag{3.5}$$

che si ottiene approssimando la funzione $f(t, y(t))$ in (3.1) con il polinomio d'interpolazione di grado 3 costruito sulle coppie $(t_{n+1}, f_{n+1}), \ldots, (t_{n-2}, f_{n-2})$. Un altro esempio è il `metodo di Simpson`

$$u_{n+1} = u_{n-1} + \frac{h}{3}(f_{n+1} + 4f_n + f_{n-1}),\ n \geq 1\,. \tag{3.6}$$

3.1.2 Metodi BDF

Si ottengono approssimando la y' al tempo t_{n+1} con `formule alle differenze finite all'indietro` di ordine elevato. Una generica Backward Differentiation Formula (BDF) puó scriversi

$$\sum_{k=0}^{s} \alpha_k u_{n+1-k} = h\,\beta\, f_{n+1}\,, \tag{3.7}$$

dove h é il passo d'integrazione e i coefficienti α_i e β sono scelti in modo da ottenere l'ordine s (massimo possibile).

nome	formula	ordine
BDF1	$u_{n+1} - u_n = hf_{n+1}$	1 (*)
BDF2	$u_{n+1} - \frac{4}{3}u_n + \frac{1}{3}u_{n-1} = \frac{2}{3}hf_{n+1}$	2
BDF3	$u_{n+1} - \frac{18}{11}u_n + \frac{9}{11}u_{n-1} - \frac{2}{11}u_{n-2} = \frac{6}{11}hf_{n+1}$	3

Tabella 3.2: I metodi BDF[1,2,3]

(*) il BDF1 corrisponde al metodo di Eulero implicito che é A-stabile e si applica anche a problemi *stiff*. Si vedrá piú oltre, che per $s \geq 7$ i metodi BDF non sono zero-satbili e quindi sono inutilizzabili.

3.2 Consistenza e zero-stabilità dei metodi multistep

Per comprendere al meglio come stabilire se un metodo multistep è consistente vediamo dapprima la sua formulazione generale

$$u_{n+1} = \sum_{j=0}^{p-1} a_j u_{n-j} + h\sum_{j=0}^{p-1} b_j f_{n-j} + hb_{-1}f_{n+1}\,, \tag{3.8}$$

dove p indica il numero di passi. I coefficienti a_j e b_j individuano il metodo multistep. L'unica assunzione che faremo è che

$$a_{p-1} \neq 0 \ \vee \ b_{p-1} \neq 0 \, .$$

Inoltre, se $b_{-1} \neq 0$ il metodo è implicito altrimenti sarà esplicito. Osserviamo che per innescare un metodo a p passi, servono p condizioni iniziali $u_0, \ldots, u_{n-p+1}$ che potremo determinare ad esempio con un metodo di R-K.

Associato al metodo multistep (3.8) vi è il `primo polinomio caratteristico`

$$\pi(r) = r^{p+1} - \sum_{j=0}^{p} a_j r^{p-j} \, , \tag{3.9}$$

e il `secondo polinomio caratteristico`

$$\sigma(r) = b_{-1} r^{p+1} + \sum_{j=0}^{p} b_j r^{p-j} \, . \tag{3.10}$$

Definiamo quindi il `polinomio caratteristico` associato al metodo multistep (3.8) il polinomio

$$\rho(r) = \pi(r) - h\lambda\sigma(r) \, , \tag{3.11}$$

Le radici r_j (che dipendono dal passo h) del polinomio $\rho(r)$ si chiamano `radici caratteristiche` del metodo. Se esse sono tali che $|r_j(h)| < 1$, allora il metodo si dice `consistente`. Vale il seguente *criterio delle radici*

Proposizione 2. *Ai fini della consistenza del metodo multistep (3.8), le seguenti condizioni sono equivalenti.*

1. $\displaystyle\sum_{j=0}^{p} a_j = 1 - \sum_{j=0}^{p} j\, a_j + \sum_{j=-1}^{p} b_j = 1 \, .$
2. *$r = 1$, è radice del primo polinomio caratteristico $\pi(r)$.*

Vale il seguente risultato

Teorema 3. *Per un metodo multistep, la condizione sulle radici è equivalente alla* `zero-stabilità`.

Ad esempio, nei metodi di Adams,

$$\pi(r) = r^{p+1} - r^p = r^p(r-1)$$

le cui radici sono $r_1 = 0$ (multipla di ordine p) e $r_2 = 1$. Pertanto i metodi di Adams sono zero-stabili.

Vale anche il seguente risultato

Teorema 4. *Un metodo multistep consistente è convergente* `se e solo se` *è zero-stabile e l'errore sul dato iniziale tende a zero per* $h \to 0$.

Purtroppo vale un risultato, similmente a metodi di Runge-Kutta, che è noto con il nome di `prima barriera di Dalquist`.

Proposizione 3. *Non esistono metodi multistep lineari zero-stabili a* `p` *passi con ordine di convergenza maggiore di* `p+1` *se* `p` *dispari e di ordine* `p+2` *quando* `p` *è pari.*

3.2.1 Assoluta stabilità dei metodi multistep

Iniziamo con una definzione.

Definizione 9. *Un metodo multistep soddisfa la condizione di assoluta stabilità sulle radici se esiste un* $h_0 > 0$ *tale che*

$$|r_j(h\lambda)| < 1, \quad j = 1, \ldots, p, \ \ \forall\, h \le h_0\,.$$

Questa condizione è necessaria e sufficiente affinchè ci sia assoluta stabilità. Vale anche in questo caso una `barriera`, detta `seconda barriera di Dalquist`.

Proposizione 4. *Un metodo multistep lineare esplicito non può essere A-stabile. Non esistono multistep lineari epliciti A-stabili di ordine* > 2.

Vediamo ora la consistenza dei metodi multistep che abbiamo visto nella sezione precedente.

- `AB3` e `AM4` sono consistenti e zero stabili. Infatti

  $$\pi(r) = r^3 - r^2 = r^2(r-1)\,.$$

  Inoltre, `AB3` è assolutamente stabile se $h < 0.54/|\lambda|$, mentre `AM4` è assolutamente stabile se $h < 3/|\lambda|$

- `BDF3` ha primo polinomio caratteristico uguale ad

  $$\pi(r) = r^3 - \frac{18}{11}r^2 + \frac{9}{11}r - \frac{2}{11}\,,$$

  le cui radici sono $r_1 = 1$ e $r_{2,3} = 0.32 \pm 0.29i$ e tutt'e tre sono in modulo ≤ 1. Questo metodo è anche `incondizionatamente assolutamente` stabile per ogni $\lambda \in \mathbb{R}_-$ ma non per $\lambda \in \mathbb{C}$, `Re`$\lambda < 0$. Ovvero `BDF3` non è A-stabile.

3.2.2 Metodi predizione-correzione

Osserviamo che nei metodi di tipo implicito, ad ogni passo dobbiamo risolvere un'equazione algebrica per determinare u_{n+1} (cosa che in Matlab/Octave potremo fare ad esempio con `fzero`). Alternativamente si possono fare delle iterazioni di punto fisso. Ad esempio, usando il metodo di Crank-Nicolson

$$u_{n+1}^{(k)} = u_n + \frac{h}{2}\left[f_n + f(t_{n+1}, u_{n+1}^{(k)})\right] .$$

Si può dimostrare che se $u_{n+1}^{(0)}$ è scelto opportunamente, basta un'unica iterazione per ottenere $u_{n+1}^{(1)}$ la cui accuratezza è dello stesso ordine di u_{n+1} calcolata con il metodo implicito.

Ecco allora la "ricetta" del metodo. Supponiamo che il metodo abbia ordine $p \geq 2$.

1. determino il valore iniziale $u_{n+1}^{(0)}$ con un metodo esplicito di ordine almeno $p-1$ (passo di `predizione`);
2. useremo un metodo implicito per raffinare (passo di `correzione`).

Due esempi di metodi predizione-correzione sono: il metodo di Heun

$$\begin{aligned} u_{n+1}^{*} &= u_n + hf_n \\ u_{n+1} &= u_n + \frac{h}{2}\left[f_n + f(t_{n+1}, u_{n+1}^{*})\right] \end{aligned}$$

oppure la coppia (AB3)+(AM4).

L'ordine di accuratezza è quello del metodo di correzione. Essendo poi metodi espliciti, la stabilità dipende dalla scelta del passo h e quindi non sono adeguati per la soluzione di problemi di Cauchy su intervalli illimitati.

In Matlab/Octave, la funzione `ode113` del toolbox `funfun`, implementa un metodo di Adams/Bashforth con passo h variable.

Capitolo 4

Metodi numerici per problemi con valori al bordo

4.1 Problemi con valori al bordo

Consideriamo il problema del secondo ordine

$$\begin{cases} y'' = f(x, y, y')\,, & a < x < b \\ y(a) = \alpha\,, & y(b) = \beta\,. \end{cases} \tag{4.1}$$

Consideriamo poi punti equispaziati $x_n = a + nh,\ \ n = 0, \ldots, N,\ x_0 = a,\ x_N = b$. Pertanto il nostro problema consiste nel determinare la soluzione numerica nei punti `interni` $x_1, \ldots, x_{N-1}$.

Come approssimazioni di y' e y'', prendiamo le differenze finite centrali

$$\begin{aligned} y'(x_n) &= \frac{1}{2h}\left[y(x_{n+1}) - y(x_{n-1})\right] + \mathcal{O}(h^2) \\ y''(x_n) &= \frac{1}{h^2}\left[y(x_{n+1}) - 2y(x_n) + y(x_{n-1})\right] + \mathcal{O}(h^2) \end{aligned}$$

Sostituendo in (4.1) otteniamo

$$\begin{cases} y_{n-1} - 2y_n + y_{n+1} = h^2 f(x_n, y_n, \frac{y_{n+1} - y_{n-1}}{2h}),\ \ n = 1, \ldots, N-2 \\ y_0 = \alpha\,,\ \ y_N = \beta\,. \end{cases} \tag{4.2}$$

Facciamo alcune osservazioni.

1. Il sistema (4.2), è in generale non lineare e pertanto per la soluzione dovremo applicare metodi iterativi per sistemi, quali il metodo di Newton o delle secanti (cfr. [7, Cap. 3]).

2. Ogni equazione in (4.2), coinvolge 3 incognite, $y_{n-1},\ y_n,\ y_{n+1}$ per cui lo Jacobiano del sistema è tridiagonale.

3. Le condizioni al bordo in (4.2) si chiamano *condizioni di Dirichlet* e coinvolgono solo la funzione incognita y. Condizioni al bordo che coinvolgono invece i valori di y' sono note come *condizioni di Neumann*. Esistono anche delle condizioni che coinvolgono sia valori di y che y' e sono dette *condizioni di tipo misto* (vedasi oltre).

Se $f(x, y, y')$ è lineare, come ad esempio

$$y'' - p(x)y' - q(x)y = r(x)$$

allora il sistema (4.2) è lineare e tridiagonale della forma

$$\begin{cases} \left(1 + p_n \frac{h}{2}\right) y_{n-1} - (2 + q_n h^2) y_n + \left(1 - p_n \frac{h}{2}\right) y_{n+1} = h^2 r_n, \ \ n = 1, \ldots, N-1 \\ \\ y_0 = \alpha\,, \ \ y_N = \beta\,. \end{cases} \tag{4.3}$$

Per questo esempio, vale anche la seguente Proposizione.

Proposizione 5. *Se $p(x)$, $q(x)$, $r(x) \in \mathcal{C}[a,b]$, $q(x) \geq 0, \forall x \in [a,b]$ e*

$$h < \frac{2}{\|q\|_\infty}$$

allora il sistema (4.2) è diagonalmente dominante in senso stretto e quindi non singolare. Se inoltre $y(x) \in \mathcal{C}^4[a,b]$, allora lo schema alle differenze finite centrali è convergente per $h \to 0$ e si ha

$$\max_{1 \leq n \leq N-1} |y(x_n) - y_n| = \mathcal{O}(h^2)\,.$$

Esempio 9. Prendiamo l'equazione

$$\begin{cases} y''(x) - y(x) = x, \ \ x \in (0,1) \\ y(0) = 0\,, \ \ y(1) = 0\,. \end{cases} \tag{4.4}$$

la cui soluzione analitica è $y(x) = \frac{e}{e^2-1}(e^x - e^{-x}) - x$. Vogliamo risolverla con differenze finite centrali con passo $h = 0.2$. Il sistema lineare corrispondente (dopo aver opportunamente sostituito i valori)

$$\begin{pmatrix} -2.4 & 1 & 0 & 0 \\ 1 & -2.4 & 1 & 0 \\ 0 & 1 & -2.4 & 1 \\ 0 & 0 & 1 & -2.4 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 0.8 \\ 1.6 \\ 2.4 \\ 3.2 \end{pmatrix}\,,$$

risulta diagonalmente dominante in senso stretto e quindi non singolare.

Si verifichi, che per $h \to 0$, l'errore relativo decresce come $\mathcal{O}(h^2)$. Per questa verifica, basterà calcolare $e_n = |y(x_n) - y_n| / |y(x_n)|$.

$\diamond\diamond$

Se le condizioni al contorno sono di `tipo misto`, ovvero del tipo

$$\begin{cases} y'(a) + \gamma y(a) = 0 \\ y'(b) + \delta y(b) = 0 \end{cases} \tag{4.5}$$

per mantenere l'ordine di convergenza $\mathcal{O}(h^2)$, dovremo approssimare le derivate con formule alle differenze centrali

$$\begin{cases} y'(a) = \frac{1}{2h}(y(x_1) - y(x_{-1}) + \mathcal{O}(h^2) \\ \\ y'(b) = \frac{1}{2h}(y(x_{N+1}) - y(x_{N-1}) + \mathcal{O}(h^2) \end{cases}$$

dove x_{-1} e x_{N+1} sono nodi *ad hoc*. Il sistema a cui si perviene è

$$\begin{cases} y_{n-1} - 2y_n + y_{n+1} = h^2 f(x_n, y_n, \frac{y_{n+1}-y_{n-1}}{2h}), \quad n = 1, \ldots, N-1 \\ \\ y_{-1} = y_1 + 2h\gamma y_0 \\ \\ y_{N+1} = y_{N-1} - 2h\delta y_N \,. \end{cases} \tag{4.6}$$

Consideriamo infine un sistema differenziale lineare di ordine m

$$\begin{cases} Y'(x) = A(x)Y(x) + R(x)\,, \quad a < x < b \\ B_a Y(a) + B_b Y(b) = \alpha\,. \end{cases} \tag{4.7}$$

Presi i punti $x_0 = a < x_1 < x_2 < \ldots < x_N = b$, usando la formula dei trapezi

$$\begin{cases} Y_{n+1} = Y_n + \frac{h_n}{2}\left[A(x_{n+1})Y_{n+1} + A(x_n)Y_n + R(x_{n+1}) + R(x_n)\right]\,, \quad n = 0, 1, \ldots, N-1 \\ \\ B_a Y_0 + B_b Y_N = \alpha\,. \end{cases}$$

si perviene al sistema a blocchi

$$\begin{pmatrix} S_1 & R_1 & & & \\ & S_2 & R_2 & & \\ & & \ddots & & \\ & & & S_N & R_N \\ B_a & & & & B_b \end{pmatrix} \begin{pmatrix} Y_0 \\ Y_1 \\ \\ Y_{N-1} \\ Y_N \end{pmatrix} = \begin{pmatrix} Q_0 \\ Q_1 \\ \vdots \\ Q_{N-1} \\ \alpha \end{pmatrix}\,,$$

dove

$$\begin{aligned} S_n &= -\frac{1}{h_n} I - \frac{1}{2} A(x_n)\,, \\ R_n &= \frac{1}{h_n} I - \frac{1}{2} A(x_{n+1})\,, \\ Q_n &= \frac{R(x_{n+1}) + R(x_n)}{2}\,, \end{aligned}$$

con I ed A matrici di ordine $m \times m$.

4.1.1 Metodo di collocazione

Consideriamo ancora il problema del secondo ordine con condizioni al bordo che riscriviamo come segue

$$\begin{cases} y'' = f(x, y, y'), & a < x < b \\ g_1(y(a), y(b)) = 0, \\ g_2(y(a), y(b)) = 0. \end{cases} \tag{4.8}$$

Sia $\{\varphi_i(x),\ i = 0, 1, \ldots, N\}$ un sistema di funzioni linearmenti indipendenti, quali polinomi algebrici, polinomi ortogonali, polinomi trigonometrici, funzioni splines (per maggiori informazioni su questi insiemi di funzioni si rimanda a [7]), lo scopo è di deteminare un'approssimazione

$$y_N(x) = \sum_{i=0}^{N} c_i \varphi_i(x) \in \mathtt{span}\{\varphi_i\} = \mathbb{F}_n .$$

Sostituiamo $y_N(x)$ in (4.8) e consideriamo il residuo

$$R_N(x) = y_N''(x) - f(x, y_N, y_N')\ \ 0 \le x \le b .$$

Fissati $N-1$ punti distinti $x_n \in (a, b)$, chiederemo che

$$R_N(x_n) = 0,\ \ n = 1, \ldots, N-1 , \tag{4.9}$$

con l'aggiunta delle condizioni al bordo

$$\begin{cases} g_1(y_N(a), y_N(b)) = 0, \\ g_2(y_N(a), y_N(b)) = 0 \end{cases} . \tag{4.10}$$

Quindi (4.9) e (4.10) producono un sistema di $N+1$ equazioni in $N+1$ incognite $c_0, \ldots, c_N$.

Osserviamo che ai fini della determinazione dei coefficienti c_i, che consentono di definire univocamente la soluzione discret y_N nonché del condizionamento del sistema, è di particolare importanza la scelta della base $\{\varphi(x)\}$ e dei nodi $\{x_n\}$ allo scopo che

$$\lim_{n \to +\infty} \|y(x) - y_N(x)\| = 0 .$$

Talvolta, invece che considerare $y_N(x)$ come sopra si preferisce

$$y_N(x) = y_0(x) + \sum_{i=0}^{N} c_i \varphi(x) ,$$

con $y_0(x)$ che soddisfa alle condizioni al bordo.

Esempio 10. Riconsideriamo l'Esempio 9 e prendiamo $y_N(x) = \sum_{i=0}^{N} c_i p_i(x)$, dove p_i sono i polinomi di Legendre (sono un insieme ortogonale in [-1,1]!). Fissati $N-1$ punti distinti

$\{x_n\}$ in $(0,1)$ (es. gli $N-1$ zeri del polinomio $p_{N-1}(x)$ opportunamente scalati in $(0,1)$) imponiamo le condizioni

$$\begin{cases} y_N(0)=0, \\ y_N(1)=0, \\ y_N''(x_n)=y_N(x_n)+x_n \;\; n=1,\dots,N-1 \end{cases}$$

Ora, essendo $y_N'(x)=\sum_{i=0}^{N} c_i p_i'(x)$ e $y_N''(x)=\sum_{i=0}^{N} c_i p_i''(x)$ e ricordando che $p_0(x)=1$, $p_1(x)=x$, il sistema da risolvere avrà la forma

$$\begin{pmatrix} p_0(0) & p_1(0) & p_2(0) & \cdots & p_{N-1}(0) & p_N(0) \\ -p_0(x_1) & -p_1(x_1) & [p_2''(x_1)-p_2(x_1)] & \cdots & [p_{N-1}''(x_1)-p_{N-1}(x_1)] & [p_N''(x_1)-p_N(x_1)] \\ \vdots & \vdots & & & \vdots & \\ \vdots & \vdots & & & \vdots & \\ -p_0(x_{N-1}) & -p_1(x_{N-1}) & & \cdots & [p_{N-1}''(x_{N-1})-p_{N-1}(x_{N-1})] & [p_N''(x_{N-1})-p_N(x_{N-1})] \\ p_0(1) & p_1(1) & & & p_{N-1}(1) & p_N(1) \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ \vdots \\ c_{N-1} \\ c_N \end{pmatrix} = \begin{pmatrix} 0 \\ x_1 \\ \vdots \\ \vdots \\ x_{N-1} \\ 1 \end{pmatrix}$$

Se il sistema risulterà non singolare allora potremo determinare i c_i e la soluzione y_N.

Vediamo altri due semplici esempi

Esempio 11. Consideriamo il problema

$$\begin{cases} y''+y=x^2 & 0<x<1 \\ y(0)=0,\; y(1)=1 \end{cases}$$

Possiamo considerare $y_N(x)=x+\sum_{j=1}^{N} c_j \sin(j\pi\, x)$ che vediamo subito soddisfa alle condizioni al bordo, $y_N(0)=0$ e $y_N(1)=1$. Prendiamo quindi N punti equispaziati $x_j=(2j-1)/2N,\; j=1,\dots,N$, pertanto per determinare i coefficienti $c_j,\; j=1,...,N$, dovremo risolvere il sistema

$$y_N''(x_j)+y_N(x_j)=x_j^2,\;\; j=1,\dots,N\,.$$

Esempio 12. Consideriamo ora il problema

$$\begin{cases} y''-p(x)y'-q(x)=r(x) & a<x<b \\ y(a)=\alpha,\; y(b)=\beta \end{cases}$$

Come y_N prendiamo una *spline cubica* costruita sulla partizione $\Delta=\{a=x_0<x_1<\cdots<x_N=b\}$ con $x_j=a+jh,\; j=0,\dots,N$. Ricordando che splines hanno una base formata da B-splines e che nel caso cubico le B-splines si costruiscono usando 5 nodi (vedi figura 4.1), allora la dimensione dello spazio è $N+3$. Pertanto la nostra soluzione approssimata si scriverà come segue:

$$y_N(x)=\sum_{i=-1}^{N+1} c_i B_{3,i}(x)\;.$$

Il sistema risultante sarà

$$\begin{aligned} &y_N(x_0) = \alpha \\ &y_N''(x_j) - p_j\, y_N'(x_j) - q_j\, y_N(x_j) = r_j, \quad j = 0, \ldots, N \\ &y_N(x_N) = \beta \end{aligned} \tag{4.11}$$

che in totale danno le richieste $N+3$ condizioni. Ora, le Bsplines hanno supporto compatto. Le cubiche sono diverse da zero solo in 3 nodi, pertanto il sistema precedente è tridiagonale nelle incognite c_j.

Vale inoltre il seguente risultato di convergenza. Se p, q e r sono funzioni continue in $[a,b]$, $q(x) > 0$, $\forall\, x \in [a,b]$ e $y(x) \in \mathcal{C}^4[a,b]$ allora per h sufficientemente piccolo il sistema è non singolare e si ha

$$\|y - y_N\|_\infty = \mathcal{O}(h^2)\ .$$

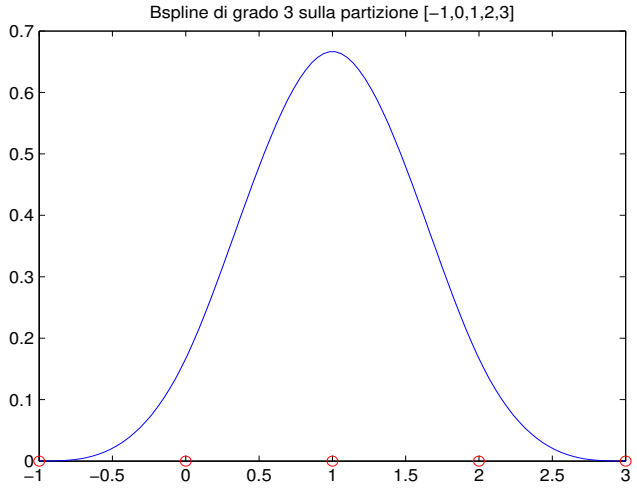

Figura 4.1: Bspline cubica

Un ultimo esempio risolve l'equazione differenziale con una *serie troncata di Fourier*. Per approffondimenti sulle serie di Fourier si rimanda all'Appendice B.

Esempio 13.

$$\begin{cases} -\dfrac{d^2u}{dx^2} = f(x) & -1 < x < 1 \\ \\ u(-1) = \alpha,\ u(1) = \beta \end{cases} \tag{4.12}$$

Consideriamo l'approssimazione $u_N = \sum_{k=0}^{N} c_k T_k(x)$ dove T_k indica il k-esimo polinomio di Chebyshev di prima specie. Essendo f continua, possiamo calcolare lo sviluppo in serie di

Fourier di f

$$f(x) = \frac{1}{2}\hat{f}_0 + \sum_{k=1}^{\infty} \hat{f}_k T_k(x) \tag{4.13}$$

con i coefficienti $\hat{f}_k$ dati dalle espressioni

$$\hat{f}_k = \frac{2}{\pi}\int_{-1}^{1} f(x)\frac{T_k(x)}{\sqrt{1-x^2}}dx\ .$$

Pertanto l'equazione differenziale (4.12) diventa

$$-u_N''(x) = \sum_{k=0}^{N-2}$$

con d_k coefficienti opportuni. Si può dimostrare (vedi ad es. [17])

$$d_k = \begin{cases} -\frac{1}{2}\sum_{j=2}^{N} j^3 c_j & k = 0 \\ -\sum_{j=k+2,\, j+N \text{ pari}}^{N} j(j^2-k^2)c_j & 1 \le k \le N-2 \end{cases}$$

Ricordando che $T_k(1) = 1$ e $T_k(-1) = (-1)^k$, essendo

$$\sum_{k=0}^{N-2} d_k T_k = \frac{1}{2}\hat{f}_0 + \sum_{k=1}^{N-2} \hat{f}_k T_k(x) + \sum_{k=N-1}^{\infty} \hat{f}_k T_k(x)$$

, i coefficienti d_k dell'approssimazione della u_N'' di (4.12) sono

$$\begin{cases} d_0 = \frac{1}{2}\hat{f}_0 & d_k = \hat{f}_k \ \ 1 \le k \le N-2 \\ \sum_{k=0}^{N}(-1)^k c_k = \alpha & \sum_{k=0}^{N} c_k = \beta \end{cases} \tag{4.14}$$

Quest'ultimo è un sistema lineare nelle incognite c_k la cui soluzione consentirà di determinare u_N. Si può dimostrare che $\lim_{N\to\infty} u_N = u$ limite che vale uniformemente in $[-1, 1]$. La velocità di convergenza è controllata dal resto

$$\left(\sum_{k=N-1}^{\infty} \hat{f}_k\right)^{1/2}\ .$$

Se $f(x) = |x|,\ x \in (-1, 1)$ e $\alpha = \beta = 0$, la soluzione analitica è

$$u(x) = \begin{cases} \frac{(1+x^3)}{6} & x \in [-1, 0] \\ \\ \frac{(1-x^3)}{6} & x \in [0, 1] \end{cases}$$

Per determinare u_N dobbiamo calcolare dapprima i coefficienti $\hat{f}_k$ dello sviluppo in serie di Fourier di $f(x) = |x|$. Questi si trovano in molti libri e sono:

$$\begin{cases} \hat{f}_k = \frac{4}{\pi}\frac{(-1)^{k+1}}{4k^2-1} & k \geq 2, k \text{ pari} \\ \\ \hat{f}_k = 0 & k \text{ dispari}\,. \end{cases}$$

Pertanto il sistema (4.14) ha come soluzione $c_k = 0$ se k dispari (perchè $d_k = \hat{f}_k = 0,\ k \geq 1$).

Nel caso $N = 2$, $u_2(x) = c_0T_0(x) + c_1T_1(x) + c_2T_2(x) = c_0 + c_2(2x^2 - 1)$. Le condizioni al bordo forniscono $c_0 = c_2$. Ora,

$$d_0 = \frac{1}{2}\hat{f}_0 = \frac{1}{\pi}\int_{-1}^{1}\frac{|x|}{\sqrt{1-x^2}}dx = \frac{2}{\pi}\,.$$

Ma $d_0 = -\frac{1}{2}2^3\,c_2$, da cui si deduce che $c_2 = -\frac{1}{2\pi}$. Infine avremo

$$u_2(x) = \frac{1}{\pi}(1 - x^2)\,.$$

4.1.2 Un problema non lineare risolto con differenze finite

In questa sottosezione presentiamo un esempio di un problema non lineare che, risolto con differenze finite del second'ordine, ci porta alla soluzione di un sistema non lineare.

Il problema è il seguente:

$$\begin{cases} y''(x) = \frac{1}{2}(x + y(x) + 1)^3 & 0 < x < 1 \\ y(0) = y(1) = 0 \end{cases} \tag{4.15}$$

Osserviamo che $f(x, y(x)) = \frac{1}{2}(x+y(x)+1)^3$ è tale che $\frac{\partial f}{\partial y} = \frac{3}{2}(x+y(x)+1)^2 \geq 0$. Pertanto esiste unica la soluzione del problema (4.15).

Discretizziamo come al solito nei punti

$$x_i = ih,\ \ h = 1/(n+1),\ \ i = 0, \ldots, n+1\,,$$

e con u_i i valori approssimati in x_i della soluzione $y(x)$:

$$\begin{aligned} &u_0 = 0 \\ &-\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} + (1 + x_i + u_i)^3 = 0 \\ &u_{n+1} = 0 \end{aligned} \tag{4.16}$$

Il sistema 4.16 può scriversi in forma matriciale

$$A\,U + h^2\,B(U) = 0 \tag{4.17}$$

dove $U = (u_1, \ldots, u_n)^T$ e

$$A = \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & & \\ & & \ddots & -1 \\ & & -1 & 2 \end{bmatrix},$$

e $B(U) = \texttt{diag}(f(x_i, u_i)),\ i = 1, ..., n$. Il sistema (4.17) è non lineare e pertanto, lo potremo risolvere, ad esempio, con il metodo di Newton (cfr. ad esempio [7, Cap. 3]) la cui matrice jacobiana è

$$J = A + h^2 B_u, \quad B_u = \frac{\partial f}{\partial u}.$$

Poiché, $f_u \geq 0$, J risulta a predominanza diagonale e definita positiva, pertanto il metodo di Newton risulterà convergente per ogni possibile scelta dei valori iniziali. Da un punto di vista implementativo, il metodo di Newton corrispondente al sistema (4.17), a partire da U_0 (stima iniziale della soluzione), si descriverà

$$\begin{cases} (A + h^2 B_u(U^{(r)}))\delta = -A\,U^{(r)} - h^2 B(U^{(r)}) \\ U^{(r+1)} = U^{(r)} + \delta \end{cases}$$

Esercizio 1. *Implementare il metodo di Newton per il problema appena illustrato, la cui soluzione analitica è* $y(x) = \frac{2}{2-x} - x - 1$. *Inoltre, riportare su una tabella* $E_\infty = \|y - u\|_\infty$ *per diversi valori di* n. *Si osserverà che per* $n \to \infty$ *(ovvero* $h \to 0$*),* $E_\infty \to 0$ *pur rimanendo inalterato il numero di iterazioni.*

Capitolo 5

Equazioni alle derivate parziali

5.1 Preliminari sulle equazioni alle derivate parziali

Un'equazione alle derivate parziali, che d'ora in avanti chiameremo con la sigla inglese di PDE (da Partial Differential Equation), è una relazione del tipo

$$F(x, y, \ldots, u, u_x, u_y, u_{xx}, u_{xy}, u_{yy}, \ldots) = 0 \tag{5.1}$$

dove x, y indicano le variabili indipendenti,mentre $u(x, y, \ldots)$ la funzione incognita e

$$u_x, u_y, \ldots, u_{xx}, \ldots$$

le sue derivate parziali prime, seconde,....

Vediamo un pò di terminologia fondamentale.

- La derivata di ordine più elevato indica l'*ordine* della PDE. Ad esempio $F(x, y, u_x, u_y) = 0$ è di *ordine 1*, $F(x, y, u_x, u_y, u_{xx}) = 0$ e $F(x, y, u_x, u_{xy})$ sono di *ordine 2*.

- $u(x, y, \ldots)$ è detta *soluzione classica* su una regione aperta $\mathcal{R}$, se è $\mathcal{C}^m(R)$ (continua con tutte le sue derivate parziali fino all'ordine m su $\mathcal{R}$) e soddisfa l'equazione (5.1). Soluzioni non classiche, cioè meno regolari, sono dette *soluzioni deboli* (non soddisfano la (5.1) puntualmente) e sono associate alla cosidetta `formulazione variazionale` (o integrale) di (5.1), che coinvolge derivate di ordine inferiore a m (definite nel senso delle distribuzioni). Parleremo di ciò più oltre nel contesto del Metodo degli Elemeti Finiti.

- L'equazione (5.1) di ordine m si dirà *lineare* se F è lineare in u e nelle sue derivate parziali $u_x, u_y, \ldots$ con coefficienti che dipendono univocamente dalle variabili indipendenti $x, y, \ldots$.

L'equazione (5.1) di ordine m si dirà *quasi lineare* se F è lineare in u e nelle sue derivate parziali $u_x, u_y, \ldots$ con coefficienti che dipendono dalle variabili indipendenti $x, y, \ldots$, da u e dalle sue derivate di ordine fino a $m-1$.

Ad esempio

$$u_y = d(x,y)u_{xx} - v(x,y)u_x + a(x,y)u + f(x,y)$$

è lineare di ordine 2, mentre

$$u_y u_{xx} - u_x^2 - u_y^2 + u = 1$$

è quasi-lineare di ordine 2, in quanto il coefficiente di u_{xx} dipende da x, y, u_y (u_y è di ordine 1).

I problemi fisici associati alle PDE sono sostanzialmente di due tipi

1. *Problemi di propagazione* (non stazionari). Si tratta di problemi ai valori iniziali, dove i dati sono assegnati al tempo $t = 0$. Pertanto il problema consiste nello studiare il comportamento del fenomeno per $t > 0$. Esempi sono: la propagazione di calore in un mezzo e la propagazione di pressione in un fluido.
2. *Problemi di equilibrio* (stazionari). Si tratta di problemi indipendenti dal tempo e sono problemi con valori al bordo (o contorno). Esempi sono: il flusso viscoso stazionario, la distribuzione stazionaria di temperature in un mezzo, l'equilibrio di tensioni in srutture elastiche.
3. *Problemi di autovalori*. Si possono pensare come un'estensione di problemi d'equilibrio nei quali si chiede di determinare dei *valori critici* di un parametro (autovalore) invece che studiare configurazioni di stazionarietà. Tipici esempi sono: deformazioni e stabilità di strutture, fenomeni di risonanza in circuiti elettrici/acustici, ricerca di frequenze naturali in problemi di vibrazione.

5.1.1 Alcuni problemi fisici e loro formulazione matematica

1. *Problema di propagazione*. Dato un filo metallico, omogeneo, di lunghezza L, densità ρ, calore specifico c_p, conduttività termica κ, termicamente isolato lungo la sua lunghezza e all'estremo L. Supponiamo che all'inizio il filo sia a temperatura T_1. L'estremo O viene quindi immerso in un mezzo a temperatura $T_0 \ll T_1$ (vedasi Fig. 5.13).

   Si chiede si studiare la conduzione di calore del filo ovvero la temperatura $u(x,t)$ nel punto x del filo all'istante $t > 0$ e capire come evolve.

   Le equazioni fisiche che sovrintendono al problema sono

$$\begin{cases} u_{xx} - \frac{\rho c_p}{\kappa} u_t = 0 & 0 < x < L,\ t > 0 \\ u(x,0) = T_1 & 0 \le x \le L \\ u(0,t) = T_0 & t > 0 \\ u_x(L,t) = 0 & t > 0 \end{cases} \tag{5.2}$$

Questo è un tipico problema di propagazione ai valori iniziali.

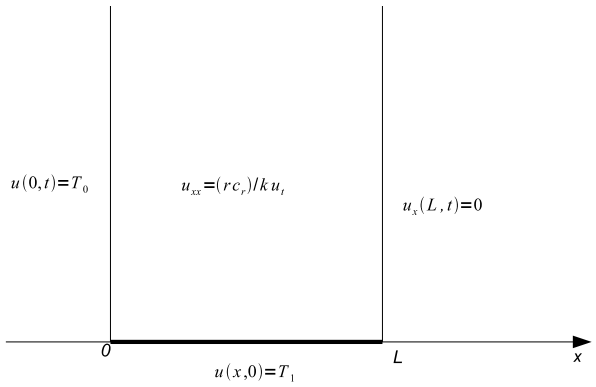

Figura 5.1: Problema di conduzione in un filo metallico

2. *Problema di equilibrio.* Consideriamo un supporto rigido $D \in (x, y)$ che possiamo assumere come il contorno di una membrana elastica ideale, con densità uniforme, sottoposta a una tensione unforme T lungo il bordo ed a una pressione verticale uniforme P. La soluzione $u(x, y))$ è soluzione del *problema di Poisson*

$$\begin{cases} -(u_{xx} + u_{yy}) = P/T & (x, y) \in D\backslash\partial D \\ u(x, y) = 0 & (x, y) \in \partial D \end{cases} \tag{5.3}$$

3. *Problema di autovalori.* Il "prototipo" di tale classe di problemi è quello della `membrana vibrante` la cui formulazione è la seguente:

$$\begin{cases} -(u_{xx} + u_{yy}) = \lambda\, u(x, y) & (x, y) \in D\backslash\partial D \\ u(x, y) = 0 & (x, y) \in \partial D \end{cases} \tag{5.4}$$

   Pertanto si tratta di determinare i valori di λ ai cui corrispondono delle autosoluzioni (non nulle) u, soluzioni u da determinarsi separatamente.

Si noti come i primi due problemi sono *ben posti* nel senso che esiste un'unica soluzione u e che questa dipende con continuità dai dati iniziali.

5.1.2 Classificazione delle PDEs

In base alla loro formulazione matematica le PDE si classificano in *ellittiche, paraboliche* e *iperboliche*. Se ci limitiamo a quelle del secondo ordine lineari a coefficienti costanti, indicheremo con $L\,u = 0$ l'equazione differenziale

$$L\,u := A\,u_{xx} + Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu - G = 0\,, \tag{5.5}$$

dove la costanti A, B, C, D, E, F, G sono numeri reali. Associato all'equazione (5.5) c'è il *discriminante*

$$\Delta = B^2 - 4AC\,. \tag{5.6}$$

Definizione 10. *In base al segno del discriminante Δ, un'equazione alle derivate parziali si dirà* `ellittica` *se $\Delta < 0$,* `parabolica` *se $\Delta = 0$ e* `iperbolica` *se $\Delta > 0$.*

Le PDE ellittiche sono di solito equazioni che descrivono problemi indipendenti dal tempo, mentre le equazioni paraboliche e iperboliche descrivono problemi tempo-dipendenti.

Esempio 14. 1. L'`equazione delle onde` monodimensionale

$$u_{xx} - u_{yy} = 0 \tag{5.7}$$

è una PDE iperbolica. Infatti per essa $\Delta = 1 > 0$.

2. L'`equazione del potenziale o di Poisson`

$$-\left(u_{xx} + u_{yy}\right) = f \tag{5.8}$$

ha discriminante $\Delta = -4 < 0$.

3. L'`equazione del calore` (o di diffusione)

$$u_t - u_{xx} = f \tag{5.9}$$

è un'equazione parabolica in quanto $\Delta = 0$. È di tipo parabolico anche l'equazione di `diffusione-trasposto`

$$u_t - \mu u_{xx} + \nabla \cdot (\beta u)$$

dove $\nabla \cdot v = \displaystyle\sum_{j=1}^{d} \frac{\partial v_j}{\partial x_j}$ indica l'operatore *divergenza* del vettore v, β indica un campo vettoriale, ovvero una funzione a valori vettoriali (es. per $d = 3,\ f(x) = (f_1(x), f_2(x), f_3(x))^T,\ x \in \mathbb{R}^3$.)

All'equazione (5.5) si può associare il cosidetto *simbolo principale* S^p definito come segue. Siano $\mathbf{x}$ e $\mathbf{q}$ vettori 2-dimensionali, allora

$$S^p(\mathbf{x}, \mathbf{q}) = -A(\mathbf{x})q_1^2 - B(\mathbf{x})q_1q_2 - C(\mathbf{x})q_2^2\,, \tag{5.10}$$

è una forma quadratica in q_1, q_2 che possiamo scrivere in termini matriciali come

$$S^p(\mathbf{x}, \mathbf{q}) = \mathbf{q}^T \begin{pmatrix} -A(\mathbf{x}) & -\frac{1}{2}B(\mathbf{x}) \\ -\frac{1}{2}B(\mathbf{x}) & -C(\mathbf{x}) \end{pmatrix} \mathbf{q}\,. \tag{5.11}$$

Ricordando che una forma quadratica è definita se la matrice associata è definita positiva o negativa; è indefinita quando la matrice ha autovalori di entrambi i segni e degenere quando la matrice è singolare, possiamo allora classificare le PDE usando la forma quadratica (5.11).

Definizione 11. *Un'equazione alle derivate parziali si dirà* `ellittica` *se la forma quadratica (5.11) è definita positiva o negativa;* `parabolica` *se degenere e infine* `iperbolica` *se indefinita.*

Ritornando all'esempio 14, vediamo come sono fatte le corrispondenti matrici

1. `Equazione di Poisson`.
   $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ da cui $A(\mathbf{x}) = 1$ (è il coefficiente di u_{xx}), $C(\mathbf{x}) = 1$ (è il coefficiente di u_{yy}). La matrice è definita positiva, quindi, in base alla definizione appena data, l'equazione è ellittica.

2. `Equazione del calore`.
   $\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$ da cui $C(\mathbf{x}) = 1$ (è il coefficiente di u_{yy}). La matrice è degenere. L'equazione è quindi parabolica.

3. `Equazione delle onde`.
   $\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$ da cui $A(\mathbf{x}) = 1$ e $C(\mathbf{x}) = -1$. La matrice è indefinita. L'equazione è di tipo iperbolico.

Infine, le equazioni alle derivate parziali si possono classificare in base all'ordine alla (non)linearità. Lo vediamo con un paio di esempi.

- $\rho\, u_{yy} + \kappa\, u_{xxxx} = f$ è lineare del quart'ordine. Questa equazione descrive il comportamento della verga vibrante di densità ρ e $\kappa > 0$ che descrive le caratteristiche geometriche della verga stessa.

- $u_x^2 + u_y^2 = f$ è un'equazione non lineare del primo ordine.

5.2 Metodi alle differenze per PDE

Il problema che affrontiamo è di risolvere una PDE su un dominio $D \subset \mathbb{R}^2$. Il dominio verrà "sostituito" da un reticolo di forma rettangolare di punti di D, punti su cui poi andremo a collocare l'equazione differenziale (vedi figura 5.2).

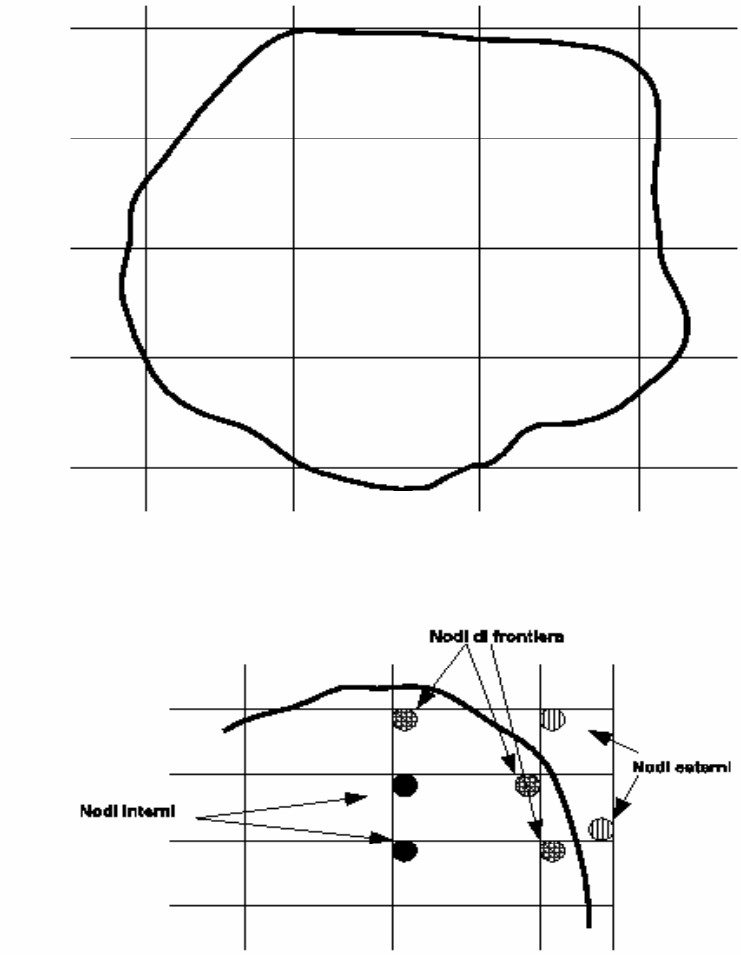

Figura 5.2: (Sopra) Dominio bidimensionale e sua discretizzazione. (Sotto) Discretizzazione lungo il bordo.

Definizione 12. *Un punto di un reticolo è detto* `interno` *se i suoi 4 vertici vicini sono interni (contorno incluso) al dominio D. Altrimenti il punto si dice di* `frontiera`.

Consideriamo il reticolo di figura 5.2 (Sotto). I nodi interni sono quelli indicati con cerchi con colore uniforme, mentre quelli di frontiera con una retinatura. Sono stati indicati anche alcuni nodi esterni.

5.2.1 Formule per le derivate parziali

Per ora consideriamo le derivate parziali lungo le due direzioni ortogonali x e y. Sia h la spaziatura tra due nodi consecutivi lungo x e k la spaziatura lungo y. Valgono le seguenti formule

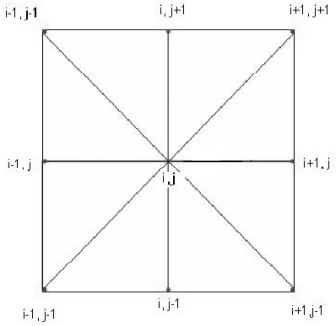

Figura 5.3: Reticolo di discretizzazione della soluzione attorno al punto $(i, j) \in (x, y)$

- *Formule per le derivate parziali prime.* Iniziamo lungo la direzione x.

$$\begin{aligned} \frac{\partial u(x_i, y_j)}{\partial x} &= \frac{u_{i+1,j} - u_{i,j}}{h} + \mathcal{O}(h) \quad \text{in avanti} \\ &= \frac{u_{i,j} - u_{i-1,j}}{h} + \mathcal{O}(h) \quad \text{all' indietro} \\ &= \frac{u_{i+1,j} - u_{i-1,j}}{2h} + \mathcal{O}(h^2) \quad \text{centrali} \end{aligned}$$

  a cui corrispondono gli schemi o *stencils* di figura 5.4

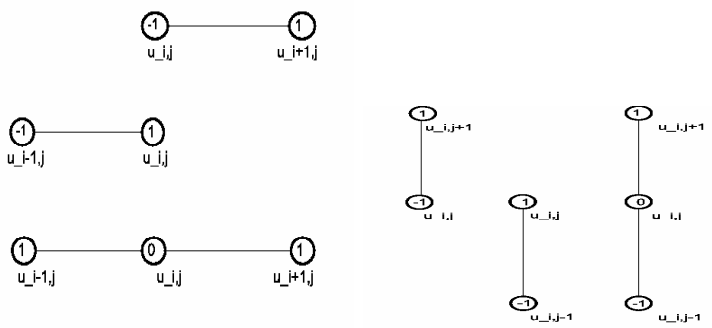

Figura 5.4: Stencils per derivate prime lungo x e y

Analogamente lungo y

$$\begin{aligned}\frac{\partial u(x_i,y_j)}{\partial x} &= \frac{u_{i,j+1}-u_{i,j}}{k}+\mathcal{O}(k) \quad \text{in avanti}\\ &= \frac{u_{i,j}-u_{i,j-1}}{k}+\mathcal{O}(k) \quad \text{all' indietro}\\ &= \frac{u_{i,j+1}-u_{i,j-1}}{2k}+\mathcal{O}(k^2) \quad \text{centrali}\end{aligned}$$

a cui corrispondono gli schemi o *stencils* di figura 5.4

- *Formule per le derivate parziali seconde.*

$$\begin{aligned}\frac{\partial^2 u(x_i,y_j)}{\partial x^2} &= \frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{h^2}+\mathcal{O}(h^2) \quad \text{lungo x}\\ \frac{\partial^2 u(x_i,y_j)}{\partial y^2} &= \frac{u_{i,j+1}-2u_{i,j}+u_{i,j-1}}{k^2}+\mathcal{O}(k^2) \quad \text{lungo y}\end{aligned} \tag{5.12}$$

con stencils che avranno coefficienti 1 in corrispondenza di $u_{i+1,j}, u_{i-1,j}$ lungo x e $u_{i,j+1}, u_{i,j-1}$ lungo y. Mentre in corrispondenza di $u_{i,j}$ gli stencils avranno coefficiente 2.

Per quanto concerne le derivate parziali seconde miste useremo l'approssimazione

$$\frac{\partial^2 u(x_i,y_j)}{\partial x\partial y} = \frac{u_{i+1,j+1}-u_{i-1,j+1}-u_{i+1,j-1}+u_{i-1,j-1}}{4hk}+\mathcal{O}((h+k)^2)\,, \tag{5.13}$$

che nel reticolo di figura 5.3 ha i 4 vertici del rettangolo con coefficienti diversi da zero.

5.2.2 Schemi alle differenze per il laplaciano in 2D

Ricordando che il gradiente è il l'operatore vettoriale che in 2 dimensioni si scrive come $\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial x}\right)^T$, per il laplaciano vale

$$\Delta = \mathrm{div}(\nabla) = \nabla\cdot\nabla = \nabla^2 = \frac{\partial^2}{\partial x}+\frac{\partial^2}{\partial y}\,.$$

Supponiamo che $h = k$. Usando le formule (5.12), per approssimare le derivate parziali secondo lungo x e y rispettivamente, otteniamo i seguenti schemi.

1. *Schema a 5 punti.*

$$\nabla^2 u(x_i,y_j) = \frac{u_{i+1,j}+u_{i-1,j}+u_{i,j+1}+u_{i,j-1}-4u_{i,j}}{h^2}+\mathcal{O}(h^2)$$

che si ottiene sommando membro a membro le (5.12). Questo schema lo indichiamo con ∇^2_5 (vedi figura 5.5).

2. *Schema a croce.*

$$\nabla^2 u(x_i, y_j) = \frac{u_{i+1,j+1} + u_{i-1,j-1}}{2h^2} + \frac{u_{i+1,j-1} + u_{i-1,j+1} - 4u_{i,j}}{2h^2} + \mathcal{O}(h^2)\,.$$

   Questo schema lo indichiamo con ∇^2_X.

3. Infine, è possibile combinare gli schemi precedenti ottenendo

$$\nabla^2_9 = \frac{2}{3}\nabla^2_5 + \frac{1}{3}\nabla^2_X$$

   che consente di raggiungere un'approssimazione di ordine h^4 (vedi figura 5.5).

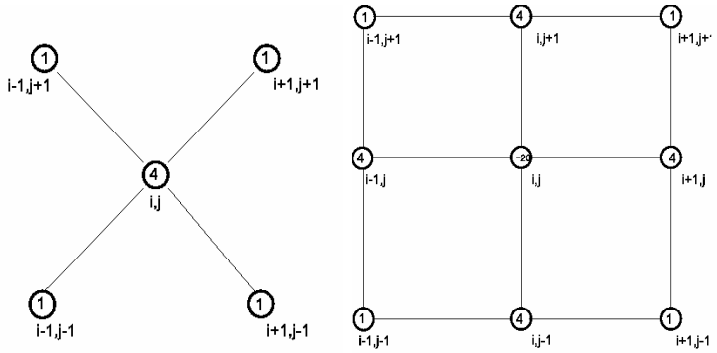

Figura 5.5: I reticoli degli schemi ∇^2_X (sx) e ∇^2_9 (dx).

5.2.3 Condizioni al bordo

Quando il dominio D ha una topologia tale per cui il reticolo che lo approssima non copre tutti i punti della curva di bordo $\Gamma = \partial D$, allora si può procedere come segue. Dato il reticolo di figura 5.6, siano A, B, C, D i punti che appartengono alla curva Γ e $1, 2, 3$ punti di frontiera (cfr. Def. 12 per punto di frontiera). Possiamo operare in 2 modi.

(a) Assegnamo ad u_1 (ovvero la soluzione nel punto 1) il valore che $u(x, y)$ assume in un qualsiasi punto di Γ che disti dal nodo 1 meno di h. Ad esempio $u_1 = u_A$ o $u_1 = u_B$.

(b) Approssimiamo le derivate nel punto 1 con formule costruite sui nodi *non equispaziati* `2, 1, B` (lungo la direzione x) e `3, 1, A` (lungo la direzione y). Ad esempio, per il laplaciano avremo

$$\nabla^2 u_1 = (u_{xx}+u_{yy})_1 = \frac{2}{h^2}\left\{\frac{u_2}{b+1} + \frac{u_3}{a+1} + \frac{u_A}{a(a+1)} + \frac{u_B}{b(a+1)} - \frac{(a+b)u_1}{ab}\right\} + \mathcal{O}(h)\,,$$

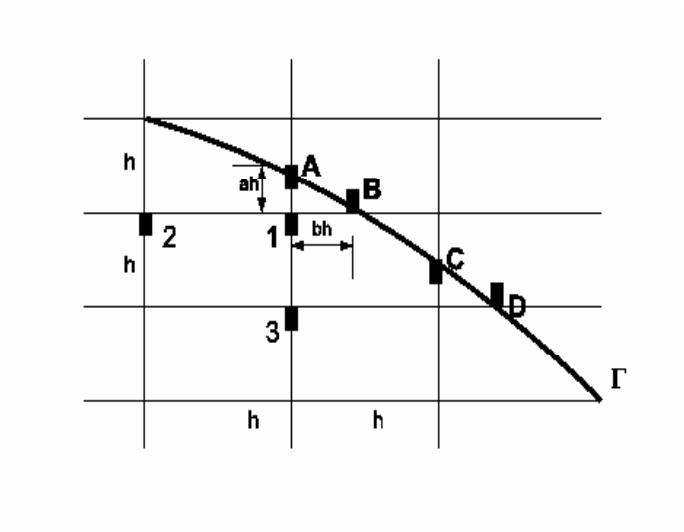

Figura 5.6: Reticolo lungo il bordo Γ del dominio.

formule che si ottengono applicando quelle per le derivate parziali seconde lungo x e y quando i punti non sono equispaziati. L'unico problema di questo approccio è che il metodo diventa del primo ordine invece del secondo. In tal caso, vedremo oltre, conviene usare metodi quali il *metodo agli elementi finiti*.

5.2.4 Approssimazione delle derivate direzionali

Consideriamo un vettore $\mathbf{t}$ di norma unitaria. La derivata della funzione u nel punto P lungo la direzione $\mathbf{t}$ è

$$\frac{du(P)}{d\mathbf{t}} = \lim_{h\to 0} \frac{u(P+h\mathbf{t}) - u(P)}{h} \tag{5.14}$$

che possiamo anche scrivere come

$$\frac{du(P)}{d\mathbf{t}} = \nabla u(P)\cdot \mathbf{t} = \|\nabla u(P)\|_2 \cos\theta \,, \tag{5.15}$$

con θ che indica l'angolo tra il gradiente di u in P e la direzione $\mathbf{t}$.

Per approssimare la *derivata normale* di u in un punto $R \in \Gamma$ la cui normale (uscente) in R di coordinate (x, y), $\mathbf{n}_R$, possiamo usare l'espansione di Taylor rispetto ad un punto Q (di frontiera) del reticolo (vedi figura **??**):

$$\begin{aligned} u(x,y) &= u(Q) + (x-x_Q)u_x(Q) + (y-y_Q)u_y(Q) + \\ &+ \frac{(x-x_Q)^2}{2}u_{xx}(Q) + \frac{(x-x_Q)(y-y_Q)}{2}u_{xy}(Q) + \frac{(y-y_Q)^2}{2}u_{yy}(Q) + \cdots \end{aligned}$$

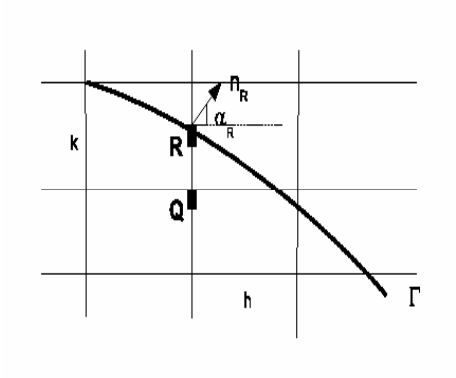

Figura 5.7: Derivata lungo la normale n_R nel punto $R \in \Gamma$.

da cui,

$$\begin{aligned} u_x(x,y) &= u_x(Q) + (x-x_Q)u_{xx}(Q) + (y-y_Q)u_{xy}(Q) + \cdots \\ u_y(x,y) &= u_y(Q) + (y-y_Q)u_{yy}(Q) + (x-x_Q)u_{xy}(Q) + \cdots \end{aligned}$$

Pertanto

$$\begin{aligned} \frac{du(R)}{d\mathbf{n}} &= \nabla u(R)\cdot \mathbf{n}_R = u_x(R)\cos\alpha_R + u_y(R)\sin\alpha_R \\ &= [u_x(Q) + \lambda_R\, k\, u_{xy}(Q)]\cos\alpha_R + [u_y(Q) + \lambda_R\, h\, u_{yy}(Q)]\sin\alpha_R + \mathcal{O}(h^2) + \mathcal{O}(k^2) \end{aligned}$$

Applicando poi i metodi per approssimare $u_x(Q)$, $u_y(Q)$, $u_{xx}(Q)$, $u_{yy}(Q)$ riusciremo ad approssimare la derivata lungo la normale alla curva Γ nel punto R.

Il metodo risulta particolarmente efficiente quando il reticolo che approssima il dominio D è *regolare*.

Definizione 13. *Un reticolo si dice regolare quando per ogni h e k esso non ha punti du frontiera che non appartengono anche al bordo Γ di D.*

Detto altrimenti, un reticolo è regolare quando tutti i punti di frontiera sono sul bordo di D.

5.3 Problemi di tipo iperbolico del primo ordine

Consideriamo l'equazione lineare di ordine 1 a coefficienti costanti sulla striscia $D = \{0 < x < 1,\ t > 0\}$

$$\begin{cases} u_t + a\, u_x = 0 & a \in \mathbb{R}_+ \\ u(x,0) = u_0(x) & \text{(valori iniziali)} \\ u(0,t) = f(t) & \end{cases} \tag{5.16}$$

Useremo questo problema come prototipo per lo studio di stabilità dei metodi numerici per PDE basati su differenze finite.

Consideriamo, al solito, una griglia equispaziata formata da $N+1$ punti, $x_0, \ldots, x_N$ di passo h lungo x e una equispaziata di passo k lungo il tempo t. Per la discretizzazione delle derivate consideriamo gli schemi

$$u_t(x_i, t_j) = \frac{u_{i,j+1} - u_{i,j}}{k} + \mathcal{O}(k)$$
$$u_x(x_i, t_j) = \frac{u_{i,j} - u_{i-1,j}}{h} + \mathcal{O}(h)$$

dove $\mathcal{O}(h)$ e $\mathcal{O}(k)$ indicano gli errori locali di trocamento in (x_i, t_j).

Sostituendo in (5.16) otteniamo lo schema

$$\begin{cases} \frac{u_{i,j+1}-u_{i,j}}{k} + a\frac{u_{i,j}-u_{i-1,j}}{h} = 0 & i = 1, \ldots, N, \;\; j \geq 0 \\ u_{0,j} = f_j & j = 1, 2, \ldots \\ u_{i,0} = u_{0,i} & i = 0, 1, \ldots, N \end{cases}$$

ovvero

$$\begin{cases} u_{i,j+1} = (1-\lambda)u_{i,j} + \lambda u_{i-1,j} & i = 1, \ldots, N, \;\; j \geq 0 \\ u_{0,j} = f_j & j = 1, 2, \ldots \\ u_{i,0} = u_{0,i} & i = 0, 1, \ldots, N \end{cases} \tag{5.17}$$

dove abbiamo definito $\lambda = \frac{ak}{h}$.

Definizione 14. *Lo schema numerico (5.17) si dirà* `stabile`*, se fissato $\tau > 0$, posto $M = \tau/k$, per h e k sufficientemente piccoli, le perturbazioni sono limitate in $(0, \tau]$ uniformemente rispetto ad h e k. Nel caso contrario lo schema sarà* `instabile`*.*

In figura 5.8 si è evidenziato il comportamento dello schema (5.17) per $\lambda = 1/2$ e $\lambda = 2$, nel caso di una perturbazione ϵ sul solo dato x_1 lungo $t = 0$. Come si nota, per $\lambda = 1/2$ la perturbazione si dissipa mentre, al contrario, per $\lambda = 2$ si amplifica.

La domanda che ci poniamo è: per quali valori di λ lo schema (5.17) risulta essere stabile?

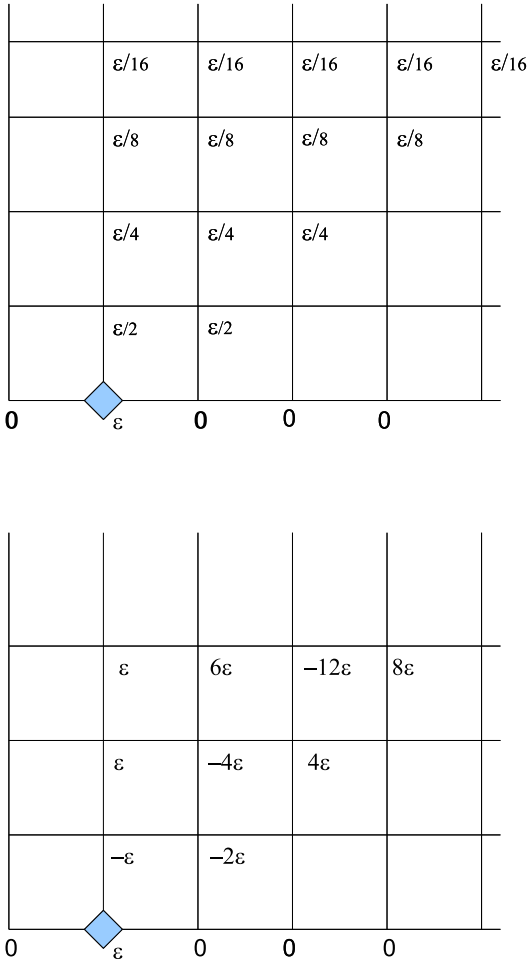

Figura 5.8: Propagazione dell'errore per lo schema (5.17), nel caso $\lambda = 1/2$ (sopra), schema `stabile` e nel caso di $\lambda = 2$ (sotto),schema `instabile`

Consideriamo i nodi $(x_i, t_j),\ i = 1, \ldots, N,\ j \geq 0$ e di perturbare solo i dati iniziali $u_{i,0}$. Siano poi

$$U_j = (u_{1,j}, \ldots, u_{N,j})^T, \quad E_j = (\epsilon_{1,j}, \ldots, \epsilon_{N,j})^T$$

il vettore soluzione e perturbazione, rispettivamente, al passo j. Possiamo allora riscrivere le equazioni (5.17) in forma matriciale

$$U_{j+1} = AU_j + \lambda\, v_j\,, \quad U_0 = (u_{0,j}, \ldots, u_{0,j})^T \tag{5.18}$$

dove

$$A = \begin{bmatrix} 1-\lambda & & & 0 \\ \lambda & 1-\lambda & & \\ & \ddots & \ddots & \\ 0 & & & \\ & & \lambda & 1-\lambda \end{bmatrix}, \quad v_j = \begin{pmatrix} u_{0,j}(= f_j) \\ 0 \\ \vdots \\ \\ 0 \end{pmatrix}.$$

L'errore E_j soddisfa allora la relazione

$$E_{j+1} = AE_j, \;\; j \geq 0$$

ovvero

$$E_{j+1} = A^{j+1}E_0 \,,$$

perchè la perturbazione è solo sui dati iniziali. Pertanto, la domanda a cui rispondere sarà: per quali valori del parametro λ, A^{j+1} risulta limitata in norma, uniformemente rispetto a j e N?

Una condizione sufficiente è che $\|A\|_1 \leq 1$, che nel nostro caso coincide anche con $\|A\|_\infty$ e che si traduce nella disequazione

$$|\lambda| + |1-\lambda| \leq 1\,. \tag{5.19}$$

Questa disequazione è certamente verificata quando $0 < \lambda \leq 1$. In conclusione lo schema numerico (5.17) è stabile, quando i passi h e k soddisfano quella che viene chiamata la *condizione CFL* (da Courant, Friedrichs e Lewy)

$$\frac{a\,k}{h} \leq 1\,. \tag{5.20}$$

In particolare, se $a = 1$, si ha che $k \leq h$. Lo schema (5.17) è quindi *condizionatamente stabile*. Inoltre, osserviamo, che questa analisi è stata fatta per la propagazione di errori presenti nei dati iniziali, ovvero per $t = 0$. Si procede analogamente per la propagazione di dati al bordo $x = 0$.

$\diamond\diamond$

Consideriamo ora il problema iperbolico (5.16), senza condizioni al bordo

$$\begin{cases} u_t + a\,u_x = 0 & a \in \mathbb{R}\backslash\{0\} \\ u(x,0) = u_0(x) & -\infty < x < \infty, \,, t > 0 \end{cases} \tag{5.21}$$

La soluzione è nota analiticamente

$$u(x,t) = u_0(x - at)\,.$$

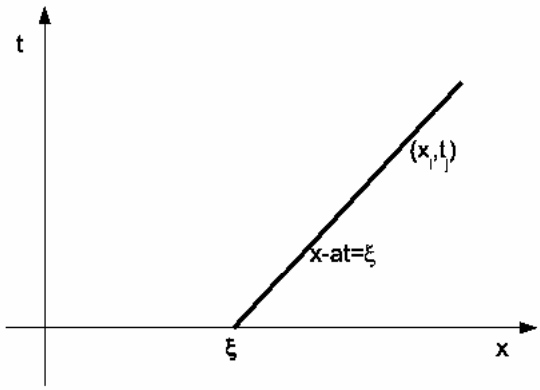

Figura 5.9: Retta caratteristica per problema iperbolico (5.19)

Ritornando al problema (5.21), soluzione u in (x_i, t_j) dipende solo dal dato iniziale u_0 nel punto $\xi = x_i - at_j$. Anzi, $u(x,t) = u_0(\xi)$ per ogni punto della *retta caratteristica* (vedasi figura 5.9)

$$x - at = \xi\,.$$

Il punto iniziale ξ viene detto anche punto di *inflow* o *dominio di dipendenza* del punto $x_i, t_j)$, mentre la retta caratteristica $t(x) = \dfrac{x-\xi}{a}$ è il *dominio d'influenza* del punto $x = \xi$.

Perchè *rette, linee* caratteristiche? La soluzione di (5.21) rappresenta un'onda 1-dimensionale che si sposta alla velocità a. Se consideriamo nel piano (x,t), le curve $x(t)$ soluzioni di

$$\begin{cases} \dfrac{dx}{dt} = a \qquad t > 0 \\ x(0) = x_0 \end{cases}$$

al variare di x_0, esse rappresentano delle rette, appunto le *rette caratteristiche*, e lungo tali rette, l'equazione (5.21) è costante. Infatti, usando la (5.21), si ottiene

$$\frac{du}{dt} = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\frac{dx}{dt} = 0\,.$$

Esempio 15. Se considerassimo il problema

$$\begin{cases} u_t + a\,u_x + a_0 u = f \\ u(x,0) = u_0(x) \end{cases} \tag{5.22}$$

con $-\infty < x < \infty$, , $t > 0$ e a, a_0, f funzioni assegnate di (x,t), le rette caratteristiche sono ora soluzioni di

$$\begin{cases} \dfrac{dx}{dt} = a(x,t) \quad t > 0 \\ x(0) = x_0 \end{cases}$$

In tal caso però, le soluzioni di (5.22) soddisfano la relazione

$$\frac{du(x(t),t)}{dt} = f(x(t),t) - a_0(x(t),t)\, u(x(t),t)\,.$$

Le curve caratteristiche diventano un modo di ricavare la soluzione u: basterà risolvere una equazione differenziale ordinaria su ogni curva caratteristica. Questo è il cosidetto *metodo delle linee caratteristiche* che vedremo più oltre nel caso dei problemi parabolici.

Esempio 16. Il problema

$$\begin{cases} u_t + u_x = 0 \quad x \in (-3,3),\ 0 < t \le 1 \\ u_0(x) = \begin{cases} \sin(k\pi x) & x \in [-1,1] \\ 0 & altrove \end{cases} \\ u(-3,t) = 0 \end{cases} \tag{5.23}$$

La soluzione esatta, come si deduce da quanto appena detto, è

$$u(x,t) = \begin{cases} \sin(k\pi(x-t)) & x - t \in [-1,1] \\ 0 & altrove \end{cases}$$

Quest'ultimo esempio ci permette di comprendere un altro aspetto. Nel caso del problema

$$\begin{cases} u_t + a\, u_x = 0\,, \quad 0 < x < 1,\ t > 0 \\ u(x,0) = u_0(x) \\ u(0,t) = f \end{cases} \tag{5.24}$$

Alcune osservazioni.

1. I dati iniziali assegnati su $0 \le x < 1$ definiscono univocamente $u(x,t)$ solo sul triangolo individuato dalle rette $t = x/a,\ t = 0,\ x = 1$.

2. Il dato $u(0,t) = f(t)$ è indispensabile per poter definire la soluzione su tutta la striscia $0 \le x \le 1,\ t > 0$.

3. Se $a < 0$, per definire univocamente $u(x,t)$ sulla striscia, è necessario assegnare il dato al bordo su $x = 1$.

I dati assegnati, ad esempio, nell'intervallo $[x_n, 1]$ definiscono la soluzione solo nel triangolo Δ_n delimitato dalle rette $t = (x - x_n)/a,\ x = 1$ e $t = 0$. Pertanto, la condizione CFL, $\lambda \le 1$, in questo caso implica la conoscenza della soluzione nel triangolo. Se $\lambda > 1$ allora lo schema numerico (5.17) darebbe delle approssimazioni anche fuori dal triangolo Δ_n.

$\diamond\diamond$

Vediamo ora uno schema numerico `stabile`:

$$\begin{cases} u_{i,j+1} = u_{i,j} + \dfrac{a\,k}{h} u_{i-1,j+1} - u_{i,j+1} & i = 1, \ldots, N, \;\; j \geq 0 \\ u_{0,j} = f_j & j = 1, 2, \ldots \\ u_{i,0} = u_{0,i} & i = 0, 1, \ldots, N \end{cases} \tag{5.25}$$

che in (x_i, t_{j+1}) da un errore locale di troncamento, $\mathcal{O}(h) + \mathcal{O}(k)$. Posto ancora $\lambda = \frac{a\,k}{h}$, lo studio di stabilità di questo schema conduce all'analisi dell'errore

$$AE_{j+1} = E_j$$

con

$$A = \begin{bmatrix} 1+\lambda & & & 0 \\ -\lambda & 1+\lambda & & \\ & \ddots & \ddots & \\ 0 & & & \\ & & -\lambda & 1+\lambda \end{bmatrix} = (1+\lambda) \begin{bmatrix} 1 & & & 0 \\ -\beta & 1 & & \\ & \ddots & \ddots & \\ 0 & & & \\ & & -\beta & 1 \end{bmatrix}$$

con $\beta = \dfrac{\lambda}{1+\lambda}$. Essendo A invertibile, possiamo scrivere $E_{j+1} = A^{-1} E_j$ dove

$$A^{-1} = \frac{1}{1+\lambda} \begin{bmatrix} 1 & & & 0 \\ \beta & 1 & & \\ \vdots & \ddots & \ddots & \\ \beta^{N-2} & \ldots & & \\ \beta^{N-1} & \ldots & \beta & 1 \end{bmatrix}.$$

Quindi

$$\|A^{-1}\|_1 = \frac{1}{1+\lambda} \sum_{k=0}^{\infty} \beta^k = 1\,.$$

In conclusione, lo schema (5.25), per ogni scelta dei passi h e k è incondizionatamente stabile (e ciò è vero anche in qualsiasi norma $p, 1 \leq p \leq \infty$ e norma pesata).

5.4 Problemi di tipo iperbolico del secondo ordine

L'equazione protitipo per questo tipo di problemi è quella delle onde 1-dimensionale:

$$\frac{\partial^2 u}{\partial x^2} - c^2 \frac{\partial^2 u}{\partial x^2} = 0\,, \tag{5.26}$$

ove $c > 0$ è la velocità di propagrazione dell'onda.

L'integrale generale (o soluzione generale) dell'equazione (5.26)

$$u(x,t) = f_1(x + ct) + f_2(x - ct) \tag{5.27}$$

con f_1, f_2 funzioni arbitrarie differenziabili. In particolare, nel caso $u(x,t)$ sia soluzione classica, richiederemo che $f_1,\ f_2 \in \mathcal{C}^2(\mathbb{R})$. Alunce osservazioni

- Se associamo inoltre le condizioni iniziali sullo spostamento e sulla velocità

$$\begin{cases} u(x,0) = f(x) \\ u_t(x,0) = g(x) \end{cases} \tag{5.28}$$

la soluzione diventa

$$u(x,t) = \frac{1}{2}\left[f(x+ct) + f(x-ct)\right] + \frac{1}{2c}\int_{x-ct}^{x+ct} g(\xi)d\xi \ , \tag{5.29}$$

detta *formula di D'Alambert*. In particolare, se $f \in \mathcal{C}^2(\mathbb{R})$ e $g \in \mathcal{C}^1(\mathbb{R})$ allora $u(x,t)$ in (5.29) viene detta *soluzione classica* dell'equazione delle onde.

- La soluzione $u(x,t)$ nel generico punto (x_0,t_0) dipende solo dai dati iniziali sul segmento $[x_0-ct_0, x_0+ct_0]$ (vedi Figura 5.10 (sopra)). Tale segmento viene detto *dominio di dipendenza* (o anche intervallo di dipendenza) del (x_0,t_0). Le rette $x = x_0+c(t-t_0)$ e $x = x_0 - c(t-t_0)$ sono le due rette caratteristiche. Pertanto, la conoscenza della soluzione nell'intervallo $[x_0 - ct_0, x_0 + ct_0]$, consente di definire *univocamente* la soluzione solo sul triangolo di vertici i punti $(x_0 - ct_0, 0)$, $(x_0 + ct_0, 0)$, (x_0, t_0).

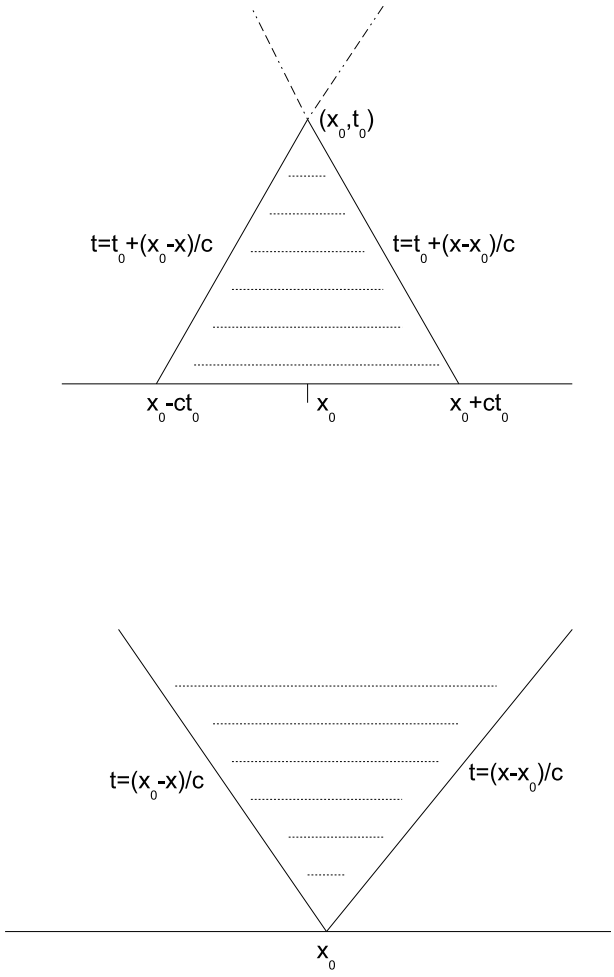

Figura 5.10: Dominio di dipendenza per equazioni iperboliche.

Domanda: il punto $x = x_0$ quali valori influenza della soluzione $u(x,t)$? Quelli della regione infinita individuata dalle semirette $x = x_0 - ct$ e $x = x_0 + ct$ (vedasi Figura 5.10 (sotto)). Osserviamo che per $t \to \infty$ la soluzione in generale non va verso lo stato stazionario.

5.4.1 Uno schema numerico alle differenze per l'equazione delle onde

Il problema che risolveremo è il seguente

$$\begin{cases} u_{tt} - c^2 u_{xx} = 0 \quad \text{con} \quad 0 < x < 1, \;\; t > 0 \\ u(x,0) = f(x) \\ u_t(x,0) = 0 \end{cases} \tag{5.30}$$

con l'aggiunta delle condizioni al bordo $u(0,t) = \alpha(t), \quad u(1,t) = \beta(t)$ che consentono di trovare univocamente la soluzione su tutta la striscia $D = \{(x,t) : \ 0 < x < 1, \, t > 0\}$.

Per avere la soluzione classica, chiederemo che $f, \alpha, \beta \in \mathcal{C}^2(D)$, $g \in \mathcal{C}^1(D)$ e siano soddisfatte le condizioni (ovvie!)

$$\alpha(0) = f(0), \ \alpha'(0) = g(0), \ \alpha^{"}(0) = c^2 f^{"}(0) \ ,$$

$$\beta(0) = f(1), \ \beta'(0) = g(1), \ \beta^{"}(0) = c^2 f^{"}(1) \ .$$

Inoltre, se $\alpha(0) \neq f(0)$ allora la soluzione u avrà una discontinuità lungo tutta la retta caratteristica $x = ct$. Questo è un fatto generale per le equazioni di tipo iperbolico: a dati iniziali regolari corrispondono soluzioni regolari mentre a dati iniziali discontinui soluzioni discontinue lungo le caratteristiche.

Nel caso che $u \in \mathcal{C}^4(D)$, posto $y = ct$ possiamo riscrivere il problema come segue

$$\begin{cases} u_{xx}(x,y) - u_{yy}(x,y) = 0 \\ u(x,0) = f(x) \\ u_y(x,0) = \dfrac{1}{c} g(x) := g_1(x) \\ u(0,y) = \alpha\left(\dfrac{y}{c}\right) := \alpha_1(y) \\ u(1,y) = \beta\left(\dfrac{y}{c}\right) := \beta_1(y) \end{cases} \tag{5.31}$$

con D che diventa la regione $R = \{0 < x < 1, \ \ y > 0\}$.

Ora, come fatto in altri casi, reticoliamo R prendendo N punti equispaziati di passo h lungo x e punti equispaziati di passo k lungo y ottenendo

$$\begin{cases} u_{xx}(x_i, y_j) = u_{yy}(x_i, y_j), \quad i = 1, \ldots, N-1, \ \ j = 1, 2, \ldots \\ u(x_i, 0) = f(x_i) \, , \ \ i = 0, \ldots, N \\ u_y(x_i, 0) = g_1(x_i) \\ u(0, y_j) = \alpha_1(y_j) \, , \ \ j = 0, 1, \ldots \\ u(1, y_j) = \beta_1(y_j) \end{cases} \tag{5.32}$$

Quindi, approssimiamo le derivate presenti in (5.32) con le formule alle differenze finite centrali viste nel paragrafo **??**, che riportiamo nuovamente per semplicità di trattazione

$$\begin{aligned} u_{xx}(x_i, y_j) &= \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2} + \mathcal{O}(h^2) \\ u_{yy}(x_i, y_j) &= \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{k^2} + \mathcal{O}(k^2) \\ u_y(x_i, 0) &= \frac{u_{i,1} - u_{i,-1}}{2k} + \mathcal{O}(k^2) \ . \end{aligned}$$

Il sistema (5.32) diventa

$$\begin{cases} \dfrac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2} = \dfrac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{k^2} \\ u_{i,0} = f_i \\ \dfrac{u_{i,1} - u_{i,-1}}{2k} = g_{1,i} \\ u_{0,j} = \alpha_{1,j} \\ u_{N,j} = \beta_{1,j} \ . \end{cases} \tag{5.33}$$

commettendo un errore di troncamento locale in (x_i, y_j) che è $\mathcal{O}(h^2) + \mathcal{O}(k^2)$.

Si noti che $u_{i,-1}$ (-1 è un livello fittizio) si determina usando la quarta relazione in (5.33)

$$u_{i,-1} = u_{i,1} - 2kg_{1,i} \ .$$

Detto quindi $\lambda = k/h$, il sistema (5.33) può scriversi come

$$\begin{cases} u_{i,j+1} = \lambda^2(u_{i-1,j} + u_{i+1,j}) + 2(1-\lambda^2)u_{i,j} - u_{i,j-1} \\ u_{i,0} = f_i \\ u_{i,-1} = u_{i,1} - 2kg_{1,i} \\ u_{0,j} = \alpha_{1,j} \\ u_{N,j} = \beta_{1,j} \ . \end{cases} \tag{5.34}$$

pervenendo allo schema esplicito

$$\begin{cases} u_{i,0} = f_i \\ u_{i,1} = \dfrac{\lambda^2}{2}(f_{i-1} + f_{i+1}) + (1-\lambda^2)f_i + kg_{1,i} \ , \quad i = 1, \ldots, N-1 \\ u_{0,j} = \alpha_{1,j} \\ u_{i,j+1} = \lambda^2(u_{i-1,j} + u_{i+1,j}) + 2(1-\lambda^2)u_{i,j} - u_{i,j-1} \ , \ \ i = 1, \ldots, N-1 \ , \ j = 1, 2, \ldots \\ u_{N,j} = \beta_{1,j} \ . \end{cases} \tag{5.35}$$

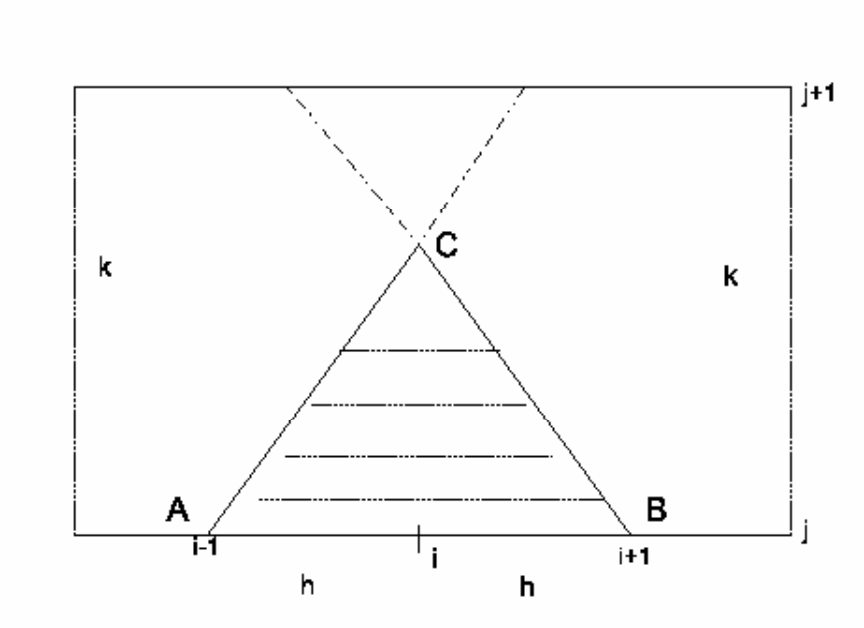

Figura 5.11: Dominio di dipendenza per equazioni iperboliche.

Come si nota in Figura 5.11, se $\lambda > 1$ lo schema è impreciso (leggi, instabile), perchè la conoscenza della soluzione sul segmento (A,B) determina univocamente la soluzione solo sul triangolo ABC, mentre la soluzione numerica, $u_{i,j+1}$ al passo $j+1$-esimo, non appartiene al triangolo ABC.

Consideriamo, come prima, il vettore $U_j = (u_{1,j}, \ldots, u_{N-1,j})^T$ che indica la soluzione nei nodi interni $1, \ldots, N-1$ al tempo t_j. Lo schema (5.35) si può allora riscrivere vettorialmente come

$$\begin{cases} U_0 = [f(x_i)] \\ \\ U_1 = \left[\dfrac{\lambda^2}{2}(f_{i-1} + f_{i+1}) + (1-\lambda^2)f_i + kg_{1,i}\right] \\ \\ U_{j+1} = AU_j - U_{j-1} + \lambda^2 b_j, \quad j = 1, 2, \ldots \end{cases} \tag{5.36}$$

con

$$A = \begin{bmatrix} 2(1-\lambda^2) & \lambda^2 & & 0 \\ \lambda^2 & \ddots & \ddots & \\ & \ddots & \ddots & \\ 0 & & \lambda^2 & 2(1-\lambda^2) \end{bmatrix} \in \mathbb{R}^{(N-1)\times(N-1)} \quad b_j = \begin{bmatrix} \alpha_{1,j} \\ 0 \\ \vdots \\ \beta_{1,j} \end{bmatrix} \in \mathbb{R}^{(N-1)}.$$

La relazione tra U_{j+1}, U_j e U_{j-1} si può anche riscrivere

$$U_{j+1} = (2I - \lambda^2 T)U_j - U_{j-1} + \lambda^2 b_j \tag{5.37}$$

con $T = \texttt{tridiag}(-1, 2, -1)$ (la "solita" matrice tridiagonale) di ordine $N-1$.

Posto poi,

$$V_j = \begin{pmatrix} U_j \\ U_{j-1} \end{pmatrix}, \quad B = \begin{pmatrix} A & -I \\ I & 0 \end{pmatrix},$$

cosicché

$$\begin{pmatrix} U_{j+1} \\ U_j \end{pmatrix} = \begin{pmatrix} A & -I \\ I & 0 \end{pmatrix} \begin{pmatrix} U_j \\ U_{j-1} \end{pmatrix} + \lambda^2 \begin{pmatrix} b_j \\ 0 \end{pmatrix} = \begin{pmatrix} AU_j - U_{j-1} + \lambda^2 b_j \\ U_j \end{pmatrix}$$

da cui otteniamo il sistema

$$V_{j+1} = BV_j + \lambda_j \,. \tag{5.38}$$

In definitiva per studiare la stabilità del metodo dobbiamo studiare gli autovalori μ_i della matrice B.

Ora,

$$\begin{pmatrix} A & -I \\ I & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \mu \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$$

equivale all'equazione $\mu^2 u_2 - A\mu u_2 + u_2 = 0$ ovvero

$$(\mu^2 I - \mu A + I)u_2 = 0 \,. \tag{5.39}$$

Ricordando poi che$A = 2I - \lambda^2 T$, otteniamo

$$[(\mu^2 - 2\mu + 1)I + \mu\lambda^2 T]u_2 = 0 \,.$$

Posto

$$\alpha = -\frac{(\mu - 1)^2}{\mu\lambda^2}$$

il problema della ricerca degli autovalori μ_i di B consiste dapprima nella ricerca degli autovalori α_i della matrice T. Infatti la (5.39) diventa

$$(T - \alpha I)u_2 = 0 \,. \tag{5.40}$$

Per studiare gli autovalori della matrice tridiagonale T, osserviamo che se y è autovettore di T associato all'autovalore λ, varrà

$$y_i + (\lambda - 2)y_{i+1} + y_{i+2} = 0, \quad i = 0, \ldots, N-2, \quad y_0 = y_N = 0 \,. \tag{5.41}$$

Essendo T simmetrica e diagonalmente dominante (non in senso stretto!), essa ha autovalori reali e positivi. Per il teorema di Gerschgorin essi staranno nell'intervallo $|\lambda-2| \le 2$ ovvero $\lambda \in [0,4]$. Ma, l'equazione alle differenze (5.41) ha equazione caratteristica

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

con soluzioni

$$\mu_{1,2} = 1 - \frac{\lambda}{2} \pm i\sqrt{\lambda\left(1 - \frac{\lambda}{4}\right)}\,.$$

Ma $|\mu_{1,2}| = 1$, essendo $\lambda \in [0,4]$, posto

$$\begin{aligned} \cos\varphi &= 1 - \tfrac{\lambda}{2} \\ \sin\varphi &= \sqrt{\lambda\left(1 - \tfrac{\lambda}{4}\right)} \end{aligned}$$

allora $\mu_{1,2} = \mathrm{e}^{\pm i\varphi}$.

Ora, dalla teoria sulle equazioni alle differenze, l'equazione (5.41) ha la soluzione esprimibile come

$$y_k = a\cos k\varphi + b\sin k\varphi\ .$$

Le costanti a, b si determinano dal fatto che $y_0 = y_N = 0$, ottendo $a = 0$ e $b\sin(N\varphi) = 0$. Il parametro b deve essere diverso da zero, pertanto l'angolo φ si ottiene risolvendo

$$\sin(N\varphi) = 0$$

le cui soluzioni sono $\varphi = k\pi/N,\ \ k = 1, \ldots, N-1\,.$

In definitiva gli autovalori della matrice T sono

$$\lambda(T) = 2(1 - \cos\varphi) = 4\sin^2\left(\frac{k\pi}{2N}\right) := \alpha_k, \quad k = 1, \ldots, N-1\,. \tag{5.42}$$

Per completezza, essendo $A = 2I - \lambda^2 T$, gli autovalori di A saranno $\ell_k = 2 - \lambda^2\alpha_k,\ \ k = 1, \ldots, N-1$.

Gli autovalori di B sono

$$\mu_i = \left(1 - \frac{\lambda^2\alpha_i}{2}\right) \pm \sqrt{\left(1 - \frac{\lambda^2\alpha_i}{2}\right)^2 - 1}\,, \quad \alpha_i \text{ autovalore di } T \tag{5.43}$$

Osserviamo che μ_i non deve essere reale, altrimenti una delle radici risulta certamente essere > 1 (essendo il loro prodotto uguale ad 1). Infatti, se $\mu_i \in \mathbb{C}$ allora necessariamente $|\mu_i| = 1$.

Pertanto, poichè gli autovalori sono tutti distinti, con $|\mu_i| = 1$, ciò garantisce che i moduli di tutte le componenti del vettore perturbazione E_{j+1} sono limitate uniformemente rispetto a j ed a N

$$E_{j+1} = B^j\ E_1$$

e ciò equivale a chiederci *quando lo schema è stabile?* Ma essendo $|\mu_i|$ distinti, B è diagonalizzabile

$$B = H\Lambda H^{-1}$$

con H matrice degli autovettori di B e Λ matrice diagonale contenente gli autovalori. Essendo $B^j = H\Lambda^j H^{-1}$ segue che

$$E_{j+1} = H\Lambda^j H^{-1} E_1 \, .$$

Posto $\tilde{E}_{j+1} = H^{-1}E_{j+1}$, perveniamo al sistema disaccoppiato

$$\tilde{E}_{j+1} = \Lambda^j \tilde{E}_1 \, ,$$

o per componenti

$$(\tilde{E}_{j+1})_i = \mu_i^j (\tilde{E}_1)_i, \;\; i = 1, \ldots, N-1 \, .$$

In modulo

$$\left|(\tilde{E}_{j+1})_i\right| = \left|(\tilde{E}_1)_i\right|, \;\; i = 1, \ldots, N-1 \, .$$

Se $0 < \dfrac{\lambda^2 \alpha_i}{2} < 2$ abbiamo $\mu_i \in \mathbb{C}$, ovvero

$$\lambda^2 < \frac{1}{\sin^2\left(\frac{(N-1)\pi}{2N}\right)} \tag{5.44}$$

con denominatore in modulo < 1. Ma essendo $\lim\limits_{N\to\infty} \sin^2\left(\dfrac{(N-1)\pi}{2N}\right) = 1$ potremo senz'altro dire che lo schema numerico è stabile se $\lambda \leq 1$, ovvero quando $k \leq h$.

5.5 Equazioni di tipo parabolico

L'equazione del calore, come visto, è l' equazione di evoluzione che fa, per così dire, da rappresentante di questa classe di equazioni alle derivate parziali.

Nella sua formulazione più semplice essa è

$$u_t - u_{xx} = 0 \tag{5.45}$$

che rappresenta la diffusione di calore in un mezzo.

Facciamo ora vedere che le curve caratteristiche di questa equazione sono le rette $t = cost$. Questo implica, che non sarà possibile prescrivere arbitrariamente sull'asse x il valore di u e di u_t.

Consideriamo l'equazione quasi-lineare del primo ordine

$$au_x + bu_y = c \, . \tag{5.46}$$

Supponiamo di conoscere $u(x,y)$ lungo una curva γ di equazioni parametriche

$$\begin{cases} x = \varphi(\tau)\,, & \tau \in I \\ y = \psi(\tau)\,, & \varphi, \psi \in \mathcal{C}^1(I) \end{cases}$$

cosicché $u(\varphi(\tau), \psi(\tau)) = f(\tau)$. Ad esempio, in $\mathbb{R}^2$, se γ fosse il cerchio unitario, avremmo $x = \cos\tau,\ y = \sin\tau\,.$

Su γ i coefficienti a, b, c sono pure funzioni di τ ovvero

$$\begin{array}{l} a = a(\varphi(\tau), \psi(\tau), f(\tau)) \\ b = b(\varphi(\tau), \psi(\tau), f(\tau)) \\ c = c(\varphi(\tau), \psi(\tau), f(\tau)) \end{array} .$$

Inoltre

$$\frac{d}{d\tau}u(x,y) = u_x\frac{dx}{d\tau} + u_y\frac{dy}{d\tau} = f'(\tau)\,.$$

Domanda: la conoscenza di u, quindi di $\frac{du}{d\tau}$ su γ, è sufficiente per definire $u_x = u_x(\varphi(\tau), \psi(\tau))$ e $u_y = u_y(\varphi(\tau), \psi(\tau))$ su γ?

Ciò è equivalente a chiedere quando il sistema

$$\begin{cases} au_x + bu_y = c \\ \frac{dx}{d\tau}u_x + \frac{dy}{d\tau}u_y = f'(\tau) \end{cases} \tag{5.47}$$

ha unica soluzione. Ma ciò è equivalente a chiedere che il determinante sia diverso da zero. Ovvero che

$$a\frac{dy}{d\tau} - b\frac{dx}{d\tau} \neq 0\,,$$

che implica

$$\frac{dy}{dx} \neq \frac{b}{a}\,, \tag{5.48}$$

ovvero quando il coefficiente angolare della tangente alla curva γ in $(x,y) = (\varphi(\tau), \psi(\tau)$ è diverso da b/a.

Definizione 15. *Le curve γ del piano (x,y) che hanno coefficiente angolare della retta tangente uguale a b/a sono dette* `curve caratteristiche` *dell'equazione differenziale data.*

Esempio 17. Consideriamo l'equazione

$$\begin{cases} u_x + bu_y = -ku \quad x > 0, y > 0 \\ u(x,0) = u_0, \quad x > 0 \\ u(0,y) = 0 \end{cases} \tag{5.49}$$

con b, k, u_0 costanti assegnate. Il determinante relativo a (**??**) è

$$\begin{vmatrix} 1 & b \\ \frac{dx}{d\tau} & \frac{dy}{d\tau} \end{vmatrix} = 0$$

da cui $\dfrac{dy}{dx} = b$ e le linee caratteristiche sono quindi le rette $y(x) = bx + cost$, come disegnato in figura **??**.

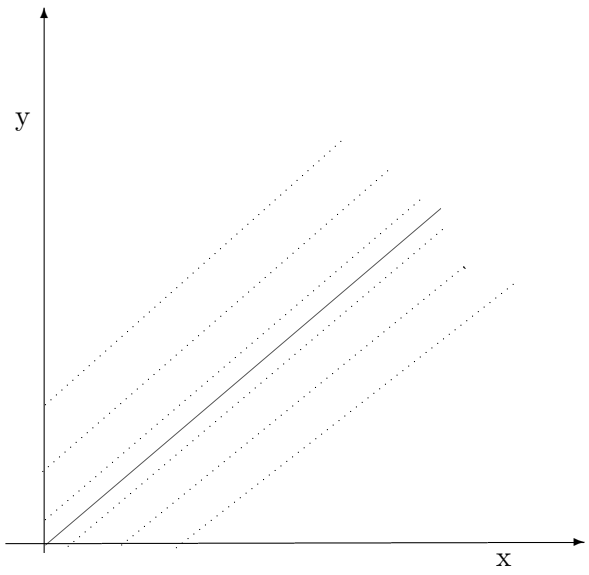

Figura 5.12: Le curve caratteristiche dell'esempio **??**.

Esempio 18. Consideriamo il sistema 2×2

$$\begin{cases} u_x - v_y = 0 \\ v_x - u_y = 0, \end{cases} \tag{5.50}$$

Prima di discutere chi sono le linee caratteristiche per il sistema dato, discutiamo come operare nel caso di sistemi di 2 equazioni quasi-lineari del primo ordine. Consideriamo infatti

$$\begin{cases} a_1 u_x + b_1 u_y + c_1 v_x + d_1 v_y = f_1 \\ a_2 u_x + b_2 u_y + c_2 v_x + d_2 v_y = f_2 \end{cases} \tag{5.51}$$

il cui sistema derivato associato è

$$\begin{cases} \frac{du(x,y)}{d\tau} = u_x \frac{dx}{d\tau} + u_y \frac{dx}{d\tau} \\ \frac{dv(x,y)}{d\tau} = v_x \frac{dx}{d\tau} + v_y \frac{dx}{d\tau} \end{cases} \tag{5.52}$$

Pertanto, per determinare le caratteristiche dovremo risolvere il sistema

$$\begin{pmatrix} a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ \frac{dx}{d\tau} & \frac{dy}{d\tau} & 0 & 0 \\ 0 & 0 & \frac{dx}{d\tau} & \frac{dx}{d\tau} \end{pmatrix} \begin{pmatrix} u_x \\ u_y \\ v_x \\ v_y \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ \frac{du}{d\tau} \\ \frac{dv}{d\tau} \end{pmatrix}$$

se e solo se il suo determinante sarà diverso da zero. Il determinante è

$$A\left(\frac{dy}{d\tau}\right)^2 - B\left(\frac{dx}{d\tau}\right)\left(\frac{dy}{d\tau}\right) + C\left(\frac{dx}{d\tau}\right)^2$$

ove $A = a_1c_2 - a_2c_1$, $B = a_1d_2 - a_2d_1 + b_1c_2 - b_2c_1$ e $C = b_1d_2 - b_2d_1$. Pertanto le linee caratteristiche, dy/dx, che rendono nullo il determinante, (dividendo per $(dx/d\tau)^2$) sono le soluzioni di

$$A\left(\frac{dy}{dx}\right)^2 - B\left(\frac{dy}{dx}\right) + C = 0 \tag{5.53}$$

A seconda del segno del discriminante, il sistema risulterà, al solito, ellittico (se negativo), parabolico (se nullo) o iperbolico (se positivo).

$\diamond\diamond$

Il sistema dato nell'esempio, è di tipo iperbolico. Come è facile provare, l'equazione delle caratteristiche è

$$\left(\frac{dy}{dx}\right)^2 - 1 = 0$$

le cui soluzioni sono $dy/dx = \pm 1$. Pertanto, il sistema ammette due famiglie di caratteristiche: $y_1(x) = x + cost$, $y_2(x) = -x + cost$ che sono rette parallele rispettivamente alle due bisettrici dei quadranti 1,3 e 2,4, rispettivamente, del piano (x, y).

$\diamond$

Nel caso di un'equazione del secondo ordine

$$au_{xx} + bu_{xy} + cu_{yy} = f$$

la domanda equivalente al caso uni-dimensionale consiste nel chiederci sotto quali condizioni la conoscenza di u sulla curva parametrica γ

$$\gamma \; : \; \begin{cases} x = \varphi(\tau)\,, \\ y = \psi(\tau)\,, \end{cases} \quad \varphi, \psi \in \mathcal{C}^2(I)$$

per qualche intervallo I, e la conoscenza della derivata lungo la normale alla curva γ, $\frac{\partial u}{\partial n_y}$, consente di definire su γ le derivate u_{xx}, u_{xy} e u_{yy}? La risposta è, se e solo se il sistema

$$\begin{pmatrix} a & b & c \\ \frac{dx}{d\tau} & \frac{dy}{d\tau} & 0 \\ 0 & \frac{dx}{d\tau} & \frac{dy}{d\tau} \end{pmatrix} \begin{pmatrix} u_{xx} \\ u_{xy} \\ u_{yy} \end{pmatrix} = \begin{pmatrix} f \\ \frac{du_x}{d\tau} \\ \frac{du_y}{d\tau} \end{pmatrix} \tag{5.54}$$

risulta non singolare. Ovvero, chiederemo

$$a\left(\frac{dy}{dx}\right)^2 - b\frac{dy}{dx} + c \neq 0$$

Come già osservato, a seconda del segno del discriminante $\Delta = b^2 - 4ac$, l'equazione risulterà essere ellittica se $\Delta < 0$, parabolica se $\Delta = 0$ oppure iperbolica se $\Delta > 0$.

Tornando all'equazione del calore $u_t - u_{xx} = 0$, che è un'equazione del secondo ordine, ha caratteristiche che si ottengo risolvendo

$$\frac{dt}{dx} = 0$$

che hanno soluzione $t = cost$. Questo spiega perchè, come detto sopra, le caratteristiche sono rette costanti.

Risolvendo l'equazione

$$u_t = u_{xx} \tag{5.55}$$
$$u(x,0) = f(x) \quad -\infty < x < +\infty \tag{5.56}$$

si dimostra che la soluzione esatta è

$$u(x,t) = \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{+\infty} \mathrm{e}^{-(x-s)^2/4t} f(s) ds \,. \tag{5.57}$$

Vale la pena osservare, che il valore della soluzione $u(x,t)$ in (x_0,t_0) dipende dal dato iniziale sull'intero asse $f(x)$. Detto altrimenti, il `dominio di dipendenza` di (x_0,t_0) è tutto $\mathbb{R}$. Inoltre, per $t \to +\infty$ la soluzione decade esponenzialmente.

`Osservazione.` L'equazione (5.57), ha un effetto regolarizzante. Ovvero, anche se $f(s)$ è solo limitata e continua a tratti su x, la soluzione $u(x,t) \in \mathcal{C}^\infty([-\infty,+\infty], t > 0)$. Inoltre,

$$\lim_{x \to \xi,\, t \to 0} u(x,t) = f(\xi) \,.$$

Questo effetto, come si vedrà nelle applicazioni, è tipico delle equazioni paraboliche.

5.5.1 Schemi numerici per l'equazione del calore

Riconsideriamo il problema del filo metallico, lungo 1, termicamente isolato, con distribuzione iniziale di temperatura nota, $f(x)$, che agli estremi $0, 1$ è mantenuto a temperature note, $g_0(t)$ e $g_1(t)$, rispettivamente.

Formalmente, si tratta di risolvere l'equazione,

$$\begin{cases} u_t = u_{xx} & 0 < x < 1,\ t > 0 \\ u(x,0) = f(x) & 0 \le x \le 1 \\ u(0,t) = g_0(x) & t > 0 \\ u(1,t) = g_1(x) & t > 0 \end{cases} \tag{5.58}$$

Questo è un tipico problema di propagazione ai valori iniziali (vedi Fig. 3).

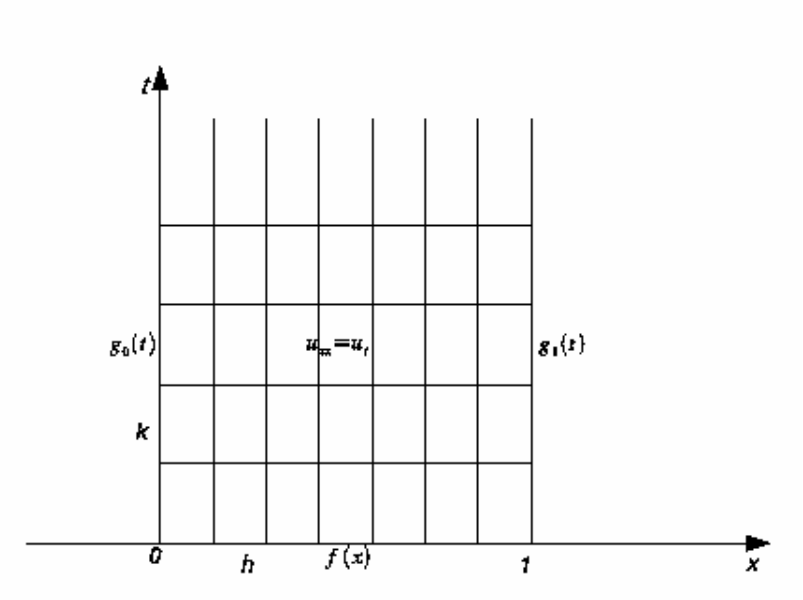

Figura 5.13: Discretizzazione del problema del filo metallico

1. **Schema alle differenze.** Prendiamo su $[0, 1]$, i punti equispaziati di passo $h = 1/N$, x_i, $i = 0, ..., N$ e discretizziamo u_t e u_{xx} come segue:

$$u_t(x_i, t_j) = \frac{u_{i,j+1} - u_{i,j}}{k} + \mathcal{O}(k), \quad u_{xx}(x_i, t_j) = \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2} + \mathcal{O}(h^2)\,.$$

Posto $\lambda = k/h^2$, l'equazione differenziale al tempo t_{j+1} si può scrivere come

$$u_{i,j+1} = \lambda u_{i-1,j} + (1 - 2\lambda)u_{i,j} + \lambda u_{i+1,j}, \ \ i = 1, \ldots, N-1\,. \tag{5.59}$$

L'errore di troncamento locale in (x_i, t_j) sarà $\mathcal{O}(h^2) + \mathcal{O}(k)$.

Otteniamo così uno schema `esplicito`, il cui 'stencil' è facilmente individuabile dal reticolo di Fig. 5.3, si ottiene prendendo in esame, al livello j, i punti (x_k, t_j), $k = i-1, i, i+1$ per determinare il valore della soluzione nel punto (x_i, t_{j+1}).

Complessivamente avremo,

$$\begin{cases} u_{i,0} = f_i & i = 0, \dots, N \\ u_{0,j+1} = g_{0,j+1}, & j = 0, 1, \dots \\ u_{i,j+1} = \lambda u_{i-1,j} + (1-2\lambda)u_{i,j} + \lambda u_{i+1,j}\,, & i = 1, \dots, N-1 \\ u_{N,j+1} = g_{1,j+1}, \end{cases} . \tag{5.60}$$

Definendo quindi i vettori $U_j = (u_{1,j}, \dots, u_{N-1,j})^T$, $v_j = (u_{0,j}, \dots, u_{N,j})^T$ e la matrice

$$A = \begin{bmatrix} 1-2\lambda & \lambda & 0 & \dots & \\ \lambda & 1-2\lambda & \lambda & & \\ & \ddots & & \ddots & 0 \\ 0 & \ddots & \ddots & & \lambda \\ & & & \lambda & 1-2\lambda \end{bmatrix}$$

possiamo riscrivere lo schema in forma vettoriale

$$U_{j+1} = AU_j + \lambda v_j, \quad j = 0, 1, \dots . \tag{5.61}$$

Indichiamo con E_0 il vettore delle perturbazioni sui dati iniziali U_0 (non sui valori al bordo!). Al passo j, l'errore sarà governato dall'equazione

$$E_{j+1} = AE_j, \;\; j \geq 0\,.$$

Per la stabilità dello schema numerico chiederemo, come al solito, che per ogni norma matriciale indotta $\|A\| \leq 1$. Nel caso in cui, gl'autovalori di A siano distinti, la condizione si raffina chiedendo che $\rho(A) \leq 1$. Pertanto, per studiare la stabilità dello schema (5.61), dobbiamo studiare lo spettro di A. Ma è facile osservare che $A = I - \lambda T$ con la matice T che è la solita matrice tridiagonale di ordine $N-1$ che, con notazione Matlab/Octave possiamo scrivere semplicemente come

$$T = \texttt{diag}(-\texttt{ones}(\texttt{N}-2,1),-1) + \texttt{diag}(2*\texttt{ones}(\texttt{N}-1,1)) + \texttt{diag}(-\texttt{ones}(\texttt{N}-2,1),1).$$

Gl'autovalori di T, come visto nella sessione precedente, sono $\alpha_i = 4\sin^2\left(\dfrac{i\pi}{2N}\right)$, $i = 1, \dots, N-1$, quindi gl'autovalori di A sono $\mu_i = 1 - 4\lambda\sin^2\left(\dfrac{i\pi}{2N}\right)$, $i = 1, \dots, N-1$ che sono distinti. Pertanto, per la stabilità chiederemo

$$\rho(A) \leq \max\{1, 4\lambda - 1\}\,.$$

In conclusione, lo schema è stabile se e solo se

$$\lambda \leq \frac{1}{2} \Longleftrightarrow k \leq \frac{h^2}{2}\,. \tag{5.62}$$

2. **Schema di Crank-Nicolson.** La condizione (5.62) è troppo restrittiva. Possiamo ovviare a ciò collocando l'equazione differenziale nel punto $(x_i, t_{j+\frac{1}{2}})$ con $t_{j+\frac{1}{2}} = t_j + \frac{k}{2}$. Siccome i nodi del reticolo sono (x_i, t_j) dobbiamo estendere u a questi nodi *fittizi* nel seguente modo

$$u_{xx}(x_i, t_{j+\frac{1}{2}}) = \frac{1}{2}\left[u_{xx}(x_i, t_{j+1}) + u_{xx}(x_i, t_j)\right] + \mathcal{O}(k^2)\,.$$

Inoltre, approssimeremo le derivate parziali seconde $u_{xx}(x_i, t_{j+1})$, $u_{xx}(x_i, t_j)$ con le usuali approssimazioni del secondo ordine .

Per u_t useremo

$$u_t(x_i, t_{j+\frac{1}{2}}) = \frac{u_{i,j+1} - u_{i,j}}{2\left(\frac{k}{2}\right)} + \mathcal{O}(k^2)\,.$$

Otteniamo così uno schema di tipo `implicito`

$$-\lambda u_{i-1,j+1} + 2(1+\lambda)u_{i,j+1} - \lambda u_{i+1,j+1} = \lambda u_{i-1,j} + 2(1-\lambda)u_{i,j} + \lambda u_{i+1,j}\,. \quad (5.63)$$

In forma matriciale,

$$AU_{j+1} = BU_j + \lambda(v_{j+1} + v_j), \quad j = 0, 1, \ldots \quad (5.64)$$

con

$$A = \begin{bmatrix} 2(1+\lambda) & -\lambda & 0 & \ldots & \\ -\lambda & 2(1+\lambda) & -\lambda & & \\ & \ddots & & \ddots & 0 \\ 0 & \ddots & \ddots & & -\lambda \\ & & & -\lambda & 2(1+\lambda) \end{bmatrix}$$

$$B = \begin{bmatrix} 2(1-\lambda) & \lambda & 0 & \ldots & \\ \lambda & 2(1-\lambda) & \lambda & & \\ & \ddots & & \ddots & 0 \\ 0 & \ddots & \ddots & & \lambda \\ & & & \lambda & 2(1-\lambda) \end{bmatrix}.$$

Gl'autovalori μ_i di $A^{-1}B$ (che si ottengono da quelli della matrice tridiagonale T) sono

$$\mu_i = \frac{2 - 4\lambda \sin^2\left(\frac{i\pi}{2N}\right)}{2 + 4\lambda \sin^2\left(\frac{i\pi}{2N}\right)}, \quad i = 1, 2, \ldots, N-1 \quad (5.65)$$

con $\mu_i < 1$, $\forall\, i$.

Lo schema (5.63) è pertanto `incondizionatamente stabile`. Questo è lo *schema di Crank-Nicolson* per l'equazione del calore.

Osservazione. Se le condizioni iniziali e quelle al bordo non coincidono nei vertici $(0,0)$ e $(1,0)$, la soluzione $u(x,t)$ dovrebbe risultare discontinua in questi punti. Però, nei problemi parabolici le discontinuità non si propagano. Ponendo

$$\begin{aligned} u_{0,0} &= \frac{1}{2}\left\{\lim_{x\to 0} f(x) + \lim_{t\to 0} g_0(t)\right\} \\ u_{1,0} &= \frac{1}{2}\left\{\lim_{x\to 1} f(x) + \lim_{t\to 0} g_1(t)\right\} \end{aligned}$$

si verificherà che lo schema funziona ugualmente. In alternativa, si ignorano le discontinuità e si sceglie uno solo dei valori, ovvero $u(0,0) = f(0) \vee u(0,0) = g_0(0)$ oppure $u(1,0) = f(1) \vee u(1,0) = g_1(0)$.

3. **Metodo delle linee.** L'idea è di discretizzare tutte le variabili dell'equazione del calore tranne una. Anticipiamo che questo ci farà pervenire ad un sistema di equazioni differenziali ordinarie.

   Ad esempio, se consideriamo $u_t(x,t) = u_{xx}(x,t)$ e discretizziamo solo lungo x. In corrispondenza al punto x_i, la funzione sarà solo funzione di t, cioè $u(x_i,t) = u_i(t)$. Pertanto potremo discretizzare ognuna delle $u_i(t)$ come di consueto:

$$u_t(x_i,t) = \frac{du_i(t)}{dt} \approx \frac{1}{h^2}\left[u_{i-1}(t) - 2u_i(t) + u_{i+1}(t)\right], \quad i = 1,\ldots,N-1\,. \tag{5.66}$$

   Così procedendo, l'equazione del calore (5.45) più le condizioni iniziali e al bordo, equivarrà al sistema di ODEs ai valori iniziali:

$$\begin{cases} u_1'(t) = \frac{1}{h^2}\left[u_0(t) - 2u_i(1) + u_2(t)\right], & u_1(0) = f(h) \\ u_2'(t) = \frac{1}{h^2}\left[u_1(t) - 2u_i(2) + u_3(t)\right], & u_2(0) = f(2h) \\ & \ldots \\ u_{N-1}'(t) = \frac{1}{h^2}\left[u_{N-2}(t) - 2u_i(N-1) + u_N(t)\right], & u_1(0) = f((N-1)h) \\ u_0(t) = g_0(t) \\ u_N(t) = g_1(t) \end{cases} \tag{5.67}$$

   dove ora le incognite sono le $N-1$ funzioni $u_i(t)$, $i = 1,\ldots,N-1$, vedasi Fig. 5.14.

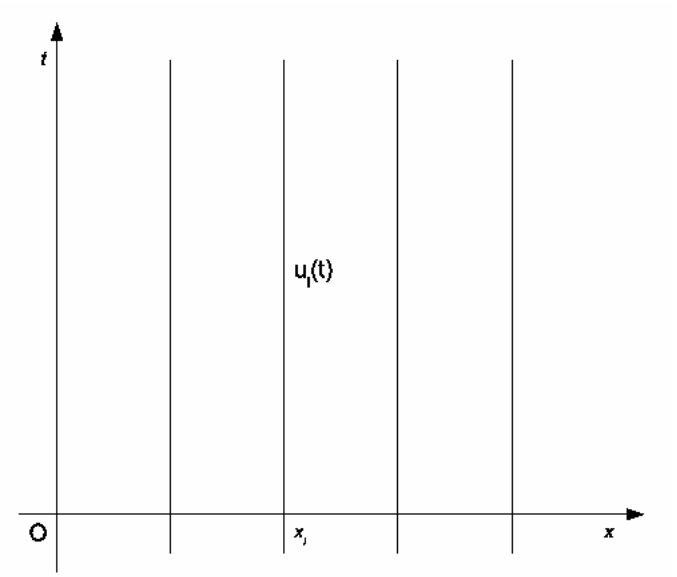

Figura 5.14: Reticolo per il metodo delle linee

Vettorialmente il sistema (5.67) diventa

$$U' = \frac{1}{h}(TU + v) \tag{5.68}$$

con $U = (u_1(t), \ldots, u_{N-1}(t))^T$, $v = (u_0(t), 0, \ldots, 0, u_N(t))^T$ e T è l'`opposta` della solita matrice tridiagonale T (introdotta precedentemente).

Il sistema (5.67) potrà poi essere risolto con uno dei metodi visti nel capitoli precedenti per sistemi di equazioni differenziali ordinarie, quali il metodo di Eulero o quello di Crank-Nicolson.

Circa gli autovalori di $\frac{1}{h^2}T$, questi sono noti e sono

$$\alpha_i = -\frac{4}{h^2}\sin\frac{i\pi}{2N}, \quad i = 1, ..., N-1$$

e sono *reali e negativi*. Il più piccolo è $\alpha_1 \approx -\left(\frac{\pi}{2N}\right)^2$ (si ottiene per $i = 1$ e poi approssimando il seno al primo ordine) mentre il più grande in modulo vale $\alpha_{N-1} \approx -\left(\frac{2}{h^2}\right)^2$ (si ottiene per $i = N-1$).

Risolvendo il sistema per valori di t che superano il transitorio ed arrivare così allo stato stazionario, la quantità

$$\frac{\alpha_{N-1}}{\alpha_1} \approx \frac{4N^2}{\pi^2} \tag{5.69}$$

rappresenta l'*indice di stiffness* del sistema. Pertanto, più N sarà grande e più stiff il sistema sarà.

5.5.2 Equazione del calore in due dimensioni spaziali

Nel corso di questa sessione indicheremo con $\mathbf{x} = (x, y)$ il generico punto del piano $\mathbb{R}^2$. L'equazione del calore corrispondente sarà

$$u_t = \gamma \Delta u(\mathbf{x}, t) \tag{5.70}$$

con γ costante positiva.

Ad esempio, il problema

$$\begin{cases} u_t = \gamma(u_{xx} + u_{yy}), \;\; 0 \le x, y \le 1 \\ u(x, y, 0) = 1 \\ u(x, 0, t) = u(x, 1, t) = 0 \\ u(0, y, t) = u(1, y, t) = 0 \end{cases} \tag{5.71}$$

è la tipica equazione in cui si chiede di determinare la soluzione $u(x, y, t)$ sulla "striscia" $[0, 1]^2 \times [0, \infty)$.

Presentiamo la soluzione mediante il `metodo delle linee`. Discretizziamo solo le variabili x ed y. Poniamo,

$$u(x_i, y_i, t) = u_{ij}(t)\,.$$

L'equazione (5.71) diventa

$$\begin{cases} \dfrac{du_{i,j}(t)}{dt} = \gamma \left(\dfrac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2} + \dfrac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{k^2} \right), \;\; 0 \le x, y \le 1 \\ u_{i,j}(0) = 1, \;\; i = 1, \ldots, N-1 \\ u_{0,j}(t) = u_{N,j}(t) = 0, \;\; j = 1, \ldots, M-1 \\ u_{i,0}(t) = u_{i,M}(t) = 0, \;\; t > 0 \end{cases} \tag{5.72}$$

dove si è assunto una discretizzazione con $N+1$ punti equispaziati di passo h lungo x e una discretizzazione con $M+1$ punti equispaziati di passo k lungo y.

Definiamo

$$\mathtt{u}(t) = \begin{array}{l} [u_{11}(t), \ldots, u_{1,M-1}(t), \\ u_{2,1}(t), \ldots, u_{2,M-1}(t) \\ \ldots\ldots \\ u_{N-1,1}(t), \ldots, u_{N-1,M-1}(t)] \end{array}$$

e la matrice

$$A = \begin{bmatrix} c & 1/k^2 & 0 & \ldots & 1/h^2 & 0 & \ldots \\ 1/k^2 & c & 1/k^2 & 0 & \ldots & 1/h^2 & \ldots \\ 0 & \ddots & \ddots & \ddots & & \ddots & \\ 1/h^2 & 0 & \ddots & & 0 & \ddots & \\ 0 & \ddots & \ddots & & \ddots & \ddots & 1/k^2 \\ 0 & & 1/h^2 & & & 1/k^2 & c \end{bmatrix}$$

con $c = \left(\frac{-2}{h^2} + \frac{-2}{k^2}\right)$. A è simmetrica di ordine $K = (M-1)\times(N-1)$ con *solo* 5 diagonali di elementi non nulli.

Il sistema (5.72), riscritto in forma matriciale diventa:

$$\begin{cases} \dfrac{d\mathtt{u}(t)}{dt} = \gamma A\mathtt{u}(t) + \mathtt{v}(t) \\ \mathtt{u}(0) = 1 \end{cases} \tag{5.73}$$

dove v(t) indica il vettore dei termini noti che contiene i dati assegnati al contorno ad ogni istante $t > 0$.

Il sistema (5.73), è un sistema di equazioni differenziali ordinarie lineari a coefficienti costanti del prim'ordine a valori iniziali che ha (circa) lo stesso grado di stiffness del corrispondente problema unidimensionale. Si devono pertanto preferire *metodi di tipo implicito* o l'*esponenziale di matrice*.

Se usassimo come metodo risolutivo il *metodo dei trapezi* (o di Crank-Nicolson),ad oagni step temporale dovremmo risolvere un sistema lineare di odine K (la cui matrice è pentadiagonale con 2 diagonali "distanti" dalle 3 principali). La soluzione di sifatto sistema con Gauss sarebbe molto costosa (proprio per la sparsità della matrice del sistema).

C'è un metodo alternativo, noto col nome di `metodo delle direzioni alternate` o `ADI`.

5.5.3 Metodo delle direzione alternate (ADI)

L'idea fondamentale è di ridurre il costo computazionale, ad ogni step temporale, mantenendo le caratteristiche di stabilità degli schemi impliciti. L'algoritmo ADI funziona come segue.

(a) Usiamo dapprima *Eulero implicito all'indietro* lungo x e *Eulero esplicito* lungo y per passare da $t = t_n$ e $t = t_{n+1/2} = t_{n+k/2}$, ovvero

$$u_{i,j}^{n+1/2} = u_{i,j}^{n} + \gamma\frac{k}{2}\left(\frac{u_{i+1,j}^{n+1/2} - 2u_{i,j}^{n+1/2} + u_{i-1,j}^{n+1/2}}{h^2} + \frac{u_{i,j+1}^{n} - 2u_{i,j}^{n} + u_{i,j-1}^{n}}{k^2}\right), \quad \begin{matrix} i = 1,\ldots,N-1, \\ j = 1,\ldots,M-1 \end{matrix}. \tag{5.74}$$

(b) Successivamente passiamo da $t_{n+1/2}$ a $t_{n+1} = t_{n+k}$ usando Eulero all'indietro lungo y ed Eulero esplicito lungo x

$$u_{i,j}^{n+1} = u_{i,j}^{n+1/2} + \gamma\frac{k}{2}\left(\frac{u_{i+1,j}^{n+1/2} - 2u_{i,j}^{n+1/2} + u_{i-1,j}^{n+1/2}}{h^2} + \frac{u_{i,j+1}^{n+1} - 2u_{i,j}^{n+1} + u_{i,j-1}^{n+1}}{k^2}\right), \quad \begin{matrix} i = 1,\ldots,N-1, \\ j = 1,\ldots,M-1 \end{matrix}. \tag{5.75}$$

Si dimostra che (5.74) e (5.75) è un metodo di ordine 2 (cioè $u_{i,j}^{n+1}$ è un'approssimazione di ordine 2) *incondizionatamente stabile* nonostante i metodi di Eulero usati siano di ordine 1.

La cosa interessante del metodo ADI è il costo computazionale. Ad ogni passo k lungo t, il sistema (5.74) si può interpretare come un insieme di $M-1$ sistemi di ordine $N-1$

$$Au_j^{n+1/2} = v_j^n, \quad j = 1, \ldots, M-1\,, \tag{5.76}$$

nelle incognite

$$u_j^{n+1/2} = \left(u_{1,j}^{n+1/2}, u_{2,j}^{n+1/2}, \ldots, u_{N-1,j}^{n+1/2}\right)^T\,,$$

con $A = I + \gamma\frac{k}{2h^2}T$ (T la solita matrice tridiagonale) che è simmetrica, tridiagonale e diagonalmente dominante. Pertanto, ciascun sistema (5.76) si può risolvere con il MEG senza pivoting con solo $3(N-2)$ moltiplicazioni e addizioni e $2N-3$ divisioni (perchè si può usare l'*Algoritmo di Thomas*). Analogo discorso possiamo fare per i sistemi (5.75).

Complessivamente il metodo ADI costa

$$(M-1)(3(N-2)+2N-3) = (M-1)(5N-9) \approx 5MN$$

Infine, ci ricordiamo che dobbiamo risolvere i sistemi (5.74). In definitiva, la complessità di ADI è $10\,MN$ flops.

5.6 Equazioni di tipo ellittico

Consideriamo, come prototipo, l'*equazione di Poisson*

$$-\Delta u = f\,, \tag{5.77}$$

con f funzione assegnata. Lungo il bordo Γ del dominio D possiamo associare le condizioni

$$\begin{aligned} u &= g \text{ Dirichlet} \\ \frac{du}{d\mathbf{n}} &= g \text{ Neumann} \\ au + b\frac{du}{d\mathbf{n}} &= g_1 \text{ miste}\,. \end{aligned}$$

1. Se f in (5.77) è di classe $\mathcal{C}^n(D)$ (o analitica), allora lo saranno anche le soluzioni u in ogni punto del dominio aperto D. Questo è il comportamento tipico delle equazioni di tipo ellittico.
2. Se aggiungiamo le condizioni di Dirichlet, la soluzione u non manterrà in generale le stesse proprietà di regolarità nel chiuso $D \cup \Gamma$ a meno che la funzione g e la curva Γ non siano esse stesse sufficientemente regolari (vedi Esempio :..).

Esempio 19.

5.6.1 Schemi alle differenze per il problema di Dirichlet

Appendice A

Integratori esponenziali

A.1 Esponenziale di matrice

Questa appendice è stata scritta dal Dott. Marco Caliari a cui va il mio personale ringraziamento per la fattiva collaborazione alla stesura di questi appunti.

Data una matrice quadrata $A \in \mathbb{R}^{N\times N}$, si definisce

$$\exp(A) = \sum_{j=0}^{\infty} \frac{A^j}{j!} .$$

Tale serie converge per qualunque matrice A, essendo A un operatore lineare tra spazi di Banach e avendo la serie esponenziale raggio di convergenza ∞. Se A e B sono *permutabili* (cioè $AB = BA$), allora

$$\exp(A+B) = \exp(A)\exp(B) .$$

A.2 Sistemi di ODEs

Data l'equazione differenziale

$$\begin{cases} u'(t) = au(t) + b(u(t)), \quad t > 0 \\ u(t_0) = u_0 \end{cases} \tag{A.1}$$

la soluzione può essere scritta analiticamente mediante la formula delle *variazioni delle costanti*

$$u(t) = e^{(t-t_0)a}u_0 + \int_{t_0}^{t} e^{(t-\tau)a} b(u(\tau))\mathrm{d}\tau . \tag{A.2}$$

Infatti, si ha

$$u'(t) = ae^{(t-t_0)a}u_0 + a\int_{t_0}^{t} e^{(t-\tau)a}b(u(\tau))\mathrm{d}\tau + e^{(t-t)a}b(u(t)) = au(t) + b(u(t)) \ .$$

Si osservi che

$$\begin{aligned}\int_{t_0}^{t} e^{(t-\tau)a}\mathrm{d}\tau &= -\frac{1}{a}\int_{t_0}^{t} -ae^{(t-\tau)a}\mathrm{d}\tau = -\frac{1}{a}e^{(t-\tau)a}\Big|_{t_0}^{t} = \\ &= -\frac{1}{a}\left(1 - e^{(t-t_0)a}\right) = (t-t_0)\frac{e^{(t-t_0)a}-1}{(t-t_0)a} = \\ &= (t-t_0)\varphi_1((t-t_0)a) \ ,\end{aligned}$$

ove

$$\varphi_1(z) = \frac{e^z - 1}{z} = \sum_{j=0}^{\infty}\frac{z^j}{(j+1)!} \ . \tag{A.3}$$

Consideriamo ora un sistema differenziale

$$\begin{cases} u'(t) = Au(t) + b(u(t)), \quad u \in \mathbb{R}^{N\times 1}, \ A \in \mathbb{R}^{N\times N} \\ u(t_0) = u_0 \end{cases} \tag{A.4}$$

Ancora, la soluzione esplicita può essere scritta come

$$u(t) = \exp((t-t_0)A)u_0 + \int_{t_0}^{t}\exp((t-\tau)A)b(u(\tau))\mathrm{d}\tau \ .$$

Consideriamo l'approssimazione $b(u(\tau)) \approx b(u(t_0)) = b(u_0)$. Allora

$$\begin{aligned}u(t) &\approx \exp((t-t_0)A)u_0 + \int_{t_0}^{t}\exp((t-\tau)A)b(u_0)\mathrm{d}\tau = \\ &= \exp((t-t_0)A)u_0 + (t-t_0)\varphi_1((t-t_0)A)b(u_0) \ .\end{aligned} \tag{A.5}$$

A.3 Integratori esponenziali

Il metodo *Eulero esponenziale* per la risoluzione di (A.1) è

$$u_{n+1} = \exp(kA)u_n + h\varphi_1(kA)b(u_n) \tag{A.6}$$

ove $u_n \approx u(t_n)$, $t_{n+1} = t_n + k$.

Proposizione 6. *Per il metodo di Eulero esponenziale* (A.6)*, se* $u(t_n) = u_n$*, si ha*

$$\|u(t_{n+1}) - u_{n+1}\| = \mathcal{O}(k^2)$$

e

$$u(t_{n+1}) = u_{n+1}, \quad \textit{se } b(u(t)) = b(u_0) \equiv b \ .$$

Dunque il metodo è esatto *per problemi lineari autonomi a coefficienti costanti, di ordine 1 altrimenti.*

Proof. Si ha

$$u_{n+1} = \exp(kA)u_n + \int_{t_n}^{t_{n+1}} \exp((t_{n+1}-\tau)A)g(t_n)\mathrm{d}\tau \ ,$$

ove si è posto $g(t) = b(u(t))$. Per la formula di variazioni delle costanti (A.2)

$$\begin{aligned} u(t_{n+1}) &= \exp(hA)u_n + \int_{t_n}^{t_{n+1}} \exp((t_{n+1}-\tau)A)g(\tau)\mathrm{d}\tau = \\ &= \exp(kA)u_n + \int_{t_n}^{t_{n+1}} \exp((t_{n+1}-\tau)A)(g(t_n) + g'(\tau_n)(\tau - t_n)\mathrm{d}\tau = \\ &= u_{n+1} + \int_{t_n}^{t_{n+1}} \exp((t_{n+1}-\tau)A)g'(\tau_n)(\tau - t_n))\mathrm{d}\tau = \\ &= u_{n+1} + k^2\varphi_2(kA)g'(\tau_n) \ , \end{aligned}$$

ove

$$\varphi_2(z) = \frac{e^z - 1 - z}{z^2} = \sum_{j=0}^{\infty} \frac{z^j}{(j+2)!} \tag{A.7}$$

□

Proposizione 7. *Per un problema* non autonomo

$$\begin{cases} u'(t) = au(t) + b(t), \quad t > 0 \\ u(t_0) = u_0 \end{cases}$$

il metodo esponenziale–punto medio

$$u_{n+1} = \exp(kA)u_n + k\varphi_1(kA)b(t_n + k/2)$$

è di ordine 2.

Proof. Procedendo come sopra, si arriva a

$$\begin{aligned} u(t_{n+1}) &= u_{n+1} + \int_{t_n}^{t_{n+1}} \exp((t_{n+1}-\tau)A)g'(\tau_n + k/2)(\tau - (t_n + h/2))\mathrm{d}\tau = \\ &= u_{n+1} + \int_{t_n}^{t_{n+1}} \exp((t_{n+1}-\tau)A)g'(\tau_n + k/2)(\tau - t_n - k/2)\mathrm{d}\tau = \\ &= u_{n+1} + (k^2\varphi_2(kA) - k/2k\varphi_1(kA))g'(\tau_n + k/2) = \\ &= u_{n+1} + \left(\frac{k^2 I}{2} + \frac{k^3 A}{6} + o(k^3) - \frac{k^2 I}{2} - \frac{k^3 A}{2} + o(k^3)\right) g'(\tau_n + k/2) \ . \end{aligned}$$

□

A.4 Calcolo di $\exp(A)$

Come per la risoluzione di sistemi lineari, non esiste *il* modo per calcolare $\exp(A)$, ma diversi modi, ognuno adatto a particolari situazioni.

A.4.1 Matrici piene, di modeste dimensioni

Questi metodi si applicano, in pratica, a quelle matrici per le quali si usano i metodi diretti per la risoluzione di sistemi lineari.

Decomposizione spettrale

Se la matrice è diagonalizzabile, cioè $A = VDV^{-1}$, allora $\exp(A) = V\exp(D)V^{-1}$, ove $\exp(D)$ è la matrice diagonale con elementi $e^{d_1}, e^{d_2}, \ldots, e^{d_N}$. Basta infatti osservare che

$$A^2 = (VDV^{-1})^2 = (VDV^{-1})(VDV^{-1}) = VD^2V^{-1}$$

e scrivere $\exp(A)$ come serie di Taylor. La decomposizione spettrale di una matrice costa, in generale, $\mathcal{O}(N^3)$. Si ottiene in GNU Octave con il comando `eig`.

Approssimazione razionale di Padé

Si considera un'approssimazione razionale della funzione esponenziale

$$e^z \approx \frac{a_0 + a_1 z + \ldots + a_p z^p}{b_0 + b_1 z + \ldots + b_q z^q} ,$$

ove $b_0 = 1$ per convenzione. Essa è chiamata *diagonale* quando $p = q$. Si può dimostrare che le approssimazioni diagonali sono le più efficienti. Fissato il grado di approssimazione, si sviluppa in serie di Taylor la funzione esponenziale e si fanno coincidere quanti più coefficienti possibile. Per esempio, fissiamo $p = q = 1$. Si ha allora

$$\left(1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \ldots\right)(b_0 + b_1 z) = a_0 + a_1 z$$

$$b_0 + b_1 z + b_0 z + b_1 z^2 + b_0 \frac{z^2}{2} + o(z^2) = a_0 + a_1 z$$

da cui

$$\begin{cases} b_0 = 1 \\ b_0 = a_0 \\ b_1 + b_0 = a_1 \\ b_1 + \dfrac{b_0}{2} = 0 \end{cases}$$

L'approssimazione di Padé si estende banalmente al caso matriciale. Considerando sempre il caso $p = q = 1$, si ha

$$\exp(A) \approx B = (b_0 I + b_1 A)^{-1}(a_0 I + a_1 A) ,$$

cioè B è soluzione del sistema lineare $(b_0 I + b_1 A)B = a_0 I + a_1 A$.

L'approssimazione di Padé è accurata solo quando $|z| < 1/2$ (o, nel caso matriciale, $\|A\|_2 < 1/2$). Per la funzione esponenziale esiste una tecnica, chiamata *scaling and squaring* che permette di aggirare il problema. Si usa infatti la proprietà

$$e^z = \left(e^{z/2}\right)^2 = \left(e^{z/2^j}\right)^{2^j} .$$

Se $|z| > 1/2$, allora $|z|/2^j < 1/2$ per $j > \log_2(|z|) + 1$. Si calcola dunque l'approssimazione di Padé di $e^{z/2^j}$ e poi si eleva al quadrato j volte. Per la funzione φ_1 vale

$$\varphi_1(z) = \varphi_1\left(\frac{z}{2}\right)\left(\frac{z}{4}\varphi_1\left(\frac{z}{2}\right) + 1\right) .$$

Anche l'approssimazione di Padé matriciale ha costo $\mathcal{O}(N^3)$. In GNU Octave si usa una variante di questa tecnica nel comando `expm`.

A.4.2 Matrici sparse, di grandi dimensioni

I metodi visti nel paragrafo precedente ignorano l'eventuale sparsità delle matrici. Inoltre, negli integratori esponenziali, non è mai richiesto di calcolare esplicitamente funzioni di matrice, ma solo funzioni di matrice applicate a vettori, cioè $\exp(A)v$ (è l'analoga differenza tra calcolare A^{-1} e $A^{-1}v$). Si possono allora usare dei metodi *iterativi*.

Metodo di Krylov

Mediante la *tecnica di Arnoldi* è possibile, tramite prodotti matrice-vettore, decomporre A in $A \approx V_m H_m V_m^{\mathrm{T}}$, ove $V_m \in \mathbb{R}^{N\times m}$, $V_m^{\mathrm{T}} V_m = I$, $V_m e_1 = v$ e H_m è matrice di Hessenberg di ordine m (con $m \ll N$). Allora $AV_m \approx V_m H_m$ e

$$\exp(A)v \approx V_m \exp(H_m) e_1 .$$

Il calcolo di $\exp(H_m)$ è fatto mediante l'approssimazione di Padé. Il costo della tecnica di Arnoldi è $\mathcal{O}(Nm^2)$ se A è matrice sparsa. È necessario inoltre memorizzare la matrice V_m.

Interpolazione su nodi di Leja

Se il polinomio $p_m(z)$ interpola e^z nei nodi $\xi_0, \xi_1, \ldots, \xi_m$, allora $p_m(A)v$ è una approssimazione di $\exp(A)v$. È una *buona* approssimazione se i nodi sono buoni (*non* equispaziati, per esempio) e se sono contenuti nell'involucro convesso dello spettro di A. È difficile stimare a priori il grado di interpolazione m necessario. È conveniente usare la formula di interpolazione di Newton

$$p_m(z) = d_0 + d_1(z-\xi_0) + d_2(z-\xi_1)(z-\xi_2) + \cdot\ldots\cdot + d_m(z-\xi_1)\cdots(z-\xi_{m-1})$$

ove $\{d_i\}_i$ sono le differenze divise. Tale formula si può scrivere, nel caso matriciale,

$$p_m(A)v = p_{m-1}v + d_m w_m, \quad w_m = \left(\prod_{i=0}^{m-1}(A - \xi_i I)\right) v = (A - \xi_{m-1})w_{m-1}$$

Dunque, la complessità è $\mathcal{O}(Nm)$ è richiesta la memorizzazione di un solo vettore w.

Quali nodi usare? I nodi di Chebyshev, molto buoni per l'interpolazione, non possono essere usati, in quanto non permettono un uso efficiente della formula di interpolazione di Newton (cambiano tutti al cambiare del grado). I *nodi di Leja* sono distribuiti asintoticamente come i nodi di Chebyshev e, dati i primi m, ξ_m è il nodo per cui

$$\prod_{i=0}^{m-1} |\xi_m - \xi_i| = \max_{\xi \in [a,b]} \prod_{i=0}^{m-1} |\xi - \xi_i| \,,$$

ove l'intervallo $[a, b]$ è in relazione con lo spettro di A, per esempio $[a, b] = \sigma(A) \cap \{y = 0\}$. Il primo nodo coincide, solitamente, con l'estremo dell'intervallo $[a, b]$ di modulo massimo. È chiaro che l'insieme dei primi $m+1$ nodi di Leja coincide con l'unione di $\{\xi_m\}$ con l'insieme dei primi m nodi di Leja.

A.5 Esercizi

1. Si risolva la PDE

$$\begin{cases} \dfrac{\partial}{\partial t}u(x,t) = \dfrac{\partial^2}{\partial x^2}u(x,t) + (1+\pi^2)e^t \sin(\pi x), & (x,t) \in (0,1) \times (0,+\infty) \\ u(x,0) = \sin(\pi x) & x \in (0,1) \\ u(0,t) = u(1,t) = 0 & t \in (0,+\infty) \end{cases}$$

usando il metodo delle linee. Si confrontino gli integratori esponenziali Eulero esponenziale e esponenziale—punto medio e il metodo Eulero implicito per l'integrazione fino al tempo $T = 1$. Si mostrino gli ordini degli integratori temporali, sapendo che la soluzione esatta è $u(x,t) = e^t \sin(\pi x)$.

Appendice B

Espansioni di Fourier

L'argomento delle serie di Fourier ha importanza per il suo legame con le approssimazioni ai minimi quadrati nonché nella ricerca di soluzioni di equazioni differenziali con la tecnica di collocazione (vedi sezione 4.1.1).

B.1 Espansioni di Fourier

Sia X uno spazio vettoriale, finito o infinito, dotato di prodotto scalare.

Definizione 16. *Dato X, consideriamo una sequenza finita o infinita di vettori ortonormali $x_1^*, x_2^*, \ldots$. La serie di Fourier di un vettore $y \in X$ è*

$$\sum_{k=1}^{\infty} (y, x_k^*) x_k^* \,, \tag{B.1}$$

dove (w, z) indica il prodotto scalare dei vettori w, z.

I numeri (y, x_k^*) sono detti *coefficienti di Fourier* di y. D'ora in poi, per indicare che (B.1) è la serie di Fourier di y useremo

$$y \sim \sum_{k=1}^{\infty} (y, x_k^*) x_k^* \,.$$

Ricordando che $(y, x_k^*)x_k^*$ indica la proiezione di y su x_k^*, allora la serie di Fourier di y è la somma delle proiezioni di y su elementi ortonormali.

Se, $x_k,\ k \geq 0$ sono vettori non nulli e ortogonali, la serie di Fourier si scriverà come

$$y \sim \sum_{k=1}^{\infty} \left(y, \frac{x_k}{\|x_k\|} \right) \frac{x_k}{\|x_k\|} = \sum_{k=1}^{\infty} \frac{(y, x_k)}{(x_k x_k)} x_k \,. \tag{B.2}$$

ESEMPIO 20. Sia $X = \mathbb{R}^3$ dotato del prodotto scalare discreto $(x, y) = \sum_{i=1}^{3} x_i y_i$. Come vettori ortonormali possiamo prendere $x_i^* = e_i$ ovvero i versori della base canonica. Pertanto, dato un generico vettore $z = (z_1, z_2, z_3)$ lo possiamo scrivere come

$$z = z_1\, e_1 + z_2\, e_2 + z_3\, e_3\,.$$

In questo caso i coefficienti di Fourier di z coincidono con le componenti z_i del vettore z.

ESEMPIO 21. Prendiamo $X = \mathcal{C}[-\pi, pi]$ oppure $X = L_2[-\pi, \pi]$ con prodotto scalare $(f, g) = \int_{-\pi}^{\pi} f(x)g(x)\,dx$. In questo caso un sistema ortonormale è rappresentato dalle funzioni

$$\left\{ \frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}}\sin x, \frac{1}{\sqrt{\pi}}\cos x, \frac{1}{\sqrt{\pi}}\sin 2x, \frac{1}{\sqrt{\pi}}\cos 2x\,, \dots \right\}\,.$$

Pertanto la serie di Forurier di f è

$$f(x) \sim \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos kx + b_k \sin kx$$

dove $a_k = \dfrac{1}{\pi}\displaystyle\int_{-\pi}^{\pi} f(x)\cos kx\,dx$ e $b_k = \dfrac{1}{\pi}\displaystyle\int_{-\pi}^{\pi} f(x)\sin kx\,dx$.

ESEMPIO 22. Sia $X = \mathcal{C}[-1, 1]$ con prodotto scalare

$$(f, g) = \int_{-1}^{1} \frac{f(x)g(x)}{\sqrt{1-x^2}}dx\,.$$

Un sistema ortonormale è costituito dai polinomi di Chebyshev di prima specie

$$\frac{1}{\sqrt{\pi}}T_0(x), \sqrt{\frac{2}{\pi}}T_0(x), \sqrt{\frac{2}{\pi}}T_1(x), \dots$$

La serie di Fourier di f è

$$f(x) \sim \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k T_k(x)\,,$$

dove $a_k = \dfrac{2}{\pi}\displaystyle\int_{-1}^{1} \frac{f(x)T_k(x)}{\sqrt{1-x^2}}\,dx\,.$

Se lo spazio X è finito dimensionale, l'espansione di Fourier di un suo elemento coincide con l'elemento stesso. Vale infatti il seguente teorema

Teorema 5. *Se $x_1, \dots, x_n$ sono n elementi linearmente indipendenti ed x_i^* i corrispondenti normalizzati, se $w = \sum_{i=1}^{n} a_i x_i$ allora*

$$w = \sum_{i=1}^{n} (w, x_i^*)x_i^*\,.$$

A verifica del Teorema, vediamo un paio d'esempi.

ESEMPIO 23. Siano $p_n^*(x) = \sum_{j=0}^n k_{n,j} x^j, \quad k_{n,n} \neq 0$ una famiglia di polinomi ortogonali rispetto al prodotto scalare (f, g). Allora

$$p(x) = \sum_{j=0}^{n} (p, p_j^*) p_j^*, \quad \forall p \in \mathbb{P}_n \, .$$

ESEMPIO 24. Siano $x_1, \ldots, x_n$ vettori non nulli e ortogonali in $\mathbb{R}^n$ o $\mathbb{C}^n$. Allora per ogni $y \in \mathbb{R}^n$(o $\mathbb{C}^n$)

$$y = \sum_{j=0}^{n} (y, x_j) x_j, \; .$$

Proprietà di minimo delle espansioni di Fourier

Teorema 6. *Siano $x_1^*, x_2^*, \ldots$ un sistema ortonormale di uno spazio X e y un arbitrario elemento. Allora*

$$\|y - \sum_{k=1}^{N} (y, x_k^*) x_k^*\| \leq \|y - \sum_{k=1}^{N} a_k \, x_k^*\| \tag{B.3}$$

per ogni scelta delle costanti $a_1, \ldots, a_N$.

Dim. vedi [5, p. 170] □

Il problema dei *minimi quadrati* nell'analisi numerica classica si può formulare come il seguente problema

$$\min_{a_k} \|y - \sum_{k=1}^{N} a_k \, x_k\|$$

in un opportuno spazio dotato di prodotto scalare.

Corollario 1. *Siano $x_1, \ldots, x_N$ elementi indipendenti di uno spazio X. Il problema*

$$\min_{a_k} \|y - \sum_{k=1}^{N} a_k \, x_k\|$$

ha soluzine unica data da

$$\sum_{k=1}^{N} (y, x_k^*) x_k^* \, ,$$

con x_k^* elementi normalizzati degli x_k.

In sostanza, il corollario dice che l'unica soluzione del problema ai minimi quadrati è data da un'opportuna serie troncata di Fourier.

Corollario 2.

$$\min_{a_k} \|y - \sum_{k=1}^{N} a_k\, x_k\|^2 = \|y\|^2 - \sum_{k=1}^{N} |(y, x_k^*)|^2 \ .$$

Dim. Basta inserire $a_k = (y, x_k^*)$ nella dimostrazione del Teorema precendenete (cfr. [5, p. 172]). □

Bibliografia

[1] K. E. Atkinson, *An Introduction to Numerical Analysis*, Second Edition, Wiley, New York, 1989.

[2] R. Bevilacqua, D. Bini, M. Capovani e O. Menchi *Metodi Numerici*, Zanichelli, 1992.

[3] V. Comincioli, *Analisi numerica: metodi, modelli, applicazioni. E-book*, Apogeo, 2005.

[4] V. Comincioli, *Analisi numerica. Complementi e problemi*, McGraw-Hill Companies, 1991.

[5] P. J. Davis, *Interpolation & Approximation*, Dover Publications Inc., New York, 1975.

[6] C. de Boor, *A Practical Guide to Splines*, Springer-Verlag, New York, 1978.

[7] S. De Marchi, *Appunti di Calcolo Numerico, Parte I*, Dispense disponibili in rete http://profs.sci.univr.it/ demarchi/CN2006-07/diarioBook.pdf, 2007 (con esercitazioni in Matlab/Octave).

[8] S. De Marchi, *Funzioni splines univariate*, Forum Ed. Udinese, Seconda ed., 2001 (con floppy).

[9] G. Farin, *Curves and Surfaces for CAGD: A Practical Guide*, Third Edition, Academic Press, San Diego, 1993.

[10] D. Greenspan, V. Casulli *Numerical Analysis for Applied Mathematics, Science and Engineering*, Addison-Wesley, 1988.

[11] E. Isaacson e H. Bishop Keller, *Analysis of Numerical Methods*, John Wiley & Sons, New York, 1966.

[12] J. Lambert, *Numerical methods for Ordinary Differential Equations*, Weley, 1991.

[13] Lax, P. D. e Richtmyer, R. D. Survey of the stability of linear finite difference equations. Comm. Pure Appl. Math. 9 (1956), 267–293.

[14] G. G. Lorentz, *Bernstein Polynomials*, Chelsea Publishing Company, New York, 1986.

[15] G. Monegato *Elementi di Calcolo Numerico*, Levrotto&Bella, Torino, 1995.

[16] A. Quarteroni e F. Saleri *Introduzione al Calcolo Scientifico*, Esercizi e problemi risolti in Matlab, Terza Ed., Springer-Verlag, Milano, 2006.

[17] A. Quarteroni, R. Sacco e F. Saleri *Matematica Numerica*, Seconda Ed., Springer-Verlag, Milano, 2004.

[18] R. D. Richtmyer e K. W. Morton, *Difference Methods fo Initial-value Problems*, Wiley-Interscience, 1967.

[19] T. J. Rivlin, *An Introduction to the Approximation of Functions*, Dover Publications Inc., New York, 1969.

[20] J. Stoer, Bulirsch *Introduction to Numerical Analysis* Ed. Springer-Verlag, Berlin, 1980.