Metodi numerici per problemi differenziali

Alberto Tibaldi

22 gennaio 2010

Indice

Capitolo 1

Equazioni differenziali ordinarie

Al fine di introdurre l'argomento principe della trattazione, ossia la soluzione, in forma numerica, di equazioni alle derivate parziali, può essere utile presentare alcuni richiami dall'Analisi Matematica e dai precedenti corsi di Analisi Numerica. Verranno dunque presentati i suddetti richiami, soprattutto mediante alcune definizioni e alcuni esempi applicativi, più o meno semplici.

1.1 Problemi ai valori iniziali e problemi ai limiti

Nei corsi di Analisi Numerica molto spesso vengono presentati, di tutti i problemi differenziali scalari, soltanto un sottoinsieme: i problemi ai valori iniziali. Data un'equazione differenziale (o un sistema di equazioni differenziali), e le **condizioni iniziali** sulla funzione e su tutte le derivate fino all'ordine $n-1$, dove n è l'ordine dell'equazione differenziale[1]; problemi di questo tipo solitamente attribuiscono alla variabile, un significato di tipo *temporale*, dunque, stabilite tutte le condizioni iniziali, il problema è sostanzialmente un problema di evoluzione: date le caratteristiche iniziali del problema differenziale, si individuano, di tutte le soluzioni solo quelle (o quella, nel caso sia garantita l'esistenza e unicità della soluzione del problema ai valori iniziali) che rispettino le suddette condizioni.

Quando si hanno problemi con condizioni agli estremi, ai limiti, generalmente si definiscono situazioni di tipo stazionario: la definizione di onde stazionarie, o comunque fenomeni che possono essere di questo tipo (ad esempio, il vibrare di una corda con gli estremi saldi); anche in questo caso il sig-

[1]Si ricorda che l'ordine di un'equazione differenziale coincide con il massimo ordine di derivazione nell'equazione differenziale.

nificato della variabile indipendente potrebbe essere quello di un tempo, ma in questo caso è meno probabile (dipende sostanzialmente dalla modellistica che si intende introdurre).

Un'ulteriore definizione: per *equazione differenziale ordinaria* si intende un'equazione modellante fenomeni che dipendono da *una sola variabile indipendente*, o un sistema di equazioni che porti a un vettore di soluzioni, ma ciascuna funzione di una sola variabile.

Si consideri il seguente esempio:

$$y''(x) = y(x), \qquad 0 \leq x \leq 1$$

Si considererà, nella trattazione, generalmente l'uso di domini chiusi. Questo comporta, nel formalismo matematico, alcune differenziazioni rispetto al caso di domini aperti, tuttavia si eviterà di trattare questi dettagli.

La soluzione del problema differenziale è la funzione $y(x)$: essa è una funzione tale per cui, se introdotta in questa equazione differenziale, soddisfi l'eguaglianza. Per problemi di questo tipo esistono infinite soluzioni, che dipendono, in questo caso, da due coefficienti arbitrari. Questa osservazione deriva semplicemente dallo studio dell'ordine dell'equazione differenziale: l'ordine dell'equazione differenziale (in questo caso, 2), coincide con il numero di parametri liberi che differenzierà le varie curve integrali tra loro, ottenendo diverse soluzioni.

Le condizioni sono utili al fine di determinare una particolare soluzione del problema differenziale; il problema differenziale completo, dunque deve avere anche un numero di condizioni (iniziali o ai bordi) pari all'ordine dell'equazione, in modo da permettere l'individuazione di una sola particolare curva, tra tutte quelle che permettono la soluzione dell'equazione differenziale.

Per quanto riguarda il problema ai valori iniziali, come già accennato, devono essere forniti i valori della funzione e di tutte le derivate, fino all'ordine $(n-1)$-esimo, per la funzione $y(x)$; nel caso dell'equazione differenziale appena presentata, dunque, sarà necessario definire il valore della funzione e della pendenza della funzione, nel punto $x = 0$; il problema ai valori iniziali dunque risulta essere definito, se ha forma:

$$\begin{cases} y''(x) - y(x) = 0 \\ y(0) = \alpha \\ y'(0) = \beta \end{cases}$$

Per tutta la trattazione si supporrà (a meno che non si specifichi diversamente) che le equazioni differenziali interessate siano **lipschitziane** rispetto alla funzione $y(x)$, ossia qualcosa di poco maggiore alla continuità: si chiede

che il rapporto incrementale della suddetta funzione sia limitato. Supponendo che questa condizione sia dunque sempre valida, il teorema di esistenza e unicità della soluzione del problema differenziale afferma che, soddisfatta la suddetta specifica, si ha per l'appunto la certezza di avere una soluzione esistente e unica, se sono introdotte tutte le condizioni iniziali sul problema.

Per quanto riguarda i problemi ai limiti, non è così semplice introdurre qualcosa di questo genere. Si consideri il seguente problema differenziale:

$$\begin{cases} u''(x) - u(x) = 0 \ , & 0 \leq x \leq 1 \\ u(0) = \alpha \\ u(1) = \beta \end{cases}$$

In questo caso abbiamo sempre due condizioni, una per bordo, ma non abbiamo particolari vincoli sulle condizioni: possono essere sulla funzione, sulle derivate, una sulla funzione una sulle derivate; abbiamo in questo caso due condizioni, ma, nel caso ve ne siano più di due, non è detto che siano in egual numero dedicate a estremo inferiore o superiore (possono esserci due condizioni per un estremo e una per un altro estremo, e così via). L'equazione differenziale è uguale alla precedente, ma, in questo caso, non si hanno garanzie nè sull'esistenza nè sull'unicità della soluzione del problema, dal momento che il teorema di esistenza e unicità è valido solo per problemi ai valori iniziali. In questo caso infatti non si richiede il fatto che vengano rispettate tutte le condizioni su un certo punto (quello iniziale), bensì che la curva integrale, soluzione della sola equazione differenziale, soddisfi tutte le specifiche sui bordi; non esistono teoremi *semplici* in grado di garantire l'esistenza e l'unicità di una curva di questo tipo, poichè sarebbero fortemente dipendenti dal tipo di condizioni fornite dalle specifiche del problema.

Si supponga ad esempio che un'equazione differenziale sia associata al moto di un proiettile: dato il punto di partenza (condizione al bordo iniziale) e il bersaglio (condizione al limite finale), si vuole vedere se è possibile che il proiettile passi per questi due punti. Ciò non è detto! Non sappiamo quale energia verrà attribuita al proiettile, e altre eventuali caratteristiche della curva, dunque non è banale dire che la soluzione esista di sicuro.

Consideriamo un altro esempio: data la seguente equazione differenziale

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left[E \cdot I(x) y''(x)\right] + Ky(x) = q(x), \qquad x \in [-L, L]$$

Dati E e K coefficienti costanti, $I(x)$ funzione di x nota, è possibile attribuire due tipi di insiemi di condizioni; condizioni iniziali:

$$\begin{cases} y(-L) = ... \\ y'(-L) = ... \\ y''(-L) = ... \\ y'''(-L) = ... \end{cases}$$

In questo tipo di problema, come già detto, vi sono tutte le derivate. Un esempio di condizioni ai bordi può essere:

$$\begin{cases} y''(-L) = 0 \\ y''')(-L) = 0 \\ y''(L) = 0 \\ y'''(L) = 0 \end{cases}$$

In questo caso, la fenomenologia propone, come specifiche, queste; ciò, tuttavia, non permette di aver la sicurezza di avere soluzioni e di averle uniche. I vincoli sono sempre quattro (ovvio: se non si hanno almeno quattro vincoli, a meno di casi particolari, diventa problematico trovare un'unica soluzione a un'equazione), ma, non essendo *sistematici* come quelli dei problemi ai valori iniziali, possono non portare a una sola soluzione (o neanche a una soluzione).

Consideriamo un altro esempio, che permetterà di comprendere meglio questa cosa; data la seguente equazione differenziale:

$$y''(x) + q^2 y(x) = 0, \quad 0 \leq x \leq 1, \quad q \neq 0$$

La soluzione dell'equazione differenziale, come noto dall'Analisi Matematica, ha forma:

$$y(x) = c_1 \cos(qx) + c_2 \sin(qx)$$

A questa soluzione, dunque, vanno aggiunte le condizioni (iniziali o al bordo); consideriamo per incominciare le seguenti condizioni iniziali:

$$\begin{cases} y(0) = 0 \\ y'(0) = \gamma \end{cases}$$

Si può ricavare:

$$y(0) = 0 \Longrightarrow c_1 \cos(0) = 0$$

Dunque la forma sarà:

$$y(x) = c_2 \sin(qx)$$

Ora, determiniamo c_2 come:

$$y'(0) = \gamma \Longrightarrow c_2 \cdot q \cos(0) = \gamma$$

Dunque:

$$c_2 \cdot q = \gamma \Longrightarrow c_2 = \frac{\gamma}{q}$$

Questo problema differenziale, come soluzione unica ammette:

$$y(x) = \frac{\gamma}{q} \sin(qx)$$

E se associamo delle condizioni ai limiti, cosa capita? Beh, vediamo:

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

Come prima, il coseno va annullato, per quanto riguarda l'altra condizione, tuttavia, si avrà:

$$c_2 \sin(q) = \beta$$

Dunque:

$$c_2 = \frac{\beta}{\sin(q)}$$

Questa soluzione dipende da q, o meglio dal seno di q. Escludendo $q = 0$, tuttavia, si sa che il seno vale zero per tutti i multipli di π, dunque, se $q = n\pi$, $n \in \mathbb{Z} \setminus \{0\}$, si ha:

$$c_2 \cdot 0 = \beta$$

A questo punto, se β può essere zero, abbiamo infinite soluzioni valide: qualsiasi multiplo di $n\pi$ può essere un buon valore per q; dualmente se $\beta \neq 0$, non esiste nessuna soluzione valida per il problema.

1.2 Descrizione generale del problema: forma canonica

In generale, si intende avere a che fare con equazioni differenziali di primo ordine, singole; nel caso di problemi ai valori iniziali, dunque, si vuole avere qualcosa del tipo:

$$\begin{cases} y'(x) = f(x, y(x)), & x > a \\ y(a) = \alpha \end{cases}$$

Questa è la rappresentazione canonica di un generico problema differenziale, ai valori iniziali. Ciò che si intende fare, al fine di sviluppare una per la risoluzione di problemi differenziali, è riportare sempre un sistema o un'equazione di ordine superiore al primo a questa forma (eventualmente, introducendo una notazione di tipo vettoriale, come verrà fatto presto). $y(x)$ generalmente dunque sarà un vettore, dunque $\underline{y}(x)$, nato da un insieme di cambi di variabili; l'espressione viene dunque riportata a un insieme di eguaglianze, raggruppabili in un'unica eguaglianza vettoriale. Data la forma canonica, dunque, è possibile applicarvi i vari metodi numerici per la soluzione dei vari problemi differenziali. Qualsiasi modello, formulato in una generica forma, dunque, andrà riportato nella forma canonica, in modo da utilizzare le formule dei testi. La forma generale alla quale si intende arrivare è:

$$\begin{cases} \underline{y}'(x) = \underline{f}(x, \underline{y}(x) \\ \underline{g}(\underline{y}(a), \underline{y}(b) = 0 \end{cases}$$

Si ha a che fare, in questo caso, con una generica funzione vettoriale $\underline{f}()$, funzione comprendente la variabile indipendente x (in questo caso essa non è vettoriale, dal momento che si stanno trattando equazioni differenziali ordinarie), e $\underline{g}()$ come funzione rappresentante le condizioni iniziali. In questo caso si ha a che fare con relazioni qualunque sia all'interno dell'equazione differenziale, sia tra le condizioni iniziali/ai bordi: relazioni anche non lineari. Nel caso si decida di trattare esclusivamente equazioni lineari, si può utilizzare il formalismo matriciale applicabile delle applicazioni lineari, ottenendo:

$$\begin{cases} \underline{y}'(x) = \underline{\underline{A}}(x) \cdot \underline{y}(x) + \underline{r}(x) \\ \underline{\underline{B}}_a \cdot \underline{y}(a) + \underline{\underline{B}}_b \cdot \underline{y}(b) = \underline{\alpha} \end{cases}$$

Dal momento che ci limitiamo a considerare applicazioni lineari, avremo semplicemente bisogno di matrici, per rappresentarle; i relativi prodotti tra vettori e matrici, ovviamente, sono i classici prodotti *riga per colonna*, prodotti matriciali. Sono state introdotte diverse matrici e vettori:

- $\underline{\underline{A}}$: si tratta della matrice modellante le equazioni differenziali; in ambito di teoria dei sistemi dinamici, questa è nota anche come *matrice di stato*;
- $\underline{r}(x)$ rappresenta il vettore dei cosiddetti *coefficienti*, ossia degli elementi (variabili o ostanti rispetto a x), non coinvolgenti operatori di derivazione;

- $\underline{\underline{B}}_a$ e $\underline{\underline{B}}_b$ sono le matrici delle condizioni iniziali / al contorno dei problemi differenziali, ossia ciò che racchiude in sè le caratteristiche delle condizioni al contorno.

Si noti che le condizioni ai bordi non sono separate: esse sono espresse come equazioni comprendenti sia termini sul bordo iniziale, sia sul bordo finale (a e b), a differenza di come normalmente successo (condizioni separate); questa espressione rappresenta dunque una casistica più generale.

1.2.1 Introduzione ai problemi di shooting

Esiste un teorema che afferma che se $\underline{f}$, definita come funzione della variabile indipendente x e della soluzione dell'equazione y è lipschitziana, la soluzione del problema alle condizioni iniziali esiste ed è unica, come già detto. Ci si pone a questo punto una domanda: quando si può dire che la soluzione esista e sia unica, anche per problemi ai bordi?

Si proverà a rispondere, fornendo un esempio teorico/pratico: si consideri il seguente problema differenziale:

$$\begin{cases} u'(x) = f(x, u(x)), & x \geq a \\ u(a) = s \end{cases}$$

Si sa che la soluzione $u(x)$ dipenderà da s: infatti, di tutte le soluzioni dell'equazione differenziale, ne verrà scelta proprio una in grado di tenere conto della condizione su s; si può dunque dire che u sia funzione sia di x, sia di s: $u(x, s)$. Data $y(s)$ la soluzione della stessa equazione differenziale, ma con una condizione iniziale, si sa che, se si sceglie un s arbitrario:

$$u(x, s) \neq y(x)$$

Ovvio: si tratta sempre di soluzioni dell'equazione, ma non del problema differenziale!

La nostra idea potrebbe essere la seguente: conosciamo un risultato interessante per determinare l'esistenza e l'unicità di problemi ai valori iniziali; è possibile determinare un s tale per cui si può fare in modo che le soluzioni del problema differenziale ai valori iniziali e di quello ai valori al bordo coincidano?

Per quanto riguarda il problema differenziale ai valori, abbiamo detto che, in forma scalare:

$$\begin{cases} y'(x) = f(x, y(x)), & x > a \\ g(y(a), y(b) = 0 \end{cases}$$

Dobbiamo lavorare, su questo problema, con u al posto di y: se u soddisfa la condizione, infatti, deve essere in grado di soddisfarla mediante applicazione sulla funzione g; quello che si fa, dunque, è definire una funzione $\Phi(s)$ come:

$$\Phi(s) = g\left(u(a,s); u(b,s)\right) = 0$$

Dal momento che i bordi a e b sono fissati, l'unica variabile che interesserà la funzione Φ sarà s, ossia il punto scelto come *condizione iniziale equivalente*, al fine di riportarci al teorema di esistenza e unicità.

In cosa consiste la determinazione della risposta alla nostra domanda? Beh, innanzitutto, un risultato a bruciapelo: **sono ammesse tante soluzioni al problema differenziale quante sono le soluzioni di** $\Phi(s)$, dal momento che si ammettono tante soluzioni quante sono le soluzioni del sistema lineare. I passi da seguire, al fine di determinare ciascuna soluzione, sono:

1. Risolvere il problema di Cauchy iniziale, proposto all'inizio della sottosezione; in questo modo, si determina la funzione $u(x)$ per un certo s fissato all'inizio;

2. Mediante la funzione $\Phi(s)$ precedentemente definita, applicando su di essa uno dei metodi di soluzione delle equazioni non lineari (bisezione, secanti, Newton...), si determini con maggior precisione il valore di s;

3. Si risolva nuovamente il problema differenziale ai valori iniziali associato al nuovo valore di s;

4. Si iteri da 1 a 3 fino ad aver ottenuto la soluzione desiderata.

Questo metodo è abbastanza complicato, ed è alla base dei problemi detti **shooting**. Per quanto riguarda la trattazione, tutto ciò che è stato detto può essere utile al fine di discutere l'esistenza di soluzioni per un problema ai limiti. Si tratta di un metodo sostanzialmente basato sull'associazione di un problema ai limiti a uno equivalente ai valori iniziali, mediante lo studio di s.

1.3 Riconduzione di un problema differenziale alla forma canonica

Precedentemente è stata presentata la forma canonica di un'equazione differenziale ordinaria, ossia:

$$\begin{cases} \underline{y}'(x) = \underline{f}(x, \underline{y}(x)) \\ \underline{y}(a) = \underline{y}_0 \end{cases}$$

Questa è la forma canonica più generale di un'equazione differenziale ordinaria associata a condizioni sui valori iniziali, dunque di un problema di Cauchy; generalmente, si ha a che fare con grandezze di questo tipo:

$$\underline{y}(x) = \begin{bmatrix} y_1(x) \\ y_2(x) \\ \vdots \\ y_m(x) \end{bmatrix}$$

In questo modo, è possibile rappresentare la funzione $\underline{f}(x, \underline{y}(x))$ come il vettore di componenti:

$$\underline{f}(x, \underline{y}(x)) = \begin{bmatrix} f_1(x, y_1(x), y_2(x), ..., y_m(x)) \\ f_2(x, y_1(x), y_2(x), ..., y_m(x) \\ \vdots \\ f_m(x, y_1(x), y_2(x), ..., y_m(x) \end{bmatrix}$$

Anche le condizioni iniziali sono ovviamente rappresentate mediante una notazione vettoriale, ottenendo:

$$\underline{y}_0 = \begin{bmatrix} y_{1,0} \\ y_{2,0} \\ \vdots \\ y_{m,0} \end{bmatrix}$$

Forniti tutti questi dati, si può semplicemente ricondurre il sistema di problemi differenziali in un unico problema differenziale, in forma più compatta. Se $m = 1$, ovviamente, ci si riconduce al caso scalare.

Questo trucco è stato ora utilizzato per quanto concerne un insieme di problemi differenziali, che sono stati ricondotti a un unico problema; questa non è, di fatto, la cosa più interessante: quello che spesso ci si trova ad affrontare, di fatto, è un unico problema, con però ordine superiore al primo. Utilizzando una tecnica (già accennata nelle sezioni precedenti) è tuttavia possibile, in maniera analoga a quanto appena fatto, riportare un'unica equazione differenziale di ordine m in m equazioni differenziali, di ordine 1. Questo viene fatto, considerando l'introduzione di funzioni aggiuntive, definite ad hoc: $z_i(x)$. Generalmente, si fa qualcosa del genere:

$$z_1(x) = y(x); \quad z_2(x) = y'(x); \quad z_3(x) = y''(x); \quad ... \quad ; z_m(x) = y^{(m-1)}(x)$$

La stessa cosa si può fare, ovviamente, per quanto riguarda le condizioni iniziali: esse sono infatti semplicemente le derivate valutate nel punto iniziale, dunque:

Dunque, è possibile riportare il sistema complessivo nella forma canonica del problema di Cauchy:

$$\begin{cases} \underline{z}'(x) = \underline{f}(x, \underline{z}(x)) \\ \underline{z}(a) = \underline{z}_0 \end{cases}$$

Esempio con problema ai limiti

Introduciamo un esempio pratico, proponendo da subito un caso particolare; spesso capita di aver a che fare anche con problemi con valori ai limiti; l'equazione è sempre dello stesso tipo, ma la condizione avrebbe forma:

$$\begin{cases} y'(x) = f(x, y(x)), \quad x \in [a, b] \\ g(y(a), y(b)) = 0 \end{cases}$$

Nel nostro esempio:

$$\begin{cases} y''(x) - xy(x) = f(x), \quad x \in [0, 1] \\ y(0) + y(1) = c_1 \\ -y'(0) + y'(1) = c_2 \end{cases}$$

Come si procede? Beh, seguendo semplicemente la teoria precedentemente studiata:

$$z_1(x) = y(x)$$

$$z_2(x) = y'(x)$$

Dunque:

$$z_1'(x) = z_2(x)$$

$$z_2'(x) = xz_1(x) + f(x)$$

Dal momento che, come faremo spesso, stiamo considerando equazioni lineari rispetto a $y(x)$, possiamo esprimere tutto in termini di un problema differenziale lineare equivalente, di forma:

$$\begin{cases} \underline{z}'(x) = \underline{\underline{A}}(x)\underline{z}(x) + \underline{r}(x) \\ \underline{\underline{B}}_0\underline{z}(0) + \underline{\underline{B}}_1\underline{z}(1) = \underline{c} \end{cases}$$

Come si può vedere semplicemente studiando le già presenti equazioni, si può vedere che i parametri del modello canonicizzato sono:

$$\underline{\underline{A}}(x) = \begin{bmatrix} 0 & 1 \\ x & 0 \end{bmatrix}$$

$$\underline{r}(x) = \begin{bmatrix} 0 \\ f(x) \end{bmatrix}$$

Riprendendo le equazioni rappresentanti le condizioni al contorno, si può vedere che:

$$z_1(0) + z_1(1) = c_1$$

$$-z_2(0) + z_2(1) = c_2$$

Dunque:

$$\underline{\underline{B}}_0 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad \underline{\underline{B}}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \underline{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$$

Il problema di Cauchy dunque è stato riportato alla forma interessata, e tutti i parametri del modello sono stati determinati.

1.4 Metodi numerici per la soluzione di equazioni differenziali ordinarie

Verranno a questo punto richiamati, per quanto rapidamente, alcuni metodi numerici utilizzabili per la soluzione dei problemi differenziali appena proposti; essi, come già detto, andranno applicati sul problema, una volta che esso è stato ricondotto alla forma canonica.

A questo punto, considerando $x \in [a, b]$, il punto di partenza per la definizione di un metodo numerico è l'introduzione di una partizione dell'intervallo, ossia una suddivisione dell'intervallo di esistenza (o comunque considerato) della x in un certo numero di sottointervalli. Si introducono dei *nodi*, identificati con x_i:

$$x_0 = a$$

$$x_n = x_0 + n \cdot h$$

Dove h è detto *mesh*, della partizione:

$$h = \frac{b-a}{N}$$

E dove N è il numero di sottointervalli presenti nella partizione:

$$n = 0 \div N$$

In realtà, non è detto che h sia fisso; nella maggior parte dei casi che considereremo tuttavia considereremo l'uniformità degli intervalli.

Lo scopo del metodo numerico è quello di ottenere un'approssimazione della soluzione esatta, nei nodi del reticolo che formiamo. Ovviamente, quanto l'approssimazione sia valida, ossia quanto essa sia simile alla soluzione reale, dipende da alcuni parametri. Un parametro che può migliorare la validità dell'approssimazione è ad esempio h: più si riduce h, più si può sperare di avere un'approssimazione prossima alla realtà.

Esistono alcune classificazioni dei metodi numerici:

- Metodi a 1 passo: si tratta di metodi in cui il valore della soluzione al passo n dipende solo dal passo precedente, ossia $n-1$;
- Metodi a k passi: si tratta di metodi in cui il valore della soluzione al passo n dipende da k passi precedenti a n.

Di questi metodi, esistono due sottocategorie:

- Metodi espliciti, ossia metodi in cui si esplicita il passo successivo a partire da soli elementi appartenenti a passi precedenti;
- Metodi impliciti, ossia metodi in cui, al fine di portare avanti l'evoluzione del modello (calcolare i valori delle funzioni in passi successivi), è necessario disporre già di alcuni elementi appartenenti ai passi successivi, elementi da calcolare mediante altri metodi.

Verrà a questo punto proposta una panoramica comprendente cinque dei più famosi metodi a un passo.

Metodo di Eulero esplicito

Il metodo di Eulero esplicito può essere semplicemente rappresentato mediante la seguente equazione:

$$y_{n+1} = y_n + h \cdot f(x_n, y_n)$$

Il metodo è esplicito, dal momento che la forma dipende esplicitamente solo dagli elementi dello step n.

Metodo di Eulero implicito

Il metodo è riassumibile così:

$$y_{n+1} = y_n + h \cdot f(x_{n+1}, y_{n+1})$$

Abbastanza simile a prima, se non fosse che in questo caso si introduce la dipendenza dal valore della f dello step $n+1$; questo metodo è più complicato del precedente, dal momento che è necessario valutare questa f.

Metodo dei trapezi

Si tratta di un altro metodo implicito; l'equazione che lo rappresenta è la seguente:

$$y_{n+1} = y_n + \frac{1}{2}h\left[f(x_n, y_n) + f(x_{n+1}, y_{n+1})\right]$$

Metodo di Heun

$$y_{n+1} = y_n + \frac{1}{2}h\left[f(x_n, y_n) + f(x_{n+1}, y_n + h \cdot f(x_n, y_n))\right]$$

Si tratta di un metodo come al solito a 1 passo, ma in questo caso è esplicito (non bisogna farsi ingannare dal fatto che è presente x_{n+1}: esso è semplicemente il valore del nodo, il punto in cui si vuole valutare la funzione, dunque esso è noto a priori, e non è critico come informazione; per il resto, è completamente determinato dal passo n. Si può osservare che questo metodo è semplicemente uguale a quello dei trapezi, applicando a y_{n+1} il metodo di Eulero esplicito, esplicitando dunque il metodo dei trapezi.

Metodo di Eulero modificato

Si presenta infine il metodo di Eulero modificato:

$$y_{n+1} = y_n + h \cdot f\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2} \cdot h \cdot f(x_n, y_n)\right)$$

1.5 Convergenza e stabilità dei metodi numerici

Abbiamo proposto una panoramica dei vari metodi, senza però porci una domanda: quale di essi è il migliore? A partire da questa domanda, dovremmo porci un'altra domanda: cosa significa *il migliore*? Beh, provando a

rispondere, potremmo dire che esso sia il più *semplice*, e *meno costoso*; sicuramente, di essi, il più semplice è il metodo di Eulero esplicito: è esplicito, dunque semplice da utilizzare, e poco costoso, dal momento che esso richiede semplicemente la determinazione (il calcolo) della funzione valutata nello step precedente. Il problema è il seguente: non è prima di tutto detto che il metodo sia *sempre* applicabile, come d'altra parte non è detto che esso sia il più *rapido* a convergere: nei vari discorsi fatti, infatti, non si è parlato di velocità di convergenza al metodo a una soluzione. A tal proposito, verranno ora proposti alcuni concetti.

1.5.1 Concetto di convergenza in un metodo numerico

Si vuole a questo punto introdurre il concetto di convergenza in un metodo numerico. Si è detto che è stato introdotto un reticolo e che, presumibilmente, più esso è piccolo, più la curva deve diventare ideale. Un metodo a 1 passo si dice convergente nell'intervallo $[a, b]$ se, data f, dotata di derivate parziali prime continue e limitate in $[a, b] \times \mathbb{R}^n$, si fa in modo che:

$$\lim_{h \to 0} ||y(x_n) - y_n|| = 0 \quad \forall x_n$$

Nei nostri problemi, di solito, considereremo funzioni di classe infinita, dunque con infinite derivate continue. Oltre alla definizione di convergenza, si introduce la definizione di **ordine** di convergenza: se

$$||y(x_n) - y_n|| = O(h^p) \quad \forall x_n$$

Si dice che il metodo converga con ordine p. Si ricorda che la notazione *o grande*, o $O()$, indica il fatto che si vada a 0 al più con la stessa velocità dell'argomento.

Si può dimostrare che i metodi a un passo precedentemente introdotti hanno i seguenti ordini di convergenza:

- Metodo di Eulero esplicito, $p = 1$;
- Metodo di Eulero implicito, $p = 1$;
- Metodo dei trapezi, $p = 2$;
- Metodo di Heun, $p = 2$;
- Metodo di Eulero modificato, $p = 2$.

Questo può portare a una riflessione: prima si parlava di *costo* dei metodi, in termini di difficoltà di utilizzo; ora possiamo dire che il metodo dei trapezi, sebbene più costoso di Eulero esplicito, sia anche più rapido: la soluzione tende a convergere più rapidamente.

A questo punto, un'osservazione: dato il nostro attuale livello teorico, potremmo pensare che il metodo di Eulero implicito non serva a niente: complicato e lento. Purtroppo non è così: nessuno assicura, ora come ora, che un errore esploda o meno: si può convergere, ma ottenendo errori elevatissimi. Per gli stessi motivi, non è detto che, fissato h, la soluzione sia buona, valida: per diversi problemi potrebbe capitare che serva un h piccolo. Avere h piccoli è brutto: di fatto la soluzione converge meglio, ma al prezzo di avere metodi molto più pesanti, molto più costosi.

1.5.2 Problemi stiff

Un esempio di problemi molto complicati da trattare sono i problemi detti **stiff** (rigidi). Si vuole a questo punto proporre una definizione di problemi classificabili come "stiff".

Dato un problema differenziale del tipo:

$$\begin{cases} \underline{y}'(x) = \underline{f}(x, \underline{y}(x)) \\ \underline{y}(0) = 0 \end{cases}$$

Esso può essere ovviamente ricondotto al caso lineare (se il modello lo permette):

$$\underline{f}(x, \underline{y}(x)) = \underline{\underline{A}} \cdot \underline{y}(x) + \underline{b}$$

Volendo comunque utilizzare questo tipo di formalismo anche nel caso non lineare, è possibile farlo, considerando uno sviluppo in serie di Taylor del sistema non lineare, ottenendo dunque $\underline{\underline{A}}$ semplicemente come la matrice jacobiana rispetto al punto scelto per la linearizzazione.

Supponendo di avere in qualche modo ottenuto una matrice $\underline{\underline{A}}$, e supponendo che essa sia diagonalizzabile, con autovalori λ_i, il problema differenziabile espresso è detto **stiff** se:

- Agli eventuali autovalori λ_i con parte reale positiva corrispondono quantità

  $$(b-a)\mathbb{R}\mathrm{e}\left\{\lambda_i\right\}$$

  che siano **non grandi**.

- Esiste almeno un autovalore λ_j con parte reale negativa e tale per cui

$$(b-a)\mathbb{R}\mathrm{e}\{\lambda_j\} \ll -1$$

Per i problemi stiff, non tutti i metodi numerici sono validi! Innanzitutto si dovrà avere un h piccolo, al fine di garantire la convergenza, dunque il metodo risulterà essere piuttosto costoso. I metodi più utilizzati per la soluzione dei problemi stiff sono:

- Metodo di Eulero implicito;
- Metodo dei trapezi;
- Metodo BDF : Backward Differentiation Formula (non analizzato).

1.5.3 Esempio di soluzione di problema stiff

Si vuole a questo punto proporre un esempio di soluzione per un problema tipo stiff; dato il seguente problema:

$$\begin{cases} y'(x) = \lambda y(x), \quad x \geq 0 \\ y(0) = 1 \end{cases}$$

Supponendo che λ sia un numero reale negativo con modulo molto elevato. La soluzione sarà:

$$y(x) = \mathrm{e}^{\lambda x}$$

A questo punto, proviamo a vedere come *reagiscono* due diversi metodi numerici: Eulero esplicito e implicito.

Soluzione con metodo di Eulero esplicito

Si sa che:

$$y_{n+1} = y_n + hf(x_n, y_n)$$

Identifichiamo f: essa sarà, semplicemente, λ ! Questo significa che, a ogni step, dovremo semplicemente moltiplicare per λ. Proviamo a questo punto a operare alcuni step:

$$y_1 = y_0 + h \cdot f(x_0, y_0) = y_0 + h \cdot \lambda \cdot y_0 = (1 + h \cdot \lambda) y_0$$

Andiamo avanti:

$$y_2 = y_1 + h \cdot f(x_1, y_1) = (1 + h \cdot \lambda)y_1 = (1 + h \cdot \lambda)^2 y_0$$

Si continuerà ad avere termini tra parentesi elevati al quadrato, man mano che si va avanti. Ci chiediamo ora: per $n \to +\infty$, funziona il nostro metodo? Beh, come si può vedere, asintoticamente, il limite sarebbe semplicemente:

$$\lim_{n \to +\infty} |1 + h \cdot \lambda|^{n+1}$$

Come si può vedere dall'espansione per induzione dei calcoli precedenti (il valore assoluto è stato messo per evitare incomprensioni e studi coi segni). Quando questo metodo converge? Quando cioè i valori delle funzioni non tendono a esplodere? Beh, semplicemente quando:

$$|1 + h \cdot \lambda| \leq 1 \Longrightarrow -1 \leq 1 + h \cdot \lambda \leq 1$$

Dunque:

$$\begin{cases} h \cdot \lambda \geq -2 \\ h \cdot \lambda \leq 0 \end{cases}$$

Dunque:

$$h \leq \frac{2}{|\lambda|}$$

L'errore tende a zero solo se λ rispetta questo vincolo, utilizzando il metodo di Eulero esplicito.

Soluzione con metodo di Eulero implicito

Vediamo cosa capiterebbe, utilizzando il metodo di Eulero implicito; è noto che:

$$y_1 = y_0 + h \cdot \lambda \cdot y_1$$

(recuperando le osservazioni su f precedentemente fatte). Dunque:

$$(1 - h \cdot \lambda)y_1 = y_0$$

Dunque:

$$y_1 = \frac{1}{1 - h \cdot \lambda} y_0$$

Consideriamo ancora una volta i valori assoluti, ottenendo:

$$|y_1| = \frac{1}{|1 - h \cdot \lambda|} |y_0|$$

Ora, iterando su y_2, otterremmo (basta rifare i conti):

$$|y_2| = \frac{1}{|1 - h \cdot \lambda|} |y_1| = \frac{1}{|1 - h \cdot \lambda|^2} |y_0|$$

Considerando un n molto elevato:

$$y_{n+1} = \frac{1}{|1 - h \cdot \lambda|^{n+1}} |y_n|$$

Dunque, guardando il limite:

$$\lim_{n \to \infty} \frac{1}{|1 - h \cdot \lambda|^{n+1}} = 0$$

Quando vale? Beh, sicuramente, quando il denominatore tende a divergere, dunque quando $|1 - h \cdot \lambda| > 1$; garantire questa condizione non è difficile, dal momento che λ in modulo è molto grande, dunque permette di far valere l'espressione per qualsiasi h. Si è dimostrato che, usando questo metodo, non si hanno vincoli su h (a differenza del metodo di Eulero esplicito).

1.6 Metodi per la soluzione di problemi differenziali

Supponendo come faremo sempre di trovarci nella forma canonica, dunque con una equazione differenziale e una condizione al contorno (anche in forma vettoriale), si considerino problemi ai limiti; ai fini di risolverli, verranno introdotti tre metodi numerici molto utilizzati, e spesso esportabili anche in ambito di equazioni alle derivate parziali:

- Metodo delle differenze finite
- Metodo di collocazione
- Metodo di Galerkin (nella fattispecie, verrà analizzato il metodo degli elementi finiti)

Gli ultimi due sono noti come casi particolari del **metodo dei residui pesati**.

1.6.1 Metodo delle differenze finite

Verrà a questo punto presentato il primo dei tre metodi: il metodo delle differenze finite. Esso è un metodo utilizzato per discretizzare le equazioni differenziali, basandosi su di un'approssimazione delle derivate mediante rapporti incrementali generalizzati (come verrà presentato presto). Questo metodo è applicabile in diverse forme e per problemi differenziali di diversi ordini; nella trattazione verrà presentato esclusivamente un metodo di risoluzione per problemi di ordine 2; nel caso si voglia risolvere, con questo metodo, problemi di ordine superiore a 2, si possono trasformare, mediante le tecniche precedentemente introdotte, in problemi del primo o del secondo ordine.

Al fine di introdurre questo metodo, verranno presentati alcuni esempi pratici.

Esempio pratico

Si consideri il seguente problema differenziale (non lineare) di esempio:

$$\begin{cases} y''(x) = f(x, y(x), y'(x) \quad a \leq x \leq b \\ y(a) = \alpha \\ y(b) = \beta \end{cases}$$

La formulazione classica del metodo delle differenze finite si basa su di un'ipotesi: che $y(x)$ abbia un comportamento molto liscio e regolare, senza cambi bruschi di derivata nè tantomeno derivate con valori eccessivamente alti (evitando dunque discese rapide e oscillazioni rapide).

Il primo passo da seguire si basa sulla suddivisione dell'intervallo $[a, b]$ in N sotto-intervalli uguali, di dimensione h; metodi non-classici sviluppati negli anni 80 prevedono la suddivisione in parti diverse, infittendo la suddivisione dove sono presenti pendenze più elevate; si evita di utilizzare metodi di questo genere in questa trattazione poichè complicati e che non portano a risultati strepitosi.

Il secondo passo si basa sulla formulazione di una richiesta: si chiede che la soluzione dell'equazione differenziale sia esattamente soddisfatta nei nodi del reticolo, ossia nei punti $x_0 \div x_N$. Questo fatto permette di definire un sistema di equazioni, ciascuna della quale richiede l'imposizione del passaggio della soluzione per il punto desiderato; si avrà a che fare, dunque, con $N-1$ identità numeriche, a differenza di infinite identità: precedentemente (Analisi Matematica) si chiedeva infatti che la soluzione dell'equazione differenziale fosse valida in **ogni** punto dell'intervallo, ora solo in alcuni punti; questo fatto introduce una prima perdita di informazione: da dominio continuo si

considera solo un dominio discreto, e ci si aspetta che la soluzione sia *reale* solo nei suddetti punti.

A questo punto vi è un problema: abbiamo $N-1$ equazioni, ciascuna delle quali nasconde tre incognite: la funzione, la sua derivata prima, la sua derivata seconda: troppe incognite rispetto al numero di equazioni presenti. Entra a questo punto in gioco l'idea nascosta dietro al metodo alle differenze finite: anzichè considerare tre incognite, si considera il fatto che esse non sono indipendenti tra loro! Ciò che si fa, dunque, è cercare di approssimare un legame tra le varie funzioni, ottenendo un analogo discreto della funzione di derivazione. Esistono (e verranno ora presentate) tre formule dette **formule delle differenze finite**:

$$y'(x_n) = \frac{y(x_{n+1}) - y(x_n)}{h} + O(h), \quad y \in \mathrm{C}^{(2)}(x)$$

$$y'(x_n) = \frac{y(x_{n+1}) - y(x_{n-1})}{2h} + O(h^2), \quad y \in \mathrm{C}^{(3)}(x)$$

$$y'(x_n) = \frac{y(x_n) - y(x_{n-1})}{h} + O(h), \quad y \in \mathrm{C}^{(2)}(x)$$

In questo ambito, $O(h)$ o $O(h^2)$ rappresentano gli errori commessi nelle rispettive approssimazioni: nella prima e nella terza si commette un errore che cresce (o decresce) linearmente con la crescita (o decrescita) dell'ampiezza del sottointervallo h; nella seconda, questa relazione è di tipo quadratico; se dunque si riduce di un decimo l'ampiezza dell'intervallo, l'errore decrescerà 100 volte più rapidamente rispetto al caso senza modifica del h. Queste relazioni (e le relative espressioni degli errori) derivano dai metodi di interpolazione. Verrà utilizzata soprattutto la formula centrale: nonostante essa sia quella più *pretenziosa* in termini di regolarità della funzione, essa sarà anche quella che permetterà una maggiore attenuazione dell'errore.

Per quanto riguarda la derivata seconda, esiste la seguente formula di approssimazione (derivabile applicando iterativamente una delle formule precedenti):

$$y''(x_n) = \frac{y(x_{n+1}) - 2y(x_n) + y(x_{n-1})}{h^2} + O(h^2), \quad y \in \mathrm{C}^{(4)}(x)$$

Stiamo considerando, in questa trattazione, problemi che ammettano almeno soluzione classica; il che significa che almeno tutte le derivate siano continue fino all'ordine del problema differenziale (esempio: per un problema di ordine 2, almeno 2 derivate continue). Soluzione in senso non-classico

può riguardare la derivazione in senso distribuzionale, ossia l'esistenza delle derivate ma nel senso della teoria delle distribuzioni.

Precedentemente è stato proposto un problema; ora ci sono tutti i dati per risolverlo, introducendo le formule delle differenze finite (quelle che portano a una maggior attenuazione dell'errore):

$$\begin{cases} \frac{y(x_{n+1})-2y(x_n)+y(x_{n-1})}{h^2} + O(h^2) = f\left(x_n, y(x_n), \frac{y(x_{n+1})-y(x_{n-1})}{2h} + O(h^2)\right), \quad n : 1 \div N-1 \\ y(x_0) = \alpha \\ y(x_N) = \beta \end{cases}$$

A questo punto abbiamo $N-1$ equazioni in $N-1$ incognite, o quasi: in realtà, se studiamo formalmente il sistema, vediamo che esso è leggermente diverso.

Introduciamo due elementi: la differenziazione delle equazioni intermedie dalla prima e dall'ultima, e il troncamento degli errori: in questo momento stiamo approssimando la funzione, dal momento che trascuriamo la presenza degli errori, ossia degli $O(h^2)$; in tal senso, otteniamo, al posto degli $y(x_n)$, ossia dei valori dell'equazione differenziale valutata nel nodo, y_n, ossia il valore della soluzione nel nodo n-esimo approssimata in seguito agli errori di troncamento; il sistema di equazioni risultante sarà:

$$\begin{cases} \frac{y_{n+1}-2y_n+y_{n-1}}{h^2} = f\left(x_n, y_n, \frac{y_{n+1}-y_{n-1}}{2h}\right), \quad n : 0 \div N \\ y_0 = \alpha \\ y_N = \beta \end{cases}$$

Le condizioni ai limiti possono essere introdotte rispettivamente nella prima e nell'ultima equazione, ottenendo:

$$\begin{cases} y_2 - 2y_1 + \alpha = h^2 \cdot f\left(x_1, y_1, \frac{y_2-\alpha}{2h}\right) \\ \frac{y_{n+1}-2y_n+y_{n-1}}{h^2} = f\left(x_n, y_n, \frac{y_{n+1}-y_{n-1}}{2h}\right), \quad n : 2 \div N-2 \\ \beta - 2y_{N-1} + y_{N-2} = h^2 \cdot f\left(x_{N-1}, y_{N-1}, \frac{\beta-y_{N-2}}{2h}\right) \end{cases}$$

A questo punto bisognerebbe risolvere questo sistema non lineare, tridiagonale, con metodi quali quello di Newton: si calcola la matrice jacobiana, che risulterà essere tridiagonale, dunque si potrebbe verificare che essa ha altre proprietà, e risolvere il sistema.

Esempio pratico 2 - problema lineare

Al fine di presentare una nuova applicazione, si intende presentare un esempio di problema differenziale lineare:

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

Dove $q(x) \geq 0$, e $q(x)$, $p(x)$, $r(x)$, sono funzioni note e di classe almeno 1 (almeno una derivata continua). Queste funzioni saranno dette *coefficienti* del problema differenziale in quanto, anche se non costanti, sono note.

Si può dire che:

$$f(x, y(x), y'(x)) = r(x) + q(x)y(x) + p(x)y'(x)$$

La linearità va vista su f: per quanto variabili, i coefficienti restano coefficienti, dunque, dal momento che f è una funzione data dalla somma di funzioni lineari, si può dire che il problema differenziale sia lineare. Talvolta potrebbe servire (come in questo caso) qualcosa di più della differenziabilità delle funzioni: ci potrebbe interessare che la soluzione sia più che classica, ossia avere più garanzie rispetto a quelle fornite. In questo caso infatti, si ha che:

$$y(x) \in \mathrm{C}^{(m+2)}[a, b]$$

Vogliamo dunque, perchè ci sia classe 4 (in modo che tutte le formule siano applicabili), che i coefficienti siano di classe 2. Partendo da questa ipotesi, si passa al discreto:

$$\begin{cases} y''(x_n) - p(x_n)y'(x_n) - q(x_n)y(x_n) = r(x_n) \\ y(x_0) = \alpha \\ y(x_N) = \beta \end{cases}$$

Usando ancora una volta le formule più accurate, si ottiene:

$$\frac{y(x_{n+1}) - 2y(x_n) + y(x_{n-1})}{h^2} + O(h^2) - p(x_n)\left[\frac{y(x_{n+1}) - y(x_{n-1})}{2h} + O(h^2)\right] - q(x_n)y(x_n) = r(x_n)$$

Questa, vale per $n = 1 \div N - 1$. A questo punto tronchiamo, eliminando gli errori, costruendo dunque un'approssimazione del secondo ordine rispetto al parametro di discretizzazione; si ha che:

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} - p(x_n)\frac{y_{n+1} - y_{n-1}}{2h} - q(x_n)y_n = r(x_n)$$

Passando al sistema, si ottiene, imponendo le condizioni $y_0 = \alpha$, $y_N = \beta$:

$$\begin{cases} -(2+h^2q(x_1))y_1 + \left(1-\frac{h}{2}p(x_1)\right)y_2 = h^2r(x_1) - \alpha\cdot\left(1+\frac{h}{2}p(x_1)\right) \\ \left(1+p(x_n)\frac{h}{2}\right)y_{n-1} - (2+h^2q(x_n))\,y_n + \left(1-p(x_n)\frac{h}{2}\right)y_{n+1} = h^2r(x_n) \\ \left(1+p(x_{N-1})\frac{h}{2}\right)y_{N-2} - (2+h^2q(x_{N-1}))\,y_{N-1} = h^2r(x_{N-1}) - \left(1-p(x_{N-1})\frac{h}{2}\right)\beta \end{cases}$$

Anche in questo caso, si ha a che fare con un sistema tridiagonale; se inoltre:

$$h \leq \frac{2}{||p(x)||_\infty}$$

Si può dimostrare che la matrice è tridiagonale, e a diagonale dominante, dunque non singolare, e diagonalizzabile. In questo caso, si può dimostrare[2] che l'errore complessivo del sistema sia:

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

Si può dire, dato che l'errore va come $O(h^2)$, che questo metodo alle differenze finite sia del secondo ordine; esso è *ben costruito*: se si discretizzasse male, si potrebbero ottenere errori veramente gravi, tali da rendere insensato il sistema o addirittura tali da non farlo proprio convergere. Si noti che l'errore coincide con gli errori di troncamento effettuati: questo è un caso! Infatti, non è assolutamente detto che la relazione sia così semplice e, come già accennato, non verrà presentato alcun metodo per la determinazione dell'errore commesso.

Schema di soluzione alternativo

Verrà a questo punto proposto uno schema di soluzione alternativo, per applicare il metodo delle differenze finite. Si consideri il seguente problema differenziale di esempio:

$$\begin{cases} y''(x) - p(x)y'(x) - q(x)y(x) = r(x), \quad a \leq x \leq b \\ y'(a) + \gamma y(a) = 0 \\ y'(b) + \delta y(b) = 0 \end{cases}$$

In questo caso, dovremo valutare le y_n, ma con $n \in 0 \div N$; questo è problematico, dal momento che è necessario introdurre delle condizioni anche per quanto riguarda gli estremi. Quello che si può fare al termine di risolvere problemi di questo tipo è **prolungare** la soluzione (almeno temporaneamente), richiedendo qualcosina in più sulla funzione, allargando il

[2]Nella trattazione non verrà mai spiegato nè richiesto come fare.

dominio, in modo da utilizzare il metodo vecchio; le soluzioni aggiuntive potranno o essere scartate o, come si vedrà, proprio non calcolate. Riciclare il metodo già utilizzato in precedenza è molto importante per un particolare motivo: esso permette di avere un errore che decresce come $O(h^2)$; questo è fondamentale perchè, se introduciamo errori di tipo diverso, otterremo errori che decrescono più lentamente, influenzando pesantemente il risultato finale. Dunque:

$$y'(x_0) = \frac{y(x_1) - y_(x_{-1})}{2h} + O(h^2)$$

$$y'(x_N) = \frac{y(x_{N+1}) - y_(x_{N-1})}{2h} + O(h^2)$$

Con questo trucco, introducendo questi nodi fittizi, è possibile riutilizzare il metodo generale. Approssimando direttamente, eliminando gli errori, si ottiene:

$$\begin{cases} \frac{y_{n+1}-2y_n+y_{n-1}}{h^2} - p(x_n)\frac{y_{n+1}-y_{n-1}}{2h} - q(x_n)y_n = r(x_n), \quad n = 0 \div N \\ \frac{y_1-y_{-1}}{2h} + \gamma y_0 = 0 \\ \frac{y_{N+1}-y_{N-1}}{2h} + \delta y_n = 0 \end{cases}$$

Potremmo a questo punto scartare le incognite sovrabbondanti, risolvendo i nostri problemi. Si può anche però eliminare immediatamente le soluzioni, ricordando le formule introdotte all'inizio, e ribaltandole:

$$y_{-1} = y_1 + 2h\gamma y_0$$

$$y_{N+1} = y_{N-1} - 2h\delta y_N$$

A questo punto, sostituendo nel sistema precedente, si trova la soluzione anche agli estremi, ma senza tenere per forza conto del prolungamento del dominio introdotto. Oltretutto, l'errore di troncamento va come $O(h^2)$, dunque abbiamo ottenuto il nostro obiettivo.

Schema di soluzione basato sul metodo dei trapezi

Si sta parlando di metodi di soluzione alle differenze finite, ossia metodi basati sull'approssimazione delle derivate della funzione mediante rapporti incrementali. Sono state proposte alcune formule in grado di approssimare, a meno di un errore, le derivate. Si potrebbe tuttavia utilizzare un approccio di tipo diverso, basato sui metodi di soluzione numerica di equazioni differenziali: essi, oltre a risolvere i suddetti problemi, sono in grado di effettuare

uno degli step fondamentali per la soluzione di generici sistemi differenziali: la discretizzazione.

Si consideri il seguente sistema:

$$\begin{cases} \underline{y}'(x) = \underline{\underline{A}}(x)\underline{y}(x) + \underline{r}(x), & a \leq x \leq b \\ \underline{\underline{B}}_a \underline{y}(a) + \underline{\underline{B}}_b \underline{y}(b) = \alpha \end{cases}$$

Quello che si potrebbe fare è utilizzare il metodo dei trapezi: si richiede infatti che il metodo di discretizzazione scelto sia **implicito**. Si ricordi che:

$$y(x_{n+1}) = y(x_n) + \frac{h_n}{2}\left[f(x_n, y(x_n)) + f(x_{n+1}, y(x_{n+1}))\right] + O(h^2)$$

Trascurando $O(h^2)$ e approssimando direttamente, si ottiene:

$$y_{n+1} = y_n + \frac{h_n}{2}\left[f(x_n, y_n) + f(x_{n+1}, y_{n+1})\right]$$

A questo punto, essa può esser utilizzata per discretizzare la nostra equazione differenziale; in forma vettoriale:

$$\begin{cases} \underline{y}'(x_n) = \underline{\underline{A}}(x_n)\underline{y}(x_n) + \underline{r}(x_n), & n = 0 \div N-1 \\ \underline{\underline{B}}_a \underline{y}_a + \underline{\underline{B}}_b \underline{y}_b = \alpha \end{cases}$$

Raccogliendo, si ottiene:

$$\left[\underline{\underline{\mathcal{I}}} + \frac{h_n}{2}\underline{\underline{A}}(x_n)\right]\underline{y}_n - \left[\underline{\underline{\mathcal{I}}} - \frac{h_n}{2}\underline{\underline{A}}(x_{n+1})\right] y_{n+1} - \frac{h_n}{2}\left[\underline{r}(x_n) + \underline{r}(x_{n+1})\right]$$

Si divide per $-h_n$:

$$-\frac{1}{h_n}\left[\underline{\underline{\mathcal{I}}} + \frac{h_n}{2}\underline{\underline{A}}(x_n)\right]\underline{y}_n + \frac{1}{h_n}\left[\underline{\underline{\mathcal{I}}} - \frac{h_n}{2}\underline{\underline{A}}(x_{n+1})\right] y_{n+1} + \frac{1}{2}\left[\underline{r}(x_n) + \underline{r}(x_{n+1})\right]$$

A questo punto, il sistema è in forma:

$$\begin{cases} \underline{\underline{S}}_{n+1}\underline{y}_n + \underline{\underline{R}}_{n+1}\underline{y}_{n+1} = \underline{q}_n \\ \underline{\underline{B}}_a \underline{y}_a + \underline{\underline{B}}_b \underline{y}_N = \alpha \end{cases}$$

Si può osservare, scrivendo il sistema per esteso, che esso risulterebbe essere bidiagonale, se non fosse per un termine nell'ultima riga. Il fatto è comunque interessante, dal momento che il sistema è *quasi bidiagonale a blocchi*. Si dice *a blocchi* dal momento che si può suddividere la matrice in altre matrici, ottenendo dunque come soluzione del sistema altre sottomatrici. Esistono

metodi rapidi (iterativi) per la soluzione di sistemi di questo tipo, quali il metodo di Jacobi o di Gauss-Seidel in versione *a blocchi*. Questo metodo alternativo **non richiede** che i nodi siano equispaziati, dunque anche sotto questo punto di vista esso rappresenta una variante piuttosto interessante. L'ordine di convergenza dell'errore è:

$$\max_{0\leq n\leq N} \left|\left| \underline{y}(x_n) - \underline{y}_n \right|\right| = O(h^2)$$

1.6.2 Metodo di collocazione

Un altro metodo (le cui caratteristiche *intuitive* verranno presentate in seguito) per la risoluzione di problemi differenziali è il cosiddetto **metodo di collocazione**; anch'esso, come il precedente, richiede quantomeno l'esistenza di una soluzione classica, al fine di poter essere applicato e utilizzato con successo.

Si consideri il seguente problema generale:

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

Considerando una funzione (per ora scalare, ma ovviamente è tutto generalizzabile al caso vettoriale) $y(x)$, essa va approssimata; si ha tuttavia la certezza che essa è soluzione del problema, che esiste e che è unica. Quello che si sceglie è dunque una certa funzione approssimante, $y_N(x)$, tale per cui:

$$y(x) \sim y_N(x)$$

$y_N(x)$ è una funzione approssimante, tale da essere realizzata mediante una combinazione lineare di altre funzioni, ottenendo dunque qualcosa nella forma:

$$y_N(x) = \sum_{i=0}^{N} c_i \varphi_i(x)$$

Dove $\varphi_i(x)$ può essere un monomio, una funzione goniometrica, un polinomio di Chebishev, un polinomio di Legendre, o qualsiasi altra cosa. Fondamentale è però che la funzione approssimante *assomigli* alla funzione da approssimare, in termini di proprietà: questo metodo, di fatto, si basa su di uno sviluppo in una base più o meno arbitraria di una funzione, a patto però di ottenere, dallo sviluppo, un andamento simile a quello desiderato.

Si avranno, su $[a, b]$, $N + 1$ nodi totali, e $N - 1$ nodi interni; si consideri dunque l'equazione, approssimata:

$$y_N(x_n) \simeq f(x_n, y_N(x_n), y'_N(x_n))$$

Si sceglie, di tutti i punti possibili, di soddisfare l'equazione solo nei nodi scelti; si sceglierà poi di avere N molto grandi, in modo da soddisfare l'equazione in più punti possibile. Come per le equazioni, anche per le derivate dovrà sussistere una relazione del tipo:

$$y^{(k)} \sim y_N^{(k)} = \sum_{i=0}^{N} c_i \varphi_i^{(k)}(x)$$

A questo punto, è dunque possibile riscrivere il sistema iniziale in un sistema approssimato, che avrà forma:

$$\begin{cases} \sum_{i=0}^{N} c_i \varphi_i''(x_n) = f(x_n, \sum_{i=0}^{N} c_i \varphi_i(x_n), \sum_{i=0}^{N} c_i \varphi_i'(x_n) \\ g_1\left(\sum_{i=0}^{N} c_i \varphi_i(a), \sum_{i=0}^{N} c_i \varphi_i(b)\right) = 0 \\ g_2\left(\sum_{i=0}^{N} c_i \varphi_i(a), \sum_{i=0}^{N} c_i \varphi_i(b)\right) = 0 \end{cases}$$

Per quanto riguarda la funzione φ, si è dunque trovato un vincolo: tutte le sue derivate dovranno esistere, almeno fino all'ordine di integrazione richiesto. Utilizzare una polinomiale dunque sarà impossibile, dal momento che potrebbe capitare che nessuna delle sue derivate sia continua nei punti richiesti.

Se tutto è lineare, il sistema risulta essere quadrato: $N - 1$ equazioni, $N - 1$ incognite. Il comportamento del sistema (la sua eventuale singolarità e la sua definizione) dipendono sostanzialmente da φ_i e dai **nodi di collocazione**, ossia dall'insieme dei punti scelti per l'applicazione del **metodo di collocazione**: i nodi in cui si **colloca** l'equazione differenziale, in cui imponiamo che tutto funzioni. Come si scelgono i nodi? Beh, generalmente, non equispaziati, a meno di casi particolari, quali la presenza di periodicità.

Vogliamo a questo punto provare a comprendere meglio questo metodo, almeno sotto il punto di vista pratico, e a questo fine ci poniamo una domanda: qual è la differenza tra il metodo delle differenze finite e il metodo di collocazione? Beh, innanzitutto, partiamo dai punti in comune: entrambi richiedono una discretizzazione! la differenza è molto semplice: il metodo di collocazione si basa su di una approssimazione delle derivate, e dunque, mediante le formule delle differenze finite, si finisce per ottenere tutto il problema in una sola variabile. Con il metodo di collocazione, le soluzioni del problema differenziale sono i c_i: scelta (si spera in maniera intelligente, in modo da rendere il sistema non singolare e ben definito), e la spaziatura tra i nodi,

al fine di risolvere il sistema è necessario semplicemente calcolare i valori dei c_i. Ora è la soluzione stessa ad essere approssimata nel senso che non si introduce un'approssimazione della derivata nel senso di *funzione derivata*, ma si sceglie una funzione ben nota per approssimare un certo comportamento, imponendo condizioni, dunque le derivate delle funzioni esisteranno, come derivate della combinazione lineare di elementi della base scelta. Con il metodo di collocazione l'operatore di derivazione va applicato alle funzioni approssimate, ma l'operatore è quello tradizionale; con il metodo alle differenze finite, quello che si fa è introdurre una sorta di approssimazione dell'operatore di derivazione.

Esercizio di esempio

Si consideri il seguente problema differenziale:

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

Come approssimante, si ipotizzi di considerare la seguente:

$$y_N(x) = \sin\left(\frac{\pi}{2}x\right) + \sum_{i=1}^{N} c_i \sin(i\pi x)$$

Questa approssimante è scelta molto bene: essa è in grado intrinsicamente di soddisfare le condizioni al contorno del sistema: $y_N(0) = 0$, $y_N(1) = 1$ (come si può dimostrare provando a sostituire!).

Come spaziatura, si sceglie:

$$x_n = \frac{2n-1}{2N}, \quad n = 1 \div N$$

Di solito, con il metodo di collocazione, l'unica cosa che non si riesce a soddisfare è l'equazione differenziale (triste ma vero); calcoliamo, al fine di sostituire nel sistema, le espressioni delle derivate delle funzioni approssimanti:

$$y'_N(x) = \frac{\pi}{2}\cos\left(\frac{\pi}{2}x\right) + \pi \sum_{i=1}^{N} c_i \cos(i\pi x)$$

$$y''_N(x) = -\frac{\pi^2}{4}\sin\left(\frac{\pi}{2}x\right) - \pi^2 \sum_{i=1}^{N} c_i \sin(i\pi x)$$

Approssimando direttamente e calcolando il risultato finale, sostituendo nel sistema, si ottiene:

$$\left(1 - \frac{\pi^2}{4}\right) \sin\left(\frac{\pi}{2}x\right) + \sum_{i=1}^{N} c_i \left(1 - \pi^2 i^2\right) \sin(i\pi x_n) = x_n^2$$

Si risolve a questo punto il sistema differenziale derivante da questa equazione, a meno che non si veda che esso è singolare o non ben definito; il risultato sarà l'insieme dei c_i!

Capitolo 2

Metodi alle differenze finite per le equazioni alle derivate parziali

2.1 Introduzione

Le equazioni alle derivate parziali sono la naturale generalizzazione, in più variabili, delle già analizzate equazioni alle derivate parziali ordinarie. Dal momento che queste funzioni possono dipendere dunque da più variabili indipendenti, ciascuna derivata dovrà essere identificata rispetto a una delle variabili di derivazione; data u la soluzione del problema alle derivate parziali (quella che prima chiamavamo y), essa avrà forma:

$$u(x_1, x_2, ..., x_n)$$

Nel caso di un oggetto tridimensionale, per esempio potrebbero essere considerate tre o quattro variabili: le tre variabili spaziali e il tempo, o solo le tre spaziali, o magari neanche. Si è detto che le derivate vanno espresse specificando la variabile di derivazione; si definisce *derivata parziale* rispetto ad esempio a x_2 di u l'espressione:

$$\frac{\partial}{\partial x_2} u(x_1, x_2, ..., x_n)$$

Un'equazione alle derivate parziali è un'equazione in cui si ha qualcosa di questo tipo:

$$F\left(u, \frac{\partial}{\partial x} u, ..., \frac{\partial^m}{\partial x_3} u, ...\right)$$

Ossia che coinvolge la soluzione del problema u e un certo numero di sue derivate parziali. Si definisce, anche in questo caso, l'*ordine* dell'equazione alle derivate parziali come l'ordine massimo di derivazione all'interno dell'equazione.

La funzione $u(x_1, x_2, ..., x_n)$, è soluzione se, sostituita in F, soddisfa l'equazione; noi, come soluzioni, considereremo solo soluzioni almeno classiche, dunque u dovrà essere continua su tutte le derivate parziali fino all'ordine del problema e all'interno dell'intervallo di definizione.

Un'equazione alle derivate parziali può essere lineare o non lineare; in questo secondo caso, esiste una sottocategoria, ossia quella delle equazioni quasi-lineari: si tratta di equazioni lineari, nelle derivate di ordine massimo.

Introducendo (e usando di qui in poi) la notazione:

$$u_{xx} = \frac{\partial^2 u}{\partial x^2}(x, y)$$

Ricordando che, valendo la continuità, vale il teorema di Schwartz, dunque l'ordine di derivazione è ininfluente, la forma più generica di un'equazione alle derivate parziali del secondo ordine sarà:

$$au_{xx} + bu_{xy} + cu_{yy} + du_x + eu_y + fu = g$$

I vari coefficienti possono essere costanti o funzioni, o addirittura funzioni di u, rischiando di rendere non lineare l'equazione. Consideriamo rapidamente due esempi:

$$u_{xx} - uu_{yy} = 1$$

Questa è un'equazione non lineare.

$$au_{xx} + bu_{xy} + cu_{yy} + F(u, u_x, u_y) = g$$

Questa è un'equazione quasi-lineare: isolando le derivate di ordine massimo rispetto alle altre, se esse formano una propria combinazione lineare (non c'è moltiplicazione per una funzione di u), allora l'equazione è detta quasi-lineare. Questo è importante perchè, come vedremo, queste funzioni hanno proprietà interessanti.

Oltre a equazioni alle derivate parziali di ordine n esistono sistemi di equazioni alle derivate parziali; noi considereremo prevalentemente problemi base, semplici, e metodi numerici classici, applicati su equazioni del primo o del secondo ordine; si noti che questo fatto non è assolutamente riduttivo, dal momento che i problemi di ordine superiore vanno riportati a problemi di questo tipo, dunque finiscono per essere risolti nello stesso modo.

Analizzeremo più nel dettaglio a questo punto tre tipi di problemi, rappresentanti di fatto i tre più importanti problemi alle derivate parziali, su cui si basano poi tutti gli altri.

2.1.1 Problema delle onde (problema iperbolico)

Data una corda di sezione puntiforme (unidimensionale), si supponga di fissarla ai due estremi; a questo punto si prende la corda, la si pizzica, tenendola tesa, dunque per $t = 0$ la si lascia andare, e questa inizierà a oscillare. La nostra domanda è: per ogni istante di tempo, quale sarà la posizione assunta dalla corda, in ciascun punto dello spazio considerato? Beh, questo sarebbe un problema non lineare, ma, considerando piccole oscillazioni, si può supporre che esso sia linearizzabile. Dunque, l'equazione differenziale, che terrà conto della densità ϱ e della tensione T della fune, sarà:

$$u_{xx}(x,t) - \frac{\varrho}{T} u_{tt}(x,t) = 0, \qquad 0 < x < L, \quad t > 0$$

Ora, considereremo i domini aperti.
Si definiscono a questo punto due regioni:

- La regione di definizione dell'equazione differenziale; in questo caso:

$$\mathcal{R} = \{(x,t), 0 < x < L, t > 0\}$$

- Il dominio spaziale di $\mathcal{R}$:

$$D = (0, L)$$

  Nel caso non ci sia dipendenza dal tempo t, si può dire che le due definizioni siano coincidenti.

Ora, come ben noto, l'equazione differenziale ha infinite soluzioni; imponendo delle condizioni al contorno e iniziali, è possibile, di tutte le soluzioni, solo quella che ci interessa, supponendo al solito che essa esista e sia unica. Si vuole fare presente che, per questo tipo di problema, non sarà banale dire che la soluzione esista e sia unica.

Abbiamo detto che gli estremi devono essere fissati; questo significa che, per ogni istante di tempo:

$$u(0,t) = 0; \quad u(L,t) = 0$$

Supponiamo dunque che stiamo tirando la corda; si avrà:

$$u(x,0) = u_0(x)$$

Ossia, imponiamo, per una certa x, il fatto che, al tempo $t = 0$, la corda sia tesa a una certa posizione.

In questo caso, manca ancora una cosa: mancano informazioni sulla velocità iniziale! Sembra un'osservazione stupida, ma in realtà va imposta: dal momento che abbiamo due derivate spaziali e due derivate temporali, è necessario, per ciascuna derivazione, introdurre una condizione iniziale (nel caso di derivate temporali) o al contorno (nel caso di derivate spaziali); dunque:

$$u_t(x,0) = v_0(x)$$

Ossia si impone che, quando si lascia la corda, essa abbia una certa velocità (per esempio, 0).

La soluzione, almeno speriamo, dovrebbe essere una sola; è la soluzione classica? Beh, beh, in questo caso, è necessario soddisfare delle condizioni di raccordo: qualsiasi sia t, devono essere rispettate le funzioni sui vincoli, nel senso che, quando $x \to 0$, u_0 deve tendere alla condizione iniziale, dunque devono essere, oltre alle condizioni iniziali e ai bordi, essere rispettate le condizioni ai limiti verso i vincoli; il raccordo va creato perchè tutte le funzioni, tendendo ai vincoli, devono dare lo stesso valore al fine di avere soluzione classica. Nel caso di problemi iperbolici come questo, le eventuali discontinuità vengono propagate nel dominio secondo delle *linee caratteristiche*, introducendo la non-classicità nella soluzione. Bisogna essere coscienti del fatto che dati iniziali e al bordo devono essere continui o comunque garantire queste condizioni.

2.1.2 Problema della conduzione del calore (problema parabolico)

Dato un filo metallico di lunghezza l, densità ϱ, calore specifico C_p, conduttività termica K, si han due casi:

- Filo termicamente isolato;
- Filo termicamente isolato tranne nell'estremo L.

A questo punto, come varia in ogni istante la temperatura sul filo? Vediamo:

$$\begin{cases} u_{xx}(x,t) - \frac{\varrho C_p}{K} u_t(x,t) = 0, \quad 0 < x < L, \quad t > 0 \\ u(x,0) = T_1, \quad 0 \leq x \leq L \\ u(0,t) = T_0 \quad t > 0 \\ u(L,t) = T_2, \quad t > 0 \end{cases}$$

Ossia, abbiamo imposto che, per $t = 0$, ci sia una certa temperatura T_0 iniziale su tutto il filo, dunque imposte le temperature agli estremi.

Si noti che vi sarà un salto, una discontinuità, nelle condizioni al contorno; questo nei problemi iperbolici sarebbe critico, dal momento che le irregolarità viaggiano sulle cosiddette linee caratteristiche; nei problemi parabolici (ed ellittici) questo problema non sussiste: le discontinuità restano confinate al singolo punto, senza dare problemi nel resto del dominio. Si noti inoltre che l'equazione del calore è del secondo ordine ma solo rispetto allo spazio, dunque non sarà necessario introdurre la seconda condizione iniziale.

A questo punto, parlando di condizioni da imporre al bordo, ve ne sono di due-tre tipi:

- Se si impone come condizione il valore assunto, sul bordo, dalla soluzione desiderata dell'equazione differenziale, quella che si sta imponendo è una **condizione di Dirichlet**;
- Se si impone come condizione il valore della derivata della soluzione desiderata sul bordo, questa condizione sarà detta di **condizione di Neumann**. Questa spesso è utilizzata proprio nei problemi di propagazione del calore, dove i modelli di questo tipo vanno per l'appunto modellati mediante derivate (in modo da tenere conto delle differenze di calore nelle varie zone).
- Se si hanno, come condizioni, combinazioni di condizioni di Neumann e di Dirichlet, si ha a che fare con **condizioni di Robin**, o **miste**.

Con condizioni di tipo Neumann, c'è da stare attenti, dal momento che l'unicità della soluzione non è garantita.

2.1.3 Problema della membrana elastica (problema ellittico

L'ultimo problema, in questo caso un problema stazionario, verrà presentato immediatamente: si considera, anzichè la variabile temporale, due variabili spaziali, considerando i due punti spaziali della membrana che si trovano in diverse posizioni. Definito l'operatore laplaciano come:

$$\Delta u = \nabla^2 u$$

Si può definire l'equazione della membrana elastica come:

$$-\Delta u(x, y) = \frac{P}{T}$$

Dove P è la pressione, e T la tensione della membrana. Questa equazione è detta *equazione di Laplace non omogenea*, o più comunemente **equazione di Poisson**. In questo caso:

$$x, y \in \mathbb{C}$$

Se si vincola la cornice alla membrana, al bordo della membrana, si fa in modo da ottenere una condizione del tipo:

$$u(x, y) = g(x, y)$$

Anche questa è una condizione di tipo Dirichlet; anche in questo caso si può dimostrare che i problemi concernenti le cosiddette *linee caratteristiche* non sono fondamentali, e che l'unico punto in cui è necessario veramente prestare attenzione è l'uso delle condizioni di Neumann, che non garantiscono l'unicità della soluzione. Anche in questo caso, dunque, come prima, la regolarità dipende esclusivamente dalla regolarità dei coefficienti dell'equazione alle derivate parziali.

2.2 Linee caratteristiche e classificazione dei problemi differenziali

Si considereranno, come già accennato, solo modelli per il primo e per il secondo ordine, generalmente lineari, per quanto la teoria che si sta introducendo sia assolutamente valida anche per quanto concerne i sistemi non lineari. Verrà evidenziata l'importanza delle già citate *linee caratteristiche* nell'ambito dello studio dei problemi differenziali, elementi introducibili solo se sussiste almeno la condizione di **quasi linearità** (dunque se l'equazione è lineare o quasi lineare).

2.2.1 Problemi del primo ordine

Si considerino in questo primo momento solo problemi del primo ordine, considerando equazioni nella forma quasi-lineare; si avrà dunque generalmente a che fare con un'equazione del tipo:

$$au_x + bu_y = c$$

Dove:

$$u = u(x, y)$$

$$a = a(x, y, u)$$

$$b = b(x, y, u)$$

$$c = c(x, y, u)$$

Ossia, dove u, soluzione dell'equazione differenziale, è generalmente funzione solo delle variabili indipendenti x e y, mentre i coefficienti a, b e c possono essere funzione anche della soluzione dell'equazione differenziale, u. Nella fattispecie, se il coefficiente a è funzione solo di x e y l'equazione è quasi lineare, e se anche b lo è l'equazione risulta essere lineare.

Detto ciò, parliamo di linee caratteristiche: come ben noto, data la regione di definizione del problema differenziale $\mathcal{R}$, solitamente una regione limitata dell'intero piano, esistono infinite curve che lo attraversano; le linee caratteristiche sono quelle curve del piano tali per cui il coefficiente angolare della curva sul piano vale:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{b}{a}$$

Se quindi associamo alle equazioni differenziali una condizione, si ha un problema di Cauchy; esso, generalmente, va definito su di una curva, Γ, ottenendo il seguente problema di Cauchy:

$$\begin{cases} au_x + bu_y = c \\ u = u_a(x, y), \quad (x, y) \in \Gamma \end{cases}$$

A questo punto, una definizione (che in realtà sarebbe un teorema, ma noi lo considereremo come un assunto): se la curva Γ sulla quale definiamo le condizioni iniziali del problema è una curva caratteristica, viene meno l'esistenza e unicità del problema. Questo significa che le curve caratteristiche sono quelle che fanno cadere l'esistenza e unicità del problema di Cauchy, nel caso in cui la condizione al contorno sia data su di essa.

Stiamo per ora parlando solo del problema del primo ordine; come si può determinare, in un problema del primo ordine, quale sia o quali siano le linee caratteristiche? Siamo interessati a questa domanda dal momento

che, se sapessimo come determinare le linee caratteristiche, sapremmo anche come evitare di utilizzarle, al fine di non definire su di esse le condizioni iniziali, in modo da garantire l'esistenza e unicità della soluzione del problema differenziale. Come si trovano dunque? Beh, abbiamo detto che:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{b}{a}$$

Sappiamo però che b e a sono funzione delle variabili e della soluzione dell'equazione differenziale, dunque:

$$y'(x) = \frac{b(x, y, u)}{a(x, y, u)}$$

Un'osservazione: dal momento che, nel caso più generale, u è funzione di x e y, si può dire che il rapporto di b e a sia generalmente una funzione f delle sole variabili x e y. Determinare a priori tuttavia quale sia l'espressione del rapporto non è banale, dal momento che si dovrebbe determinare la soluzione del problema differenziale prima ancora di aver trovato le linee caratteristiche:

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

Le curve caratteristiche sono dunque le soluzioni di questa equazione differenziale. Dal momento che, generalmente, al fine di determinare le linee caratteristiche critiche per un problema, è necessario vedere quali siano quelle che passano per un punto, è sufficiente scegliere, di tutte le infinite soluzioni di questo problema differenziale, quella che passa per il punto in cui si vuole definire la condizione iniziale, per esempio:

$$y(x_0) = y_0$$

Due sottocasi: nel caso in cui a e b non abbiano dipendenza da u, si può dire che il rapporto dei coefficienti sia determinabile semplicemente come:

$$f(x, y) = \frac{b(x, y)}{a(x, y)}$$

In questo modo, si può integrare questo rapporto (e risolvere l'equazione differenziale) mediante integrali doppi. Caso ancora più semplificato è quello in cui a e b sono costanti; in tal caso, l'equazione differenziale si riduce a:

$$y'(x) = \frac{b}{a} \Longrightarrow y(x) = \frac{b}{a}x + k$$

Dove k è la costante di integrazione a meno di cui si definisce la primitiva.

2.2.2 Problemi del secondo ordine

Si vuole a questo punto trattare i problemi del secondo ordine, partendo dall'espressione generale assumibile dalle equazioni di questo tipo (in modo da avere almeno la quasi-linearità):

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

Dove:

$$a = a(x, y, u, u_x, u_y)$$

$$b = b(x, y, u, u_x, u_y)$$

$$c = c(x, y, u, u_x, u_y)$$

$$f = f(x, y, u, u_x, u_y)$$

Dal momento che l'equazione è lineare nelle derivate di ordine massimo, questa equazione sarà almeno quasi lineare.

A questo punto, domanda: come è possibile classificare le diverse equazioni derivabili da questa? Beh, anche in questo caso, ci vengono incontro le linee caratteristiche, nella fattispecie fornendoci informazioni legate al numero di linee caratteristiche esistenti. Le linee caratteristiche, come già detto, sono delle curve Γ funzioni di y e x, in cui $y = y(x)$; è anche possibile esprimere la curva in termini parametrici, usando un parametro temporale fittizio τ. Si può dimostrare che, al fine di determinare le linee caratteristiche per un problema differenziale del secondo ordine, è necessario risolvere il seguente trinomio di secondo grado:

$$a\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2 - b\frac{\mathrm{d}y}{\mathrm{d}x} + c = 0$$

Data dunque una $y'(x)$, ossia il coefficiente angolare della curva, se esso inserito nell'equazione soddisfa l'identità, la curva è detta *linea caratteristica*. Essendo questo un trinomio, esso dovrà avere radici, in numero al più uguale a 2; quello che si fa per caratterizzare le diverse equazioni differenziali è dunque studiare, mediante il discriminante (Δ), il numero di radici (e il loro campo di esistenza); come è noto:

$$\Delta = b^2 - 4ac$$

A questo punto, vi sono tre casistiche:

- Se $\Delta > 0$, ossia se esistono due coefficienti angolari reali e distinti che soddisfano l'equazione, l'equazione è detta **iperbolica**;
- Se $\Delta = 0$, esiste un solo coefficiente angolare (con molteplicità doppia) in grado di soddisfare l'equazione; in tal caso, l'equazione è detta **parabolica**;
- Se $\Delta < 0$, esistono nuovamente due coefficienti angolari distinti, ma in questo caso complessi coniugati, dunque non reali; in tal caso l'equazione è detta **ellittica**.

Come è stato accennato precedentemente, è possibile caratterizzare qualsiasi equazione, anche lineare, mediante le linee caratteristiche; unica differenza sta nella *difficoltà* del problema: se il problema è lineare, la soluzione è abbastanza semplice; se il problema è quasi-lineare, significa che si ha dipendenza dalla u nei coefficienti, cosa che provocherebbe la necessità di conoscere la soluzione del problema a priori. Si provi dunque a caratterizzare le equazioni già note:

- Data l'equazione della corda vibrante:

$$u_{xx} - \frac{1}{\mathrm{c}^2} u_{yy} = 0$$

  Si ha $a = 1$, $b = 0$, $c = -\frac{1}{\mathrm{c}^2}$

  In questo caso:

$$\Delta = 0 + \frac{4}{\mathrm{c}^2} > 0$$

  Dunque l'equazione è iperbolica.

- Data l'equazione del calore:

$$u_y + k u_{xx} = 0$$

  Si ha: $a = k$, $b = 0$, $c = 0$

  Dunque:

$$\Delta = 0$$

  Indipendentemente da k, l'equazione risulterà sempre essere parabolica.

- Data l'equazione della membrana elastica:

$$u_{xx} + u_{yy} = 0$$

  Si ha che $a = 1$, $b = 0$, $c = 1$:

$$\Delta = 0 - 4 < 0$$

  Si potrebbe considerare una variante: se si considerasse la seguente equazione:

$$u_{xx} + cu_{yy} = 0$$

  In questo caso la natura dell'equazione dipenderebbe dal segno di c.

- Equazione di Tricomi:

$$yu_{xx} + u_{yy} = 0$$

  In questo caso, si ha $a = y$, $b = 0$, $c = 1$; dunque:

$$\Delta = -4y$$

  Questo significa che, a seconda di dove ci si trova, l'equazione può essere sia ellittica sia parabolica sia iperbolica. Dipende sostanzialmente dalla curva scelta.

Un generico problema di Cauchy associato a equazioni del secondo ordine avrà dunque forma del tipo:

$$\begin{cases} au_{xx} + bu_{xy} + cu_{yy} = f \\ u(x_0) = u_0, \quad (x_0, v_0) \in \Gamma \\ \frac{\partial u}{\partial \vec{n}} = v_0, \quad (x_0, v_0) \in \Gamma \end{cases}$$

In questo caso, se Γ è una delle linee caratteristiche prima descritte, questo problema non avrà soluzione univoca, altrimenti sì. Per questo motivo è fondamentale capire se Γ è una curva caratteristica: è fondamentale verificare l'esistenza e unicità della soluzione del problema! Riassumendo:

- Nel caso dell'equazione del primo ordine, nel caso il rapporto $\frac{b}{a}$ esista, è come se l'equazione avesse discriminante positivo, dunque l'equazione è iperbolica; nel caso non sia definito il rapporto, non si può dire nulla sulla natura dell'equazione;

- Nel caso delle equazioni del secondo ordine, è necessario guardare il segno del discriminante, al fine di caratterizzare l'equazione.

Si ricordi che nel caso del problema iperbolico, i dati iniziali devono essere assegnati su di una linea non caratteristica, ma oltre a ciò non devono presentare discontinuità, perchè in problemi di questo tipo accade che le irregolarità ai bordi nel tempo si propagano sulle linee caratteristiche che escono dai punti della discontinuità; è dunque possibile prevedere la propagazione delle irregolarità, semplicemente determinando la soluzione delle equazioni differenziali che permettono di determinarle, dunque imponendo come condizione per determinarla il passaggio per il punto sfortunato.

2.2.3 Esempio teorico/pratico

Data una funzione u = u(x, t), definita su $0 < x < 1$, $t > 0$, soluzione dell'equazione differenziale:

$$u_t + au_x = 0, \quad a > 0, \quad 0 < x < 1, t > 0$$

Impostiamo come condizione iniziale, dunque per $t = 0$, il fatto che:

$$u(x, 0) = u_0(x)$$

Ovviamente, definendo per $0 \leq x \leq 1$; in questo caso, si includono anche i bordi all'interno dell'insieme di definizione.

Ricordando la teoria, si sa che:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{b}{a} = \frac{1}{a}$$

Dal momento che consideriamo a costante.Dunque, le linee caratteristiche sono le curve tali per cui:

$$t(x) = \int \frac{1}{a}\mathrm{d}x + k = \frac{1}{a}x + k$$

Cosa si può dire a questo punto? Beh, se $a > 0$, si hanno rette a pendenza positiva, dunque che, a partire dal bordo sinistro, si propagano verso il bordo destro. Si può dimostrare che i dati assegnati si propagano lungo la direzione delle linee caratteristiche, sulla loro direzione, a velocità costante. Quello che serve, dunque, al fine di avere un problema ben definito, è avere una sorta di *sorgente* in grado di garantire che il segnale che parte da quello che per l'appunto deve essere il nostro punto iniziale continui a esistere. Nel caso in cui $a > 0$, dunque, il segnale deve essere messo sul bordo **sinistro** del

dominio considerato, ossia su $x = 0$. In questo caso, dunque, avremo un problema differenziale del tipo:

$$\begin{cases} u_t + au_x = 0 \quad 0 < x < 1, t > 0 \\ u(x,0) = u_0(x), \quad 0 \leq x \leq 1 \\ u(0,t) = f(x), \quad t > 0 \end{cases}$$

In questo modo si impone la presenza di una sorta di generatore di segnale, che continua a far perdurare l'esistenza del segnale.

Nel caso in cui $a < 0$, al contrario, si avrà propagazione in senso opposto, dunque conviene mettere la condizione sul bordo destro, ottenendo dunque:

$$u(1,t) = f(t)$$

In questo modo, si può dunque risolvere il problema dapprima studiando le linee caratteristiche che, come si sta vedendo, forniscono un grosso numero di informazioni interessanti, dopodichè si può partire con la soluzione del problema (o con la sua definizione). Questo esempio pratico era collegato a un'equazione molto importante per la fisica matematica: l'equazione del trasporto.

2.3 Schema di soluzione: metodo delle differenze finite

Questo schema di soluzione deriva da quello precedentemente studiato per le equazioni differenziali ordinarie. La prima differenza che si può incontrare è il fatto che vanno discretizzate entrambe le variabili, ottenendo dunque:

$$h = \frac{1}{N}; \qquad k = \frac{1}{N}$$

Si approssima dunque il dominio con un reticolo; indicando con gli indici i e j, si ottengono i punti (x_i, t_j):

$$1 \leq i \leq N - 1 \qquad j = 1, 2...$$

Dunque, il sistema risultante diventa:

$$\begin{cases} u_t(x_i, t_j) + au_{xx}(x_i, t_j) = 0 \\ u(x_i, t_j) = u_0(x_i), i = 0 \div N \\ u(x_0, t_j) = f(t_j), j > 0 \end{cases}$$

Ci sono troppe incognite nel problema rispetto alle equazioni fornite; quello che si deve fare è dunque sfruttare la discretizzazione e introdurre le formule

approssimate, in modo da costruire lo schema alle differenze finite, approssimando le derivate delle funzioni nel punto con queste funzioni fittizie. Si hanno:

$$u_t(x_i, t_j) = \frac{u(x_i, t_{j+1} - u(x_i, t_j)}{k} + O(k) = \frac{u(x_i, t_j - u(x_i, t_{j-1})}{k} + O(k)$$

E

$$u_x(x_i, t_j) = \frac{u(x_{i+1}, t_j - u(x_i, t_j)}{h} + O(h) = \frac{u(x_i, t_j - u(x_{i-1}, t_j)}{h} + O(h)$$

In questo caso le formule migliorate, simmetriche, non sono presentate e non è necessario utilizzarle, dal momento che servirebbero due condizioni ai bordi, cosa che non intendiamo fornire in problemi di questo tipo, dunque nel caso delle equazioni alle derivate parziali utilizzeremo sempre queste relazioni.

Si provi a questo punto a combinare queste equazioni: utilizzando la prima formula per approssimare il tempo, e la prima formula per approssimare lo spazio, si potrebbero coinvolgere questi tre nodi:

Dunque:

$$\frac{u(x_i, t_{j+1} - u(x_i, t_j)}{k} + O(k) + a\frac{u(x_{i+1}, t_j - u(x_i, t_j)}{h} + O(h) = 0$$

Trascurando gli $O()$, commettendo dunque gli errori di troncamento, si definisce un termine α in modo da alleggerire la notazione, come:

$$\alpha = a\frac{k}{h}$$

Dunque, si ottiene, approssimando:

$$u_{i,j+1} - u_{i,j} + \alpha(u_{i,j} - u_{i-1,j}) = 0$$

In questo modo, livello dopo livello, si reitera questa espressione, passando sempre a livelli successivi; si ricava dunque:

$$u_{i,j+1} = \alpha u_{i-1,j} + (1 - \alpha)u_{i,j}$$

A questo punto, è possibile verificare se, con questo schema, si possa ottenere, almeno qualitativamente, il risultato desiderato, ossia l'approssimazione delle soluzioni per tutta la griglia.

Per costruire un punto è necessario, come si può leggere dall'espressione ricavata, sono necessari il punto sottostante e il punto *sotto a sinistra*; non si avranno dunque problemi nel generare la prima riga. Nel caso non siano state correttamente definite le condizioni al bordo sinistro, tuttavia, questo metodo non sarà in grado di generare tutti gli elementi della seconda riga, della terza, e così via, costruendo una sorta di *triangolo*. Questa informazione ci era già nota, di fatto: questo *triangolo* segue la direzione della linea caratteristica! Dunque, studiare la linea caratteristica permette di determinare a priori queste informazioni, posizionare correttamente le condizioni e dunque ottenere un problema ben posto.

Discussione

Allora, una volta terminato il problema, riprendiamolo dal principio, in modo da enfatizzare alcuni aspetti fondamentali. Si sapeva che:

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

Le condizioni che garantiscono l'esistenza e l'unicità della soluzione classica sono:

$$u_0 \in \mathrm{C}^{(1)}([0,1]), \in \mathrm{C}^{(1)}([0,1])$$

$$f(0) = u_0(0)$$

$$f'(0) + au_0'(0) = 0$$

$$u \in \mathrm{C}^{(1)}([0,1] \times [0,\infty))$$

Ossia, rispettivamente, si richiede che la soluzione e la funzione f abbiano derivate continue fino al primo ordine, che le condizioni di raccordo sulla funzione e sulle derivate siano rispettate, e che dunque la soluzione sia classica nel chiuso.

Quello che abbiamo fatto ora, ossia applicare uno schema alle differenze finite, richiede soluzione più che classica; questo significa che, per essere formali, andrebbero aggiunte condizioni aggiuntive:

$$u \in \mathrm{C}^{(2)}([0,1])$$

E inoltre, una condizione di raccordo in più, sulle derivate seconde di u e f.

Una nota: utilizzando lo schema dell'altra volta, di fatto si suppone di collocare l'equazione differenziale su i, j, prima al livello 0, dunque per $j = 0$, poi andare avanti; questo significa che bisogna supporre che l'equazione sia continua su tutto il chiuso, dunque pretendere non solo $t > 0$, ma $t \geq 0$! Se non si collocasse qui l'equazione, infatti, lo schema non potrebbe funzionare; supponiamo dunque che l'equazione differenziale sia soddisfatta anche sul bordo iniziale, in modo da poterla collocare anche su $t = 0$: questo fatto è dovuto al fatto che lo schema funziona solo se il sistema è sufficientemente regolare. Come noto, poi è necessario fare:

$$u_{i,j} \sim u(x_i, t_j)$$

Dunque, si approssimano i valori della soluzione su ciascun nodo (si vuole ricordare che ora il dominio è un piano, dunque bidimensionale; le soluzioni dell'equazione differenziale staranno su un terzo asse, z, formando una *superficie*. Il metodo delle differenze finite può essere schematizzato nel seguente algoritmo:

```
for j = 0 : J
for i = 1 : N
u(i, j+1) = alpha u(i-1, j) + (1 - alpha) u(i, j)
```

A questo punto, un'osservazione: facendo tendere a 0 sia h sia k, ci si potrebbe aspettare che l'approssimazione $u_{i,j}$ tenda a degenerare nella soluzione *esatta* dell'equazione differenziale; questo in realtà capita esclusivamente se lo schema in questione è *convergente*: fissata una certa tolleranza, in queste condizioni, si potrebbe scegliere dei passi h e k sufficientemente piccoli, in modo da avere un discostamento al più pari alla tolleranza. Il *problema* del metodo però non è tanto uno dei due parametri prima citati, quanto α: sperimentalmente si potrebbe vedere che lo schema converge solamente se:

$$0 \leq \alpha \leq 1$$

Altrimenti, se $\alpha > 1$, lo schema diverge, ossia gli errori tendono a esplodere.

Bisogna dunque scrivere questo schema, al fine di dimostrarne le proprietà utilizzando l'algebra lineare, in forma matriciale, *sintetizzando* l'algoritmo. Esso funziona così: nota la soluzione al livello, al tempo t_j, si può produrre quella al tempo t_{j+1}. Si ha dunque che:

$$\underline{u}_0 = (u_{0,1}, u_{0,2}, ..., u_{0,N})$$

Da qui, si possono calcolare una a una le varie componenti:

$$u_{1,j+1} = (1-\alpha)u_{1,j}$$

$$u_{2,j+1} = \alpha u_{1,j} + (1-\alpha)u_{2,j}$$

$$u_{3,j+1} = \alpha u_{2,j} + (1-\alpha)u_{3,j}$$

...

$$u_{N,j+1} = \alpha u_{N-1,j} + (1-\alpha)u_{N,j}$$

Data questa notazione, ora è possibile fare l'analisi della stabilità dello schema con maggiore facilità; si introduca una perturbazione del vettore dei dati iniziali:

$$\overline{\underline{u}}_0 = \underline{u}_0 + \underline{E}_0$$

Dunque, la successione a questo punto diventerà:

$$\overline{\underline{u}}_{j+1} = \underline{\underline{A}}\,\overline{\underline{u}}_j + \alpha \underline{u}_j$$

Dunque, definito il vettore di perturbazione, per ogni j:

$$\underline{E}_j = \overline{\underline{u}}_j - \underline{u}_j$$

Si ha:

$$\underline{E}_{j+1} = \underline{\underline{A}}\,\underline{E}_j, \quad j = 0, 1, ...$$

Dunque, considerando la norma del vettore:

$$||\underline{E}_{j+1}|| = ||\underline{\underline{A}}\,\underline{E}_j|| \leq ||\underline{\underline{A}}||\,||\underline{E}_j||$$

Si avrà ciò; normalmente, come norme, si considerano o la norma 1, o la norma 2, o la norma ∞; dal momento che per ora stiamo però considerando vettori di dimensione finita, tutte le norme sono equivalenti, ossia producono lo stesso risultato.

La condizione sufficiente per garantire la stabilità dell'algoritmo, ossia il fatto che gli errori non divergano, è:

$$\left|\left|\underline{\underline{A}}\right|\right| < 1$$

In questo modo, se questa condizione è soddisfatta, si avrà:

$$\left|\left|\underline{E}_{j+1}\right|\right| < \left|\left|\underline{E}_j\right|\right| < ... < ||\underline{E}_0||$$

Questo significa che, se la norma della matrice $\underline{\underline{A}}$ è minore di 1, la norma dell'errore presente è sempre minore o uguale di quella introdotta al livello (al tempo) iniziale.

Prima si parlava di **convergenza**, ora si parla di **stabilità**: esse non sono la stessa cosa:

- La convergenza per un metodo numerico è la garanzia che esso prima o poi tenda a un valore (si spera sensato);
- La stabilità per un metodo numerico è il fatto che gli errori non divergano, ma tendano ad annullarsi.

Noi saremmo in realtà interessati alla convergenza, ma essa da studiare è estremamente delicata; quello che possiamo fare dunque è ribaltare le carte in tavola, come un matematico ha dimostrato, implicando la convergenza a partire dallo studio della stabilità del metodo.

A questo punto: se si fanno tendere h e k (soprattutto h) a 0, il numero di nodi cresce, tendendo a ∞, dunque anche J, ossia il valore finale di j, tenderà a ∞; in questo modo, le norme dei vettori diventano serie, dunque non si ha più equivalenza delle varie norme: si potrà studiare la stabilità a partire dalla convergenza delle serie dei vettori di errore, ma ora si dovrà definire secondo quale norma: 1, 2, ∞, p. Lo schema sarà dunque detto stabile se, per $(h,k) \to 0$, $\left|\left|\underline{E}_j\right|\right| < c, \forall j$; dipendendo dalla norma scelta, dunque, si può verificare se essa sia uniformemente limitata.

Vi è però un risultato, sulla matrice $\underline{\underline{A}}$, che può essere molto interessante:

$$||A||_1 = \left|\left|A^{\mathrm{T}}\right|\right|_\infty$$

Questo può essere utilizzato in congiunzione al teorema che afferma che, se le norme p, per $p = 1$ e $p = \infty$ sono uguali, allora si può anche dire che lo siano tutte le norme *in mezzo*.

Studiamo dunque la norma 1: essa è la somma degli elementi colonna per colonna, e si prende il maggiore risultato per ciascuna colonna; questo sarà come dire:

$$|\alpha| + |1 - \alpha| = 1$$

Questo, ovviamente, solo se $\alpha \leq 1$, come si può vedere: il modulo tende a prendere la parte positiva, il valore assoluto, dunque si finisce per ottenere ciò.

A questo punto, si ha una condizione aggiuntiva, che può tornare molto utile: se lo schema è stabile indipendentemente dai valori di h e k, la stabilità è detta **stabilità incondizionata**; altrimenti, se il metodo risulta essere stabile solo dati dei vincoli su h e k, la stabilità è detta **stabilità condizionata**. Una condizione potrebbe essere ad esempio:

$$k \leq \frac{h}{a}$$

Ossia, k non può essere più grande di questo valore.

Questo vincolo può essere trovato anche per altre equazioni, quale quella del calore. Uno schema di soluzione, generalmente:

- Può essere instabile, dunque non convergere, dal momento che la stabilità implica la convergenza, e fornisce informazioni anche sulla velocità di convergenza;
- Può essere condizionatamente stabile, cosa che capita in metodi generalmente più semplici e veloci rispetto ai successivi, dunque da utilizzare in caso si disponga di specifiche leggere;
- Può essere incondizionatamente stabile; questo si usa se il vincolo sulla scelta del passo è troppo pesante, dunque se la condizione di stabilità è troppo pesante per essere applicata.

Introduzione di uno schema di soluzione incondizionatamente stabile

A partire dallo stesso problema, cambiando *l'incastro delle formule*, è possibile costruire uno schema di soluzione incondizionatamente stabile. Questo significa usare la seguente *molecola di calcolo*:

Ora l'equazione si colloca su di un nodo diverso da prima, considerando:

$$i = 1 \div N, \qquad j = 0, 1, ...$$

A questo punto, si riprendono le equazioni già precedentemente viste:

$$\frac{u(x_i, t_{j+1} - u(x_i, t_j)}{k} + O(k) + a\frac{u(x_{i+1}, t_j - u(x_i, t_j)}{h} + O(h) = 0$$

Si introducono gli errori di troncamento eliminando gli $O()$; si consideri:

$$\frac{f(x_1) - f(x_0)}{x_1 - x_0} + O(h), \quad \forall x_0 \leq x \leq x_1$$

Si può dimostrare che, per questa espressione, la differenza tra valore esatto e approssimato sia $O(h) + O(k)$, dunque una somma degli errori di troncamento. Si avrà, per lo schema:

$$u_{i,j+1} - u_{i,j} + \alpha(u_{i,j+1} - u_{i-1,j+1}) = 0$$

Dunque:

$$-\alpha u_{i-1,j+1} + (1 + \alpha)u_{i,j+1} = u_{i,j}$$

Provando graficamente la molecola di calcolo, è possibile vedere che essa sfrutta tutti i valori assegnati; se non fosse così, il metodo tenderebbe a divergere, dunque sarebbe un problema, dal momento che questo schema non funzionerebbe.

Dunque:

$$u_{i,0} = u_0(x_i), \quad i = 0 \div N$$

$$u_{0,j} = f(t_j), \quad j = 1, 2, ...$$

```
for i = 0, 1, ...
for i = 1 : N
u(i, j+1) = 1/(1 + alpha) [ alpha u(i-1, j+1) + u(i, j) ]
```

Questo schema non impone vincoli di alcun tipo, ottenendo, per questo caso fortunato, qualcosa di molto simile, in termini di difficoltà, allo schema precedente.

Considerazioni finali

Si vuole a questo punto proporre delle considerazioni finali, in modo da fissare meglio i concetti. Quello che si sta facendo è costruire schemi alle differenze finite; ciò non è difficile, dal momento che è sufficiente discretizzare, costruire la griglia, dunque *incastrare* le varie formule in modo da ottenere lo schema.

Sono stati costruiti due schemi: uno che *guarda avanti*, uno che *guarda indietro*. Gli schemi devono essere convergenti, nel senso che l'approssimazione $u_{i,j}$ deve essere tale da avere un errore che tenda a zero.

$$|u_{i,j} - (x_i, t_j)| \to 0, \quad h, k \to 0$$

Per verificare che ciò sia vero, è sufficiente verificare che i termini di errore vadano a 0, per h e k tendenti a 0: se ciò si verifica, si dice che il metodo numerico è **consistente**. Se lo schema è consistente, condizione necessaria e sufficiente per la convergenza è dunque la stabilità dello schema: data una perturbazione E_0, si crea un vettore di perturbazioni per ciascun i-esimo step; per $h \to 0$, il numero di punti tende a crescere: si prende dunque la norma del vettore, e si dice che:

$$\forall j, \quad ||E_j|| < c$$

Ossia, questo vincolo, data una costante c, deve essere sempre verificato, indipendentemente da h o k; se inoltre lo schema è stabile:

$$|u_{i,j} - (x_i, t_j)| = O(h) + O(k)$$

Si è inoltre parlato di stabilità condizionata e incondizionata: queste definizioni sono valide esclusivamente nel caso di schemi consistenti, e nel primo caso si hanno condizioni su h e k che, solo se soddisfatte, permettono di avere la stabilità del metodo; nel caso di stabilità incondizionata, non ci sono ulteriori condizioni necessarie.

Si noti che, spesso, i metodi che approssimano *tornando indietro*, come l'ultimo metodo analizzato, sono quelli che non richiedono condizioni (che garantiscono la stabilità incondizionata, se ben progettati); *guardando in avanti*, **spesso** (ma non sempre), si hanno schemi condizionatamente stabili.

2.4 Equazione delle onde

Verrà a questo punto costruito uno schema di soluzione alle differenze finite per quanto riguarda l'equazione delle onde; si ha a che fare con un'equazione del tipo:

$$\frac{\partial^2 \varphi}{\partial t^2} - \mathrm{c}^2 \frac{\partial^2 \varphi}{\partial \mathrm{x}^2} = 0$$

Con condizioni:

$$\mathrm{c} > 0, \quad 0 < \mathrm{x} < 1, \quad \mathrm{t} > 0$$

A questo punto verrà costruito lo schema, in seguito a un certo numero di osservazioni.

Osservazioni sulle linee caratteristiche

Ricordando come calcolare le linee caratteristiche, si intende studiarle, al fine di essere sicuri che le condizioni che verranno associate all'equazione per costituire il problema differenziale non ricadano su di esse. Si sa che:

$$a = -\mathrm{c}^2 \quad \mathrm{b} = 0 \quad \mathrm{c} = 1$$

Dunque, si ha che:

$$-\mathrm{c}^2 \left(\frac{\mathrm{dt}}{\mathrm{dx}}\right)^2 + 1 = 0$$

Dunque:

$$\frac{\mathrm{d}t}{\mathrm{d}x} = \pm\frac{1}{\mathrm{c}}$$

Dunque, le curve caratteristiche avranno, integrando, forma del tipo:

$$t = \pm\frac{1}{\mathrm{c}}x + \text{costante}$$

Vi sono due famiglie di curve caratteristiche, ossia due fasci di rette parallele (fasci impropri). Positivo il fatto che il livello iniziale (ossia $t = 0$) non sia una linea caratteristica: possiamo tranquillamente assegnare le condizioni al bordo su di esso.

Definizione del problema differenziale

Il problema differenziale avrà dunque la seguente forma:

$$\begin{cases} \frac{\partial^2\varphi}{\partial t^2} - \mathrm{c}^2\frac{\partial^2\varphi}{\partial \mathrm{x}^2} = 0 \\ \varphi(x,0) = f(x), 0 < x < 1 \\ \varphi_t(x,0) = g(x), 0 < x < 1 \\ \varphi(0,t) = \alpha(t), t \geq 0 \\ \varphi(1,t) = \beta(t), t \geq 0 \end{cases}$$

Due osservazioni:

- Si richiede, per le condizioni al bordo, che $t \geq 0$, e non solo che $t > 0$; il motivo verrà evidente in seguito, ma per ora si sappia che non è un errore, bensì che è assolutamente necessario; per quanto riguarda invece le condizioni iniziali, non è fondamentale che esse siano comprese.

- Si chiede, al fine di determinare una soluzione più che classica al problema, utilizzando correttamente il metodo delle differenze finite, che $\varphi \in \mathrm{C}^{(4)}$; si potrebbe avere anche di meno, ma al fine di avere la certezza conviene avere *ben soddisfatte* le condizioni richieste.

Si può effettuare, al fine di avere a che fare con equazioni più semplici, un cambio di variabili; si introduce una variabile y definita come:

$$y = \mathrm{ct}$$

In questo modo, si ha una nuova funzione u definita come:

$$u(x, y) = \varphi\left(x, \frac{y}{\mathrm{c}}\right)$$

L'equazione differenziale si riconduce dunque a:

$$u_{yy}(x, y) = u_{xx}(x, y)$$

A questo punto, sostituendo nelle condizioni, il problema differenziale diventa:

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

Applicazione del metodo delle differenze finite

Si vuole a questo punto applicare il metodo delle differenze finite, collocando l'equazione nei nodi interni; avremo:

$$h = \frac{1}{N}$$

Inoltre:

$$0 < i < N, j = 0, 1, 2...$$

Contando dunque anche $t = 0$. Si ha:

$$\begin{cases} u_{yy}(x_i, y_j) = u_{xx}(x_i, y_j) \\ u(x_i, 0) = f(x_i) \\ u_y(x_i, 0) = g_1(x_i) \\ u(0, y_j) = \alpha_1(y_j) \\ u(1, y_j) = \beta_1(y_j) \end{cases}$$

Si ricordi dunque l'espressione per l'approssimazione della derivata seconda:

$$u_{xx} = \frac{u(x_{i-1}, y_j) - 2u(x_i, y_j) + u(x_{i+1}, y_j)}{h^2} + O(h^2)$$

$$u_{yy} = \frac{u(x_i, y_{j-1}) - 2u(x_i, y_j) + u(x_i, y_{j+1})}{k^2} + O(k^2)$$

Trascurando gli errori di troncamento, del secondo ordine, sostituendo nell'equazione differenziale, si ottiene:

$$\frac{u(x_{i-1}, y_j) - 2u(x_i, y_j) + u(x_{i+1}, y_j)}{h^2} = \frac{u(x_i, y_{j-1}) - 2u(x_i, y_j) + u(x_i, y_{j+1})}{k^2}$$

Ciò va approssimato: si otterrà:

$$\frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h^2} = \frac{u_{i,j-1} - 2u_{i,j} + u_{i,j+1}}{k^2}$$

Questo è l'inizio dello schema alle differenze finite: se h e k tendono a zero, i due schemi convergono alle formule originali di derivazione, dal momento che gli errori di troncamento vanno come $O(h^2)$ e $O(k^2)$, dunque si può dire che questo schema sia consistente con l'equazione differenziale: si tratta di una corretta discretizzazione. In questo caso, la molecola di calcolo è a tre livelli: lo schema infatti richiede anche il livello superiore!

Riorganizzando l'equazione, si ottiene:

$$u_{i,j+1} = \frac{k^2}{h^2}\left[2u_{i-1,j} - 2u_{i,j} + u_{i+1,j}\right] + 2u_{i,j} - u_{i,j-1}$$

Definendo un parametro λ come:

$$\lambda = \frac{k}{h}$$

Si ottiene:

$$u_{i,j+1} = \lambda^2\left(u_{i-1,j} + u_{i+1,j}\right) + 2\left(1 - \lambda^2\right)u_{i,j} - u_{i,j-1}$$

Questa è una formula esplicita di $u_{i,j+1}$. A questo punto si ha però un problema: è necessario conoscere la u non solo in un livello, ma nei due livelli precedenti, dal momento che questo è uno schema a tre livelli. C'è anche da dire un'altra cosa: dal momento che questa equazione è del secondo ordine (cosa che ha creato questa differenza rispetto agli schemi precedentemente introdotti), abbiamo un'equazione in più: la seconda condizione

iniziale! Questa sarà il punto di partenza che permetterà di terminare la costruzione dello schema.

Quello che bisogna dunque a questo punto fare è determinare il valore y_1 della funzione, a partire dalla condizione aggiuntiva. Si potrebbero fare due cose:

- Approssimare la derivata con il rapporto incrementale, mediante una delle formule *semplici*; questa non è una via corretta, dal momento che gli errori qua vanno come $O(k)$, non come $O(k^2)$ come desidereremmo; questa cosa è estremamente negativa, dal momento che questo errore in realtà non si propaga neanche come $O(k)$, nella formula: esso tende a divergere, dopo un certo valore di j, provocando grossi problemi, esplodendo.

- Utilizzare la formula *simmetrica*, più complicata (sotto un certo punto di vista) ma anche superiore:

$$u_y(x_i, y_0) = \frac{u(x_i, y_1) - u(x_i, y_{-1})}{2k} + O(k^2)$$

Si decide di utilizzare la seconda formula, dal momento che ha un errore decrescente più rapidamente; si noti che c'è un nodo -1, non presente nell'attuale definizione del testo; esso è la causa dei problemi nell'utilizzo di questa funzione. Si supponga, per ora, che la soluzione sia **prolungabile** per tempi negativi, ottenendo dunque, in seguito all'errore di troncamento:

$$u_{i,1} = \frac{u_{i,1} - u_{i,-1}}{2k} = g(x_i)$$

Si ricava dunque il livello fittizio da questa espressione:

$$u_{i,-1} = u_{i,1} + 2kg(x_i)$$

Si fa, come già detto, partire j da 0; si consideri dunque lo schema finora ricavato (che verrà ora riscritto) per $j = 0$:

$$u_{i,j+1} = \lambda^2 \left(u_{i-1,j} + u_{i+1,j}\right) + 2\left(1 - \lambda^2\right) u_{i,j} - u_{i,j-1}$$

Dunque:

$$u_{i,1} = \lambda^2 \left(u_{i-1,0} + u_{i+1,0}\right) + 2\left(1 - \lambda^2\right) u_{i,0} - u_{i,-1}$$

Però, a questo punto, abbiamo un'espressione per la $u_{i,-1}$! Sostituiamola, ottenendo:

$$u_{i,1} = \lambda^2 \left(u_{i-1,0} + u_{i+1,0}\right) + 2\left(1 - \lambda^2\right) u_{i,0} - u_{i,1} - 2kg(x_i)$$

Si vadano dunque a riprendere le definizioni delle condizioni al bordo, per ottenere:

$$u_{i,1} = \frac{\lambda^2}{2}\left[f(x_{i-1}) + f(x_{i+1})\right] + (1 - \lambda^2) f(x_i) - kg(x_i)$$

Il nodo fittizio è stato cancellato dalla formula: ora essa è sempre del secondo ordine, senza avere dipendenze particolari; a questo punto possono essere utilizzati per costruire la prima riga soltanto gli elementi del livello zero, mentre per le altre lo schema alle differenze finite precedentemente utilizzato. Lo schema finale sarà dunque:

$$u_{i,0} = f(x_i), i = 0 \div N$$

$$u_{i,1} = \frac{\lambda^2}{2}\left[f(x_{i-1}) + f(x_{i+1})\right] + (1 - \lambda^2) f(x_i) - kg(x_i)$$

Per quanto riguarda il resto, si può applicare il seguente algoritmo:

```
for i = 1, 2, ...
u_{0, j} = alpha_1(y_i)
u_{N, j} = beta_1(y_i)
for i = 1 to N-1
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}
```

Si noti che, ovviamente, tutto questo schema è valido e funziona dal momento che l'equazione differenziale è stata definita anche sul valore iniziale; se non fossero presenti le condizioni al bordo, si può vedere, come le linee caratteristiche già suggerivano all'inizio della discussione, che la regione di definizione sarebbe un triangolo, contenuto nelle linee caratteristiche passanti per le origini dei bordi.

Si può dimostrare che questo schema di soluzione è condizionatamente stabile, per $\lambda \leq 1$, dunque solo se il passo k è minore di h.

2.5 Equazione del calore

Si consideri il seguente problema differenziale, con equazione già normalizzata:

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

La teoria garantisce il fatto che $u(x,t) \in \mathrm{C}^{(\infty)}$, nell'aperto, indipendentemente dai dati assegnati; a noi in realtà serve che essa sia $\mathrm{C}^{(4)}$. ma nel chiuso, garantendo inoltre che le condizioni di raccordo siano soddisfatte.

Procediamo con l'ormai tradizionale analisi.

Linee caratteristiche

Si ha che $a = 1$, e che $b = c = 0$: l'equazione è di tipo parabolico. Dunque:

$$\left(\frac{\mathrm{d}t}{\mathrm{d}x}\right)^2 = 0$$

Dunque, si ha come soluzione $t = \mathrm{c}$, dove c è una costante arbitraria.

Si noti un fatto: stiamo assegnando le condizioni anche su $t =$c, ossia sulla linea caratteristica; si tenga presente che questo problema **non è un problema di Cauchy**, di conseguenza, dal momento che non stiamo assegnando **tutte le condizioni iniziali** (ossia sia la condizione della funzione sia della sua derivata per $t = 0$) possiamo dire che il problema abbia soluzione esistente e unica.

Discretizzazione

Consideriamo a questo punto il problema, discretizzato:

$$\begin{cases} u_t(x_i,t_j) = u_{xx}(x_i,t_j), \quad i = 1 \div N-1, j = 1,2,... \\ u(x_i,t_0) = f(x_i), i = 1 \div N-1 \\ u(x_0,t_j) = g_0(t_j), j = 1,2,... \\ u(1,t) = g_1(t_j), j = 1,2,... \end{cases}$$

Discretizziamo a questo punto le derivate, utilizzando le espressioni più semplici (quella di primo ordine che approssima con $O(k)$, e l'unica nota di secondo ordine. Si ottiene:

$$\frac{u(x_i,t_{j+1}) - u(x_i,t_j)}{k} + O(k) = \frac{u(x_{i-1},t_j) - 2u(x_i,t_j) + u(x_{i+1},t_j)}{h^2} + O(h^2)$$

Questo metodo è esplicito: esso infatti presenta solo una incognita al livello superiore; effettuando gli errori di troncamento, e utilizzando la definizione

$$\lambda \triangleq \frac{k}{h^2}$$

Si ottiene:

$$u_{i,j+1} = u_{i_j} + \lambda \left(u_{i-1,j} - 2u_{i,j} + u_{i+1,j}\right)$$

Dunque:

$$\begin{cases} u_{i,j+1} = \lambda \left(u_{i-1,j} + u_{i+1,j}\right) + (1 - 2\lambda)u_{i,j}, i = 1 \div N - 1, j = 0, 1, ... \\ u_{i,0} = f(x_i), i = 1 \div N - 1 \\ u_{0,j} = g_0(t_j), j = 0, 1, ... \\ u_{N,j} = g_1(t_j), j = 0, 1, ... \end{cases}$$

Questo schema funziona, ed è esplicito. Se non ci fossero i bordi, come si può vedere, si avrebbe una regione triangolare in cui è definita la struttura. Questo **non** è visualizzabile dalle linee caratteristiche, dal momento che in questo caso il problema è di tipo parabolico, dunque la relazione non è più valida come nel caso dei problemi iperbolici.

Volendo fare l'analisi dell'errore, perturbando solo il dato iniziale, si troverebbe il vettore delle perturbazioni; verificando il fatto che:

$$|u(x_i, t_j) - u_{i,j}| = O(k) + O(h^2)$$

Si può scoprire/dimostrare che:

$$\lambda \leq \frac{1}{2}$$

Questo è un vincolo abbastanza stringente; esso infatti dice che:

$$j \leq \frac{1}{2}h^2$$

Si noti che quando non si è interessati allo studio del fenomeno per molto tempo, ma per esempio solo durante un intervallo limitato, per un tempo molto breve, allora questa condizione non è restrittiva. Si noti inoltre che k e h sono vincolati. A seconda di quello che si desidera, ossia uno studio approfondito piuttosto che uno studio del solo comportamento qualitativo della soluzione, si possono avere valori sia di h sia di k poco stringenti.

2.5.1 Metodo di Crank-Nicholson per l'equazione del calore

Quello che sarà ora introdotto è un metodo implicito, incondizionatamente stabile: h e k possono assumere qualsiasi valore senza che si abbia alcun problema sulla stabilità dell'algoritmo.

Qual è il motivo che ci spinge allo studio di questo metodo? Beh, semplice: il metodo precedente ha una precisione sulla derivata temporale abbastanza scadente: si vorrebbe avere una formula del secondo ordine anche nel tempo, non solo nello spazio; il problema è che esiste solo un rapporto incrementale con errore $O(h^2)$: la cosiddetta *formula simmetrica*. Il motivo per cui non è stato utilizzato è il fatto che servono tre livelli, cosa che, in questo problema, non è ammissibile: nel caso dell'equazione delle onde si aveva infatti la seconda condizione iniziale, che permetteva l'uso del livello fittizio. Ora non si hanno seconde condizioni iniziali (cosa che peraltro garantisce che il problema abbia soluzione, dal momento che, con un'altra condizione iniziale, si avrebbe un problema di Cauchy con condizioni definite su di una linea caratteristica), dunque si dovrà fare qualcos'altro.

Si tenga ben presente un fatto: esiste un caso particolare, in cui la formula normale di approssimazione permette di introdurre un errore pari a $O(k^2)$: quando l'equazione risulta essere collocata nel nodo intermedio tra due nodi temporali:

$$u_t(x_i, t_{j+\frac{1}{2}}) = \frac{u(x_i, t_{j+1}) - u(x_i, t_j)}{j} + O(k^2)$$

Solo e soltanto quando il livello temporale è quello intermedio, si può dire che l'errore vada come $O(k^2)$; questa cosa è particolarmente comoda dal momento che, per motivi di simmetria, si ha un'eliminazione di errori di vario tipo. Si noti che, ovviamente, anche la derivata spaziale andrà collocata nel nodo temporale intermedio.

Qual è il problema? Stiamo approssimando il problema in un punto medio; questo non è positivo, dal momento che i valori di u devono solo essere considerati sui nodi del reticolo, altrimenti sarebbero necessarie altre incognite. Quello che si può fare dunque è introdurre una diversa approssimazione delle derivate, sbarazzandoci di questi nodi intermedi. Sviluppando mediante Taylor l'espressione, si può dire che:

$$u_{xx}(x_i, t_{j+\frac{1}{2}}) = \frac{1}{2}\left[u_{xx}(x_i, t_j) + u_{xx}(x_i, t_{j+1})\right] + O(k^2)$$

Ossia, si considera il punto come una media aritmetica del valore della derivata nei due punti; l'errore derivante da questo tipo di approssimazione

è un errore non lineare, dal momento che l'approssimazione è valida fino al primo ordine, dunque da qui l'errore è come $O(k^2)$: più che lineare.

Con questa formula è possibile passare dal calcolo nel nodo intermedio a quello tradizionale, ottenendo:

$$\frac{u(x_i, t_{j+1}) - u(x_i, t_j)}{k} + O(k^2) = \frac{1}{2}\left[u_{xx}(x_i, t_j) + u_{xx}(x_i, t_{j+1})\right] + O(k^2)$$

Le due derivate sono approssimabili nella seguente maniera:

$$u_{xx}(x_i, t_j) = \frac{u(x_{i-1}, t_j) - 2u(x_i, t_j) + u(x_{i+1}, t_j)}{h^2} + O(h^2)$$

$$u_{xx}(x_i, t_{j+1}) = \frac{u(x_{i-1}, t_{j+1}) - 2u(x_i, t_{j+1}) + u(x_{i+1}, t_{j+1})}{h^2} + O(h^2)$$

A questo punto, troncando, si ottiene la seguente espressione:

$$2u_{i,j+1} - 2u_{i,j} = \lambda\left[u_{i-1,j} - 2u_{i,j} + u_{i+1,j} + u_{i-1,j+1} - 2u_{i,j+1} + u_{i+1,j+1}\right]$$

Dunque:

$$-\lambda u_{i-1,j+1} + 2(1+\lambda)u_{i,j+1} - \lambda u_{i+1,j+1} = \lambda u_{i-1,j} + 2(1-\lambda)u_{i,j} + \lambda u_{i+1,j}$$

Questo, per $j = 0, 1, 2, ...$, e per $i = 1 \div N - 1$.
Ricordando le condizioni iniziali:

$$u_{i,0} = f(x_i), i = 0 \div N$$

$$u_{0,j} = g_0(t_j)$$

$$u_{N,j})g_1(t_j)$$

Si può scrivere il metodo per esteso come:

$$2(1+\lambda)u_{1,j+1} - \lambda u_{2,j+1} = 2(1-\lambda)u_{i,j} + \lambda u_{2,j} + \lambda\left[g_0(t_j) + g_0(t_{j+1})\right]$$

...

$$-\lambda u_{i-1,j+1} + 2(1+\lambda)u_{i,j+1} - \lambda u_{i-1,j} = 2(1-\lambda)u_{i,j} + \lambda u_{i+1,j}$$

...

$$\lambda u_{N-2,j+1} + 2(1+\lambda)u_{N-1,j+1} = u_{N-2,j} + 2(1-\lambda)u_{N-1,j} + \lambda\left[g_1(t_j) + g_1(t_{j+1})\right]$$

La prima e l'ultima equazione hanno due incognite, le altre tre; questo significa che il sistema è tridiagonale. Volendo calcolare la matrice $\underline{\underline{A}}$ essa risulterebbe essere a diagonale dominante, tridiagonale, simmetrica, dunque semplicemente risolubile mediante Gauss senza pivoting. A ogni passo si dovrebbe risolvere un sistema di questo tipo ma, date le caratteristiche del sistema, il metodo di soluzione di esso è $O(N)$, dunque piuttosto onesto.

Si può dimostrare (non è stato fatto) che questo schema è **incondizionatamente stabile**.

Nota finale

Esiste in realtà un metodo alternativo, per la soluzione numerica di questo problema: si può discretizzare solo lo spazio, lasciando di fatto continuo il tempo; quello che si ottiene, dunque, come discretizzazione, è soltanto un insieme di linee verticali (ottenendo dunque il cosiddetto *metodo delle linee*). Dal momento che non si introduce una discretizzazione rispetto al tempo, le derivate nel tempo si mantengono, ottenendo dunque un sistema di equazioni differenziali ordinarie.

Questa spiegazione non vuole tanto introdurre un metodo, quando motivare l'eventuale presenza, nei calcoli, di sistemi di equazioni differenziali ordinarie molto grandi (anche nell'ordine del milione di equazioni presenti).

2.6 Equazione di Poisson

Data la seguente equazione alle derivate parziali, nota come **equazione di Poisson**:

$$-\Delta u = f$$

Dove $u = u(x, y)$, ossia dipende da due variabili **spaziali**. Questa nota è assai importante dal momento che provocherà alcune differenze nell'attuale metodo di risoluzione rispetto ai precedenti.

Come mai non si introduce alcun parametro temporale? Beh, la risposta è abbastanza semplice: generalmente, con l'equazione di Poisson, si considerano problemi **stazionari**, senza dunque la presenza di variazioni della fenomenologia nel tempo. Un modo di vedere l'equazione di Poisson può per esempio essere il seguente:

$$u_t + \Delta u = f$$

A partire da questa, considerando $u_t = 0$, è possibile passare dall'equazione del calore, precedentemente analizzata, all'equazione di Poisson, che siamo sul punto di introdurre. Studiando il termine in u_t, supponendo che ci siano i dati che dunque lo interessano, possiamo studiare anche le variazioni nel tempo, dunque il transitorio del fenomeno; in questa sezione tuttavia siamo solo interessati a studiare il regime, il risultato finale. Dunque, dato:

$$u(x, y), (x, y) \in D$$

Dove D è una regione limitata (tratteremo solo problemi **interni**), si consideri il problema di Dirichlet: quando il punto sta sul contorno Γ del dominio D, $u = g$, dove g è una funzione nota. Si ha dunque:

$$\begin{cases} -\Delta u = f \\ u(\Gamma) = g \end{cases}$$

Nei problemi ellittici la regolarità dipende esclusivamente dai coefficienti dell'equazione: se essi sono costanti, o comunque $C^{(\infty)}$, o con una certa regolarità, la soluzione avrà la stessa regolarità. Ciò che si saprebbe dalla teoria (non affrontata) è il fatto che ogni soluzione dell'equazione di Laplace nell'aperto è armonica; purtroppo, quando si comprende anche la frontiera, si può dire di meno. Ciò che si può dire è che, se f è molto regolare, allora si avrà la stessa regolarità o pure superiore; se f è continua, nella fattispecie, si avranno derivate seconde continue (sebbene a noi possa interessare una regolarità anche superiore, per l'applicazione del metodo delle differenze finite, già così non è male!). La regolarità dipende anche dal dominio: a seconda per esempio della forma della frontiera, ad esempio se vi è presenza di angoli, esso può introdurre discontinuità nella soluzione: più l'eventuale angolo è acuto, maggiore è la discontinuità presente nella soluzione, o nella fattispecie, nel suo gradiente.

Presenza di condizioni di Robin

Nel caso si abbia a che fare con condizioni di Robin, ossia condizioni miste Neumann-Dirichlet (o in realtà anche solo in presenza di una condizione Neumann, ossia sulle derivate), è possibile applicare il famoso trucco del **nodo fittizio**: esso può e **deve** essere applicato ogni volta che sul bordo si impone una condizione sulla derivata normale. Dunque, in questi casi, questo trucco **deve** essere usato anche su problemi ellittici o parabolici, **discretizzando le derivate al bordo** mediante il metodo delle differenze finite.

Discretizzazione

Al fine di introdurre lo schema di soluzione, è necessario introdurre una discretizzazione. A tal fine, si sceglie una geometria semplice: un rettangolo, senza irregolarità, tale per cui sia possibile definire una maglia quadrata; è possibile anche non usare maglie quadrate, ma senza dubbio è più comodo in questo modo. Si impone, dunque, che la soluzione dell'equazione differenziale valga solo nei corrispondenti nodi del dominio discreto.

Vi è una sostanziale differenza rispetto ai casi precedenti: ora non vi è dipendenza dal tempo, dunque si ha un numero finito e conosciuto di nodi, a differenza dei casi precedenti in cui si aveva un numero di nodi illimitato, dal momento che la griglia continuava ad aumentare al crescere dell'istante di tempo considerato.

Ora, non è più indispensabile, ma anzi dannoso, utilizzare due indici per l'identificazione dei nodi: dal momento che, una volta effettuata la discretizzazione, i nodi sono in numero costante, si deve scegliere un certo ordine di numerazione, e identificare ciascun nodo con un singolo indice.

La domanda a questo punto è: possiamo scegliere l'ordine che vogliamo, o no? Beh, purtroppo il risultato finale non è indifferente dall'ordine scelto: la struttura del sistema lineare finale dipende da come si ordinano i nodi; a seconda di come si fa la cosa, il sistema potrà o meno presentare una struttura particolare, che ne semplificherà la risoluzione numerica. Se si usa un ordine preciso, ben definito, il sistema risulterà essere **pentadiagonale**: data la simmetria, il sistema lineare risultante, sparso e organizzato, sarà molto più semplice da risolvere, in termini di calcoli che l'elaboratore deve fare.

Qual è la notazione, l'ordine che si deve utilizzare?

Il fatto di numerare ogni riga (o, volendo, si potrebbe fare anche con ogni colonna) da sinistra a destra permette di mantenere costante la differenza degli indici tra gli elementi adiacenti nell'altra direzione (per esempio, se si ordina per righe, gli elementi adiacenti per colonne avranno gli indici sem-

pre con la medesima differenza); questo fatto determinerà la simmetria del sistema finale.

Dunque, si procede con la collocazione dell'equazione differenziale e delle relative condizioni al bordo nei nodi; si considera, nell'esempio, una suddivisione in nodi da 1 a 12 per i punti interni, e da 13 a 26 per le condizioni al contorno; la numerazione delle condizioni al contorno non è importante come quella dei punti interni, dunque non si dia troppo peso a essa:

$$\begin{cases} -\Delta u(p_i) = f(p_i), i = 1 \div 12 \\ u(p_j) = g(p_j), j = 13 \div 26 \end{cases}$$

A questo punto, ricordando la definizione dell'operatore laplaciano, si ha:

$$\Delta u(p_i) = \left.\frac{\partial^2 u}{\partial x^2}\right|_{p_i} + \left.\frac{\partial^2 u}{\partial y^2}\right|_{p_i}$$

Ricordando a questo punto le formule delle differenze finite, tenendo presente che $h = k$ data la scelta della maglia quadrata, si può dire che:

$$\frac{\partial^2 u}{\partial x^2} = \frac{u(x_{i-1}, y_j) - 2u(x_i, y_j) + u(x_{i+1}, y_j)}{h^2}$$

$$\frac{\partial^2 u}{\partial y^2} = \frac{u(x_i, y_{j-1}) - 2u(x_i, y_j) + u(x_i, y_{j+1})}{h^2}$$

Dunque, si può riscrivere il laplaciano applicato a u come:

$$\Delta u(x_i, y_j) = \frac{1}{h^2}\left[u(x_{i-1}, y_j) + u(x_{i+1}, y_j) - 4u(x_i, y_j) + u(x_i, y_{j-1}) + u_(x_i, y_{j+1})\right] + O(h^2)$$

Si noti che non abbiamo ancora fatto intervenire il fatto di usare un solo indice; questo verrà introdotto nel sistema finale. Per ora, introduciamo l'errore di troncamento:

$$\Delta u(x_i, y_j) \sim \frac{1}{h^2}\left[ui - 1, j + u_{i+1,j} - 4u_{i,j} + u_{i,j-1} + u_{i,j+1}\right]$$

Alla formula si associa una molecola di calcolo di questo tipo:

Ora, come si può osservare, il laplaciano nel nodo centrale è approssimato, con peso -4, dai quattro nodi adiacenti; poi il fatto che ci sia $-\Delta$ fa cambiare il segno a tutta l'equazione. Il sistema di equazioni sarà del tipo:

$$4u_1 - u_2 - u_5 = u_{13} + u_{26} + f(p_1)h^2$$

$$-u_1 + 4u_2 - u_3 - u_6 = h^2 f_2 + g_{14}$$

...

E così via. Come si vede, il sistema è all'inizio tri-diagonale, poi diventa tetra-diagonale, e penta-diagonale. Essendo la molecola di calcolo composta da cinque elementi, la matrice è penta-diagonale, dunque a partire da questa semplice osservazione si poteva già determinare qualche informazione sul sistema risultante. Purtroppo non è possibile fare di meglio, neanche modificando la numerazione, dal momento che la riga di zeri è intrinsecamente presente in questo tipo di metodo.

Questa matrice è simmetrica definita positiva, a diagonale dominante; per questo motivo, utilizzando metodi iterativi quali Gauss-Seidel, si avrebbe una convergenza con una buona velocità: esso converge bene anche se la matrice è solo debolmente dominante, con dominanza non debole almeno per una riga/colonna, dunque, dal momento che questa ci garantisce pure condizioni più forti, si può usare tranquillamente.

Metodo alternativo nel caso di dominio quadrato

Si consideri, questa volta, come dominio di definizione, un quadrato di lato 2 centrato nell'origine del sistema di riferimento; dato dunque il problema differenziale così definito:

$$\begin{cases} -(u_{xx} + u_{yy}) = 1 \\ u = 0, x = \pm 1, y = \pm 1 \end{cases}$$

Ricordando che $u = u(x, y)$, e tenendo presente che le condizioni sono simmetriche, si può dire che:

$$u(-x, y) = u(x, -y) = u(-x, -y) = u(x, y)$$

La soluzione è simmetrica, sia rispetto x sia rispetto y.

Vi sono due approcci: quello brutale, e quello un po' più ragionato, che riduce a un quarto il costo computazionale: la soluzione può essere calcolata anche solo su di un quadrante, e trovare le altre sfruttando la simmetria della soluzione. Problema: noi non abbiamo condizioni sul bordo del quadratino, ossia sugli assi; quali condizioni si possono assegnare? Beh, essendo la funzione simmetrica, ci si può aspettare che lungo gli assi ci sia o un massimo o un minimo relativi, ma dunque derivata normale nulla rispetto agli assi! Solitamente questa derivata normale si prende nel verso uscente, dal momento che questa è la convenzione comunemente utilizzata per i problemi tipo Neumann. Il problema differenziale semplificato si riduce dunque a:

$$\begin{cases} -(u_{xx} + u_{yy}) = 1 \\ u = 0, x = 1, y = 1 \\ u_x = 0, x = 0 \\ u_y = 0, y = 0 \end{cases}$$

Il problema a questo punto è diventato misto, ma solo su di un quadrante; mediante il metodo del nodo fittizio è possibile risolverlo.

Capitolo 3

Metodi dei residui pesati per le equazioni alle derivate parziali

Il metodo delle differenze finite si applica solo se la soluzione è più che classica, e definita su di un dominio abbastanza regolare, in modo da permettere la costruzione di griglie regolari, per poter utilizzare le note formule delle differenze finite. Quelli che si vedranno ora sono metodi che si possono applicare nei seguenti casi:

- Uno su domini semplici e soluzioni molto regolari.
- Uno su domini con geometrie arbitrarie e soluzioni anche non classiche.

Questi metodi rientrano nella categoria dei **metodi dei residui pesati**, e se ne vedranno due: **metodo di collocazione** e **metodo di Galerkin**.

Con le differenze finite, cosa si faceva? Beh, si prendeva l'equazione differenziale, e si discretizzava tutto: il dominio veniva discretizzato mediante il reticolo, si collocava l'equazione differenziale nei nodi del reticolo, si discretizzavano e approssimavano le derivate.

Ora, le derivate **non si toccano**: si approssima la funzione, mediante altre funzioni, **note**, scelte in modo da avere una buona convergenza del metodo (si vedrà in seguito cosa si intende per convergenza in questo ambito).

Si consideri il fatto che l'equazione differenziale può essere espressa nel seguente modo: data la soluzione u del problema differenziale, applicandovi un operatore lineare $\mathcal{L}$, che sarà composto da una combinazione lineare di derivate e altri operatori lineari di questo tipo, si ottiene qualcosa del tipo:

$$\mathcal{L}u(\underline{x}) = f(\underline{x})$$

Dove $\underline{x}$ è il dominio considerato; si considera per ora un dominio spaziale, dove dunque $\underline{x}$ è un insieme di coordinate spaziali. Nel caso dell'equazione

di Poisson, per esempio, $\mathcal{L} = -\nabla^2$. Si considerino inoltre soluzioni u tali per cui:

$$u \in X \subset L^2(D)$$

Dove D è il dominio di interesse; X è uno spazio di dimensione anche infinita, vettoriale.

A questo punto: considerando il fatto che tutte queste funzioni dipendono da $\underline{x}$ (non si specificherà più per alleggerire la notazione), qual è la nostra idea? Beh, applicando l'operatore lineare a portando tutto a primo membro, data u la soluzione **vera** dell'equazione differenziale, si ha:

$$\mathcal{L}u - f = 0$$

E se questa u, anzichè la **vera** soluzione dell'equazione differenziale, fosse solo una soluzione approssimata? Beh, di fatto, sarebbe concettualmente molto simile al sostituire, in un'equazione algebrica, un numero non vero: l'identità non è confermata, a meno di un qualche scarto. Quello che si ha ora, considerando una funzione approssimante u_N, è il cosiddetto **residuo**:

$$\mathcal{L}u_N - f = \mathcal{R}$$

Ossia, l'identità non è più soddisfatta, a meno di questo residuo. Il concetto di residuo sarà la base dei metodi per l'appunto detti **dei residui pesati**, che sono sostanzialmente metodi che trattano in una qualche maniera il residuo, in modo da ottenere un'approssimazione più o meno valida della funzione.

Entriamo un po' più nei dettagli dell'idea nascosta dietro a questo metodo: data la funzione

$$u \in X \subset L^2(D)$$

X è uno spazio come detto di dimensione infinita, dunque, mediante l'analisi tradizionale, difficile da studiare. Si consideri il seguente escamotage: data una successione di sottospazi X_i, dove il pedice i indica anche la dimensione del sottospazio in questione, si ha:

$$X_1 \subset X_2 \subset X_3 \subset ... \subset X_N \subset ... \subset X$$

I sottospazi X_i hanno dimensione **finita**, dunque si possono studiare mediante l'analisi di funzioni in più variabili o mediante la geometria; u, inteso come elemento dello spazio X di dimensione infinita, non può chiaramente appartenere anche a uno dei sottospazi, dal momento che dire che un elemento appartiene a uno spazio di dimensione infinita implica anche che esso debba

essere per forza espresso mediante combinazione lineare di elementi di una base composta da infiniti vettori (o da infinite funzioni); in ciascun i-esimo sottospazio si avrà a che fare con basi composte da i funzioni, dunque che permettono un'approssimazione della funzione u a meno di un certo errore.

Vi è tuttavia un risultato piuttosto interessante: dato $u \subset X$, per qualsiasi $\varepsilon > 0$, esiste un numero intero N_0 (che detterà la dimensione del sottospazio di X considerato) tale per cui in X_{N_0} esista un elemento u_{N_0} per cui:

$$||u - u_{N_0}|| < \varepsilon$$

Ossia, una volta fissata la tolleranza con la quale si intende approssimare la soluzione dell'equazione differenziale, è possibile trovare, in un sottospazio di dimensione N_0, una funzione che sia in grado di approssimare la funzione a meno di una tolleranza ε, dove però ε è un numero **da noi scelto**!

A questo punto, fatto questo ragionamento, entrano in gioco i metodi dei residui pesati: essi sono, a tutti gli effetti, metodi di **selezione**: ciascun metodo selezionerà in un sottoinsieme un elemento tale per cui si abbia una distanza dalla u reale inferiore alla tolleranza desiderata. Le domande a questo punto sono:

- Per $\varepsilon \to 0$, il criterio sarà in grado di selezionare un elemento che disti da u meno di ε, facendo convergere l'elemento del sottospazio di dimensione finita (per quanto elevata) alla funzione reale?
- E magari, se la risposta precedente è positiva, quanto rapidamente lo farà?

Si proceda a questo punto, introducendo la notazione che verrà (almeno per ora) utilizzata. Data la funzione $u(x)$ (si rimembra che essa può essere anche funzione di $\underline{x}$, ossia di un vettore), essa può essere scomposta nel seguente modo:

$$u(x) = u_0(x) + \overline{u}(x)$$

Dove questa decomposizione è scelta in modo da avere

$$u_0(x) = g(x), x \in \Gamma$$

$$\overline{u}(x) = 0, x \in \Gamma$$

In questo modo, $u_0(x)$ stabilisce le condizioni al contorno definite dalla funzione Γ sul bordo e $\overline{u}$ è nulla sul bordo, in modo che su di esso solo u_0

possa agire. Quello che si fa è considerare u in uno spazio affine, ossia in uno spazio traslato rispetto a elemento u_0:

$$u \in u_0 + X$$

A questo punto, in maniera analoga, si consideri l'approssimante; si noti che u_0 è *nota* e non deve essere approssimata; quella che andrà approssimata sarà solo $\overline{u}$:

$$u_N(x) = u_0(x) + \overline{u}_N(x)$$

Dove $\overline{u}_N \in X_N$, ossia appartiene allo spazio vettoriale che permette di effettuare l'approssimazione; analogamente a prima, inoltre, si sceglie l'approssimante in modo da avere:

$$\overline{u}_N = 0, x \in \Gamma$$

In totale, la funzione $u_N(x)$ avrà una forma del tipo:

$$u_N(x) = u_0(x) + \sum_{i=1}^{N} c_i \varphi_i(x), x \in D \cup \Gamma$$

$$u_0(x) = g(x), x \in \Gamma$$

$$\varphi_i(x) = 0, x \in \Gamma$$

In questo modo, si ha naturalmente che:

$$u_N(x) = g(x), \forall c_i$$

Si torni a questo punto all'osservazione fondamentale: applicando a $u_N(x)$ l'operatore lineare $\mathcal{L}$, si avrà:

$$\mathcal{L}u_N(x) - f(x) = \mathcal{R}_N(x)$$

Dove $\mathcal{R}_N(x)$ è il residuo dato dall'applicazione di $\mathcal{L}$ alla approssimante. Ovviamente, $\mathcal{R}_N(x)$ non sarà mai nullo, a meno che la soluzione dell'equazione differenziale non sia esattamente coincidente con l'approssimante scelta, caso veramente troppo fortuito per essere vero. A questo punto, entrano in gioco i metodi numerici, che saranno tali da **selezionare** un criterio da applicare su $\mathcal{R}_N(x)$ in modo da garantire un certo tipo di convergenza. Si analizzeranno dunque i metodi più importanti.

Metodo di collocazione

Abbiamo detto di aver scelto un certo sottospazio di dimensione N, dunque abbiamo di fatto N coefficienti liberi: c_i, $i = 1 \div N$. Scelti dunque N punti $x_j \in D$ **distinti**, il metodo di collocazione chiede che:

$$\mathcal{R}_N(x_j) = 0, j = 1 \div N$$

Si chiede che il residuo sia nullo, per tutti i punti x_j scelti.

Questo criterio è abbastanza simile a un criterio di interpolazione, applicato tuttavia sulla funzione residuo $\mathcal{R}_N(x)$. Il residuo, come noto, è dato da:

$$\mathcal{R}_N(x) = \mathcal{L}u_N(x) - f(x) = \mathcal{L}u_0(x) + \sum_{i=1}^{N} c_i \mathcal{L}\varphi_i(x_j) - f(x_j)$$

Questo deve essere nullo, per $j \in 1 \div N$. Portando a destra i termini noti, si avrà:

$$\sum_{i=1}^{N} c_i \mathcal{L}\varphi_i(x_j) 0 f(x_j) - \mathcal{L}u_0(x_j)$$

Questo permette di fatto di scrivere il sistema lineare: gli elementi della matrice $\underline{\underline{A}}$ sono:

$$a_{i,j} = \mathcal{L}\varphi_i(x_j)$$

Mentre a destra si avrà:

$$f(x_j) - \mathcal{L}u_0(x_j)$$

In N punti.

Osservazione finale, per questa introduzione al metodo: il sistema è senza dubbio lineare, ma il suo condizionamento dipende da due elementi: dai nodi di collocazione x_j scelti, e dalla base di funzioni scelta; questo metodo richiede dunque **troppe scelte da fare**, e per questo è poco utilizzato.

Metodo di Galerkin

Il metodo di Galerkin sfrutta un'idea un poco più complicata, che verrà realizzata tra breve: dato lo sviluppo in serie di Fourier del residuo, ci si può aspettare che i primi coefficienti diano un contributo più grande rispetto

agli altri. L'obiettivo del metodo di Galerkin è quello di annullare un certo numero di coefficienti dello sviluppo, in modo da avere una convergenza sufficientemente buona.

Cosa si fa: come noto, in L^2 esiste un prodotto scalare, così definito:

$$< f, g >= \int_D f\ g \mathrm{d}x$$

Quello che si cerca di fare è introdurre una condizione di ortogonalità per $\mathcal{R}_N(x)$: si chiede che i c_i siano scelti in modo che la funzione di residuo sia ortogonale alle funzioni di base, ossia che:

$$\int_D \mathcal{R}_N(x)\varphi_j(x)\mathrm{d}x = 0, j = 1 \div N$$

Dunque, ricordando che:

$$\mathcal{R}_N(x) = \mathcal{L}u_N(x) - f(x)$$

Si avrà (ricordando, per svolgere i calcoli, il fatto che il prodotto scalare è un operatore lineare):

$$< \mathcal{L}u_n - f, \varphi_j >=< \mathcal{L}u_0 + \sum_{i=1}^{N} c_i \mathcal{L}\varphi_i - f, \varphi_j >= 0$$

Questo è esprimibile anche come:

$$< \mathcal{L}u_0, \varphi_j > + \sum_{i=1}^{N} c_i < \mathcal{L}\varphi_i, \varphi_j > - < f, \varphi_j >= 0$$

Dunque, portando i termini noti a secondo membro, si ottiene l'espressione del sistema lineare:

$$\sum_{i=1}^{N} c_i < \mathcal{L}\varphi_i, \varphi_j >=< f - \mathcal{L}u_0, \varphi_j >, j = 1 \div N$$

A questo punto bisogna di nuovo risolvere un sistema lineare, in cui:

$$a_{i,j} =< \mathcal{L}\varphi_i, \varphi_j >$$

E gli altri coefficienti (termini noti) si calcolano come visto sopra.

A questo punto, può risultare evidente il limite del metodo di Galerkin, la sua complicazione: al fine di applicarlo, è necessario calcolare N^2 integrali, e da qui si vede il costo del metodo: questi integrali vanno calcolati con tecniche efficienti, note dal corso di Calcolo Numerico, ma comunque costituiscono un onere piuttosto pesante in termini di costo computazionale.

3.0.1 Esercizio di esempio per metodo di collocazione

Si consideri l'equazione di Poisson, definita sul quadrato di lato 2 e centrato nell'origine del sistema di riferimento, avendo come condizione al contorno al nullità al brdo:

$$\begin{cases} -(u_{xx} + u_{yy}) = 1 \text{ su } D \\ u = 0 \text{su } c = \partial D \end{cases}$$

Dove D è per l'appunto il quadrato appena definito.

Come già visto in precedenza, abbiamo la simmetria rispetto agli assi; dal momento che si intende utilizzare il metodo di collocazione, sono necessarie due scelte da parte del risolutore: il tipo di approssimante (polinomiale, trigonometrica..), e in quali nodi collocare l'equazione. Si sceglie di utilizzare, a differenza di quanto fatto nel metodo alle differenze finite, una coppia di indici anzichè un indice solo: questo fatto è molto importante, in questa situazione, dal momento che il dominio considerato può essere pensato come prodotto cartesiano di due domini di dimensione 1; dal momento che più l'approssimante contiene caratteristiche del dominio e della soluzione reale, più di può dire che essa sia una buona approssimante, si sfrutta la caratteristica del dominio per introdurre come approssimante una funzione a variabili separabili, ossia in cui le due variabili sono indipendenti tra loro. Si propongono due alternative:

$$u_M(x,y) = \sum_{i=1}^{M}\sum_{j=1}^{M} c_{i,j} \cos\left[(2i-1)\frac{\pi}{2}x\right] \cos\left[(2j-1)\frac{\pi}{2}x\right]$$

Questa funzione è pari, dunque rispetta le simmetrie interessate, ed è a variabili separabili:

$$\varphi_{i,j}(x,y) = \Psi_i(x)\xi_j(y)$$

Questo è un esempio di funzione goniometrica; un esempio di funzione polinomiale che rispetti comunque le condizioni potrebbe essere:

$$u_M(x,y) = \sum_{i=1}^{M}\sum_{j=1}^{M} c_{i,j}(1-x^2)^i(1-y^2)^j$$

Si tratta, anche in questo caso, di un'approssimante abbastanza valida; la prima come vedremo è piuttosto interessante in quanto, sul quadrato, è una base ortogonale (come si spiegherà meglio dopo).

Si è detto più volte che, oltre alle funzioni, è necessario scegliere i nodi in cui collocare l'equazione; nel caso delle funzioni trigonometriche è possibile

utilizzare semplicemente nodi equispaziati; nel caso di funzioni polinomiali, i nodi equispaziati sono la peggior scelta possibile: sarà necessario utilizzare come nodi gli zeri di polinomi ortogonali, quali quelli di Legendre o di Jacobi, a seconda del caso considerato.

Si consideri per ora l'approssimante trigonometrica, e come nodi si scelgano:

$$h = \frac{2}{M+1}; \quad x_i, y_i = ih; \quad i = 1 \div M$$

Si ha dunque:

$$(u_{Mxx(x_k,y_l)} + u_{Myy}(x_k, y_l)) = -1$$

Da qui, sostituendo, si può vedere, derivando due volte le approssimanti, che:

$$\sum_{i=1}^{M}\sum_{j=1}^{M} c_{i,j}\left[(2i-1)^2 + (2j-1)^2\right]\frac{\pi^2}{4}\cos\left[(2i-1)\frac{\pi}{2}x_k\right]\cos\left[(2j-1)\frac{\pi}{2}y_l\right] = 1$$

A questo punto, si ha il sistema in forma implicita; esplicitando, introducendo il sistema lineare in forma matriciale, si può trovare il punto di partenza per la routine.

Risoluzione dello stesso esercizio mediante Galerkin

Volendo risolvere mediante il metodo di Galerkin questo esercizio, si deve imporre che:

$$< -(u_{Mxx} + u_{Myy}), \Psi_k\xi_l >=< 1, \varphi_{k,l} >$$

Ricordando la definizione di prodotto scalare di funzioni, si ottiene:

$$\sum_{i=1}^{M}\sum_{j=1}^{M} c_{i,j}\frac{\pi^2}{4}\left[(2i-1)^2 + (2j-1)^2\right]\cdot$$

$$\cdot\int_{-1}^{1}\int_{-1}^{1}\cos\left[(2i-1)\frac{\pi}{2}x\right]\cos\left[(2j-1)\frac{\pi}{2}y\right]\cos\left[(2k-1)\frac{\pi}{2}x\right]\cos\left[(2l-1)\frac{\pi}{2}y\right]\mathrm{d}x\mathrm{d}y =$$

$$= \int_{-1}^{1}\int_{-1}^{1}\cos\left[(2k-1)\frac{\pi}{2}x\right]\cos\left[(2l-1)\frac{\pi}{2}y\right]\mathrm{d}x\mathrm{d}y$$

Questi si possono calcolare come prodotto di due integrali, grazie al fatto che le variabili sono separate, ottenendo dunque qualcosa di semplificato rispetto all'integrale doppio; si ottiene dunque:

$$= \int_{-1}^{1} \cos\left[(2i-1)\frac{\pi}{2}x\right] \cos\left[(2k-1)\frac{\pi}{2}x\right] \mathrm{d}x \int_{-1}^{1} \cos\left[(2j-1)\frac{\pi}{2}y\right] \cos\left[(2l-1)\frac{\pi}{2}y\right] \mathrm{d}y$$

Dal momento che inoltre si può dimostrare che:

$$\int_0^{\pi} \cos(n\vartheta)\cos(m\vartheta)\mathrm{d}\vartheta \begin{cases} = 0, m \neq n \\ \neq 0, m = n \end{cases}$$

Si può dire che una successione di funzioni di questo tipo costituisca una base ortogonale, nell'intervallo di integrazione. Sfruttando questo fatto, si può vedere che tutti i termini con $i \neq j$ si cancellino, e valgano un certo valore α calcolabile mediante l'Analisi Matematica. Si può dimostrare che il risultato finale valga:

$$c_{i,j} = c_{j,i} = \left(\frac{8}{\pi}\right)^2 \frac{(-1)^{i+j}}{(2i-1)(2j-1)\left[(2i-1)^2 + (2j-1)^2\right]}$$

Dunque, prendendo questa approssimante, che è una base ortogonale, tutto il sistema si semplifica, ottenendo un sistema sostanzialmente diagonale, che non necessita di metodi numerici per la soluzione immediata.

3.0.2 Esempio teorico/pratico di soluzione di ODE mediante metodo di Galerkin

A questo punto, torniamo alle equazioni differenziali ordinarie con valori ai limiti, in modo da proporre un primo esempio di soluzione mediante metodo di Galerkin, evidenziando alcuni aspetti fondamentali. Si consideri il seguente problema differenziale:

$$\begin{cases} u''(x) - u'(x) + u(x) = -\cos\left(x + \frac{\pi}{4}\right), \quad 0 < x < \pi \\ u(0) = \frac{\sqrt{2}}{2} \\ u(\pi) = -\frac{\sqrt{2}}{2} \end{cases}$$

La soluzione a questo problema è esprimibile in forma chiusa, ed è nota:

$$u(x) = \sin\left(x + \frac{\pi}{4}\right)$$

Come approssimante, si sceglierà come al solito qualcosa in forma:

$$u_M(x) = u_0(x) + \sum_{i=1}^{M} c_i \varphi_i(x)$$

Dove si deve avere:

$$u_0(0) = \frac{\sqrt{2}}{2} \qquad u_0(\pi) = -\frac{\sqrt{2}}{2}$$

Un esempio di funzione potrebbe essere:

$$u_0(x) = \frac{\sqrt{2}}{2}\cos(x)$$

A questo punto, bisogna scegliere le $\varphi_i(x)$ tali da essere nulle al bordo; un esempio potrebbe essere:

$$\varphi_i(x) = \sin(ix)$$

A questo punto, si hanno condizioni automaticamente soddisfatte, ed è possibile incominciare ad applicare il metodo di Galerkin al fine di introdurre un metodo numerico per la soluzione del problema. Dunque:

$$\int_0^\pi \left[u_M''(x) - u_M'(x) + u_M(x)\right]\varphi_j(x)\mathrm{d}x = -\int_0^\pi \cos\left(x + \frac{\pi}{4}\right)\varphi_j(x)\mathrm{d}x, \;\; j = 1 \div M$$

A questo punto, sostituendo tutto, si ottiene:

$$\int_0^\pi \left[u_0''(x) + \sum_{i=1}^{M} c_i\varphi_i''(x) - u_0'(x) - \sum_{i=1}^{M} c_i\varphi_i'(x) + u_0(x) + \sum_{i=1}^{M} c_i\varphi_i(x)\right]\varphi_j(x)\mathrm{d}x = b_j$$

Dove si è definito

$$b_j = -\int_0^\pi \cos\left(x + \frac{\pi}{4}\right)\varphi_j(x)\mathrm{d}x$$

Si portano dunque a secondo membro le funzioni note (le u_0), e si definisce:

$$a_0(x) = u_0''(x) - u_0'(x) + u_0(x)$$

Dunque:

$$\int_0^\pi \left[\sum_{i=1}^{M} c_i\varphi_i''(x) - \sum_{i=1}^{M} c_i\varphi_i'(x) + \sum_{i=1}^{M} c_i\varphi_i(x)\right]\varphi_j(x)\mathrm{d}x = b_j - \int_0^\pi a_0(x)\varphi_j(x)\mathrm{d}x$$

Definendo il membro destro come d_j, e la funzione integranda come $\Psi(x)$, si ha, portando fuori il simbolo di sommatoria:

$$\sum_{i=1}^{M} c_i \int_0^{\pi} \Psi_i(x)\varphi_j(x)\mathrm{d}x = d_j, \qquad j = 1 \div M$$

A questo punto, si possono definire i coefficienti del sistema lineare a_{ji} come:

$$a_{ji} = \int_0^{\pi} \Psi_i(x)\varphi_j(x)\mathrm{d}x$$

Dunque, si può scrivere il sistema lineare in forma matriciale, in modo da avere sulle colonne i coefficienti che andran moltiplicati per i vari c_i, e sulle righe i vari j, da 1 a M.

Guardiamo i lati positivi e i lati negativi di questo metodo: aumentando M è possibile risolvere un grosso numero di problemi, al prezzo di calcolare M^2 integrali: il sistema è infatti $M \times M$, dunque, dal momento che (a meno di casualità) non è possibile vedere elementi nulli, sarà inevitabile avere a che fare con un enorme numero di integrali, da calcolare o in forma chiusa o in forma numerica, mediante le varie formule di quadratura.

Una nota positiva aggiuntiva: se, nel metodo delle differenze finite, la rapidità di convergenza alla soluzione dipendeva esclusivamente dal tipo di schema implementato (a patto ovviamente che esso fosse applicabile), ora, se lo schema di Galerkin è applicabile, è possibile avere diverse velocità di convergenza, e a seconda della regolarità della soluzione questa *velocità di convergenza* può variare: non si ha solo più dipendenza dallo schema della velocità, ma anche dal tipo di approssimante utilizzata, e dalla sua regolarità: dalla regolarità della soluzione! La regolarità della soluzione a sua volta dipende sostanzialmente dalla regolarità del dominio e dei coefficienti dell'equazione differenziale, dunque si può introdurre una prima stima a partire da una semplice osservazione del problema differenziale (non entriamo in dettaglio). Se la soluzione fosse poco regolare, magari solo classica o pure meno, la convergenza diventerebbe molto più lenta; a questo punto servirebbe un M molto più grande, ma dunque un numero molto grande di integrali da calcolare, e un metodo conseguentemente molto pesante sotto il punto di vista computazionale. Si cercherà a questo punto di proporre un'alternativa a questo fatto.

3.0.3 Formulazione debole del metodo di Galerkin

Funzioni poligonali

Consideriamo qualcosa di diverso: possiamo cercare di utilizzare delle approssimanti tali da avere molti integrali nulli; questo permetterebbe di avere sistemi anche di dimensioni enormi, ma comunque in cui solo pochi elementi per riga dovrebbero essere determinati mediante il solito calcolo integrale. L'idea è dunque quella di cercare approssimanti di forme particolari, in modo da avere sistemi sparsi, e cercare di adattare il metodo a queste approssimanti; insomma, il nostro obiettivo è quello di avere un certo numero di elementi nulli **a priori**.

L'idea è quella di applicare il metodo di Galerkin a una particolare classe di funzioni: le funzioni polinomiali a tratti; questa idea è alla base del cosiddetto **metodo degli elementi finiti**. Delle varie funzioni polinomiali esistenti, noi utilizzeremo funzioni poligonali, ossia funzioni lineari a tratti.

Ora: qual è il nostro problema? Beh, purtroppo, molti dei problemi che studiamo sono problemi del second'ordine (o superiori, anche se noi siamo stati abituati, dal resto della trattazione, a riportarli a problemi del secondo o del primo ordine); se utilizziamo funzioni lineari a tratti, ossia spezzate, derivando una prima volta si avrà una serie di funzioni costanti a tratti, con dei salti, mentre derivando due volte si avrebbe esclusivamente una combinazione di δ di Dirac, dunque funzioni non studiabili in senso classico (solo in senso distribuzionale). Per questo motivo si parla di **modifica del metodo di Galerkin**: così come lo conosciamo non è possibile applicarvi funzioni del tipo espresso.

Le funzioni poligonali di cui si sta tanto parlando sono comunemente indicate con $N_i(x)$, dove si attribuisce il seguente significato:

Per $N_0(x)$ si intende la funzione che vale 1 nel nodo x_0, scende fino a x_1, e da lì in poi è nulla; per $N_i(x)$ si intende la funzione che vale 1 nel nodo x_i, scende fino a x_{i-1} e x_{i+1}, dunque è nulla per il resto del supporto. Allo stesso modo si definisce N_{M+1} come la poligonale duale alla 0: da x_M sale, fino a x_{M+1}. Si parla di funzioni di base a supporto minimo: esse sono non nulle solo in un intervallo minimo, in cui assumono valori *utili* per la rappresentazione della funzione (ossia, per fungere da **basi**).

Si può dimostrare che la generica poligonale u_M definita come:

$$u_M(x) = \sum_{i=0}^{M+1} c_i N_i(x)$$

ammette una rappresentazione a partire dalle funzioni di base; si fa in modo da unire con continuità i vari tratti, e che in ciascun tratto si abbiano

polinomi di grado 1 (funzioni lineari); il metodo di Galerkin andrà abbinato a questo tipo di base, in modo che, quando si calcola il prodotto $N_i N_j$, si annullino quasi tutti i termini; nella fattispecie, quello che capita è che:

$$\int_0^\pi N_i(x)N_j(x)\mathrm{d}x \neq 0, \quad |i-j| \leq 1$$

Ossia, fissato i, per $j = i-1$, $j = i$, $j = i+1$, l'integrale sarà non nullo. Si noti che inoltre l'integrale esisterà in un dominio più limitato rispetto a quello di definizione, ma di questo si parlerà in seguito. Per il problema precedentemente proposto, per esempio, una scelta potrebbe essere:

$$u_M(x) = \frac{\sqrt{2}}{2}\left[N_0(x) - N_{M+1}(x)\right] + \sum_{i=1}^{M} c_i N_i(x)$$

Dove:

$$N_i(x) = \delta_{ij}$$

E δ_{ij} è il simbolo di Kronecker, c_i le incognite del problema.

Introduzione alla forma debole di Galerkin

Abbiamo parlato abbondantemente di poligonali, ma a questo punto sussiste ancora un problema: nelle funzioni poligonali la derivata seconda non esiste in senso classico, dal momento che si finisce a parlare di delta di Dirac; al fine di risolvere il problema, proponiamo una continuazione di questo esercizio teorico, in modo da presentare cosa si può fare; supponiamo per un attimo di aver a che fare con una funzione approssimante del secondo ordine, con le derivate seconde continue; applicando Galerkin, si avrebbe:

$$\int_0^\pi \left[u''_M(x) - u'_M(x) + u_M(x)\right]\varphi_j(x)\mathrm{d}x = b_j, \quad j = 1 \div M$$

A questo punto, consideriamo per un attimo separatamente il primo termine, ossia l'integrale di $u''_M(x)$; si integri per parti, ottenendo:

$$\int_0^\pi u''_M(x)\varphi_j(x)\mathrm{d}x = u'_M(x)\varphi_j(x)|_0^\pi - \int_0^\pi u'_M(x)\varphi'_j(x)\mathrm{d}x$$

Dal momento che si ipotizza che $u'_M(x)$ è limitata agli estremi, dal momento che $\varphi_j(x)$ è nulla agli estremi, i termini fuori dai segni di integrale si possono annullare, ottenendo:

$$\int_0^\pi u''_M(x)\varphi_j(x)\mathrm{d}x = -\int_0^\pi u'_M(x)\varphi'_j(x)\mathrm{d}x$$

Cosa significa ciò: siamo riusciti a *scaricare* la derivata di secondo ordine, ottenendo due derivate di primo ordine sulle funzioni; in questo modo, sostituendo nel metodo di Galerkin, si ottiene:

$$\int_0^\pi u'_M(x)\varphi'_j(x)\mathrm{d}x + \int_0^\pi \left[-u'_M(x) + u_M(x)\right]\varphi_j(x)\mathrm{d}x = b_j, \quad j = 1 \div M$$

Nel caso si abbia una soluzione con forte regolarità, ossia:

$$u_M \in \mathrm{C}^{(2)}[0,\pi]$$

Si può dire che le due formulazioni siano assolutamente coincidenti. Questa seconda formulazione del metodo di Galerkin è detta **formulazione debole del metodo di Galerkin**, ed è molto interessante dal momento che permette di risolvere problemi anche a bassa regolarità approssimando con funzioni di questo tipo.

Termine del problema

Si porta a questo punto avanti il problema precedentemente introdotto, mediante le nozioni introdotte. Si ha:

$$-\int_0^\pi \left[u'_0(x) + \sum_{i=1}^M c_i N'_i(x)\right] N'_j(x)\mathrm{d}x - \int_0^\pi \left[u'_0(x) + \sum_{i=1}^M c_i N'_i(x)\right] N_j(x)\mathrm{d}x +$$

$$+\int_0^\pi \left[u_0(x) + \sum_{i=1}^M c_i N_i(x)\right] N_j(x)\mathrm{d}x = b_j$$

Portiamo le funzioni note a secondo membro, e definiamo subito:

$$a_0(x) = u'_0(x)N'_j(x) + u'_0(x)N_j(x) - u_0(x)N_j(x)$$

Dunque, portando fuori il simbolo di sommatoria:

$$\sum_{i=1}^M c_i \left[-\int_0^\pi N'_i(x)N'_j(x)\mathrm{d}x - \int_0^\pi N'_i(x)N_j(x)\mathrm{d}x + \int_0^\pi N_i(x)N_j(x)\mathrm{d}x\right] = b_j + \int_0^\pi a_0(x)\mathrm{d}x$$

Definendo tutti gli integrali come a_{ji}, dunque il solito d_j, si ottiene:

$$\sum_{i=1}^M c_i a_{ji} = d_j, \qquad j = 1 \div M$$

Dove ciascun elemento del sistema si calcola come:

$$a_{ji} = -\int_0^\pi N_i'(x)N_j'(x)\mathrm{d}x - \int_0^\pi N_i'(x)N_j(x)\mathrm{d}x + \int_0^\pi N_i(x)N_j(x)\mathrm{d}x$$

Dal momento che tutti gli elementi per cui $|i-j| \leq 1$ si annullano, la matrice finale risulta essere tridiagonale.

Si possono fare alcune considerazioni aggiuntive, per esempio riguardo il calcolo degli integrali; si consideri il generico integrale:

$$\int_0^\pi N_i(x)N_j(x)\mathrm{d}x, \quad 1 < i, j < N$$

Fissiamo i; l'integrale è non nullo solo per $j = i-1$, $j = i$, $j = i+1$; si può dire che:

$$j = i: \quad \int_{x_{i-1}}^{x_{i+1}} N_i^2(x)\mathrm{d}x = \int_{x_{i-1}}^{x_i} N_{i-1}(x)N_i(x)\mathrm{d}x + \int_{x_i}^{x_{i+1}} N_i(x)N_{i+1}(x)\mathrm{d}x$$

In maniera simile:

$$j = i-1: \quad \int_0^\pi N_i(x)N_{i-1}(x)\mathrm{d}x = \int_{x_{i-1}}^{x_i} N_{i-1}(x)N_i(x)\mathrm{d}x$$

E ancora:

$$j = i+1: \quad \int_0^\pi N_i(x)N_{i+1}(x)\mathrm{d}x = \int_{x_i}^{x_{i+1}} N_i(x)N_{i+1}(x)\mathrm{d}x$$

3.0.4 Esempi teorico-pratici - condizioni di Dirichlet

a questo punto, verranno presentati due esempi teorici, in modo da un lato da migliorare la manualità rispetto al mezzo, dall'altro di evidenziare alcuni aspetti teorici finora trascurati.

Si consideri a questo punto il seguente problema differenziale:

$$\begin{cases} -u''(x) + \sigma(x)u(x) = f(x), \quad a < x < b \\ u(a) = h_a \\ u(b) = h_b \end{cases}$$

Dalla teoria è noto che esiste un teorema che assicura che esiste, date f e σ continue sull'aperto (a, b), una soluzione di classe $\mathrm{C}^{(2)}[a, b]$, dunque sull'intero chiuso. Le funzioni fungenti da coefficiente (σ) e da termine noto

(f) devono però essere almeno continue: devono essere ben definite per ogni punto dell'intervallo.

Si noti che grazie a questo teorema si può inoltre garantire soluzione **classica**, e non distribuzionale, ottenendo dunque un risultato molto forte.

Il nostro obiettivo è quello di risolvere il problema mediante la sua formulazione debole; il vantaggio conseguente da ciò è ottenere una formulazione semplice del problema e che permette, in un qualche senso (classico o distribuzionale), di trovare la soluzione. al fine di affrontare in modo formale il problema, dobbiamo introdurre un certo insieme di spazi di definizione. Si consideri nella fattispecie lo spazio delle funzioni a quadrato sommabile nel senso di Lebesgue, $L^2(a,b)$, come:

$$L^2(a,b) = \left\{ f : \int_a^b |f(x)|^2 \, \mathrm{d}x < \infty \right\}$$

Per questo spazio, come noto, è possibile definire una norma e un prodotto scalare indotto dalla norma:

$$||f|| = \left(\int_a^b |f(x)|^2 \, \mathrm{d}x \right)^{\frac{1}{2}}$$

$$< f, g >= \int_a^b f(x) g(x) \mathrm{d}x$$

Questo è uno spazio di Hilbert, cosa che potrebbe portarci a definire molti altri vantaggi da esso introdotti.

In un altro contesto, introduciamo a questo punto alcune nozioni concernenti gli spazi di Sobolev: si tratta di sottospazi incapsulati in L^2, definibili come:

$$H^1 \triangleq \left\{ v \in L^2(a,b) : v' \int L^2(a,b) \right\}$$

Ossia, uno spazio di vettori/funzioni tali per cui anche la derivata prima sia a quadrato sommabile. anche per questo spazio è possibile definire una norma (e sarebbe possibile definire un prodotto scalare, cosa che però non verrà fatta):

$$||v|| = \left(\int_a^b |v(x)|^2 \, \mathrm{d}x + \int_a^b |v'(x)|^2 \, \mathrm{d}x \right)^{\frac{1}{2}}$$

Si definisce a questo punto un sottospazio di H^1, ossia H_0^1, come:

$$H_0^1 = \left\{ v \in H^1 : v(a) = v(b) = 0 \right\}$$

Si noti un fatto: questa definizione torna utile esclusivamente per il problema in studio e per la sua relativa formulazione debole; vedremo, in un secondo esempio, che la definizione dello spazio H_0^1 andrà modificata in modo da essere idonea per il secondo problema.

Un'osservazione: se v appartenesse a L^2, si potrebbe non aver continuità agli estremi, dunque i valori potrebbero anche divergere, e imporli a zero non avrebbe senso; esiste un teorema che però assicura (solo per funzioni definite su di un intervallo, ossia su di un dominio monodimensionale) che:

$$H^1(a,b) \subset \mathrm{C}^{(1)}[a,b]$$

Ossia, è necessariamente continua nel chiuso $[a,b]$, ottenendo definizione e continuità anche sugli estremi.

Si definisce a questo punto lo spazio V, ossia lo spazio delle funzioni che considereremo, come:

$$V \triangleq H_0^1$$

V è anche detto *spazio delle funzioni test*, e v è anche nota come *funzione test*.

Si ritorni a questo punto al problema originale, e si consideri la seguente equazione:

$$-u''(x) + \sigma(x)u(x) = f(x)$$

a questo punto si moltiplichi a sinistra e a destra per $v(x)$, ossia per la funzione test, e si integri nel dominio:

$$\int_a^b (-u''(x) + \sigma(x)u(x))v(x)\mathrm{d}x = \int_a^b f(x)v(x)\mathrm{d}x, \quad v \in V$$

Svolgendo i conti, si ottiene banalmente per linearità:

$$-\int_a^b u''(x)\mathrm{d}x + \int_a^b \sigma(x)u(x))v(x)\mathrm{d}x = \int_a^b f(x)v(x)\mathrm{d}x, \quad v \in V$$

Dal momento che si vuole abbassare l'ordine di derivazione, si propone il solito procedimento di integrazione per parti del primo membro; lavorando dunque solo su di esso, si ottiene:

$$-v(x)u'(x)|_a^b + \int_a^b u'(x)v'(x)\mathrm{d}x$$

Prima di tutto, osserviamo che $u'(b)$ e $u'(a)$ sono valori finiti. Detto ciò, si ricordi che, per come è stato definito lo spazio V, abbiamo per certo che i valori agli estremi sono nulli, dunque che di tutto il termine avremo soltanto il secondo integrale; l'espressione si riporta dunque a:

$$+\int_a^b u'(x)v'(x)\mathrm{d}x + \int_a^b \sigma(x)u(x))v(x)\mathrm{d}x = \int_a^b f(x)v(x)\mathrm{d}x, \quad v \in V$$

a questo punto, si definisce lo spazio affine W rispetto a V come:

$$W = u_0 + V$$

Dove u_0 è una funzione che deve rispettare gli estremi, ossia:

$$u_0 : u_0(a) = h_a, \quad u_0(b) = h_b$$

Un esempio di funzione di questo tipo potrebbe essere una retta, tale per cui:

$$u_0(x) = \frac{h_b(x-a) + h_a(b-x)}{b-a}$$

Lo spazio W è dunque un insieme di funzioni tali per cui:

$$W = u_0 + V \subseteq \left\{v \in H^1 : v(a) = h_a, v(b) = h_b\right\}$$

Dunque, si può definire la generica $u(x)$ come:

$$u(x) = u_0(x) + \overline{u}(x), \overline{u} \in V$$

Svincoliamoci un momento dal problema iniziale; non considerando più il fatto che $u \in \mathrm{C}^{(2)}$, considerando incognite dunque le sue caratteristiche in termini di regolarità, si ha:

$$\int_a^b \overline{u}'(x)v'(x)\mathrm{d}x + \int_a^b \sigma(x)\overline{u}(x)v(x)\mathrm{d}x = -\int_a^b u_0'(x)v'(x)\mathrm{d}x + \int_a^b (f(x) - \sigma(x)u_0(x))v(x)\mathrm{d}x, \forall v \in V$$

Ora, noi vorremmo che questa identità sia soddisfatta per ogni $v \in V$; ciò che sappiamo è che u_0 è scelta da noi, e $\overline{u}$ è ancora incognita; di essa sappiamo solo che dipende da V.

Poniamoci una domanda: quanti elementi di V sono tali da rendere soddisfatta questa relazione? Beh, esiste un teorema che dimostra che esiste uno e un solo elemento. Quello che si fa in pratica è trovare la $\overline{u}$, aggiungere u_0

(in modo da passare nello spazio affine), dunque sostituire dentro l'equazione differenziale di partenza, scoprendo che essa è l'unica soluzione del problema differenziale. Questa unica $\overline{u}$ dunque è anche soluzione del problema differenziale. Questo risultato ci dice che non dobbiamo preoccuparci del fatto che il problema differenziale ammetta o meno soluzione classica: se f e σ sono continue la soluzione sarà certamente classica, altrimenti potremo ottenere una soluzione, comunque valida, per quanto anche non classica: esiste, in senso distribuzionale.

Ciò ci potrebbe portare a capire una cosa, che sarà la base della prosecuzione dell'esercizio: data la forma debole del problema differenziale, appena introdotta più formalmente, si può pensare al metodo di Galerkin come a una semplice discretizzazione di questa formulazione debole.

Tornando al problema originale, supponendo dunque di avere a che fare con soluzioni di tipo classico, si abbia qualcosa del genere:

quello che si fa a questo punto è introdurre un intervallo partizionato in $M+1$ punti, da 0 a M; considerando di utilizzare come base delle poligonali, avremo a che fare con il seguente spazio di funzioni test discretizzato:

$$V_M = \{N_1, N_2, N_3, ..., N_{M-2}, N_{M-1}\}$$

Dunque, possiamo definire a questo punto $\overline{u}_M$ e u_0 come:

$$u_0(x) = h_a N_0(x) + h_b N_M(x)$$

mentre le $\overline{u}_M(x)$ saranno definite come combinazioni lineari delle funzioni appartenenti all'appena definito (per questo contesto e con questo tipo di approssimanti) spazio delle funzioni test:

$$\overline{u}_M(x) = c_1 N_1(x) + c_2 N_2(x) + ... + c_{M-1} N_{M-1}(x)$$

Infine:

$$u_M(x) = u_0(x) + \overline{u}_M(x)$$

A questo punto, come detto, u_0 è nota, mentre le $\overline{u}_M$ sono incognite. Quello che bisogna fare è cercare un'espressione tale per cui:

$$\int_a^b \overline{u}'_M(x)v'(x)\mathrm{d}x + \int_a^b \sigma(x)\overline{u}_M(x)v(x)\mathrm{d}x = -\int_a^b u'_0(x)v'(x)\mathrm{d}x + \int_a^b (f(x) - \sigma(x)u_0(x))v(x)\mathrm{d}x \;\; \forall v \in$$

Abbiamo un diverso spazio delle funzioni test, ottenuto discretizzando quello di partenza; quello che abbiamo è che dunque V_M è uno spazio di

dimensione finita, nella fattispecie di dimensione $M-1$. Quello che possiamo fare a questo punto è soddisfare l'eguaglianza per gli elementi della base, facendo in modo che sia soddisfatta per l'intero V_M: una proprietà degli spazi funzionali lineari è il fatto che, se l'eguaglianza è soddisfatta per gli elementi della base, allora lo sarà certamente anche per tutte le loro combinazioni lineari, ottenendo di fatto che sia soddisfatta per tutti gli elementi appartenenti allo spazio in questione. Quello che si può dunque dire è che la condizione appena proposta sia equivalente alla seguente:

$$\forall v \in V_M \Longleftrightarrow j = 1 \div M-1$$

Sostituendo $\overline{u}_M$, dunque, si ottiene semplicemente la formulazione debole di Galerkin, ma in modo formale.

Esempio di soluzione di problema con condizioni Robin

Si consideri a questo punto il seguente problema differenziale:

$$\begin{cases} -u''(x) + \sigma(x)u(x) = f(x), \quad a < x < b \\ u(a) = h_a \\ u'(b) + \gamma u(b) = h_b, \gamma \geq 0 \end{cases}$$

A questo punto si ha a che fare con una condizione di tipo Dirichlet, e con una mista, ossia di tipo Robin. Come ci si deve comportare ora? Beh, come anticipato precedentemente, gli spazi cambieranno un pochetto; il nuovo spazio delle funzioni test, V, sarà così definito:

$$V = H_0^1 = \left\{ v \in H^1 : v(a) = 0 \right\}$$

In altre parole, si chiede che solo l'estremo a sia nullo; più generalmente, si può dire che si richieda la nullità solo per quanto riguarda la condizione di Dirichlet.

Ciò che sostanzialmente cambierà sarà l'integrazione per parti; la formulazione debole: dal momento che in b la funzione è non nulla, si avrà qualcosa di questo tipo:

$$-\int_a^b u''(x)v(x)\mathrm{d}x = -u'(x)v(x)\big|_a^b + \int_a^b u'(x)v'(x)\mathrm{d}x$$

Svolgiamo a questo punto il primo termine:

$$-u'(x)v(x)\big|_a^b = -u'(b)v(b) + u'(a)v(a)$$

Il secondo termine sarà nullo, ma il primo no. Dalla condizione al contorno, tuttavia, si può ricavare che:

$$u'(b) = h_b - \gamma u(b)$$

Ma dunque si ottiene questo termine:

$$\gamma u(b)v(b) - h_b v(b)$$

Sostituendo:

$$\left[\gamma u_0(b) + \gamma \overline{u}(b) - h_t\right] v(b)$$

Bene: $v(b)$ è da noi scelta dunque nota, $u_0(x)$ è da noi scelta quindi $u_0(b)$ nota, $\overline{u}(b)$ è incognita. Sostituendo nell'espressione:

$$\left(\gamma u_0(b) - h_b\right) v(b) + \gamma \overline{u}(b) v(b)$$

Dove si ha da scegliere una u_0 tale per cui:

$$u_0(x) : u_0(a) = h_a$$

Avendo una condizione piuttosto semplice da realizzare per u_0, dal momento che, in questo problema, l'unica condizione di Dirichlet, ossia l'unica condizione che si possa automaticamente imporre nella scelta delle funzioni di base, è possibile scegliere $u_0(x) = h_a$, ossia pari alla costante h_a della condizione al contorno. Possiamo ora presentare la formulazione debole del problema:

$$\int_a^b \overline{u}'(x)v'(x)\mathrm{d}x + \int_a^b \sigma(x)\overline{u}(x)v(x)\mathrm{d}x + \gamma\overline{u}(b)v(t) = -\int_a^b u_0'(x)v'(x)\mathrm{d}x + \int_a^b \left[f(x) - \sigma(x)u_0(x)\right] v(x$$

In pratica, la formulazione debole è abbastanza simile alla precedente, ma il fatto di aver cambiato la condizione al contorno ha non solo cambiato il problema, ma anche la risoluzione; in questo caso cambia lo spazio considerato, e cambia il metodo di risoluzione (dal momento che cambia l'integrale per parti, o meglio i suoi prodotti). Ciò che cambia è che prima si imponeva alla soluzione la condizione, ora no: si utilizza al fine di formulare la formulazione, ma non si introducono direttamente nell'approssimante. Si definisce una distinzione tra i tipi di condizioni:

- Condizioni di Dirichlet: esse sono anche dette **condizioni essenziali**;
- Condizioni miste: esse sono anche dette **condizioni naturali**: questo perchè, dalla formulazione, *naturalmente*, si derivano le varie proprietà. La formulazione risultante è anche detta **formulazione variazionale** del problema.

Si fa lo stesso passaggio di prima, discretizzando V e passando a V_M, dunque da qui al limite su j grazie alla proprietà degli spazi vettoriali.

Una piccola nota finale sul metodo: una volta introdotta la condizione sul bordo b, è un po' come fosse stato introdotto anche l'estremo nel metodo numerico, dal momento che non si ha più la condizione essenziale su di esso; questo significa che, pur avendo introdotto una condizione in più, si dovrà calcolare anche il termine associato a $x = b$, ossia il c_M. Nel caso si fossero aggiunte condizioni su $u'(a)$, avremmo dovuto anche considerare il c_0, ottenendo $j = 0 \div M$.

3.0.5 Esempio pratico: soluzione mediante Galerkin dell'equazione del calore

Si consideri infine un ultimo problema, risolto mediante il metodo di Galerkin: l'equazione del calore (dunque un problema che, a differenza dai precedenti, contiene anche una componente temporale). Si consideri dunque il seguente problema differenziale:

$$\begin{cases} u_t(x,t) - u_{xx}(x,t) = f(x,t), \quad 0 < x < 1, t > 0 \\ u(x,0) = g(x), \quad 0 \leq x \leq 1 \\ u(0,t) = \alpha(t), \quad t > 0 \\ u(1,t) = \beta(t), \quad t > 0 \end{cases}$$

Dunque procediamo; si cnsideri:

$$u_M(x,t) = u_0(x,t) + \sum_{i=1}^{M-1} c_i(t)\varphi_i(x)$$

Qui la prima nota: si considera una separazione delle variabili al fine di avere una soluzione semplificata; in tal senso, si avrà come unica funzione del tempo t i coefficienti c_i, mentre per le x si avrà la solita funzione $\varphi(x)$. Si proceda a questo punto con la formulazione debole del problema:

$$u = u_0 + \overline{u}; \quad \overline{u} \in V \quad V \triangleq H_0^1$$

Quest'ultimo passaggio è dovuto al fatto che le condizioni sono di Dirichlet, dunque escludiamo gli estremi. Dunque:

$$\int_0^1 (u_t - u_{xx})\, v(x)\mathrm{d}x = \int_0^1 f(x)v(x)\mathrm{d}x$$

Questo, data l'equazione di partenza, moltiplicati ambo i membri per $v(x)$, dunque integrando nel dominio; a questo punto, si ottiene:

$$\int_0^1 u_t(x)v(x)\mathrm{d}x - \int_0^1 u_{xx}v(x)\mathrm{d}x = \int_0^1 f(x)v(x)\mathrm{d}x$$

Operiamo sul secondo integrale:

$$-\int_0^1 u_{xx}v(x)\mathrm{d}x = \int_0^1 u_x(x)v'(x)\mathrm{d}x$$

Otteniamo dunque, sostituendo, la formulazione debole del problema:

$$\int_0^1 u_t(x)v(x)\mathrm{d}x + \int_0^1 u_x(x)v'(x)\mathrm{d}x = \int_0^1 f(x)v(x)\mathrm{d}x, \quad \forall v \in V$$

Data la formulazione debole, possiamo applicare il metodo di Galerkin: discretizziamo, ottenendo dunque:

$$\int_0^1 u_{0t}v\mathrm{d}x + \int_0^1 \overline{u}_t v\mathrm{d}x + \int_0^1 u_{0x}v'\mathrm{d}x + \int_0^1 \overline{u}_x v'\mathrm{d}x = \int_0^1 fv\mathrm{d}x$$

Ordinando:

$$\int_0^1 \overline{u}_t v\mathrm{d}x + \int_0^1 \overline{u}_x v'\mathrm{d}x = \int_0^1 (f - u_{0t})\, v\mathrm{d}x - \int_0^1 u_{0x}v'\mathrm{d}x \;\; \forall v \in V$$

Il nostro obiettivo è quello di determinare una funzione $\overline{u} \in V$ tale da rispettare ciò. Solo a questo punto facciamo intervenire la discretizzazione, considerando un sottospazio $V_M \subset V$; come funzioni di base per V_M scegliamo:

$$V_M \triangleq \{N_1, N_2, ..., N_{M-1}\}$$

A questo punto uso N_0 e N_M per definire la funzione u_0:

$$u_0(x,t) = \alpha(t)N_0(x) + \beta(t)N_M(x)$$

Dunque:

$$u \sim u_M = u_0 + \overline{u}_M$$

Dove

$$\overline{u}_M \in V_M$$

Dunque:

$$u_M(x,t) = u_0(x,t) + \sum_{i=1}^{M-1} c_i(t) N_i(x)$$

Ora dobbiamo cercare $\overline{u}_M$ invece di $\overline{u}$, in modo tale da avere non soddisfatta la condizione $\forall v \in V$, ma $\forall v \in V_M$! Questo significa che si ottiene:

$$\sum_{i=1}^{M-1} c_i'(t) \int_0^1 N_i(x) N_j(x) \mathrm{d}x + \sum_{i=1}^{M-1} c_i(t) \int_0^1 N_i'(x) N_j'(x) \mathrm{d}x = d_j(t)$$

Dove, dal momento che V_M è uno spazio lineare, si può dire che la condizione $\forall v \in V_M$ coincida con dire $j = 1 \div M-1$! Ulteriore nota: abbiamo $c_i'(t)$; questo deriva dal fatto che si aveva nell'integrale u_t, dunque l'unica parte dipendente dal tempo della sommatoria deve essere derivata. Sostituendo e scrivendo in forma matriciale, il risultato finale è un sistema di equazioni differenziali del tipo:

$$\underline{\underline{A}}\underline{c}'(t) + \underline{\underline{B}}\underline{c}'(t) = \underline{d}$$

Si deve come al solito ricordare che gli integrali, da calcolare mediante un metodo numerico, sono da calcolare solo nell'intersezione del supporto delle funzioni, e così via.

Il sistema di equazioni differenziali ordinarie risultante dal problema è molto spesso un sistema di tipo stiff; per questo motivo si parla di sistemi stiff: essi sono molto comuni, quando si applica il metodo degli elementi finiti su problemi dipendenti dal tempo. A questo punto realizzando un programma e sostituendo all'interno tutte le condizioni iniziali per quanto concerne i primi valori, si può determinare $\underline{c}(0)$ e così via.