# Modellistica Numerica

Andrea Moiola, Università di Pavia, semestre autunnale 2022
https://euler.unipv.it/moiola/T/MN2022/MN2022.html

30 settembre 2022

In questo corso ci occuperemo di metodi numerici per l'approssimazione di equazioni differenziali. Gran parte del corso è dedicata allo studio di metodi per problemi ai limiti per equazioni differenziali ordinarie.

## 1 Problemi ai limiti e metodo di shooting

### 1.1 Problemi ai valori iniziali e problemi ai limiti

Nel corso di Analisi Numerica 2 sono stati studiati diversi metodi per approssimare numericamente il **problema ai valori iniziali** (*initial value problem, IVP*) vettoriale del primo ordine sull'intervallo $(a,b)$:

$$\begin{cases} \vec{\mathbf{y}}'(x) = \vec{\mathbf{F}}\big(x, \vec{\mathbf{y}}(x)\big) & x \in (a,b), \\ \vec{\mathbf{y}}(a) = \vec{\mathbf{y}}_0, \end{cases} \tag{1}$$

per una data funzione $\vec{\mathbf{F}} : [a,b) \times \mathbb{R}^m \to \mathbb{R}^m$ e dati iniziali $\vec{\mathbf{y}}_0 \in \mathbb{R}^m$. I metodi studiati includono ad esempio quelli di Eulero, multi-step, BDF, Runge–Kutta. Se $m = 2$, $\vec{\mathbf{F}}(x, \vec{\mathbf{y}}) = (y_2, f(x, y_1, y_2))^\top$ e le condizioni iniziali sono $\vec{\mathbf{y}} = (u_0, u_1)^\top$, il problema (1) è equivalente al problema ai valori iniziali scalare del secondo ordine

$$\begin{cases} u''(x) = f\big(x, u(x), u'(x)\big) & x \in (a,b), \\ u(a) = u_0, \\ u'(a) = u_1. \end{cases} \tag{2}$$

Le soluzioni dei due problemi sono legate dalla relazione $\vec{\mathbf{y}}(x) = (u(x), u'(x))^\top$.

Qui invece ci interessiamo al **problema ai limiti**, o **problema al contorno**, o **problema al bordo**, (*boundary value problem, BVP*) del secondo ordine:

$$\boxed{\begin{cases} u''(x) = f\big(x, u(x), u'(x)\big) & x \in (a,b), \\ u(a) = \alpha, \\ u(b) = \beta. \end{cases}} \tag{3}$$

La differenza tra il problema ai valori iniziali (2) e quello ai limiti (3) è che nel secondo caso le condizioni sono imposte in punti distinti, sul bordo dell'intervallo $(a,b)$. In generale il problema ai limiti (3) non può essere ridotto a uno ai valori iniziali.

**Nota 1.1.** Una prima differenza tra problemi ai valori iniziali e ai limiti è che mentre per i primi la regolarità del dato $f$ garantisce l'esistenza e l'unicità della soluzione, per i secondi questo non basta. Consideriamo l'equazione lineare a coefficienti costanti

$$\begin{cases} u''(x) + u(x) = 0 & x \in (a,b), \\ u(a) = \alpha, \\ u(b) = \beta, \end{cases}$$

il cui integrale generale è $u(x) = c_1 \sin x + c_2 \cos x$.

- Se i dati al contorno sono ad esempio $u(0) = 0, u(\pi/2) = 1$, allora esiste un'unica soluzione $u(x) = \sin x$.

- Se i dati sono $u(0) = 0, u(\pi) = 1$, allora non esiste alcuna soluzione.
- Se i dati sono $u(0) = u(\pi) = 0$, allora esistono infinite soluzioni $u(x) = c \sin x$ per ogni $c \in \mathbb{R}$.

Per approssimare numericamente la soluzione di un problema ai limiti vogliamo ricondurla all'approssimazione di un problema che abbiamo già imparato a risolvere numericamente. Nei precedenti corsi di analisi numerica abbiamo imparato a risolvere sistemi di equazioni algebriche (lineari e non) e problemi differenziali ai valori iniziali. I metodi di shooting approssimano la soluzione di un problema ai limiti usando tecniche numeriche già studiate per problemi ai valori iniziali, mentre i metodi che studieremo in seguito (ad esempio differenze finite ed elementi finiti) riducono le equazioni differenziali a sistemi di equazioni algebriche.

## 1.2 Metodi di shooting

La prima classe di metodi numerici che consideriamo per risolvere i problemi ai limiti è quella dei metodi di shooting. Questi sono specifici per equazioni differenziali ordinarie: non si estendono facilmente a dimensioni più alte, cioè a equazioni differenziali alle derivate parziali (PDEs).

Poiché abbiamo a disposizione diversi metodi molto efficaci per risolvere problemi ai valori iniziali, vogliamo ridurre il problema ai limiti (3) a un problema ai valori iniziali:

$$\begin{cases} u''(x) = f\big(x, u(x), u'(x)\big) & x \in (a,b), \\ u(a) = \alpha, \\ u(b) = \beta, \end{cases} \quad \rightsquigarrow \quad \begin{cases} U''(x;s) = f\big(x, U(x;s), U'(x;s)\big) & x \in (a,b), \\ U(a;s) = \alpha, \\ U'(a;s) = s. \end{cases} \tag{4}$$

Affinché la soluzione $U$ del secondo problema in (4) coincida con la soluzione $u$ del primo dobbiamo scegliere il parametro $s$, che rappresenta la derivata al tempo iniziale, in modo appropriato. Definiamo la funzione

$$\varphi(s) := U(b;s) - \beta,$$

dove $U(\cdot;s)$ è la soluzione del problema ai valori iniziali in (4). Allora $U(\cdot;s) = u(\cdot)$ è soddisfatta quando $\varphi(s) = 0$: vogliamo imporre questa condizione numericamente. In altre parole vogliamo risolvere l'equazione (non lineare) $\varphi(s) = 0$. Per questo possiamo usare un metodo iterativo di ricerca di radici, ad esempio il metodo di bisezione, quello delle secanti o quello di Newton. Ogni valutazione della funzione $\varphi$ per un diverso valore di $s$ richiede la soluzione (numerica) di un problema ai valori iniziali.

Il metodo di shooting quindi è la combinazione di un metodo per problemi ai valori iniziali e di un metodo per la ricerca di zeri.

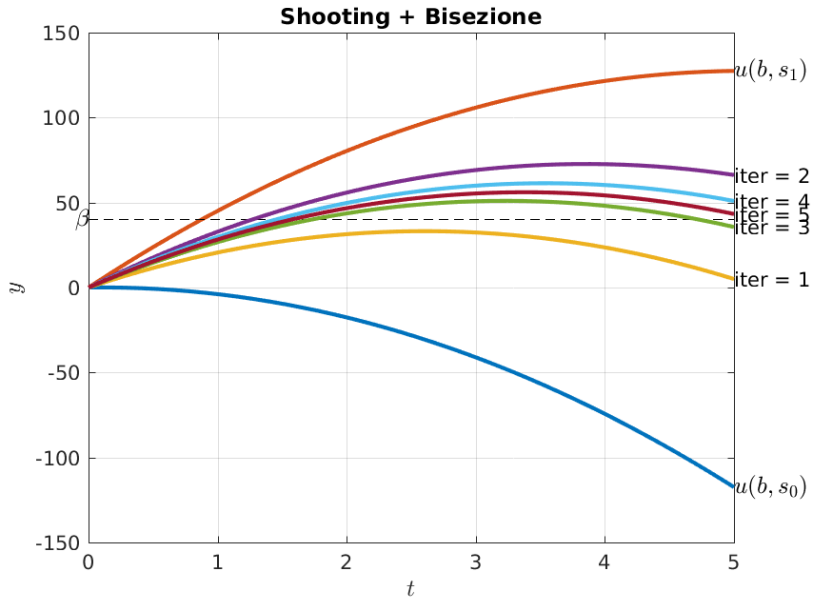



*Figura 1: Le prime 5 iterazioni del metodo di shooting nell'Esercizio 1.3. I valori per inizializzare il metodo di bisezione sono $s_0 = 1$ e $s_1 = 50$.*

**Esercizio 💻 1.2** (Imparare a usare `ode45`)**.** Il metodo di shooting richiede l'uso di un metodo numerico di risoluzione di problemi ai valori iniziali. In questo esercizio ripassiamo come usare la principale routine di Matlab per problemi ai valori iniziali, cioè `ode45`, che usa un metodo di Runge–Kutta a passi variabili. Si veda anche la pagina di supporto `help ode45`.

Per il problema ai valori iniziali del primo ordine con $m$ componenti (1), la sintassi base di `ode45` è

```
[t,y] = ode45(odefun,tspan,y0)
```

dove `odefun` rappresenta la funzione $\vec{\mathbf{F}}$ che definisce l'equazione differenziale, `tspan` è un vettore di due componenti corrispondenti al tempo iniziale e quello finale, `y0` è il vettore delle condizioni iniziali $\vec{\mathbf{y}}_0$. In particolare, `dydt = odefun(t,y)` deve accettare un input scalare `t` (la variabile indipendente) e un input vettoriale (colonna) `y` di $m$ componenti (la variabile dipendente) e deve dare in output un vettore colonna $m$-dimensionale. La funzione `odefun` può essere definita come una `function` oppure come una funzione anonima `odefun = @(t,y)...`. L'output di `ode45` è costituito da un vettore colonna `t` di $N$ tempi e da una matrice `y` di dimensione $N \times m$ dei valori approssimati delle $m$ componenti di $\vec{\mathbf{y}}$ negli $N$ istanti di tempo definiti da `t` (la $j$esima colonna è una delle $m$ componenti $y_j$, la $k$esima riga corrisponde a un tempo $t_k$).

Ad esempio, il problema con due componenti

$$\begin{aligned} y_1'(t) &= -y_2(t) - y_1(t)\big(y_1^2(t) + y_2^2(t)\big), \\ y_2'(t) &= y_1(t) - y_2(t)\big(y_1^2(t) + y_2^2(t)\big), \\ y_1(0) &= 1, \qquad y_2(0) = 0 \end{aligned}$$

può essere approssimato e plottato sull'intervallo $[0, 30]$ con i comandi

```
OdeF = @(t,y) [-y(2)-y(1)*(y(1)^2+y(2)^2); y(1)-y(2)*(y(1)^2+y(2)^2)];
[t,y] = ode45(OdeF, [0, 30], [1; 0]);
plot(t,y);
```

Il risultato è mostrato nel primo pannello di Figura 2; la linea blu e quella rossa corrispondono alle componenti $y_1$ e $y_2$, rispettivamente, cioè alle due colonne di `y`.

Approssimare i seguenti problemi ai valori iniziali usando `ode45`. Confrontare i plots ottenuti con quelli in Figura 2.

1. Equazione scalare del primo ordine:

$$y'(t) = y(t) - ty^2(t) + t, \qquad y(0) = 0, \qquad t \in [0, 5].$$

2. Sistema lineare di $m$ equazioni, ad esempio per $m = 5$:

$$\vec{\mathbf{y}}'(t) = \underline{\underline{\mathbf{A}}}\vec{\mathbf{y}}(t), \qquad A_{j,k} = \frac{-jk}{m^2}, \qquad y_1(0) = \cdots = y_m(0) = 1, \qquad t \in [0, 2].$$

3. Equazione scalare del secondo ordine (è necessario riscriverla come sistema del primo ordine):

$$y''(t) = \big(1 - y^2(t)\big)y'(t) - y(t), \qquad y(0) = 2, \qquad y'(0) = 0, \qquad t \in [0, 20].$$

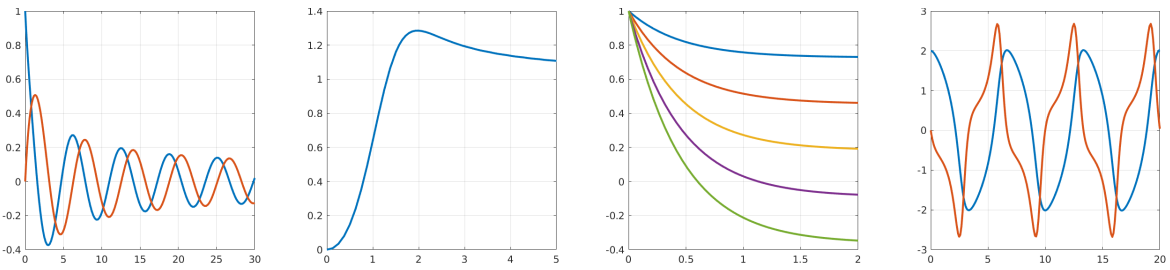


*Figura 2: Le soluzioni dei quattro problemi ai valori iniziali negli esempi dell'Esercizio 1.2. Nell'ultimo plot, la curva rossa rappresenta la derivata prima di $y$.*

**Esercizio ✎+💻 1.3.** Vogliamo lanciare da terra un fuoco d'artificio in modo tale che esploda dopo 5 secondi all'altezza di 40 metri. Se $y(t)$ rappresenta l'altezza dal suolo del petardo, la sua evoluzione è

$$\begin{cases} y''(t) = -g, \\ y(0) = 0, \\ y(5) = 40. \end{cases}$$

Vogliamo calcolare la velocità iniziale $y'(0)$ con cui dobbiamo lanciare il petardo.

- Calcolare a mano tre iterazioni del metodo di shooting combinando metodo di bisezione e soluzione esatta dei problemi ai valori iniziali. Scegliere ad esempio $s_0 = 1$ e $s_1 = 50$ per inizializzare il metodo di bisezione.
- Implementare il metodo di shooting combinando un metodo numerico per equazioni differenziali ordinarie (ad esempio `ode45`, o uno implementato ad hoc) con il metodo di bisezione. (Non dimenticarsi che `ode45` richiede di scrivere l'equazione differenziale come un problema vettoriale del primo ordine, per cui solo la prima colonna della soluzione corrisponde a $y(t)$.)
- Confrontare il risultato con il valore di $y'(0)$ ottenuto risolvendo analiticamente il problema. (Cioè $32.5\frac{m}{s}$, usando $g = 9.8\frac{m}{s^2}$.)

Plottando le prime iterazioni del metodo è possibile ottenere un grafico come in Figura 1.

Si può intuire perché questo metodo si chiami "shooting". Sparando con un'arma e volendo colpire un bersaglio a una certa distanza, un soldato/cacciatore/sportivo può procedere per tentativi regolando le condizioni iniziali (potenza, angolo di lancio) finché le condizioni finali non sono soddisfatte, cioè finché la traiettoria non colpisce il bersaglio.

**Nota 1.4** (Il metodo di shooting per equazioni lineari)**.** Nel caso di equazioni lineari il metodo di shooting non richiede iterazioni. Consideriamo il problema ai limiti

$$\begin{cases} u''(x) = p(x)u'(x) + q(x)u(x) + r(x) & x \in (a,b), \\ u(a) = \alpha, \\ u(b) = \beta. \end{cases}$$

Costruiamo due problemi ai valori iniziali

$$\begin{cases} u_1''(x) = p(x)u_1'(x) + q(x)u_1(x) + r(x), \\ u_1(a) = \alpha, \\ u_1'(a) = 0, \end{cases} \qquad \begin{cases} u_2''(x) = p(x)u_2'(x) + q(x)u_2(x) & x \in (a,b), \\ u_2(a) = 0, \\ u_2'(a) = 1. \end{cases}$$

Le soluzioni $u_1$ e $u_2$ possono essere approssimate con un metodo per problemi ai valori iniziali, ad esempio Runge–Kutta. La soluzione $U(x;s)$ del problema ai valori iniziali in (4) è loro combinazione lineare $U(x;s) = u_1(x) + su_2(x)$: si vede che questa espressione soddisfa l'equazione differenziale desiderata e le condizione $U(a;s) = \alpha$, $U'(a;s) = s$. Per soddisfare la condizione $u(b) = \beta$ dobbiamo azzerare $\varphi(s) = u_1(b) + su_2(b) - \beta$, quindi basta scegliere $s = \frac{\beta - u_1(b)}{u_2(b)}$ e

$$u(x) = U\Big(x; \frac{\beta - u_1(b)}{u_2(b)}\Big) = u_1(x) + \frac{\beta - u_1(b)}{u_2(b)} u_2(x).$$

Questo approccio è possibile solo se $u_2(b)$ è ben separata da zero e l'equazione è lineare: abbiamo usato il fatto che la combinazione lineare di due soluzioni è soluzione della stessa equazione. Se $u_2(b) = 0$ allora la soluzione del problema non può essere unica: data una soluzione $u$ (in caso questa esista), anche $u + cu_2$ sarà soluzione per ogni $c \in \mathbb{R}$.

**Esercizio ✎ 1.5.** Usare l'approccio della Nota 1.4 per risolvere il problema al bordo nell'Esercizio 1.3 risolvendo due problemi ai valori iniziali.

### 1.2.1 Il metodo di shooting combinato con il metodo di Newton

Ogni valutazione della funzione $\varphi(s)$ per un diverso valore di $s$ richiede la soluzione numerica di un problema ai valori iniziali, che può essere computazionalmente costosa. È quindi importante usare un metodo di ricerca delle radici che converga velocemente, per ridurre il numero di valutazioni. Per questo motivo, invece del metodo di bisezione, che converge solo linearmente, si può usare il metodo di Newton(–Raphson), che converge quadraticamente.

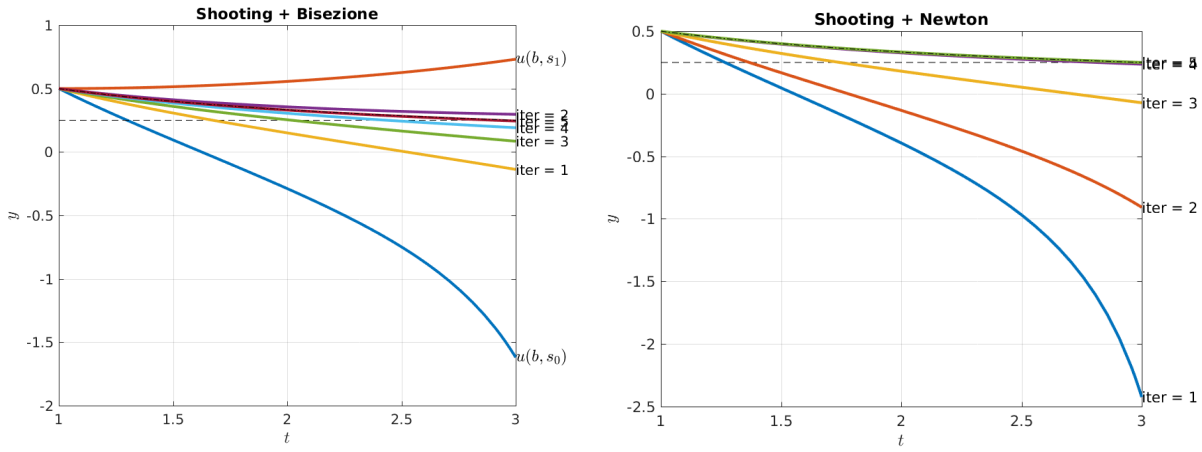



*Figura 3: Le prime 5 iterazioni del metodo di shooting per il problema nell'Esercizio 1.6, accoppiato con il metodo di bisezione (sinistra) e con quello di Newton (destra).*

Consideriamo il problema ai limiti (4). Una volta scelto un valore iniziale $s^0$, l'iterazione $(n+1)$-esima del metodo di Newton per la funzione $\varphi(s) = U(b;s) - \beta$ è

$$s^{n+1} = s^n - \frac{\varphi(s^n)}{\varphi'(s^n)} = s^n - \frac{U(b;s^n) - \beta}{U_s(b;s^n)}, \tag{5}$$

dove $U_s(b;s^n)$ è la derivata di $U(b;s)$ rispetto a $s$ valutata in $s = s^n$. Per implementare il metodo dobbiamo essere in grado di calcolare questa derivata. Derivando rispetto a $s$ l'equazione differenziale soddisfatta da $U$ e le condizioni iniziali (4), vediamo che $U_s(x;s)$ è soluzione del problema ai valori iniziali

$$\begin{cases} U_s''(x;s) = f_u\big(x, U(x;s), U'(x;s)\big)\, U_s(x;s) + f_{u'}\big(x, U(x;s), U'(x;s)\big)\, U_s'(x;s) & x \in (a,b), \\ U_s(a;s) = 0, \\ U_s'(a;s) = 1, \end{cases}$$

dove il segno $'$ denota la derivata rispetto alla variabile $x$, mentre $f_u$ e $f_{u'}$ rappresentano le derivate parziali di $f$ rispetto alla seconda e terza variabile, rispettivamente. Notiamo che il problema ottenuto è lineare. Inoltre l'equazione lineare soddisfatta da $U_s$ dipende dal valore di $U$: le due equazioni sono accoppiate e possono essere risolte simultaneamente.

In sintesi, per risolvere il problema ai limiti in (4) per $u$ usando il metodo di shooting e quello di Newton, bisogna iterativamente

- risolvere il problema ai valori iniziali per $U(x,s)$ e $U_s(x,s)$ (simultaneamente), in dipendenza dal parametro $s^n$,
- calcolare il nuovo parametro $s^{n+1}$ usando l'iterazione di Newton (5),

finché non si ottiene l'accuratezza desiderata.

**Esercizio 1.6.** Consideriamo il problema ai limiti non lineare

$$\begin{cases} u'' = u^3 - uu' & \text{in } (1,3), \\ u(1) = \frac{1}{2}, \\ u(3) = \frac{1}{4}, \end{cases}$$

la cui soluzione esatta è $u(x) = 1/(1+x)$. Approssimare la soluzione $u$ usando il metodo di shooting accoppiando `ode45` con

- il metodo di bisezione, applicato all'intervallo $s \in (-0.9, 0)$ per il parametro $s = u'(1)$;
- il metodo di Newton, partendo dal valore iniziale $s^0 = -1$ per $s$. In questo caso, dovendo risolvere simultaneamente ad ogni iterazioni due equazioni differenziali del secondo ordine, `ode45` richiede di scrivere il problema come un sistema del primo ordine con quattro componenti.

Quale metodo converge più velocemente? Come si può migliorare l'accuratezza della soluzione ottenuta?

**Nota 1.7.** Consideriamo un problema vettoriale, cioè in cui l'incognita è $\vec{\mathbf{y}} : (a,b) \to \mathbb{R}^m$, in cui $0 < \ell < m$ delle $m$ condizioni al bordo sono imposte nel punto iniziale $a$ e le restanti $m - \ell$ nel punto $b$, ad esempio:

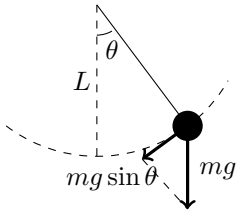
$$\begin{cases} \vec{\mathbf{y}}'(x) = \vec{\mathbf{F}}\big(x, \vec{\mathbf{y}}(x)\big) & x \in (a,b), \\ y_1(a) = \alpha_1, \quad \dots, \quad y_\ell(a) = \alpha_\ell, \\ y_{\ell+1}(b) = \beta_{\ell+1}, \quad \dots, \quad y_m(b) = \beta_m. \end{cases}$$

Se $m - \ell > 1$, la funzione $\vec{\varphi}(\vec{\mathbf{s}}) = [y_{\ell+1}(b) - \beta_{\ell+1}, \dots, y_m(b) - \beta_m]$, dove $\vec{\mathbf{s}}$ è il vettore che contiene le approssimazioni di $[y_{\ell+1}(a), \dots, y_m(a)]$, è una funzione vettoriale di variabile vettoriale, $\vec{\varphi} : \mathbb{R}^{m-\ell} \to \mathbb{R}^{m-\ell}$. In questo caso non è possibile usare il metodo di bisezione, che è definito solo per funzioni scalari. Questa è un'ulteriore motivazione per abbinare il metodo di shooting a quello di Newton, che può essere implementato in qualunque dimensione.

Il metodo di Newton è molto sensibile alla scelta del valore iniziale $s^0$: se è troppo lontano dal valore ricercato il metodo non converge. Un primo modo per aggirare questo problema consiste nell'usare alcune iterazioni del metodo di bisezione o delle secanti prima di iniziare le iterazioni di Newton. Un secondo modo consiste nel dividere l'intervallo $(a,b)$ in sotto-intervalli e applicare il metodo di shooting ad ognuno di essi simultaneamente ("*multiple shooting method*"). Notiamo comunque che l'errore commesso dal metodo di Newton dipende da quello del solutore del problema iniziale usato.

### 1.2.2 Il problema del pendolo

Consideriamo una massa $m$ fissata a un'asta di lunghezza $L$ e peso trascurabile che ruota senza attriti intorno all'origine. Denotiamo con $\theta(t)$ l'angolo tra la direzione dell'asta e la verticale verso il basso. Se $s(t) = L\theta(t)$ denota la distanza in lunghezza d'arco dal punto più basso, l'accelerazione è data dalla sua derivata seconda $a(t) = s''(t) = L\theta''(t)$. La forza di gravità agente sulla massa è $-mg$, la cui componente tangente alla circonferenza è $F = -mg\sin\theta(t)$. La legge di Newton dà $ma(t) = F = -mg\sin\theta(t)$. Uguagliando l'accelerazione in queste due espressioni abbiamo $\theta''(t) = -\frac{g}{L}\sin\theta(t)$.




Assumendo che $g$ ed $L$ siano normalizzate a 1, abbiamo l'equazione differenziale del pendolo

$$\theta''(t) = -\sin\theta(t).$$

Se $\theta$ è piccolo questa equazione viene approssimata usando $\sin\theta \approx \theta$, cioè $\theta'' = -\theta$, le cui soluzioni sono combinazioni lineari di $\sin t$ e $\cos t$ e il cui periodo è indipendente dall'ampiezza dell'oscillazione.

**Esercizio 💻 1.8.** Usare il metodo di shooting per calcolare la velocità angolare iniziale $\theta'(0)$ con cui deve muoversi un pendolo per partire da $\theta(0) = \pi/3$ e tornare nella stessa posizione esattamente dopo un tempo $T = 2\pi$. In altre parole calcolare $\theta'(0)$ dove $\theta$ è la soluzione del problema al contorno

$$\begin{cases} \theta'' = -\sin\theta, \\ \theta(0) = \frac{\pi}{3}, \qquad \theta(2\pi) = \frac{\pi}{3}. \end{cases}$$

La soluzione ottenuta dipende dal valore iniziale scelto: dalla seconda e terza immagine in Figura 5 vediamo che il pendolo può tornare a $\pi/3$ dopo un tempo $T = 2\pi$ in (almeno) due modi diversi, corrispondenti a due diverse soluzioni isolate dello stesso problema al contorno. I valori di $\theta'(0)$ ottenuti sono $-0.2121$ e $1.7205$, rispettivamente. Il metodo di Newton converge molto velocemente in pochissime iterazioni (5 nell'esempio) al livello di precisione macchina, ma l'accuratezza della soluzione dipende dal solutore dei problemi a valori iniziali (`ode45` in questo caso). Qual è il significato fisico di queste diverse soluzioni? Ne esistono altre? Quante?

In questo esempio il pendolo ritorna al punto di partenza senza avere compiuto un giro completo; se vogliamo che nel tempo $T$ il pendolo giri una volta intorno all'origine dobbiamo imporre $\theta(T) = \frac{\pi}{3} + 2\pi$.

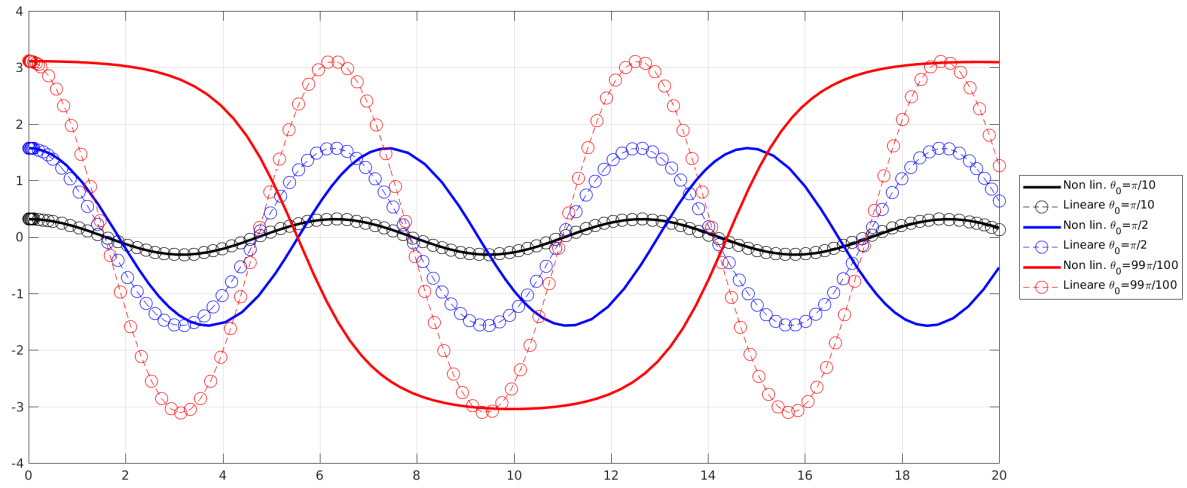



*Figura 4: Le soluzioni dell'equazione del pendolo $\theta'' = -\sin\theta$ e della sua approssimazione lineare $u'' = -\theta$ per diversi valori inziali. In tutti i casi $\theta'(0) = 0$. La soluzione dell'equazione linearizzata è vicina a quella dell'equazione non lineare solo per valori iniziali piccoli.*

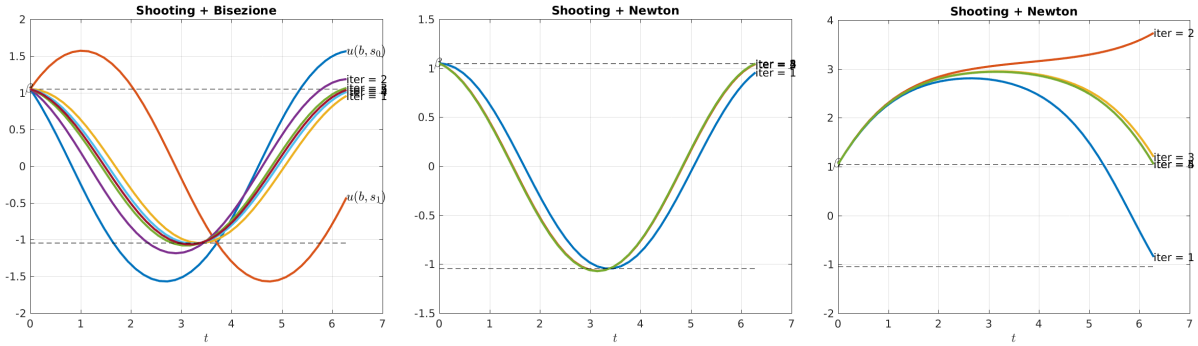



*Figura 5: La soluzione del problema nell'Esercizio 1.8. A sinistra: le prime iterazioni del metodo di bisezione con scelte iniziali $s_0 = -1$ e $s_1 = 1$. Al centro: le prime iterazioni del metodo di Newton con scelta iniziale $s_0 = 0$. A destra: lo stesso con $s_0 = 1.7$: la soluzione è qualitativamente diversa dalla precedente, abbiamo trovato due soluzioni distinte.*

# 2 Equazioni di diffusione, trasporto e reazione

## 2.1 Modelli di diffusione, trasporto e reazione

I modelli matematici che coinvolgono fenomeni di diffusione, trasporto e reazione, ad esempio di sostanze chimiche, popolazioni umane, di animali o cellule, oltre a essere rilevanti per le applicazioni, danno origine a molte equazioni differenziali estremamente importanti. Queste sono equazioni differenziali ordinarie (*ordinary differential equation, ODE*) o alle derivate parziali (*partial differential equation, PDE*), a seconda della dimensione del dominio su cui sono definite. Possono essere equazioni di evoluzione o stazionarie (cioè dipendenti o meno dalla variabile temporale), lineari o non lineari. In questa sezione deriviamo velocemente e informalmente alcune di queste equazioni differenziali, che useremo in seguito come modello per diversi metodi numerici. Le proprietà di queste equazioni e una loro derivazione più rigorosa verranno approfondite in altri corsi.

Sia $u(\mathbf{x}, t)$ la variabile che rappresenta la **densità** di una grandezza fisica in un punto $\mathbf{x}$ in spazio e all'istante $t$. Questa grandezza può essere ad esempio la massa di una sostanza chimica disciolta in un ambiente pieno d'aria o di acqua. Chiamiamo questa densità $u$ perché sarà la soluzione dell'equazione differenziale che andremo a definire, cioè è l'incognita, in inglese *unknown*. Sia $f(\mathbf{x}, t)$ il "termine di sorgente", cioè la funzione che rappresenta la produzione (o distruzione, se negativa) della stessa grandezza in $\mathbf{x}$ all'istante $t$. Assumiamo che $u$ ed $f$ siano definite e sufficientemente lisce per ogni $\mathbf{x} \in D$, dove il dominio $D$ è un insieme aperto di $\mathbb{R}^d$ rappresentante la regione spaziale d'interesse, e ogni $t \in I$, dove $I \subset \mathbb{R}$ è un intervallo. (Ovviamente nei problemi fisici i casi rilevanti sono $d = 1, 2, 3$, ma in alcune

applicazioni, ad esempio in finanza, viene considerato anche il caso $d > 3$.) Chiamiamo $\mathbf{J}$ il "**flusso**" (o più precisamente la densità di flusso) della grandezza considerata, cioè il campo vettoriale che rappresenta la quantità di questa grandezza che in un tempo unitario attraversa una superficie unitaria. In altre parole, l'integrale della componente normale di $\mathbf{J}$ su una superficie è la quantità della grandezza considerata che attraversa la superficie stessa. In particolare, se la quantità avente densità $u$ si muove con velocità $\mathbf{v}(\mathbf{x}, t)$, avremo $\mathbf{J}(\mathbf{x}, t) = u(\mathbf{x}, t)\mathbf{v}(\mathbf{x}, t)$.

### 2.1.1 Leggi di conservazione (equazioni di continuità)

La variazione di massa in una regione $\Omega \subset D$ (dal bordo liscio) tra l'istante $t_1$ e l'istante $t_2$ è

$$\delta M = \int_\Omega u(\mathbf{x}, t_2)\,\mathrm{d}\mathbf{x} - \int_\Omega u(\mathbf{x}, t_1)\,\mathrm{d}\mathbf{x} = \int_{t_1}^{t_2} \frac{\partial}{\partial t} \int_\Omega u(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t,$$

dove abbiamo usato il teorema fondamentale del calcolo. La massa totale varia a causa della produzione/distruzione dovuta a $f$ e alla quantità di materiale che attraversa la frontiera di $\Omega$:

$$\delta M = \int_{t_1}^{t_2} \int_\Omega f(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t - \int_{t_1}^{t_2} \oint_{\partial\Omega} \mathbf{J}(\mathbf{x}, t) \cdot \mathbf{n}(\mathbf{x})\,\mathrm{d}S(\mathbf{x})\,\mathrm{d}t,$$

dove $\mathbf{n}(\mathbf{x})$ è il campo vettoriale unitario (cioè tale che $\|\mathbf{n}\| = 1$) definito su $\partial\Omega$, perpendicolare a $\partial\Omega$ e che punta verso l'esterno di $\Omega$. Uguagliando $\delta M$ in queste due equazioni otteniamo l'**equazione di continuità** (in forma integrale), che lega tra loro le quantità $u, f$ e $\mathbf{J}$:

$$\int_{t_1}^{t_2} \frac{\partial}{\partial t} \int_\Omega u(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t = \int_{t_1}^{t_2} \int_\Omega f(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t - \int_{t_1}^{t_2} \oint_{\partial\Omega} \mathbf{J}(\mathbf{x}, t) \cdot \mathbf{n}(\mathbf{x})\,\mathrm{d}S(\mathbf{x})\,\mathrm{d}t. \tag{6}$$

L'equazione di continuità chiarisce il significato delle variabili coinvolte: informalmente, (6) si può leggere "la variazione in tempo ($\frac{\partial}{\partial t}$) tra gli istanti $t_1$ e $t_2$ ($\int_{t_1}^{t_2}$) della quantità di sostanza contenuta nella regione $\Omega$ ($\int_\Omega u$) è uguale alla differenza tra la quantità netta di sostanza prodotta all'interno della regione ($\int_\Omega f$) e la quantità della stessa sostanza che fuoriesce dal bordo della regione ($\oint_{\partial\Omega} \mathbf{J} \cdot \mathbf{n}$)".

Assumendo sufficiente regolarità, e usando il teorema della divergenza (o di Gauss) $\oint_{\partial\Omega} \mathbf{J} \cdot \mathbf{n}\,\mathrm{d}S = \int_\Omega \nabla \cdot \mathbf{J}\,\mathrm{d}\mathbf{x}$, l'equazione di continuità diventa

$$\int_{t_1}^{t_2} \int_\Omega \Big( \frac{\partial u}{\partial t}(\mathbf{x}, t) + \nabla \cdot \mathbf{J}(\mathbf{x}, t) - f(\mathbf{x}, t) \Big)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t = 0.$$

(Qui $\nabla\cdot$ rappresenta la divergenza nella sola variabile spaziale, cioè $\nabla \cdot \mathbf{J} = \sum_{i=1}^d \frac{\partial J_i}{\partial x_i}$.) Poiché l'integrale di questa quantità è zero per ogni aperto liscio $\Omega$ contenuto in $D$ e ogni intervallo $(t_1, t_2) \subset I$, l'integrando stesso è zero in $D$, cioè

$$\boxed{\frac{\partial u}{\partial t}(\mathbf{x}, t) + \nabla \cdot \mathbf{J}(\mathbf{x}, t) = f(\mathbf{x}, t)} \qquad \mathbf{x} \in D,\ t \in I. \tag{7}$$

Questa è l'**equazione di continuità in forma differenziale**. È la prima equazione differenziale alle derivate parziali che incontriamo: lega tra loro le derivate parziali di $u$ e $\mathbf{J}$ in $t$ e nelle $n$ variabili spaziali $\mathbf{x}$. La stessa equazione di continuità, con $f = 0$, si incontra anche nell'elettromagnetismo; in questo caso $u$ (di solito denotata $\rho$) rappresenta la densità di carica e $\mathbf{J}$ la densità di corrente elettrica.

### 2.1.2 Leggi costitutive

In molte situazioni fisiche la quantità considerata si diffonde da dove ha una concentrazione maggiore a dove ne ha una minore. Questo fatto si può modellare dicendo che il flusso $\mathbf{J}$ punta verso le regioni dove $u$ è più bassa, cioè $\mathbf{J}$ punta nella direzione opposta al gradiente $\nabla u$. Nel caso più semplice assumiamo che queste due quantità vettoriali siano proporzionali e otteniamo la **legge di Fick**:

$$\boxed{\mathbf{J}(\mathbf{x}, t) = -K \nabla u(\mathbf{x}, t),} \tag{8}$$

dove $K > 0$ è una costante di proporzionalità chiamata coefficiente di diffusione. Notiamo che mentre l'equazione di continuità è una legge "fondamentale", rappresentando il principio di conservazione della massa, l'equazione di Fick è una legge "costitutiva", cioè specifica per il materiale considerato e solitamente è derivata da un'approssimazione.

### 2.1.3 Equazioni di secondo grado

Combinando la legge di Fick e l'equazione di continuità, e ricordando che la divergenza del gradiente è il Laplaciano, otteniamo l'**equazione del calore** (*heat equation*):

$$\boxed{\frac{\partial u}{\partial t} - K\Delta u = f.} \tag{9}$$

Questa è un'equazione alle derivate parziali lineare, del secondo ordine, di evoluzione e di tipo parabolico.

L'equazione del calore ha questo nome perché viene usata anche per modellare la diffusione del calore in un corpo omogeneo. In questo caso, $u$ rappresenta la temperatura, $\mathbf{J}$ la densità di corrente termica, $K$ la conducibilità, $f$ la produzione di calore. In questa situazione, la legge di Fick prende il nome di legge di Fourier.

L'equazione del calore può essere generalizzata in molti modi. Se una sostanza distribuita uniformemente nello spazio si degrada, ad esempio attraverso una reazione chimica, la sua densità diminuisce con un tasso costante nel tempo (cioè esponenzialmente) e soddisfa l'equazione $\frac{\partial u}{\partial t} + qu = 0$, dove $q > 0$ è detto coefficiente di **reazione**. Se la stessa sostanza è sottoposta sia a processi di diffusione che di reazione allora soddisfa

$$\frac{\partial u}{\partial t} - K\Delta u + qu = f.$$

Più in generale, potremmo avere $\frac{\partial u}{\partial t} - K\Delta u + \tilde{q}(u) = f$ dove $\tilde{q}(\cdot)$ è una funzione reale non lineare.

Se inoltre la sostanza è diluita in un fluido che si muove con velocità $\mathbf{p}(\mathbf{x}, t)$, e viene trasportata da questo fluido, la legge di Fick (8) diventa $\mathbf{J}(\mathbf{x}, t) = -K\nabla u(\mathbf{x}, t) + \mathbf{p}(\mathbf{x}, t)u(\mathbf{x}, t)$. Inserendo quest'ultima nell'equazione di continuità (7) otteniamo un termine di **trasporto** (*transport*, o *advection*, o anche *convection*) del primo ordine:

$$\frac{\partial u}{\partial t} - K\Delta u + \mathbf{p} \cdot \nabla u + \hat{q}u = f,$$

dove $\hat{q} = (q + \nabla \cdot \mathbf{p})$ costituisce il nuovo termine di reazione.

Se il coefficiente di diffusione $K$ dipende dalla posizione, cioè $K(\mathbf{x}) > 0$, ad esempio perché diverse porzioni del dominio sono occupate da materiali con proprietà diverse, invece del Laplaciano otteniamo

$$\boxed{\frac{\partial u}{\partial t} - \nabla \cdot (K\nabla u) + \mathbf{p} \cdot \nabla u + qu = f.} \tag{10}$$

Se il materiale non è isotropo ma la sostanza con densità $u$ (o il calore se $u$ rappresenta la temperatura) fluisce con più facilità in una certa direzione (ad esempio se il dominio è costituito da fibre o lamine), allora il coefficiente di diffusione $K$ è una matrice $n \times n$ definita positiva. Anche $K$ può dipendere dalla densità $u$: in questo caso si ottiene un'equazione non lineare, ad esempio quella che governa il movimento di un liquido all'interno di mezzi porosi; questo modello è importante ad esempio per applicazioni all'estrazione di petrolio (*porous medium equation*). L'equazione (10) è l'equazione generale di **diffusione–trasporto–reazione** lineare e non-stazionaria.

Molte applicazioni coinvolgono la diffusione di diverse quantità che interagiscono tra loro: sostanze chimiche che partecipano a reazioni, specie animali che competono per risorse o si predano, popolazioni umane in cui si diffonde una malattia contagiosa... Queste situazioni si possono modellare con sistemi di equazioni di diffusione–trasporto–reazione (un'incognita $u_1, \ldots, u_N$ e un'equazione per ciascuna sostanza o popolazione), accoppiate tra loro attraverso il termine di reazione, che di solito è non lineare.

### 2.1.4 Equazioni stazionarie

Nel caso stazionario, cioè in cui né $u$ né le altre variabili dipendono dalla variabile temporale, perdiamo il termine con $\frac{\partial}{\partial t}$ e otteniamo l'equazione di diffusione–trasporto–reazione lineare stazionaria

$$\boxed{-\nabla \cdot (K\nabla u) + \mathbf{p} \cdot \nabla u + qu = f.} \tag{11}$$

Questo è il modello generale di equazione ellittica lineare del secondo ordine. Se $K$ è costante (e isotropa), non ci sono termini né di trasporto né di reazione ($\mathbf{p} = \mathbf{0}$ e $q = 0$) abbiamo l'**equazione di Poisson**

$$\boxed{-\Delta u = f.}$$

Quando $f = 0$ questa viene detta **equazione di Laplace** $-\Delta u = 0$, le cui soluzioni sono dette "funzioni armoniche". Lo studio delle funzioni armoniche è detto "teoria del potenziale". Nel caso 1-dimensionale le

uniche funzioni armoniche sono quelle lineari, mentre in dimensioni più alte costituiscono una classe molto più ampia: ad esempio in due dimensioni, identificando $\mathbb{R}^2$ e $\mathbb{C}$, la parte reale di una qualsiasi funzione olomorfa è armonica. (Esercizio: mostrare questo fatto usando le equazioni di Cauchy–Riemann.)

Nel caso uno-dimensionale $d = 1$, la (11) si riduce a un'equazione differenziale ordinaria

$$-\frac{\partial}{\partial x}\Big(K\frac{\partial u}{\partial x}\Big) + p\frac{\partial u}{\partial x} + qu = f. \tag{12}$$

Il caso uno-dimensionale può rappresentare la diffusione di un fluido in un tubo o del calore in una barra metallica sottile. Per buona parte del corso ci occuperemo di questa semplice equazione, spesso con ulteriori semplificazioni (come $K = 1$ e $p = 0$). La sua semplicità ci permette di introdurre agevolmente molte tecniche numeriche e analitiche che, a prezzo di complicazioni tecniche, sono applicabili a molti dei modelli più generali menzionati in questa sezione. L'importanza dell'equazione (12) sta proprio nell'essere esemplificativa di equazioni più complesse e più rilevanti per le applicazioni.

Per individuare una soluzione di un'equazione differenziale ordinaria imponiamo condizioni iniziali (2) o al contorno (3). Allo stesso modo, per le equazioni alle derivate parziali qui descritte vengono imposte condizioni al bordo sulla frontiera del dominio $D$ e, nel caso di equazioni non-stazionarie, condizioni iniziali. Descriveremo queste condizioni in seguito.

**Nota 2.1.** L'equazione (10) (e il suo caso particolare l'equazione del calore) è il prototipo delle equazioni di tipo parabolico, mentre l'equazione (11) (e il suo caso particolare l'equazione di Poisson) lo è per le equazioni di tipo ellittico. Esiste un terzo tipo di equazioni alle derivate parziali lineari del secondo ordine, cioè quelle iperboliche, il cui rappresentante più noto e importante è l'equazione delle onde

$$\frac{\partial^2 u}{\partial t^2} - \Delta u = f.$$

**Nota 2.2.** Immaginiamo di avere una barra metallica isolata e che $u(x,t)$ rappresenti la sua temperatura nel punto $x$ al tempo $t$. Intuitivamente, $u(x,t)$ è destinata ad aumentare se $u(x,t)$ è minore della media delle temperature in un intorno, cioè se $u(x,t) < \frac{1}{2}(u(x+\epsilon,t) + u(x-\epsilon,t))$ per $\epsilon$ piccolo. Questa condizione è equivalente a $\frac{\partial^2 u(x,t)}{\partial x^2} > 0$, cioè alla convessità di $u$ in $x$. Possiamo anche immaginare che la velocità di crescita di $u$ sia proporzionale a questa derivata cioè $\frac{\partial u}{\partial t} = K\frac{\partial^2 u}{\partial x^2}$. Questa è un'altra derivazione (molto) empirica dell'equazione del calore in una dimensione spaziale. In dimensione $d > 1$ vale lo stesso ragionamento, usando la formula $\Delta f(\mathbf{y}) = \lim_{\epsilon\to 0} \frac{2d}{\epsilon^2}(⨍_{\{\mathbf{y}\in\mathbb{R}^d, |\mathbf{y}-\mathbf{x}|=\epsilon\}} f(\mathbf{y})\,\mathrm{d}S(\mathbf{y}) - f(\mathbf{x}))$ (dimostrabile usando l'espansione di Taylor). Questa formula dice che il Laplaciano misura quanto il valore di una funzione in un punto differisca dalla sua media sul bordo di un intorno infinitesimo.

## 2.2 Problemi al bordo lineari in una dimensione

Consideriamo il problema ai limiti per l'equazione lineare di diffusione–trasporto–reazione ((12) con $K = 1$) con **condizioni al bordo di Dirichlet**:

$$\begin{cases} -u''(x) + p(x)u'(x) + q(x)u(x) = f(x) & x \in (a,b), \\ u(a) = \alpha, \\ u(b) = \beta, \end{cases} \tag{13}$$

dove $p, q, f \in C^0([a,b])$ e $a < b, \alpha, \beta \in \mathbb{R}$. Vogliamo studiare delle condizioni che garantiscano l'esistenza e l'unicità della soluzione $u$. Ricordiamo che nella Nota 1.1 abbiamo visto che per $p = f = 0$, $q = -1$, esistono valori di $a, b, \alpha, \beta$ per cui esiste un'unica soluzione di (13), valori per cui non ne esiste nessuna e valori per cui ne esistono infinite.

**Nota 2.3** (Problemi lineari: differenziali e algebrici)**.** Ci stiamo occupando di problemi differenziali *lineari*. Nel caso dei problemi lineari più semplici in assoluto, cioè i sistemi lineari algebrici quadrati $\underline{\underline{\mathbf{A}}}\vec{\mathbf{x}} = \vec{\mathbf{B}}$, abbiamo imparato nel corso di algebra lineare che l'esistenza e l'unicità della soluzione si studiano guardando innanzitutto il sistema omogeneo, cioè $\underline{\underline{\mathbf{A}}}\vec{\mathbf{x}}_0 = \vec{\mathbf{0}}$. Se questo ammette solo la soluzione banale $\vec{\mathbf{x}}_0 = \vec{\mathbf{0}}$ allora qualunque sistema lineare con la stessa matrice è ben posto. Seguiamo una strategia simile anche per i problemi al bordo. I dati $f, \alpha, \beta$ giocheranno il ruolo del termine noto $\vec{\mathbf{B}}$, mentre l'operatore differenziale $u \mapsto -u'' + pu' + qu$ quello della matrice $\underline{\underline{\mathbf{A}}}$.

Partiamo dal **problema omogeneo**, cioè con dati $f, \alpha, \beta$ uguali a zero:

$$\begin{cases} -u_0''(x) + p(x)u_0'(x) + q(x)u_0(x) = 0 & x \in (a,b), \\ u_0(a) = 0, \\ u_0(b) = 0. \end{cases} \tag{14}$$

La funzione nulla $u_0(x) = 0$ è chiaramente soluzione di questo problema ai limiti. Quindi per il problema (14) abbiamo sempre l'esistenza di una soluzione. Questa soluzione sarà unica o ne esisteranno altre? (Cioè esiste $u_0 \neq 0$ soluzione di (14)?)

Ricordando la Nota 2.3, seguiamo una strategia composta da due passi indipendenti: (i) in §2.2.1–§2.2.2 studiamo l'unicità del problema omogeneo (14), (ii) in §2.2.3 osserviamo come l'unicità del problema omogeneo (14) implica la buona posizione di quello non-omogeneo (13).

**Nota 2.4.** Affinché tutti i termini nell'equazione parabolica (10) abbiano la stessa dimensione fisica e quindi possano essere sommati tra loro, il parametro $q$ deve avere la dimensione dell'inverso di un tempo $[T^{-1}]$, $\mathbf{p}$ la dimensione di una velocità $[L][T^{-1}]$, $K$ quella del quadrato di una lunghezza divisa per un tempo $[L^2][T^{-1}]$. Questo perché la derivata di una quantità fisica rispetto ad un'altra ha la dimensione del rapporto tra le due quantità. Nella (13) invece abbiamo implicitamente diviso per $K$ (imponendolo uguale a 1) per cui $q$ e $p$ hanno rispettivamente le dimensioni $[L^{-2}]$ e $[L^{-1}]$.

### 2.2.1 Il metodo dell'energia

Per studiare le soluzioni del problema omogeneo usiamo il "metodo dell'energia". Data $u_0$ soluzione di (14), moltiplichiamo l'equazione differenziale per $u_0(x)$, integriamo su $(a, b)$, usiamo la formula di integrazione per parti due volte e le condizioni al bordo:

$$
\begin{aligned}
0 &= \int_a^b \Big( -u_0''(x) + p(x)u_0'(x) + q(x)u_0(x)\Big)u_0(x)\,\mathrm{d}x \\
&= \int_a^b \Big(u_0'(x)u_0'(x) + p(x)\frac{1}{2}\frac{\partial}{\partial x}u_0^2(x) + q(x)u_0^2(x)\Big)\,\mathrm{d}x - u_0'(b)\underbrace{u_0(b)}_{=0} + u_0'(a)\underbrace{u_0(a)}_{=0} \\
&= \int_a^b \Big(u_0'(x)u_0'(x) - \frac{1}{2}p'(x)u_0^2(x) + q(x)u_0^2(x)\Big)\,\mathrm{d}x + \frac{p(b)}{2}\underbrace{u_0^2(b)}_{=0} - \frac{p(a)}{2}\underbrace{u_0^2(a)}_{=0} \\
&= \int_a^b \Big(\big(u_0'(x)\big)^2 + \Big(q(x) - \frac{1}{2}p'(x)\Big)u_0^2(x)\Big)\,\mathrm{d}x.
\end{aligned}
$$

Se $q(x) - \frac{1}{2}p'(x) \geq 0$ in $(a, b)$, questa è la somma di due termini positivi. Essendo questa somma uguale a zero (per l'equazione differenziale omogenea) questo implica che $u_0(x) = u_0'(x) = 0$ in tutto l'intervallo. Abbiamo ottenuto il seguente risultato.

**Proposizione 2.5.** Dati $p \in C^1([a, b])$, $q \in C^0([a, b])$ e $[a, b] \subset \mathbb{R}$, se vale la condizione

$$q(x) - \frac{1}{2}p'(x) \geq 0 \qquad \forall x \in (a, b),$$

allora l'unica soluzione del problema omogeneo (14) è $u_0(x) = 0$.

Come previsto, il caso considerato nella Nota 1.1, cioè $p = 0$ e $q = -1$, non soddisfa la condizione nella proposizione.

### 2.2.2 Principio del massimo

Sappiamo che se $u \in C^2(a, b)$ e $x^* \in (a, b)$ (notare che escludiamo gli estremi) è un punto di massimo locale allora $u'(x^*) = 0$ e $u''(x^*) \leq 0$. Questo ci offre delle informazioni qualitative sulle soluzioni di alcuni problemi ai limiti senza bisogno di risolverli. Ad esempio, se $u$ soddisfa l'equazione $-u''(x) = f(x)$ per un dato $f < 0$ su tutto l'intervallo, allora $u$ non può avere massimi locali all'interno di $(a, b)$. Cosa possiamo dire per l'equazione differenziale in (13)?

Cominciamo dal caso $q = 0$. Se $u$ soddisfa $-u''(x) + p(x)u'(x) < 0$ per ogni $x \in (a, b)$ allora in un punto stazionario interno $x^*$ abbiamo $u'(x^*) = 0$ e $u''(x^*) > 0$, quindi $x^*$ deve essere un minimo. Nella prossima proposizione consideriamo il caso in cui la disuguaglianza non è stretta: le uniche funzioni che raggiungono il massimo all'interno dell'intervallo sono le costanti.

**Proposizione 2.6.** Sia $p \in C^0([a, b])$ e $u \in C^2([a, b])$ una funzione che soddisfa la disuguaglianza

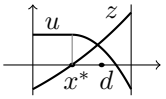
$$-u''(x) + p(x)u'(x) \leq 0 \quad \text{in } (a, b).$$

Se $u(x^*) = M := \max_{x\in[a,b]} u(x)$ per un $x^* \in (a, b)$, allora $u(x) = M$ per ogni $x \in [a, b]$.

*Dimostrazione.* L'idea alla base della dimostrazione è di mostrare che se $u$ viola la tesi allora è possibile costruire una sua piccola perturbazione $w$ che soddisfa $-w'' + pw' < 0$ ed ha un massimo locale interno, generando una contraddizione.

Per assurdo assumiamo che esista $d \in (a,b)$ tale che $u(d) < u(x^*)$. Per semplicità assumiamo che $d > x^*$. Definiamo la funzione

$$z(x) := \mathrm{e}^{\gamma(x-x^*)} - 1 \qquad \text{per qualche } \gamma > \max\{0, \max_{x\in[a,b]} p(x)\}.$$




Notiamo che $z$ è strettamente crescente e $z(x^*) = 0$, quindi $z(d) > 0$. Inoltre

$$-z''(x) + p(x)z'(x) = \gamma\big(p(x) - \gamma\big)\mathrm{e}^{\gamma(x-x^*)} < 0.$$

Ora usiamo $z$ per costruire una perturbazione di $u$: definiamo

$$w(x) := u(x) + \epsilon z(x) \qquad \text{per } 0 < \epsilon < \frac{M - u(d)}{z(d)}.$$

Poiché $-u'' + pu' \le 0$ e $-z'' + pz' < 0$ abbiamo anche $-w'' + pw' < 0$, quindi se troviamo un massimo locale di $w$ all'interno di $(a,b)$ abbiamo la contraddizione desiderata. Abbiamo

$$w(x) < M \quad \text{per } a < x < x^*, \qquad w(x^*) = M, \qquad w(d) = u(d) + \epsilon z(d) < u(d) + \frac{M - u(d)}{z(d)} z(d) < M,$$

quindi $w$ ha un massimo locale in $(a,d)$. □

**Esercizio ✎ 2.7.** Completare la dimostrazione considerando il caso $d < x^*$.

Cosa succede se aggiungiamo il termine di grado zero $q(x)u(x)$ all'equazione differenziale? In questo caso dobbiamo accontentarci di un risultato più debole: $-u'' - u = 0$ ammette la soluzione $u(x) = \sin x$, che ha massimi locali isolati. Questo esempio suggerisce che dobbiamo assumere $q(x) \ge 0$. Ma anche l'equazione $-u'' + u = 0$ ammette la soluzione $u(x) = -\cosh x$, che ha massimo locale $u(0) = -1$. Potremo quindi escludere solo i massimi locali non-negativi. Infatti, se

$$-u''(x) + p(x)u'(x) + q(x)u(x) < 0, \qquad q(x) > 0$$

e $x^*$ è un massimo locale di $u$, allora chiaramente $u(x^*) < u''(x^*)/q(x^*) \le 0$.

Per passare al caso della disuguaglianza non stretta, possiamo ripetere la dimostrazione della Proposizione 2.6 scegliendo, nella definizione di $z$, $\gamma > 0$ tale che $\gamma^2 - \gamma|p(x)| - q(x) > 0$ per ogni $x$. Otteniamo la seguente proposizione.

**Proposizione 2.8.** Sia $u \in C^2([a,b])$ una funzione che soddisfa la disuguaglianza

$$-u''(x) + p(x)u'(x) + q(x)u(x) \le 0 \qquad \text{in } (a,b), \tag{15}$$

con $p, q \in C^0([a,b])$ e $q \ge 0$. Se $u(x^*) = M := \max_{x\in[a,b]} u(x) \ge 0$ per un $x^* \in (a,b)$, allora $u(x) = M$ per ogni $x \in [a,b]$.

Notare che ora dobbiamo assumere $M \ge 0$.

**Esercizio ✎ 2.9.** Scrivere in dettaglio la dimostrazione della Proposizione 2.8.

Ne segue una forma più semplice del principio del massimo, che cercheremo di replicare nei metodi numerici che considereremo.

**Corollario 2.10.** Se $u \in C^2([a,b])$ soddisfa la disuguaglianza differenziale (15) con $q \ge 0$ e $u(a) \le 0$, $u(b) \le 0$, allora o *(i)* $u(x) = 0$ per ogni $x \in (a,b)$, o *(ii)* $u(x) < 0$ in $(a,b)$.

Da questo principio del massimo possiamo ricavare un risultato di unicità più forte di quello di Proposizione 2.5. Dal Corollario 2.10, se $u_0$ è soluzione del problema ai limiti omogeneo (14) segue che $u_0 \le 0$. Ma anche $-u_0$ soddisfa lo stesso problema quindi $-u_0 \le 0$, cioè $u_0 = 0$.

**Corollario 2.11.** Dati $p, q \in C^0([a,b])$ con $q \ge 0$, allora l'unica soluzione del problema omogeneo (14) è $u_0(x) = 0$.

### 2.2.3 Esistenza e unicità per il problema non omogeneo

La buona posizione del problema (13) segue dall'unicità della soluzione del problema omogeneo (14).

Come nella Nota 1.4, consideriamo i due problemi lineari ai valori iniziali:

$$\begin{cases} u_1''(x) = p(x)u_1'(x) + q(x)u_1(x) - f(x), \\ u_1(a) = \alpha, \\ u_1'(a) = 0, \end{cases} \qquad \begin{cases} u_2''(x) = p(x)u_2'(x) + q(x)u_2(x) \qquad x \in (a,b), \\ u_2(a) = 0, \\ u_2'(a) = 1. \end{cases}$$

Le soluzioni $u_1$ e $u_2$ esistono, sono uniche, e appartengono a $C^2([a,b])$, grazie alla teoria standard dei problemi di Cauchy. Se vale $u_2(b) \neq 0$, allora si vede facilmente che la combinazione lineare

$$u(x) = u_1(x) + \frac{\beta - u_1(b)}{u_2(b)} u_2(x).$$

soddisfa il problema al bordo (13), cioè abbiamo l'esistenza della soluzione. Come garantire $u_2(b) \neq 0$? Se così non fosse, cioè $u_2(b) = 0$, allora $u_2$ sarebbe una soluzione non nulla (poiché $u_2'(a) = 1$) del problema omogeneo (14). Se assumiamo che il problema omogeneo (14) ammette la sola soluzione nulla, allora vale $u_2(b) \neq 0$ e quindi esiste una soluzione $u$ (definita da $u = u_1 + \frac{\beta - u_1(b)}{u_2(b)} u_2$) del problema non omogeneo (13). Inoltre questa soluzione è unica: se così non fosse ed esistessero due soluzioni $u^{(1)} \neq u^{(2)}$ di (13), allora $u_0 = u^{(1)} - u^{(2)}$ sarebbe una soluzione non-nulla del problema omogeneo (14), che abbiamo assunto non essere possibile. Abbiamo ottenuto quanto segue.

**Lemma 2.12.** Se il problema omogeneo (14) ammette solo la soluzione nulla $u_0 = 0$ allora il problema non omogeneo (13) ammette una e una sola soluzione.

Combinando questo fatto con il Corollario 2.11, otteniamo il seguente teorema.

**Teorema 2.13.** Se $p, q, f \in C^0([a,b])$ e $q \geq 0$, allora il problema ai limiti (13) ammette una (e una sola) soluzione $u$.

Se $p, q, f \in C^1(a,b)$, dall'equazione differenziale abbiamo $u'' = pu' + qu - f \in C^1(a,b)$, cioè $u \in C^3(a,b)$. Ripetendo questa procedura (detta di *bootstrap*) abbiamo che, per ogni $k \in \mathbb{N}$, se $p, q, f \in C^k(a,b)$ allora $u \in C^{k+2}(a,b)$.

### 2.2.4 La funzione di Green

Dati $a, b, \alpha, \beta \in \mathbb{R}$ come sopra e $f \in C^0([a,b])$, definiamo per $x, y \in [a,b]$

$$u(x) := \frac{\alpha(b-x) + \beta(x-a)}{b-a} + \int_a^b G(x,y) f(y)\,\mathrm{d}y, \qquad G(x,y) := \begin{cases} \dfrac{(y-a)(b-x)}{b-a} & a \leq y \leq x \leq b, \\ \dfrac{(x-a)(b-y)}{b-a} & a \leq x \leq y \leq b. \end{cases} \tag{16}$$

Si vede immediatamente che $u(a) = \alpha$ e $u(b) = \beta$. Per $x \in (a,b)$, la derivata prima di $u$ è

$$\begin{aligned} u'(x) &= \frac{\beta - \alpha}{b-a} + \frac{\partial}{\partial x}\left( \int_a^x \frac{(y-a)(b-x)}{b-a} f(y)\,\mathrm{d}y + \int_x^b \frac{(x-a)(b-y)}{b-a} f(y)\,\mathrm{d}y \right) \\ &= \frac{\beta - \alpha}{b-a} + \frac{(x-a)(b-x)}{b-a} f(x) + \int_a^x \frac{a-y}{b-a} f(y)\,\mathrm{d}y - \frac{(x-a)(b-x)}{b-a} f(x) + \int_x^b \frac{b-y}{b-a} f(y)\,\mathrm{d}y. \end{aligned}$$

Derivando una seconda volta

$$u''(x) = \frac{\partial}{\partial x}\left( \int_a^x \frac{a-y}{b-a} f(y)\,\mathrm{d}y + \int_x^b \frac{b-y}{b-a} f(y)\,\mathrm{d}y \right) = \frac{a-x}{b-a} f(x) - \frac{b-x}{b-a} f(x) = -f(x).$$

Questo significa che $u$ definita in (16) è di classe $C^2$ ed è la soluzione del problema di Dirichlet con $p = q = 0$:

$$\begin{cases} -u''(x) = f(x) \quad x \in (a,b), \\ u(a) = \alpha, \\ u(b) = \beta. \end{cases} \tag{17}$$

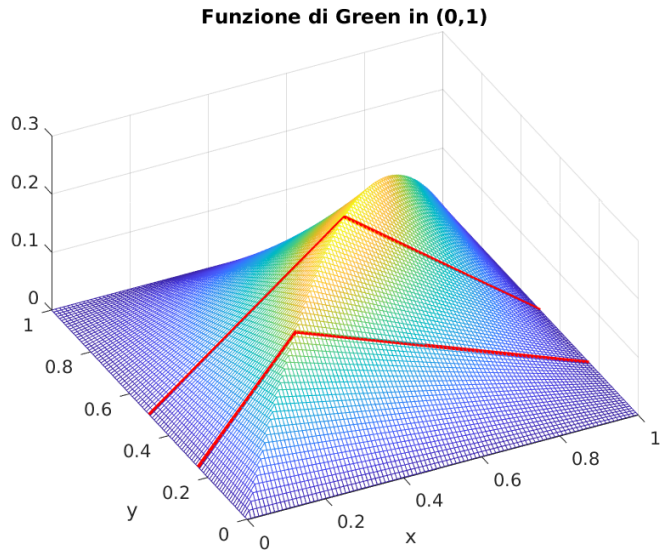



*Figura 6: La funzione di Green per $-u'' = f$ in $(0,1)$. In rosso $x \mapsto G(x,1/4)$ e $x \mapsto G(x,1/2)$.*

La funzione $G(x,y)$ è detta **funzione di Green** del problema al bordo. È una funzione continua in $[a,b]^2$, simmetrica (nel senso che $G(x,y) = G(y,x)$), affine a tratti in ciascuna variabile, non-negativa, e vale zero quando $x$ o $y$ sono uguali ad $a$ o $b$.

L'espressione (16) è un altro modo di dimostrare (costruttivamente) l'esistenza di una soluzione del problema ai limiti (17). Inoltre ci permette di dimostrare la dipendenza continua della soluzione dai dati del problema. Ricordiamo che la norma $L^\infty$ di una funzione $f$, definita sull'intervallo $(a,b)$ e continua, è definita come $\|f\|_{L^\infty(a,b)} := \sup_{x\in(a,b)} |f(x)|$.

**Proposizione 2.14** (Dipendenza continua dai dati)**.** Siano $a < b \in \mathbb{R}$, $\alpha, \beta \in \mathbb{R}$ e $f \in C^0([a,b])$. La soluzione del problema di Dirichlet (17) soddisfa

$$\|u\|_{L^\infty(a,b)} \leq \max\{|\alpha|, |\beta|\} + \frac{(b-a)^2}{8} \|f\|_{L^\infty(a,b)}\,. \tag{18}$$

*Dimostrazione.* È sufficiente usare la rappresentazione (16), $G \geq 0$ e calcolare l'integrale di $G(x,y)$ per $x$ fissato: per ogni $x \in (a,b)$,

$$\begin{aligned} |u(x)| &\leq \frac{b-x}{b-a}|\alpha| + \frac{x-a}{b-a}|\beta| + \int_a^b G(x,y)|f(y)|\,\mathrm{d}y \\ &\leq \max\{|\alpha|,|\beta|\} + \|f\|_{L^\infty(a,b)} \int_a^b G(x,y)\,\mathrm{d}y \qquad = \max\{|\alpha|,|\beta|\} + \|f\|_{L^\infty(a,b)} \frac{(b-x)(x-a)}{2}. \end{aligned}$$

□

**Esercizio ✎ 2.15.** Completare i passaggi mancanti nella dimostrazione della Proposizione 2.14.

Quando sappiamo maggiorare la norma della soluzione di un problema lineare con la norma dei dati del problema, allora segue che a piccole perturbazioni dei dati corrispondono piccole perturbazioni delle soluzioni dei corrispondenti problemi al bordo, come vediamo nell'esercizio seguente. Questo è il motivo per cui chiamiamo "dipendenza continua dai dati" una stima come quella della Proposizione 2.14.

**Esercizio ✎ 2.16.** Siano $u_1$ e $u_2$ soluzioni del problema al bordo (17) con dati $f_1, \alpha_1, \beta_1$ e $f_2, \alpha_2, \beta_2$, rispettivamente. Dimostrare che la differenza $u_1 - u_2$ è controllata (cioè una sua norma è maggiorata) dalla differenza dei dati dei due problemi.

**Esercizio ✎ 2.17.** Usare (16) per mostrare che la soluzione del problema (17) con $\alpha = \beta = 0$ soddisfa $u \geq 0$ se $f \geq 0$ e $u \leq 0$ se $f \leq 0$. (Caso particolare del Corollario 2.10.)

(Questo può essere una prima giustificazione per il misterioso segno meno davanti a $u''$: l'operazione $-\frac{\partial^2}{\partial x^2}$ "preserva il segno", come la moltiplicazione per un numero positivo.)

**Esercizio ✎ 2.18.**

- Fissiamo $y \in (a,b)$ e definiamo $g_y : [a,b] \to \mathbb{R}$ come $g_y(x) := G(x,y)$. Mostrare che la funzione $g_y$: (i) è continua, (ii) vale zero in $a$ e in $b$, (iii) soddisfa $-g_y'' = 0$ nei due intervalli $(a,y)$ e $(y,b)$, e (iv) la sua derivata prima ha salto $\lim_{\epsilon \searrow 0}[g_y'(x-\epsilon) - g_y'(x+\epsilon)] = 1$ nel punto $x = y$.
- Mostrare che $g_y$ è l'*unica* funzione che soddisfa le condizioni (i)–(iv).
- Mostrare che la funzione $u$ definita come in (16) soddisfa il problema al bordo (17) senza usare la formula esplicita di $G(x,y)$ ma (ri)definendo $G(x,y) := g_y(x)$ e usando solo le proprietà (i)–(iv).

L'unicità si può dimostrare o usando le proprietà elementari implicate dalle condizioni (i)–(iv), oppure (più interessante) usando il metodo dell'energia e passando per un opportuno "problema omogeneo" (pensare attentamente a qual è il "dato" che va messo a zero nel caso omogeneo). Questo potrebbe addirittura suggerire una definizione della funzione di Green per problemi al bordo più generali, come (13).

**Nota 2.19.** La rappresentazione di Green (16) può essere usata per calcolare la soluzione del problema al bordo (17). Se l'integrale di $G(x,y)f(y)$ non è calcolabile analiticamente, questo può essere approssimato con una formula di quadratura. Se invece l'equazione differenziale contiene coefficienti $p$ e $q$ dipendenti da $x$ come in (13), la funzione di Green corrispondente non è in generale calcolabile esplicitamente. Nelle prossime sezioni studieremo metodi numerici che possono essere immediatamente applicati a equazioni più generali di $-u'' = f$.

Una derivazione dell'espressione (16) leggermente diversa si può trovare in [TW05, §2.1], [LeVeque07, §2.11] o [QSSG14, §11.1].

### 2.2.5 Altre condizioni al bordo

Consideriamo il problema di diffusione–reazione con **condizioni al bordo di Neumann**, cioè imponiamo i valori di $u'$ agli estremi dell'intervallo $(a,b)$:

$$\begin{cases} -u''(x) + q(x)u(x) = f(x) & x \in (a,b), \\ u'(a) = \alpha, \\ u'(b) = \beta, \end{cases} \tag{19}$$

Se $q(x) = 0$ per ogni $x \in (a,b)$ vediamo che deve valere una condizione di compatibilità sui dati $f, \alpha, \beta$:

$$\int_a^b f(x)\,\mathrm{d}x = \int_a^b -u''(x)\,\mathrm{d}x = u'(a) - u'(b) = \alpha - \beta.$$

Se questa condizione non è verificata dai dati $f$, $\alpha$ e $\beta$, il problema al bordo non ha soluzione. Sempre con $q = 0$, il problema omogeneo $\alpha = \beta = f = 0$ ammette come soluzioni tutte le funzioni costanti $u_0(x) = c$. Similmente, per ogni soluzione $u$ del problema non-omogeneo, $u_c(x) = u(x) + c$ è soluzione dello stesso problema. In breve, se $q = 0$ possono verificarsi due situazioni: se la condizione di compatibilità è soddisfatta allora la soluzione del problema al bordo (19) non è unica, altrimenti non esiste soluzione.

Le cose sono più semplici se $q > 0$ almeno in parte del dominio. Usiamo metodo dell'energia di §2.2.1 per il problema con condizioni al bordo di Dirichlet. Applicato al problema (19), nel caso omogeneo $\alpha = \beta = f = 0$, ci dà

$$0 = \int_a^b \Big( -u_0''(x) + q(x)u_0(x)\Big)u_0(x)\,\mathrm{d}x = \int_a^b \Big(\big(u_0'(x)\big)^2 + q(x)u_0^2(x)\Big)\,\mathrm{d}x - u_0(b)\underbrace{u_0'(b)}_{=0} + u_0(a)\underbrace{u_0'(a)}_{=0}.$$

Se $q(x) \geq 0$ in $(a,b)$ e $q(x) > 0$ in un sotto-intervallo $(c,d) \subset (a,b)$, abbiamo $u_0(x) = 0$ in $(c,d)$ e $u_0'(x) = 0$ in $(a,b)$, da cui segue $u_0(x) = 0$ in tutto $(a,b)$.

**Esercizio ✎ 2.20.** Mostrare che sotto queste ipotesi su $q$, l'esistenza della soluzione di (19) segue dall'unicità. Seguire la dimostrazione usata nel caso del problema di Dirichlet in §2.2.3.

**Proposizione 2.21.** Se $q \in C^0([a,b])$, $q(x) \geq 0$ in $(a,b)$ e $q(x) > 0$ in qualche $(c,d) \subset (a,b)$, allora il problema di Neumann (19) ammette una e una sola soluzione $u$.

**Esercizio ✎ 2.22.** Mostrare che il problema al contorno (19) con $q = 0$ e $\int_a^b f(x)\,\mathrm{d}x = \beta - \alpha$ ammette un'unica soluzione a media nulla, cioè tale $\int_a^b u(x)\,\mathrm{d}x = 0$.

Le condizioni di Dirichlet prescrivono il valore di $u$ sul bordo, ad esempio il valore della densità o della temperatura; quelle di Neumann prescrivono il valore del flusso della quantità trasportata. Condizioni di Neumann omogenee significano che non c'è scambio di materia o temperatura con l'esterno, rappresentano un dominio "isolato".

Un terzo tipo di condizioni al bordo sono quelle di **Robin** o di impedenza: dati i parametri $\theta_a$, $\theta_b \in \mathbb{R}$,

$$-u'(a) + \vartheta_a u(a) = \alpha, \qquad u'(b) + \vartheta_b u(b) = \beta.$$

**Esercizio ✎ 2.23** (Condizioni di Robin)**.** Mostrare che se $\vartheta_a, \vartheta_b > 0$ e $q \geq 0$, l'equazione $-u'' + qu = f$ con condizioni al bordo di Robin omogenee ($\alpha = \beta = 0$) ammette un'unica soluzione.

Un'ulteriore classe di condizioni al bordo è costituita da quelle **periodiche**:

$$u(a) = u(b), \qquad u'(a) = u'(b). \tag{20}$$

**Esercizio ✎ 2.24** (Condizioni al bordo periodiche)**.**

- Mostrare che se $\int_a^b f(x)\,\mathrm{d}x \neq 0$, l'equazione $-u'' = f$ con condizioni al bordo periodiche (20) non ammette nessuna soluzione.
- Mostrare che se $\int_a^b f(x)\,\mathrm{d}x = 0$, l'equazione $-u'' = f$ con condizioni al bordo periodiche ammette infinite soluzioni, di cui una sola a media nulla.

  Suggerimento: si può costruire esplicitamente $u$ usando le primitive di $f$.
- Usare il metodo dell'energia per dimostrare l'esistenza e l'unicità della soluzione $u$ del problema $-u''+qu = f$ con $q > 0$ e condizioni al bordo periodiche.

# 3 Differenziazione numerica: le differenze finite

## 3.1 Differenze finite ed errore di troncamento

Nei precedenti corsi di analisi numerica sono state studiate le formule di quadratura, cioè le formule che permettono di approssimare numericamente l'integrale definito di una funzione. Per risolvere numericamente un problema che coinvolge un'equazione differenziale però è utile essere in grado di approssimare anche le derivate di una funzione. La tecnica più semplice è quella delle **differenze finite**.

Data una funzione $f(x)$, reale di variabile reale e differenziabile con continuità, la sua derivata prima in $x$ è il limite del rapporto incrementale: $f'(x) = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h}$. Le differenze finite sono approssimazioni di questo limite per $h$ finito. Fissato un numero $h > 0$, le **differenze finite in avanti e all'indietro** sono

$$\boxed{D_h^+ f(x) := \frac{f(x+h) - f(x)}{h},} \qquad \boxed{D_h^- f(x) := \frac{f(x) - f(x-h)}{h}.}$$

Per quantificare l'errore di queste approssimazioni usiamo l'espansione di Taylor di $f$ in $x$. Se $f$ è di classe $C^2$ nell'intervallo $[x, x+h]$, vale

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2} f''(\xi) \qquad \text{per qualche } \xi \in (x, x+h), \quad \text{quindi}$$

$$f'(x) - D_h^+ f(x) = \frac{f(x+h) - f(x) - \frac{h^2}{2} f''(\xi)}{h} - \frac{f(x+h) - f(x)}{h} = -\frac{h}{2} f''(\xi) \qquad \xi \in (x, x+h),$$

cioè l'errore di troncamento commesso dalle differenze finite in avanti converge linearmente in $h$. L'errore delle differenze finite all'indietro si comporta in modo simile.

È possibile costruire differenze finite con ordini di convergenza in $h$ più alti. Prendendo la media tra differenze in avanti e all'indietro otteniamo la **differenza finita centrata**:

$$\boxed{D_h^C f(x) := \frac{f(x+h) - f(x-h)}{2h}} = \frac{1}{2}\big(D_h^+ f(x) + D_h^- f(x)\big).$$

Per calcolare l'errore usiamo un termine in più nell'espansione di Taylor e il teorema dei valori intermedi:

$$f(x \pm h) = f(x) \pm hf'(x) + \frac{h^2}{2} f''(x) \pm \frac{h^3}{6} f'''(\xi_\pm) \quad \text{per } \xi_+ \in (x, x+h), \xi_- \in (x-h, x), \quad \text{quindi}$$

$$f'(x) - D_h^C f(x) = f'(x) - \frac{2hf'(x) + \frac{h^3}{6}f'''(\xi_+) + \frac{h^3}{6}f'''(\xi_-)}{2h} = -\frac{h^2}{6}f'''(\xi) \qquad \xi \in (x-h, x+h). \quad (21)$$

Le differenze finite centrate approssimano $f'(x)$ con ordine 2 in $h$, se $f \in C^3([x-h, x+h])$. Differenze finite di ordine arbitrario possono essere ottenute coinvolgendo il valore di $f$ in più punti, vedere ad esempio l'Esercizio 3.5.

La derivata seconda di $f$ si può approssimare con ordine 2 in $h$ usando la differenza finita centrata

$$\boxed{D_h^{2C} f(x) := \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}.}$$

**Esercizio ✎ 3.1.** Le approssimazioni $D_h^+ f(x), D_h^- f(x), D_h^C f(x)$ di $f'(x)$ corrispondono al coefficiente angolare di tre diverse rette tangenti al grafico di $f$, quali? Come interpretare in modo geometrico $D_h^{2C} f(x)$?

**Esercizio ✎ 3.2.** Mostrare che l'operatore delle differenze finite centrate del secondo ordine può essere scritto come composizione degli operatori delle differenze finite del primo ordine nei modi seguenti:

$$D_h^{2C} = D_{h/2}^C D_{h/2}^C = D_h^+ D_h^- = D_h^- D_h^+.$$

**Esercizio ✎ 3.3.** Mostrare che le differenze finite centrate del primo e del secondo ordine $D_h^C f$ e $D_h^{2C} f$ sono esatte se $f$ è un polinomio di grado minore o uguale a 2 e a 3, rispettivamente.

**Esercizio ✎ 3.4.** Data $f$ di classe $C^4$ in $[x-h, x+h]$, mostrare che esiste $\xi \in (x-h, x+h)$ tale che

$$\boxed{f''(x) - D_h^{2C} f(x) = -\frac{h^2}{12} f^{(iv)}(\xi).} \quad (22)$$

**Esercizio ✎ 3.5.** Le formule di quadratura sono spesso derivate usando gli integrali esatti di particolari interpolanti della funzione integranda. Similmente, le differenze finite possono essere interpretate come derivate di particolari interpolanti. Per aumentare l'ordine in $h$ dell'errore di troncamento è necessario aumentare l'accuratezza dell'interpolazione e quindi estendere lo *stencil*, cioè l'insieme dei punti in cui viene valutata $f$.

- Mostrare che $D_h^C f(0) = P_2'(0)$, dove $P_2$ è il polinomio di grado minore o uguale a 2 che interpola $f$ nei tre punti $-h, 0, h$.
- Definire
$$D_h^* f(x) := \frac{1}{h}\left(\frac{1}{12}f(x-2h) - \frac{2}{3}f(x-h) + \frac{2}{3}f(x+h) - \frac{1}{12}f(x+2h)\right)$$
e mostrare che $D_h^* f(0) = P_4'(0)$ dove $P_4$ è il polinomio di grado minore o uguale a 4 che interpola $f$ nei cinque punti $-2h, -h, 0, h, 2h$.

  Suggerimento: usare la formula d'interpolazione di Lagrange $P(x) = \sum_j f(x_j) \prod_{k \neq j} \frac{x - x_k}{x_j - x_k}$. Da questa formula si può scrivere $P_4'(0)$ senza calcolare esplicitamente $P_4'(x)$.
- Calcolare l'errore di troncamento di $D_h^* f(x)$ procedendo come in (21).

## 3.2 Errore di arrotondamento

Quando usiamo una formula di quadratura per approssimare un integrale, possiamo migliorarne l'accuratezza dell'approssimazione aumentando il numero di punti di quadratura, quindi aumentando il costo computazionale dell'approssimazione. Le differenze finite invece coinvolgono lo stesso numero di valutazioni della funzione $f$ da differenziare qualunque sia la scelta di $h$. Possiamo quindi scegliere $h$ piccolo a piacere per calcolare un'approssimazione di precisione arbitraria senza pagare di più? La risposta è "sì" se operiamo in aritmetica esatta, e "no" se operiamo in aritmetica floating-point, come accade normalmente quando usiamo un computer. Il motivo è che oltre all'errore di **troncamento** descritto precedentemente, anche l'errore di **arrotondamento** (*roundoff*) dovuto all'uso di aritmetica floating-point gioca un ruolo importante. Consideriamo il caso più semplice, quello della differenza finita in avanti $D_h^+ f(x) = \frac{f(x+h) - f(x)}{h}$. Se $h$ è molto piccolo questo rapporto tende a $\frac{0}{0}$, che sappiamo essere indefinito. Il numeratore $f(x+h) - f(x)$ viene calcolato in modo molto impreciso dal computer per il fenomeno della **cancellazione**, avendo $f(x+h)$ e $f(x)$ valori simili, e viene moltiplicato per $\frac{1}{h}$, che è un numero grande. Quindi l'errore di cancellazione dovuto all'arrotondamento è amplificato da questo fattore $\frac{1}{h}$, portando a risultati numerici che possono essere completamente sbagliati.

L'Esercizio 3.6 e la Figura 7 mostrano la situazione tipica. L'errore delle differenze finite del primo ordine in avanti (centrate) decresce come $h$ ($h^2$, rispettivamente) per valori non troppo piccoli di $h$. Denotando $\epsilon_M$ la precisione macchina, per $h \lesssim \epsilon_M^{1/2}$ ($h \lesssim \epsilon_M^{1/3}$, rispettivamente) l'errore di arrotondamento domina su quello di troncamento e l'errore complessivo aumenta diminuendo $h$. L'errore è minimo quando $h$ è di ordine circa $\epsilon_M^{1/2} \sim 10^{-8}$ per le differenze in avanti e $\epsilon_M^{1/3} \sim 10^{-5}$ per le differenze centrate (assumendo che tutte le altre quantità abbiano ordine $\sim 1$). Questo significa anche che l'operazione di differenziazione è numericamente instabile, al contrario di quella di integrazione. Le differenze finite centrate permettono di avere un'accuratezza migliore di quelle in avanti, grazie al maggiore ordine di convergenza.

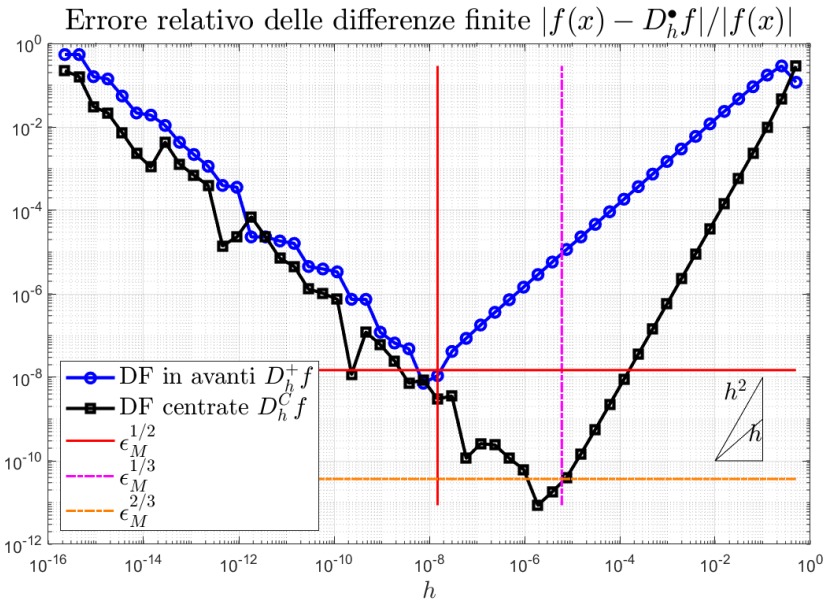



*Figura 7: L'errore commesso dalle differenze finite in avanti e centrate per la funzione $f(x) = \cos x - \sin \mathrm{e}^x$ in $x = 1$ al variare di $h$. (Esercizio 3.6.)*

**Esercizio 💻 3.6.** Calcolare l'errore relativo commesso dalle differenze finite in avanti $D_h^+$ e centrate $D_h^C$ per il calcolo della derivata di $f(x) = \cos x - \sin \mathrm{e}^x$ in $x = 1$ per $h = 2^{-1}, 2^{-2}, \ldots, 2^{-52} = \epsilon_M$. Mostrare i risultati in un grafico in scala logaritmica, cosa si osserva? Per quali valori di $h$ l'errore è dominato dall'errore di troncamento e per quali dall'errore di arrotondamento? Stimare numericamente gli ordini di convergenza per i valori di $h$ opportuni.[1]

(Attenzione a uno degli errori più comuni in Matlab: $\frac{X}{2h}$ si scrive `X/(2*h)` e non `X/2*h`!)

**Esercizio 💻 3.7.** Ripetere l'Esercizio 3.6 per la differenza finita $D_h^*$ definita nell'Esercizio 3.5. Verificare numericamente l'ordine di convergenza dell'errore di troncamento calcolato nell'Esercizio 3.5.

**Esercizio 💻 3.8.** Un modo curioso per calcolare la derivata di una funzione senza roundoff è la "*complex step differentiation*" introdotta da Squire e Trapp nel 1998. Se la funzione (reale) $f$ di classe $C^\infty$ ammette un'estensione olomorfa in un intorno complesso di $x \in \mathbb{R}$, allora $D_h^i f(x) := \mathrm{Im}\{f(x + \mathrm{i}h)\}/h$ è un'approssimazione di $f'(x)$, dove Im denota la parte immaginaria e i è l'unità immaginaria. Prendendo la parte immaginaria dell'espansione di Taylor (per l'incremento puramente immaginario $\mathrm{i}h$) $f(x + \mathrm{i}h) = f(x) + \mathrm{i}hf'(x) - \frac{1}{2}h^2 f''(x) - \frac{1}{6}\mathrm{i}h^3 f'''(\xi)$, $x \in \mathbb{R}$, si verifica che l'ordine di convergenza è quadratico[2]. La

[1] Dati due vettori `x`, `y` di uguale lunghezza, il comando `P = polyfit(x,y,n)` calcola i coefficienti del polinomio $P_1 x^n + P_2 x^{n-1} + \cdots + P_n x + P_{n+1}$ di grado $n$ che meglio approssima i dati $y$ nel senso dei minimi quadrati. Se abbiamo due vettori `h` e `v` i cui valori (positivi) sono legati da una dipendenza algebrica del tipo $v_j \approx C h_j^p$, i loro logaritmi sono legati da una relazione lineare: $\log v_j \approx p \log h_j + \log C$. Quindi il comando `P = polyfit(log(h),log(v),1)` permette di stimare i valori di $p$ e $C$. Possiamo usare questo comando ogni volta che vogliamo stimare l'ordine $p$ di convergenza algebrica di una serie di dati, ad esempio gli errori commessi da un metodo numerico.

[2] Verificando che l'errore è quadratico ci si accorge che serve $\mathrm{Im}\{f''(x)\} = 0$. Qui stiamo usando l'espansione di Taylor nel piano complesso, quindi $f'$, $f''$, $f'''$ rappresentano derivate rispetto alla variabile complessa $z = x + \mathrm{i}y$: $f' = \partial_z f = \frac{1}{2}(\partial_x - \mathrm{i}\partial_y) f$. Poiché abbiamo assunto che $f(z) \in \mathbb{R}$ se $z = x \in \mathbb{R}$, separando parte reale e immaginaria di $f$ come $f = u + \mathrm{i}v$ abbiamo $\partial_z f = \frac{1}{2}(\partial_x - \mathrm{i}\partial_y) f = \frac{1}{2}(\partial_x u + \mathrm{i}\partial_x v - \mathrm{i}\partial_y u + \partial_y v) = \partial_x u + \mathrm{i}\partial_x v$, grazie alle condizioni di Cauchy–Riemann ($\partial_x u = \partial_y v$ e $\partial_y u = -\partial_x v$). In $z = x$, abbiamo $\partial_x v = 0$ poiché $v = 0$ per ogni argomento reale, quindi $f' = \partial_z f = \partial_x u$, cioè la derivata complessa in questo punto coincide con quella reale (ed è quindi reale). Lo stesso succede per la derivata seconda quindi effettivamente l'errore è quadratico $D_h^i f(x) - f'(x) = \mathrm{Im}\{f(x+\mathrm{i}h)\}/h - f'(x) = \mathrm{Im}\{f(x+\mathrm{i}h) - \mathrm{i}hf'(x)\}/h = \mathrm{Im}\{f(x) + \frac{1}{2}h^2 f''(x) - \mathcal{O}(h^3)\}/h = \mathcal{O}(h^2)$.

cosa sorprendente è che non c'è errore di cancellazione, poiché non sottraiamo valori vicini. Quindi la scelta $h \lesssim \epsilon_M^{1/2}$ garantisce un'accuratezza di ordine pari all'errore macchina.

Verificare questo fatto ripetendo l'Esercizio 3.6 per $D_h^i f$. Per altri dettagli vedere
https://blogs.mathworks.com/cleve/2013/10/14/complex-step-differentiation/

**Esercizio 💻 3.9.** Ripetere gli esercizi precedenti per $f(x) = \sin(kx)$ e osservare cosa succede al variare di $k \in \mathbb{R}$. Come si spiegano i risultati per $k$ molto grande?

**Nota 3.10.** Nell'Esercizio 3.5 abbiamo visto che le differenze finite per l'approssimazione di $f'(x)$ possono essere interpretate come derivate di particolari interpolanti "locali", cioè che coinvolgono i valori di $f$ in alcuni punti vicino a $x$. Questo suggerisce un modo più generale e meno sensibile al roundoff per approssimare la derivata di una funzione: la derivata pseudo-spettrale. La funzione $f$ viene interpolata con un polinomio $P$ "globale" (o con un elemento di in un altro spazio finito-dimensionale di funzioni, ad esempio di funzioni trigonometriche), e $f'$ viene approssimata da $P'$, che è un altro polinomio calcolato algebricamente da $P$. Usando ad esempio l'interpolazione nei nodi di Chebyshev e le proprietà dei polinomi di Chebyshev, la derivata nei nodi può essere calcolata con un semplice prodotto matrice–vettore. L'accuratezza di questo metodo dipende da quella dell'interpolazione sottostante, quindi per funzioni analitiche l'ordine di convergenza è esponenziale. Per altri dettagli vedere [QSSG14, §9.10.3].

# 4 Il metodo delle differenze finite in una dimensione

Un'ottima descrizione del metodo delle differenze finite si trova in [LeVeque07]; la parte più rilevante per questo corso è nel capitolo 2.

## 4.1 Il metodo delle differenze finite per il problema di Dirichlet

Consideriamo il problema ai limiti lineare (di diffusione–reazione)

$$\begin{cases} -u''(x) + q(x)u(x) = f(x) & x \in (a,b), \\ u(a) = \alpha, \\ u(b) = \beta, \end{cases} \tag{23}$$

dove $q, f \in C^0([a,b])$ e $a < b, \alpha, \beta \in \mathbb{R}$. Le condizioni al bordo di questo tipo sono dette di Dirichlet. Introduciamo una griglia (*mesh*) di punti equispaziati

$$a = x_0 < x_1 < \cdots < x_n < x_{n+1} = b, \qquad x_j = a + jh, \qquad h = \frac{b-a}{n+1}, \qquad n \in \mathbb{N}.$$

Vogliamo approssimare il valore di $u$ nei nodi $x_j$ con un'approssimazione numerica denotata $U_j \approx u(x_j)$, $j = 0, \dots, n+1$. Sostituendo $u''$ con la differenza finita centrata del secondo ordine $D_h^{2C} u$ e passo $h$, e definendo $q_j := q(x_j)$, $f_j := f(x_j)$, otteniamo

$$\begin{cases} -\dfrac{U_{j+1} - 2U_j + U_{j-1}}{h^2} + q_j U_j = f_j & j = 1, \dots, n, \\ U_0 = \alpha, \\ U_{n+1} = \beta. \end{cases}$$

Questo è un sistema lineare di $n$ equazioni nell'incognita $\vec{\mathbf{U}} \in \mathbb{R}^n$:

$$\underline{\underline{\mathbf{A}}}\vec{\mathbf{U}} = \vec{\mathbf{B}}, \quad \text{dove} \tag{24}$$

$$\underline{\underline{\mathbf{A}}} = \frac{1}{h^2} \begin{pmatrix} 2+q_1h^2 & -1 & & & & \\ -1 & 2+q_2h^2 & -1 & & & \\ & -1 & 2+q_3h^2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2+q_nh^2 \end{pmatrix}, \quad \vec{\mathbf{B}} = \begin{pmatrix} f_1 + \frac{\alpha}{h^2} \\ f_2 \\ f_3 \\ \vdots \\ f_n + \frac{\beta}{h^2} \end{pmatrix}, \quad \vec{\mathbf{U}} = \begin{pmatrix} U_1 \\ U_2 \\ U_3 \\ \vdots \\ U_n \end{pmatrix}.$$

I termini non scritti nella matrice $\underline{\underline{\mathbf{A}}}$ sono uguali a zero. Notiamo che i "dati" del problema al contorno $f, \alpha, \beta$ entrano nel sistema lineare solo nel vettore $\vec{\mathbf{B}}$, mentre il coefficiente $q$ entra nella matrice $\underline{\underline{\mathbf{A}}}$. Risolvendo questo sistema lineare abbiamo un'approssimazione della soluzione del problema (23).

Per questo sistema lineare ci poniamo diverse domande:

- Esiste una soluzione? È unica? Equivalentemente, $\underline{\underline{\mathbf{A}}}$ è invertibile?
- La soluzione discreta $\vec{\mathbf{U}}$ converge a quella continua $u$ per $n \to \infty$?
- Qual è la velocità di convergenza?
- Come calcolare la soluzione in modo efficiente?
- Come estendere il metodo a problemi più generali?
  (Problemi con termini di trasporto, con altre condizioni al bordo, non lineari...)

Ricordiamo l'esempio nella Nota 1.1: se $q < 0$ il problema ai limiti (23) può non essere ben posto. Assumeremo quindi che $q \geq 0$ in $[a, b]$.

**Esercizio 💻 4.1.**

- Implementare il metodo delle differenze finite per il problema (23) con dati a piacere. Ad esempio:
  - $(a, b) = (-1, 1)$, $f(x) = 0$, $q(x) = 1$, $\alpha = \beta = 1$ (calcolare a mano $u$);
  - $(a, b) = (-1, 1)$, $f(x) = 1$, $q(x) = Q^2$ per $Q > 0$ costante, $\alpha = \beta = 0$, $u(x) = \frac{1}{Q^2}(1 - \frac{\cosh Qx}{\cosh Q})$;
  - $(a, b) = (-1, 1)$, $f(x) = 2\mathrm{e}^{-x^2} - \frac{4}{\mathrm{e}}x^2$, $q(x) = 4x^2$, $\alpha = \beta = 0$, $u(x) = \mathrm{e}^{-x^2} - \mathrm{e}^{-1}$;
  - $(a, b) = (0, 1)$, $f(x) = \frac{-1}{(1+x)^3}$, $q(x) = \frac{1}{(1+x)^2}$, $\alpha = 1$, $\beta = \frac{1}{2}$, $u = \frac{1}{1+x}$;
  - $a = 0$, $b > 0$, $f(x) = 0$, $q(x) = 1$, $\alpha = \beta = 1$, $u(x) = \frac{\sinh x + \sinh(b-x)}{\sinh b}$.
    (Come varia la soluzione al crescere della lunghezza $b$? Confrontare $u$ con la soluzione di un problema ai valori iniziali per la stessa equazione.)

  Plottare la soluzione discreta $\vec{\mathbf{U}}$ e quella continua $u$ come in Figura 8.

  Importante: la matrice $\underline{\underline{\mathbf{A}}}$ è sparsa, non assemblarla come una matrice densa!
- Implementare lo stesso metodo senza usare cicli (`for`/`while`) ma sfruttando le operazioni in forma vettoriale.

  Suggerimento: sfruttare il comando `spdiags`.

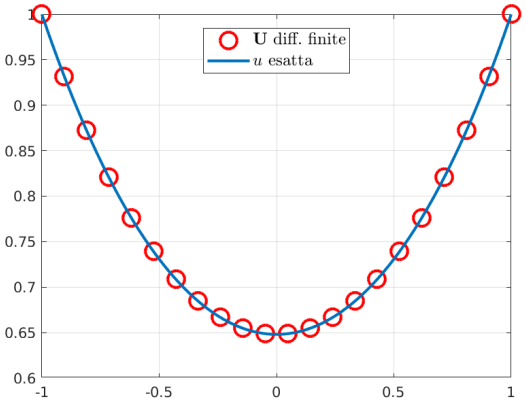



*Figura 8: La soluzione del problema al bordo (23) con $(a, b) = (-1, 1)$, $q(x) = 1$, $f(x) = 0$, $\alpha = \beta = 1$ e la sua approssimazione con il metodo delle differenze finite per $n = 20$ (Esercizio 4.1).*

## 4.2 Invertibilità della matrice delle differenze finite

L'esistenza e l'unicità del vettore $\vec{\mathbf{U}}$, soluzione del metodo delle differenze finite, seguono immediatamente se dimostriamo l'invertibilità della matrice $\underline{\underline{\mathbf{A}}}$. Cerchiamo di seguire la dimostrazione dell'esistenza e unicità della soluzione del problema al contorno usando il principio del massimo come in §2.2.2, in particolare nel Corollario 2.10. Sappiamo che, se $q \geq 0$,

$$-u'' + qu \leq 0, \quad u(a) \leq 0, \quad u(b) \leq 0 \qquad \Rightarrow \qquad u \leq 0.$$

La versione discreta di questa implicazione è la seguente.

**Lemma 4.2** (Principio del massimo discreto)**.** Dati $U_0, \dots, U_{n+1}$ e $c_1, \dots, c_n \geq 0$ per $n \in \mathbb{N}$,

$$-U_{j+1}+(2+c_j)U_j-U_{j-1} \leq 0 \quad \text{per } 1 \leq j \leq n, \quad U_0 \leq 0, \quad U_{n+1} \leq 0 \qquad \Rightarrow \qquad U_j \leq 0 \quad \text{per } 1 \leq j \leq n.$$

*Dimostrazione.* Denotiamo $j^* \in \{0, \dots, n+1\}$ l'indice tale che $U_{j^*} = \max\{U_0, \dots, U_{n+1}\}$. Assumiamo per assurdo che $U_{j^*} > 0$, quindi in particolare $1 \leq j^* \leq n$. Abbiamo

$$0 < U_{j^*} \leq \frac{U_{j^*+1} + U_{j^*-1}}{2 + c_{j^*}} \leq \frac{U_{j^*+1} + U_{j^*-1}}{2} \quad \Rightarrow \quad \text{vale (almeno) una delle due: } \begin{array}{l} U_{j^*} \leq U_{j^*-1} \text{ oppure} \\ U_{j^*} \leq U_{j^*+1}. \end{array}$$

(Attenzione: la disuguaglianza in cui sfruttiamo $c_j \geq 0$ vale perché sappiamo che il numeratore è positivo.) Assumiamo senza perdita di generalità che $U_{j^*} \leq U_{j^*-1}$; per la definizione di $j^*$ abbiamo $U_{j^*} = U_{j^*-1}$. Vale anche $U_{j^*+1} \geq 2U_{j^*} - U_{j^*-1} = U_{j^*} \geq U_{j^*+1}$, cioè $U_{j^*+1} = U_{j^*}$. Ripetendo l'operazione per $U_{j^*\pm 1}$ e procedendo verso gli estremi dell'intervallo troviamo $U_1 = U_2 = \ldots = U_n$. La stessa disuguaglianza per $j = 1$ dà $0 < U_1 \leq (U_0 + U_2)/2 \leq U_2/2 = U_1/2$ che è una contraddizione. □

Definendo $c_j := q_j h^2 \geq 0$ l'espressione $-U_{j+1} + (2 + c_j)U_j - U_{j-1}$ è l'elemento $j$esimo del vettore $h^2 \underline{\underline{\mathbf{A}}}\vec{\mathbf{U}}$. Il principio del massimo discreto dice che se $\vec{\mathbf{U}} \in \mathbb{R}^n$ soddisfa $(\underline{\underline{\mathbf{A}}}\vec{\mathbf{U}})_j \leq 0$ per $j = 1, \dots, n$ (definendo $U_0 = U_{n+1} = 0$) allora vale $\vec{\mathbf{U}}_j \leq 0$. Questo si può scrivere in forma matriciale come

$$\underline{\underline{\mathbf{A}}}\vec{\mathbf{U}} \preceq \vec{\mathbf{0}} \qquad \Rightarrow \qquad \vec{\mathbf{U}} \preceq \vec{\mathbf{0}},$$

dove per due vettori $\vec{\mathbf{v}}, \vec{\mathbf{w}} \in \mathbb{R}^n$ scriviamo che $\vec{\mathbf{v}} \preceq \vec{\mathbf{w}}$ se $v_j \leq w_j$ per ogni $j = 1, \dots, n$. Useremo la relazione d'ordine $\preceq$ anche per confrontare matrici.

**Definizione 4.3.** Una matrice quadrata $\underline{\underline{\mathbf{M}}} \in \mathbb{R}^{n \times n}$ tale che $\underline{\underline{\mathbf{M}}}\vec{\mathbf{v}} \preceq \vec{\mathbf{0}} \Rightarrow \mathbf{v} \preceq \vec{\mathbf{0}}$ è detta **matrice monotona**.

Ricordiamo che le matrici invertibili possono essere definite come quelle per cui $\underline{\underline{\mathbf{M}}}\vec{\mathbf{v}} = \vec{\mathbf{0}} \Rightarrow \mathbf{v} = \vec{\mathbf{0}}$. Elenchiamo alcune proprietà delle matrici monotone che useremo in seguito.

**Proposizione 4.4** (Proprietà delle matrici monotone)**.**

(i) Le matrici monotone sono invertibili.

(ii) Gli elementi dell'inversa di una matrice monotona sono non-negativi.

(iii) Se $\underline{\underline{\mathbf{M}}}$ e $\underline{\underline{\mathbf{N}}}$ sono matrici monotone e $\underline{\underline{\mathbf{M}}} \preceq \underline{\underline{\mathbf{N}}}$, allora $\underline{\underline{\mathbf{N}}}^{-1} \preceq \underline{\underline{\mathbf{M}}}^{-1}$.

*Dimostrazione.* (i) Data una matrice monotona $\underline{\underline{\mathbf{M}}}$ e un vettore $\vec{\mathbf{v}}$ tale che $\underline{\underline{\mathbf{M}}}\vec{\mathbf{v}} = \vec{\mathbf{0}}$, abbiamo $\underline{\underline{\mathbf{M}}}\vec{\mathbf{v}} \preceq \vec{\mathbf{0}}$ e $-\underline{\underline{\mathbf{M}}}\vec{\mathbf{v}} \preceq \vec{\mathbf{0}}$ quindi $\vec{\mathbf{v}} \preceq \vec{\mathbf{0}}$ e $-\vec{\mathbf{v}} \preceq \vec{\mathbf{0}}$, cioè $\vec{\mathbf{v}} = \vec{\mathbf{0}}$.

(ii) Se $\vec{\mathbf{c}}$ è la $j$esima colonna dell'inversa $\underline{\underline{\mathbf{M}}}^{-1}$ di una matrice monotona, $\underline{\underline{\mathbf{M}}}(-\vec{\mathbf{c}}) = -\vec{\mathbf{e}}_j \preceq \vec{\mathbf{0}}$, dove $\vec{\mathbf{e}}_j$ è il $j$esimo elemento della base naturale di $\mathbb{R}^n$. Quindi $-\vec{\mathbf{c}} \preceq \vec{\mathbf{0}}$, cioè $(\underline{\underline{\mathbf{M}}}^{-1})_{k,j} = c_k \geq 0$ per ogni $1 \leq j, k \leq n$.

(iii) Se $\underline{\underline{\mathbf{L}}} \succeq \underline{\underline{\mathbf{0}}}$ (cioè è una matrice i cui termini sono non negativi), la moltiplicazione a destra o a sinistra con $\underline{\underline{\mathbf{L}}}$ preserva la relazione $\preceq$. Usando (ii), $\underline{\underline{\mathbf{M}}} \preceq \underline{\underline{\mathbf{N}}}$ implica che $\underline{\underline{\mathbf{M}}}\underline{\underline{\mathbf{N}}}^{-1} \preceq \underline{\underline{\mathbf{N}}}\underline{\underline{\mathbf{N}}}^{-1} = \underline{\underline{\mathbf{I}}}$, e a sua volta $\underline{\underline{\mathbf{N}}}^{-1} = \underline{\underline{\mathbf{M}}}^{-1}\underline{\underline{\mathbf{M}}}\underline{\underline{\mathbf{N}}}^{-1} \preceq \underline{\underline{\mathbf{M}}}^{-1}\underline{\underline{\mathbf{I}}} = \underline{\underline{\mathbf{M}}}^{-1}$. □

Il punto (i) di questa proposizione e il principio del massimo discreto ci danno il seguente fatto.

**Teorema 4.5.** Se $q \geq 0$, la matrice $\underline{\underline{\mathbf{A}}}$ in (24) è monotona e quindi invertibile. Il metodo delle differenze finite ammette un'unica soluzione $\vec{\mathbf{U}} \in \mathbb{R}^n$.

**Esercizio ✎ 4.6.**

- Quali matrici diagonali sono monotone?
- Esiste una matrice monotona $\underline{\underline{\mathbf{M}}}$ tale che anche $-\underline{\underline{\mathbf{M}}}$ sia monotona?
- Mostrare una matrice (ad esempio $2 \times 2$) simmetrica definita positiva ma non monotona.
- Mostrare una matrice monotona e simmetrica ma non semi-definita positiva.
- Mostrare una matrice monotona la cui inversa è monotona, e una monotona la cui inversa non lo è.

(Ricordiamo che una matrice $\underline{\underline{\mathbf{M}}} \in \mathbb{R}^{n \times n}$ è detta definita positiva se $\vec{\mathbf{v}}^\top \underline{\underline{\mathbf{M}}}\vec{\mathbf{v}} > 0$ per ogni $\vec{\mathbf{v}} \in \mathbb{R}^n \setminus \{\vec{\mathbf{0}}\}$ e semi-definita positiva se $\vec{\mathbf{v}}^\top \underline{\underline{\mathbf{M}}}\vec{\mathbf{v}} \geq 0$.)

**Esercizio ✎ 4.7.** Usare la disuguaglianza di Cauchy–Schwarz per mostrare che la matrice $\underline{\underline{\mathbf{A}}}$ in (24) (con $q \geq 0$) è semi-definita positiva.

Poiché $\underline{\underline{\mathbf{A}}}$ è anche invertibile e simmetrica, segue che è definita positiva.

## 4.3 Analisi dell'errore: troncamento, consistenza, stabilità e convergenza

Vogliamo studiare la stabilità del metodo delle differenze finite e l'errore commesso. Assumiamo che la soluzione $u$ del problema al contorno sia di classe $C^4$.

L'errore che vogliamo controllare è dato dal vettore $\vec{\mathbf{e}} := \vec{\mathbf{u}} - \vec{\mathbf{U}}$, dove $(\vec{\mathbf{u}})_j := u(x_j)$ è il vettore dei valori della soluzione esatta del problema ai limiti nei nodi, e $\vec{\mathbf{U}}$ è il vettore dei valori ottenuti dal metodo delle differenze finite. Consideriamo innanzitutto l'**errore di troncamento**:

$$\vec{\mathbf{T}} := \underline{\underline{\mathbf{A}}}\vec{\mathbf{u}} - \vec{\mathbf{B}} = \underline{\underline{\mathbf{A}}}\vec{\mathbf{u}} - \underline{\underline{\mathbf{A}}}\vec{\mathbf{U}} = \underline{\underline{\mathbf{A}}}\vec{\mathbf{e}}.$$

Dalla definizione della matrice $\underline{\underline{\mathbf{A}}}$ e dall'errore di troncamento delle differenze finite del secondo ordine (22) ($D_h^{2C} u(x_j) = u''(x_j) + \frac{h^2}{12} u^{(iv)}(\xi)$ se $u$ è di classe $C^4$), per $2 \leq j \leq n-1$, l'elemento $j$esimo del vettore $\vec{\mathbf{T}}$ soddisfa

$$T_j = -D_h^{2C} u(x_j) + q(x_j)u(x_j) - f(x_j) = -u''(x_j) - \frac{h^2}{12} u^{(iv)}(\xi_j) + q(x_j)u(x_j) - f(x_j) = -\frac{h^2}{12} u^{(iv)}(\xi_j) \qquad (25)$$

per qualche $\xi_j \in (x_{j-1}, x_{j+1})$. In particolare $T_j$ converge a zero quadraticamente in $h$. Il fatto che l'errore di troncamento converge a zero è a volte chiamato "**consistenza**" (anche se "coerenza" sarebbe una traduzione più precisa di "*consistency*").

**Esercizio ✎ 4.8.** Mostrare che (25) vale per $j = 1$ e $j = n$.

Prima di analizzare l'errore commesso dal metodo, nella prossima nota e nell'esercizio successivo ricordiamo alcuni fatti utili sulle norme vettoriali e matriciali; più dettagli si possono trovare in [QSSG14, §1.10–11].

**Nota 4.9** (Norme matriciali e vettoriali)**.** Una norma matriciale $\|\cdot\|$ su $\mathbb{R}^{n\times n}$ è *compatibile* con una norma vettoriale $\|\cdot\|$ su $\mathbb{R}^n$ se $\left\|\underline{\underline{\mathbf{M}}}\vec{\mathbf{v}}\right\| \leq \left\|\underline{\underline{\mathbf{M}}}\right\| \|\vec{\mathbf{v}}\|$ $\forall \underline{\underline{\mathbf{M}}} \in \mathbb{R}^{n\times n}, \vec{\mathbf{v}} \in \mathbb{R}^n$. Data una norma vettoriale $\|\cdot\|$, la norma matriciale $\left\|\underline{\underline{\mathbf{M}}}\right\| := \sup_{\vec{\mathbf{v}}\neq\vec{\mathbf{0}}} \left\|\underline{\underline{\mathbf{M}}}\vec{\mathbf{v}}\right\| / \|\vec{\mathbf{v}}\|$ è detta norma *indotta*, ed è una norma compatibile.

Le norme vettoriali e matriciali $\|\cdot\|_p$ sono definite come

$$\|\vec{\mathbf{v}}\|_p := \left(\sum_{j=1}^n |v_j|^p\right)^{1/p} \quad 1 \leq p < \infty, \qquad \|\vec{\mathbf{v}}\|_\infty := \max_{j=1,\dots,n} |v_j|, \qquad \left\|\underline{\underline{\mathbf{M}}}\right\|_p := \sup_{\vec{\mathbf{v}}\neq\vec{\mathbf{0}}} \frac{\left\|\underline{\underline{\mathbf{M}}}\vec{\mathbf{v}}\right\|_p}{\|\vec{\mathbf{v}}\|_p} \quad 1 \leq p \leq \infty.$$

Notiamo che usiamo la stessa notazione per le norme vettoriali e quelle matriciali; a quale delle due ci stiamo riferendo è chiaro dall'argomento. I casi più importanti sono quelli per $p = 1, 2, \infty$, per cui la norma matriciale si può calcolare esplicitamente come

$$\left\|\underline{\underline{\mathbf{M}}}\right\|_1 = \max_{j=1,\dots,n} \sum_{k=1}^n |M_{k,j}|, \qquad \left\|\underline{\underline{\mathbf{M}}}\right\|_\infty = \max_{j=1,\dots,n} \sum_{k=1}^n |M_{j,k}|, \qquad \left\|\underline{\underline{\mathbf{M}}}\right\|_2 = \sqrt{\rho(\underline{\underline{\mathbf{M}}}^\top \underline{\underline{\mathbf{M}}})} \qquad (26)$$

dove $\rho(\cdot)$ denota il raggio spettrale. Poiché $\mathbb{R}^{n\times n}$ ha dimensione finita, tutte le norme matriciali sono equivalenti.

In Matlab il comando `norm(X,p)` permette di calcolare le norme 1, 2 e $\infty$ di matrici e le norme $p$ per $1 \leq p \leq \infty$ di vettori.

**Esercizio ✎ 4.10** (Norme matriciali e vettoriali)**.**

- Mostrare che data una norma vettoriale $\|\cdot\|$ su $\mathbb{R}^{n\times n}$ la norma matriciale indotta è compatibile.

  Dedurre che per la norma indotta vale $\left\|\underline{\underline{\mathbf{AB}}}\right\| \leq \left\|\underline{\underline{\mathbf{A}}}\right\| \left\|\underline{\underline{\mathbf{B}}}\right\|$ per ogni $\underline{\underline{\mathbf{A}}}, \underline{\underline{\mathbf{B}}} \in \mathbb{R}^{n\times n}$.
- Mostrare le seguenti equivalenze tra norme vettoriali

  $$\frac{1}{\sqrt{n}} \|\vec{\mathbf{v}}\|_1 \leq \|\vec{\mathbf{v}}\|_2 \leq \|\vec{\mathbf{v}}\|_1, \qquad \|\vec{\mathbf{v}}\|_\infty \leq \|\vec{\mathbf{v}}\|_2 \leq \sqrt{n}\, \|\vec{\mathbf{v}}\|_\infty.$$

  Da queste derivare le equivalenze tra norme matriciali

  $$\frac{1}{\sqrt{n}} \left\|\underline{\underline{\mathbf{M}}}\right\|_1 \leq \left\|\underline{\underline{\mathbf{M}}}\right\|_2 \leq \sqrt{n} \left\|\underline{\underline{\mathbf{M}}}\right\|_1, \qquad \frac{1}{\sqrt{n}} \left\|\underline{\underline{\mathbf{M}}}\right\|_\infty \leq \left\|\underline{\underline{\mathbf{M}}}\right\|_2 \leq \sqrt{n} \left\|\underline{\underline{\mathbf{M}}}\right\|_\infty.$$

- Dimostrare le formule (26).

  Suggerimenti: per $p=\infty$ dimostrare la disuguaglianza $\max_{j=1,\dots,n}\sum_{k=1}^{n}|M_{j,k}| \le \sup_{\vec{\mathbf{v}}\neq\vec{\mathbf{0}}}\frac{\|\underline{\underline{\mathbf{M}}}\vec{\mathbf{v}}\|_\infty}{\|\vec{\mathbf{v}}\|_\infty}$ e quella opposta. Procedere allo stesso modo per $p=1$. Per $p=2$ diagonalizzare la matrice simmetrica $\underline{\underline{\mathbf{M}}}^\top\underline{\underline{\mathbf{M}}}$.
- Mostrare che se $\underline{\underline{\mathbf{M}}}$ è simmetrica allora $\left\|\underline{\underline{\mathbf{M}}}\right\|_2 = \rho(\underline{\underline{\mathbf{M}}})$ e $\left\|\underline{\underline{\mathbf{M}}}\right\|_1 = \left\|\underline{\underline{\mathbf{M}}}\right\|_\infty$.

L'errore del metodo delle differenze finite soddisfa la stima

$$\|\vec{\mathbf{e}}\|_p = \left\|\underline{\underline{\mathbf{A}}}^{-1}\vec{\mathbf{T}}\right\|_p \le \left\|\underline{\underline{\mathbf{A}}}^{-1}\right\|_p \left\|\vec{\mathbf{T}}\right\|_p \qquad 1\le p\le\infty,$$

dove usiamo le "norme $p$" vettoriali e matriciali. Abbiamo già stimato $\vec{\mathbf{T}}$ in (25), quindi proviamo a stimare la norma di $\underline{\underline{\mathbf{A}}}^{-1}$. Sappiamo dalla §4.2 che $\underline{\underline{\mathbf{A}}}$ è una matrice monotona, quindi tutti gli elementi di $\underline{\underline{\mathbf{A}}}^{-1}$ sono non-negativi.

**Esercizio ✎ 4.11.** Dimostrare che per una matrice $\underline{\underline{\mathbf{M}}}\in\mathbb{R}^{n\times n}$ con elementi non-negativi vale $\left\|\underline{\underline{\mathbf{M}}}\right\|_\infty = \left\|\underline{\underline{\mathbf{M}}}\,\vec{\mathbf{1}}\right\|_\infty$, dove $\vec{\mathbf{1}}=(1,\dots,1)^\top\in\mathbb{R}^n$. (La prima norma nella formula è matriciale e la seconda vettoriale.)

Questo esercizio suggerisce di considerare la norma $p=\infty$. Iniziamo dal caso $q(x)=0$ e denotiamo $\underline{\underline{\mathbf{A}}}_0$ la corrispondente matrice del metodo delle differenze finite. ($\underline{\underline{\mathbf{A}}}_0$ è la matrice tridiagonale con $2/h^2$ sulla diagonale principale e $-1/h^2$ sulle due diagonali adiacenti.) Definiamo il vettore $\vec{\mathbf{W}} := \underline{\underline{\mathbf{A}}}_0^{-1}\vec{\mathbf{1}}$. Questo è l'approssimazione data dal metodo delle differenze finite della soluzione del problema al contorno

$$-w''(x)=1 \quad \text{in } (a,b), \qquad w(a)=w(b)=0,$$

cioè $w(x)=\frac{1}{2}(x-a)(b-x)$. Poiché $w$ è un polinomio di grado due, abbiamo $w^{(iv)}(x)=0$, quindi l'errore di troncamento delle differenze finite centrate è zero, cioè

$$W_j = w(x_j) = \frac{1}{2}(x_j-a)(b-x_j), \quad \text{da cui} \quad \left\|\underline{\underline{\mathbf{A}}}_0^{-1}\right\|_\infty = \left\|\underline{\underline{\mathbf{A}}}_0^{-1}\vec{\mathbf{1}}\right\|_\infty = \left\|\vec{\mathbf{W}}\right\|_\infty \le \|w\|_{L^\infty(a,b)} = \frac{1}{8}(b-a)^2.$$

Tornando al caso generale $q\ge 0$, sappiamo che $\underline{\underline{\mathbf{A}}}$ e $\underline{\underline{\mathbf{A}}}_0$ sono matrici monotone con $\underline{\underline{\mathbf{A}}}_0 \preceq \underline{\underline{\mathbf{A}}}$. Dai punti (ii)–(iii) della Proposizione 4.4, $\underline{\underline{\mathbf{0}}} \preceq \underline{\underline{\mathbf{A}}}^{-1} \preceq \underline{\underline{\mathbf{A}}}_0^{-1}$ cioè le due matrici inverse hanno elementi non-negativi e $(\underline{\underline{\mathbf{A}}}^{-1})_{j,k} \le (\underline{\underline{\mathbf{A}}}_0^{-1})_{j,k}$. Dalla formula della norma matriciale $\|\cdot\|_\infty$ segue che $\left\|\underline{\underline{\mathbf{A}}}^{-1}\right\|_\infty \le \left\|\underline{\underline{\mathbf{A}}}_0^{-1}\right\|_\infty$. Combinando le maggiorazioni per l'inversa di $\underline{\underline{\mathbf{A}}}$ e quelle per l'errore di troncamento otteniamo una stima dell'errore del metodo delle differenze finite:

$$\boxed{\left\|\vec{\mathbf{u}}-\vec{\mathbf{U}}\right\|_\infty \le \left\|\underline{\underline{\mathbf{A}}}^{-1}\right\|_\infty \left\|\vec{\mathbf{T}}\right\|_\infty \le \frac{1}{8}(b-a)^2\frac{h^2}{12}\max_{j=1,\dots,n}|u^{(iv)}(\xi_j)| \le \frac{(b-a)^2h^2}{96}\left\|u^{(iv)}\right\|_{L^\infty(a,b)}.} \tag{27}$$

Questo è un risultato di **convergenza**: la soluzione numerica $\vec{\mathbf{U}}$ ottenuta con il metodo delle differenze finite converge alla soluzione esatta $u$ del problema quando la mesh viene raffinata, cioè quando $h\to 0$. L'ordine di convergenza è quadratico in $h$, e coincide con l'ordine di consistenza, cioè l'ordine dell'errore di troncamento. Ricordiamo che (27) vale per $u\in C^4([a,b])$ e $q\ge 0$.

Se $q=0$, allora $u^{(iv)}=-f''$ e il termine a destra in (27) può essere calcolato immediatamente dai dati del problema.

**Nota 4.12.** La strategia che abbiamo seguito per dimostrare la *convergenza* del metodo delle differenze finite è costituita da due passi fondamentali:

- la stima di *stabilità* su $\|\underline{\underline{\mathbf{A}}}^{-1}\|$,
- la stima dell'errore di *troncamento* $\|\vec{\mathbf{T}}\|$.

Questi sono i due passi tipici nell'analisi di molti metodi numerici.

Una diversa tecnica per analizzare la convergenza del metodo delle differenze finite, e più vicina a quella che viene usata per il metodo agli elementi finiti, è presentata nel capitolo 3 di [Süli06].

**Esercizio 💻 4.13.** Dato un problema al bordo come in (23) risolverlo con il metodo delle differenze finite per diversi valori di $n$, ad esempio $n=2,4,8,16,\dots,2^{14}$. Plottare l'errore in dipendenza da $h$. Stimare numericamente l'ordine di convergenza in $h$ (la funzione `polyfit` può aiutare).

Ad esempio, per il problema $-u''+u=0$, $u(-1)=u(1)=1$ si ottengono il grafico in Figura 9 e gli ordini $\left\|\vec{\mathbf{u}}-\vec{\mathbf{U}}\right\|_1 \approx Ch^{0.989}$, $\left\|\vec{\mathbf{u}}-\vec{\mathbf{U}}\right\|_2 \approx Ch^{1.496}$, $\left\|\vec{\mathbf{u}}-\vec{\mathbf{U}}\right\|_\infty \approx Ch^{1.991}$.

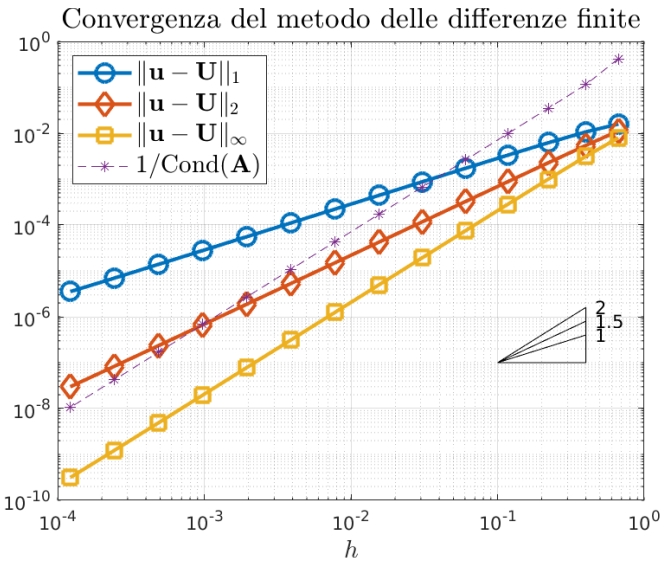



*Figura 9: La convergenza dell'errore in tre norme vettoriali per il metodo delle differenze finite applicato a* $-u''+u=0$, $u(-1)=u(1)=1$.

**Nota 4.14** (Stabilità discreta)**.** Sapendo controllare $\|\underline{\underline{\mathbf{A}}}^{-1}\|$ abbiamo anche una stima di $\vec{\mathbf{U}}=\underline{\underline{\mathbf{A}}}^{-1}\vec{\mathbf{B}}$:

$$\left\|\vec{\mathbf{U}}\right\|_\infty \le \left\|\underline{\underline{\mathbf{A}}}^{-1}\right\|_\infty \left\|\vec{\mathbf{B}}\right\|_\infty \le \frac{(b-a)^2}{8}\left(\|f\|_{L^\infty(a,b)}+\max\{|\alpha|,|\beta|\}h^{-2}\right).$$

Questo è un analogo discreto della stima di stabilità (18) per il problema continuo. Confrontando le due disuguaglianze salta all'occhio il fattore $h^{-2}$ moltiplicato per i valori al bordo $\alpha$ e $\beta$, che può diventare molto grande: possiamo liberarcene? Per trovare una stima di stabilità per $\vec{\mathbf{U}}$ indipendente da $h$ si può: (1) considerare il sistema lineare "esteso" di dimensione $(n+2)\times(n+2)$ per il vettore $(U_0,\dots,U_{n+1})^\top$, poi (2) ridursi al caso $q=0$ usando la monotonia delle matrici coinvolte, e infine (3) scrivere esplicitamente $\widetilde{\underline{\underline{\mathbf{A}}}}^{-1}$, che è una sorta di "funzione di Green discreta". I dettagli si possono trovare in [LeVeque07, §2.11].

**Nota 4.15** (Altre norme)**.** Abbiamo misurato la convergenza nella norma (vettoriale) infinito, cioè abbiamo stimato $\|\vec{\mathbf{u}}-\vec{\mathbf{U}}\|_\infty=\max_{j=1,\dots,n}|u(y_j)-U_j|$. Per misurare l'errore in altre norme vettoriali, notiamo che per ogni $\vec{\mathbf{v}}\in\mathbb{R}^n$ abbiamo $\|\vec{\mathbf{v}}\|_p \le n^{1/p}\|\vec{\mathbf{v}}\|_\infty$, usando $n=\frac{b-a}{h}-1$ troviamo

$$\left\|\vec{\mathbf{u}}-\vec{\mathbf{U}}\right\|_p \le \frac{(b-a)^{2+1/p}h^{2-1/p}}{96}\left\|u^{(iv)}\right\|_{L^\infty(a,b)} = \frac{(b-a)^4}{96\,(n+1)^{2-1/p}}\left\|u^{(iv)}\right\|_{L^\infty(a,b)}.$$

In particolare troviamo convergenza lineare in norma 1 e convergenza $\mathcal{O}(h^{3/2})$ in norma 2.

Le norme $p$ per $1\le p<\infty$ non sembrano un buon modo per misurare l'errore assoluto: raffinando la discretizzazione aumentiamo $n$, quindi anche il numero di termini sommati nel calcolo della norma. Spesso infatti $\|\vec{\mathbf{u}}\|_p \approx n^{1/p}\|\vec{\mathbf{u}}\|_\infty$ (ad esempio se tutti gli $u_j$ hanno lo stesso valore vale l'uguaglianza $\|\vec{\mathbf{u}}\|_p = n^{1/p}\|\vec{\mathbf{u}}\|_\infty$); in questi casi l'*errore relativo* converge quadraticamente per ogni $p$: $\|\vec{\mathbf{u}}-\vec{\mathbf{U}}\|_p/\|\vec{\mathbf{u}}\|_p=\mathcal{O}(h^2)$.

Possiamo pesare la norma $p$ con $h$, definendo

$$\|\vec{\mathbf{v}}\|_{p,h} := \left(h\sum_{j=1}^{n}|v_j|^p + h\frac{|v_0|^p+|v_{n+1}|^p}{2}\right)^{1/p}, \qquad p\in[1,\infty),\ \vec{\mathbf{v}}\in\mathbb{R}^{n+2}.$$

Queste sono approssimazioni delle norme $L^p(a,b)$ attraverso la regola del trapezio: se $v_j=v(x_j)$ per qualche funzione $v\in C^0([a,b])$ abbiamo $\|\vec{\mathbf{v}}\|_{p,h}\xrightarrow{h\to0}\|v\|_{L^p(a,b)}=(\int_a^b|v(x)|^p\,\mathrm{d}x)^{1/p}$. Estendendo $\vec{\mathbf{u}}$ e $\vec{\mathbf{U}}$ con $u(a)$ e $u(b)$ a vettori $(n+2)$-dimensionali, vediamo che l'errore (sia assoluto che relativo) delle differenze finite converge quadraticamente in ciascuna di queste norme: $\|\vec{\mathbf{u}}-\vec{\mathbf{U}}\|_{p,h}=\mathcal{O}(h^2)$.

**Esercizio 4.16.** Abbiamo dimostrato che per $q\ge0$ e per qualsiasi valore di $n\in\mathbb{N}$ vale $\left\|\underline{\underline{\mathbf{A}}}^{-1}\right\|_\infty \le (b-a)^2/8$, quindi il metodo delle differenze finite è stabile. Per $q<0$ invece non ci aspettiamo che lo sia.

Plottare la norma $\left\|\underline{\underline{\mathbf{A}}}^{-1}\right\|_\infty$ per il problema sull'intervallo $(a,b)=(0,1)$, discretizzato con $n=100$ nodi (ad esempio), per $q$ costante in $x$, al variare di $q\in[-100,100]$. (Cioè scegliere tanti valori di $q$, ad esempio un migliaio, e calcolare la norma della matrice corrispondente per ciascuno di essi.)

Cosa si osserva? Per quali valori di $q$ il metodo non è stabile?

**Esercizio** ✎+💻 **4.17** (Funzione di Green discreta)**.** Abbiamo visto in (16) e (17) che le operazioni "meno derivata seconda" $u \mapsto -u''$ (per una funzione $u$ che vale zero in $a$ e $b$) e l'operazione "integrale contro la funzione di Green" $f \mapsto \int_a^b G(x,y)f(y)\,\mathrm{d}y$ sono l'inverso l'una dell'altra. A livello discreto, $u$ ed $f$ sono rappresentate dai vettori $\vec{\mathbf{U}}$ e $\vec{\mathbf{B}}$ dei loro valori nei nodi. L'operazione $\vec{\mathbf{U}} \mapsto \vec{\mathbf{B}}$ è data dalla moltiplicazione per la matrice $\underline{\underline{\mathbf{A}}}$ delle differenze finite (per $q = 0$) e l'inverso $\vec{\mathbf{B}} \mapsto \vec{\mathbf{U}}$ dalla soluzione del sistema lineare $\underline{\underline{\mathbf{A}}}\vec{\mathbf{U}} = \vec{\mathbf{B}}$. Questa operazione può ovviamente essere scritta come la moltiplicazione $\vec{\mathbf{B}} \mapsto \underline{\underline{\mathbf{G}}}\vec{\mathbf{B}}$ per la matrice inversa $\underline{\underline{\mathbf{G}}} := \underline{\underline{\mathbf{A}}}^{-1}$. Quindi possiamo interpretare l'inversa $\underline{\underline{\mathbf{G}}}$ della matrice delle differenze finite come una "funzione di Green discreta".

Verificare numericamente questo fatto plottando $\underline{\underline{\mathbf{A}}}^{-1}$, ad esempio con `surf(inv(A))`, e confrontare il risultato con Figura 6. (In modo simile, è possibile visualizzare un'approssimazione della funzione di Green per i problemi al bordo con $q > 0$.)

Verificare che gli elementi della matrice $\underline{\underline{\mathbf{G}}}$ per l'intervallo $(a,b) = (0,1)$ possono essere scritti come $G_{j,k} = \frac{1}{(n+1)^3}\min\{j(n+1-k), k(n+1-j)\}$ (questo è parte dell'Esercizio 3.4 in [SM03]).

(A differenza di $\underline{\underline{\mathbf{A}}}$, la matrice $\underline{\underline{\mathbf{G}}}$ è densa. Infatti la moltiplicazione di un vettore per $\underline{\underline{\mathbf{A}}}$ approssima un'operazione "locale", cioè la derivazione, e per il nodo $j$simo coinvolge solo i nodi vicini $j-1$, $j$, $j+1$. Al contrario la moltiplicazione per $\underline{\underline{\mathbf{G}}}$ approssima un'integrazione, che è un'operazione "globale", e coinvolge tutti i nodi.)

## 4.4 Discretizzazione del problema di Neumann

Consideriamo il problema al contorno (19) con condizioni di Neumann $u'(a) = \alpha$, $u'(b) = \beta$.

La prima differenza che notiamo è che ora non conosciamo i valori di $u(a)$ e $u(b)$, quindi questi rientreranno tra le incognite. Ora il vettore da calcolare è $\vec{\mathbf{U}} = (U_0, U_1, \dots, U_{n+1})^\top \in \mathbb{R}^{n+2}$.

Approssimando le derivate presenti nelle condizioni al bordo con differenze finite del primo ordine troviamo

$$\alpha = u'(a) \approx \frac{u(a+h)-u(a)}{h} \approx \frac{U_1 - U_0}{h}, \qquad \beta = u'(b) \approx \frac{u(b)-u(b-h)}{h} \approx \frac{U_{n+1}-U_n}{h}.$$

Aggiungendo queste relazioni a quelle per i nodi interni, otteniamo un nuovo sistema lineare $\underline{\underline{\mathbf{A}}}\vec{\mathbf{U}} = \vec{\mathbf{B}}$, questa volta in $n+2$ variabili:

$$\underline{\underline{\mathbf{A}}} = \begin{pmatrix} \frac{1}{h^2} & -\frac{1}{h^2} & & & & & \\ -\frac{1}{h^2} & \frac{2}{h^2}+q_1 & -\frac{1}{h^2} & & & & \\ & -\frac{1}{h^2} & \frac{2}{h^2}+q_2 & -\frac{1}{h^2} & & & \\ & & -\frac{1}{h^2} & \frac{2}{h^2}+q_3 & -\frac{1}{h^2} & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & -\frac{1}{h^2} & \frac{2}{h^2}+q_n & -\frac{1}{h^2} \\ & & & & & -\frac{1}{h^2} & \frac{1}{h^2} \end{pmatrix}, \quad \vec{\mathbf{B}} = \begin{pmatrix} -\alpha/h \\ f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_n \\ \beta/h \end{pmatrix}, \quad \vec{\mathbf{U}} = \begin{pmatrix} U_0 \\ U_1 \\ U_2 \\ U_3 \\ \vdots \\ U_n \\ U_{n+1} \end{pmatrix}. \tag{28}$$

Ora ci ricordiamo che l'errore commesso dalle differenze finite per il problema di Dirichlet converge quadraticamente in $h$, cioè con lo stesso ordine dell'errore di troncamento. In (28) abbiamo approssimato le derivate sul bordo usando differenze finite in avanti e indietro, che hanno errore di troncamento di ordine 1. Questo porta a una perdita di velocità di convergenza per tutto il metodo. Per ovviare a questo problema dobbiamo approssimare $u'(a)$ e $u'(b)$ in un altro modo.

Per approssimare $u'(a)$ e $u'(b)$ con differenze finite centrate $D_h^C u$, che sappiamo essere accurate al secondo ordine, introduciamo due nuovi punti esterni ad $(a,b)$: $x_{-1} = a-h$ e $x_{n+2} = b+h$ (*ghost points*). Approssimiamo la condizione al bordo e l'equazione differenziale in $a$ e $b$ con le differenze finite

$$\frac{U_1 - U_{-1}}{2h} = \alpha, \qquad -\frac{U_1 - 2U_0 + U_{-1}}{h^2} + q_0 U_0 = f_0,$$
$$\frac{U_{n+2} - U_n}{2h} = \beta, \qquad -\frac{U_{n+2} - 2U_{n+1} + U_n}{h^2} + q_{n+1}U_{n+1} = f_{n+1}.$$

Eliminando $U_{-1}$ e $U_{n+2}$ da queste equazioni otteniamo

$$\frac{U_0 - U_1}{h^2} + \frac{q_0}{2}U_0 = \frac{f_0}{2} - \frac{\alpha}{h}, \qquad \frac{-U_n + U_{n+1}}{h^2} + \frac{q_{n+1}}{2}U_{n+1} = \frac{f_{n+1}}{2} + \frac{\beta}{h}.$$

Combinando queste due equazioni con quelle per i punti interni otteniamo un nuovo sistema lineare:

$$\underline{\underline{\mathbf{A}}}\vec{\mathbf{U}} = \vec{\mathbf{B}}, \quad \text{dove} \qquad \vec{\mathbf{U}} = (U_0, \dots, U_{n+1})^\top, \tag{29}$$

$$\underline{\underline{\mathbf{A}}} = \begin{pmatrix} \frac{1}{h^2}+\frac{q_0}{2} & -\frac{1}{h^2} & & & & & \\ -\frac{1}{h^2} & \frac{2}{h^2}+q_1 & -\frac{1}{h^2} & & & & \\ & -\frac{1}{h^2} & \frac{2}{h^2}+q_2 & -\frac{1}{h^2} & & & \\ & & -\frac{1}{h^2} & \frac{2}{h^2}+q_3 & -\frac{1}{h^2} & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & -\frac{1}{h^2} & \frac{2}{h^2}+q_n & -\frac{1}{h^2} \\ & & & & & -\frac{1}{h^2} & \frac{1}{h^2}+\frac{q_{n+1}}{2} \end{pmatrix}, \qquad \vec{\mathbf{B}} = \begin{pmatrix} \frac{f_0}{2}-\frac{\alpha}{h} \\ f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_n \\ \frac{f_{n+1}}{2}+\frac{\beta}{h} \end{pmatrix}.$$

**Esercizio 💻 4.18.** Come negli esercizi 4.1 e 4.13, implementare i metodi definiti da (28) e (29) per problemi al bordo di Neumann e studiarne numericamente gli ordini di convergenza.

Ad esempio si può scegliere il problema con $(a, b) = (1, 2)$, $\alpha = -1$, $\beta = -1/4$, $q(x) = x^{-2}$, $f(x) = -x^{-3}$ (e soluzione esatta $u(x) = x^{-1}$).

Si può verificare la correttezza del codice controllando se la norma $\left\|\vec{\mathbf{u}} - \vec{\mathbf{U}}\right\|_\infty$ decade come $\mathcal{O}(h)$ per il metodo in (28) e come $\mathcal{O}(h^2)$ per quello in (29). Poiché le condizioni di Neumann "non vedono" le traslazioni $u \mapsto u + c$ per $c$ costante, le soluzioni discrete spesso appaiono simili a traslate verticali della soluzione esatta.

Per $q = 0$ le matrici dei metodi (28) e (29) coincidono, mentre i vettori $\vec{\mathbf{B}}$ differiscono. Ricordiamo che se $q = 0$ il problema (19) non è ben posto. Questo è visibile anche a livello discreto: se $q_1 = q_2 = \cdots = q_n = 0$ la matrice $\underline{\underline{\mathbf{A}}}$ non è invertibile ma ha rango $n + 1$, come si deduce dal seguente esercizio.

**Esercizio ✎ 4.19.** Mostrare che per $q = 0$ il nucleo della matrice $\underline{\underline{\mathbf{A}}}$ è dato dai vettori costanti $\mathbf{v} = (v, v, \dots, v)^\top \in \mathbb{R}^{n+2}$ per ogni $v \in \mathbb{R}$.

**Definizione 4.20.** Una matrice quadrata $\underline{\underline{\mathbf{M}}} \in \mathbb{R}^{m\times m}$ tale che

$$|M_{j,j}| \geq \sum_{\substack{k=1,\dots,m \\ k\neq j}} |M_{j,k}| \quad \text{per } j = 1, \dots, m$$

si dice **a predominanza diagonale** (*diagonally dominant*). Se per ogni $j$ la disuguaglianza è stretta, si dice **a predominanza diagonale stretta** (*strictly diagonally dominant*).

Una matrice è a predominanza diagonale se ogni termine sulla diagonale è in valore assoluto maggiore o uguale alla somma dei valori assoluti degli altri termini sulla stessa riga. Per questo motivo si dice anche "predominanza diagonale per righe", quella "per colonne" è analoga. Il fatto che le matrici a predominanza diagonale stretta sono invertibili segue immediatamente dal seguente risultato.

**Teorema 4.21** (Teorema dei cerchi di Gershgorin)**.** Sia $\underline{\underline{\mathbf{M}}} \in \mathbb{C}^{m\times m}$ una matrice complessa. Allora tutti gli autovalori di $\underline{\underline{\mathbf{M}}}$ appartengono all'insieme

$$Z = \bigcup_{j=1}^{n} \left\{ z \in \mathbb{C} : \ |z - M_{j,j}| \leq \sum_{\substack{k=1,\dots,m \\ k\neq j}} |M_{j,k}| \right\}.$$

*Dimostrazione.* Sia $\lambda$ un autovalore di $\underline{\underline{\mathbf{M}}}$ e $\vec{\mathbf{v}}$ un corrispondente autovettore. Sia $j \in \{1, \dots, m\}$ un indice tale che $|v_j| = \|\vec{\mathbf{v}}\|_\infty$, e definiamo $\vec{\mathbf{w}} = \vec{\mathbf{v}}/v_j$. Il vettore $\vec{\mathbf{w}}$ è autovettore per $\underline{\underline{\mathbf{M}}}$ con autovalore $\lambda$ e soddisfa $w_j = 1 = \|\vec{\mathbf{w}}\|_\infty$. Da $\underline{\underline{\mathbf{M}}}\vec{\mathbf{w}} = \lambda\vec{\mathbf{w}}$ abbiamo $\lambda = \lambda w_j = (\underline{\underline{\mathbf{M}}}\vec{\mathbf{w}})_j = \sum_k M_{j,k}w_k = M_{j,j} + \sum_{k\neq j} M_{j,k}w_k$. Dalla disuguaglianza triangolare $|\lambda - M_{j,j}| \leq \sum_{k\neq j} |M_{j,k}||w_k| \leq \sum_{k\neq j} |M_{j,k}|$. □

Detto a parole, il teorema di Gershgorin afferma che gli autovalori di una matrice appartengono all'unione dei cerchi chiusi nel piano complesso con centro i termini diagonali della matrice e raggio la somma dei valori assoluti degli altri elementi della riga corrispondente. Di conseguenza, per una matrice a predominanza diagonale stretta, il valore 0 non appartiene a nessuno di questi cerchi, quindi non è autovalore e la matrice è invertibile.

**Proposizione 4.22.** Se tutti i $q_j$, $j = 0, \dots, n + 1$ sono positivi, il metodo definito da (29) è ben posto e stabile: $\underline{\underline{\mathbf{A}}}$ è invertibile, simmetrica e definita positiva. Inoltre $\left\|\underline{\underline{\mathbf{A}}}^{-1}\right\|_2 \leq \min\{\frac{1}{2}q_0, q_1, q_2, \dots, q_n, \frac{1}{2}q_{n+1}\}^{-1}$.

*Dimostrazione.* La matrice $\underline{\underline{\mathbf{A}}}$ in (29) è chiaramente a predominanza diagonale per $q_j \geq 0$. Se tutti i $q_j > 0$ allora è a predominanza diagonale stretta.[3] Quindi per $q > 0$ la matrice $\underline{\underline{\mathbf{A}}}$ è invertibile e il metodo delle differenze finite è ben posto.

Definiamo $q_* = \min\{\frac{1}{2}q_0, q_1, q_2, \ldots, q_n, \frac{1}{2}q_{n+1}\} > 0$. Essendo la matrice $\underline{\underline{\mathbf{A}}}$ simmetrica i suoi autovalori sono reali; il teorema di Gershgorin ci dice che $\min\{\lambda \text{ autovalore di } \underline{\underline{\mathbf{A}}}\} \geq q_*$. Poiché $\underline{\underline{\mathbf{A}}}$ è simmetrica,

$$\left\|\underline{\underline{\mathbf{A}}}^{-1}\right\|_2 = \rho(\underline{\underline{\mathbf{A}}}^{-1}) = (\max\{\mu \text{ autovalore di } \underline{\underline{\mathbf{A}}}^{-1}\}) = (\min\{\lambda \text{ autovalore di } \underline{\underline{\mathbf{A}}}\})^{-1} \leq q_*^{-1}.$$

□

**Esercizio ✎ 4.23.** Mostrare una matrice singolare a predominanza diagonale per cui la disuguaglianza stretta $|M_{j,j}| > \sum_{k=1,\ldots,n;\ k\neq j} |M_{j,k}|$ vale solo per alcuni $j$.

**Esercizio ✎ 4.24.** Il teorema dei cerchi di Gershgorin non dice che ciascun cerchio contiene un autovalore ma che ciascun autovalore sta in (almeno) un cerchio. Scrivere una matrice $\underline{\underline{\mathbf{M}}} \in \mathbb{R}^{2\times 2}$ per cui entrambi gli autovalori stanno nel cerchio centrato in $M_{1,1}$ ma non in quello centrato in $M_{2,2}$.

**Esercizio ✎+💻 4.25.** Nonostante l'ordine di convergenza del metodo definito da (29) coincida con quello (24) per il problema di Dirichlet, se la soluzione $u$ è un polinomio di grado 3 il metodo non è esatto (ignorando il roundoff). Spiegare questo fatto e verificarlo numericamente (ad esempio costruendo un problema di Dirichlet per l'equazione $-u'' + u = f$ la cui soluzione è polinomiale). Il metodo è esatto per polinomi di grado 2?

**Esercizio ✎ 4.26.** Considerare l'equazione $-u'' + qu = f$ con **condizioni al bordo miste**, cioè $u(a) = \alpha$ e $u'(b) = \beta$. Mostrare che per $q \geq 0$ (in particolare anche per $q = 0$) il problema al bordo ammette un'unica soluzione, scrivere una discretizzazione e mostrare che la matrice ottenuta è invertibile. Suggerimento: usare il metodo dell'energia.

Scrivere una discretizzazione combinando quella per il problema di Dirichlet e quella per quello di Neumann. Dimostrarne l'invertibilità ricordando l'Esercizio 4.19.

**Esercizio ✎+💻 4.27.** Considerare l'equazione $-u'' + qu = f$ con **condizioni al bordo periodiche** (20), cioè $u(a) = u(b)$ e $u'(b) = u'(a)$, e la discretizzazione determinata da

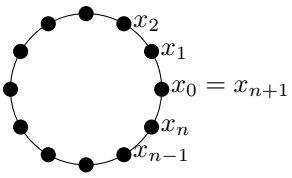
$$\underline{\underline{\mathbf{A}}} = \frac{1}{h^2}\begin{pmatrix} 2+q_1h^2 & -1 & & & & & -1 \\ -1 & 2+q_2h^2 & -1 & & & & \\ & -1 & 2+q_3h^2 & -1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -1 & 2+q_nh^2 & -1 & \\ -1 & & & & -1 & 2+q_{n+1}h^2 & \end{pmatrix},\ \vec{\mathbf{B}} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_n \\ f_{n+1} \end{pmatrix},\ \vec{\mathbf{U}} = \begin{pmatrix} U_1 \\ U_2 \\ U_3 \\ \vdots \\ U_n \\ U_{n+1} \end{pmatrix}. \tag{30}$$

Se $q$ è costante, questa è un esempio di "matrice circolante" (definite nell'Esercizio 5.20 più avanti).

Ricordiamo dall'Esercizio 2.24 che il problema al bordo è ben posto per $q > 0$ ma non per $q = 0$. Questo vale anche per il problema discreto costruito con il metodo delle differenze finite. Mostrare che $\underline{\underline{\mathbf{A}}}$ è singolare per $q = 0$ e invertibile per $q > 0$. Calcolare il nucleo di $\underline{\underline{\mathbf{A}}}$ nel caso $q = 0$ e stimare la norma di $\underline{\underline{\mathbf{A}}}^{-1}$ nel caso $q > 0$.

Implementare il metodo e studiare numericamente gli ordini di convergenza. (Ad esempio si può scegliere $-u'' + u = 2\cos x$ in $(0, 2\pi)$, che ha soluzione $u(x) = \cos x$).

Il problema al bordo periodico può essere pensato come un problema posto sulla circonferenza. Con la notazione di §4.1, i nodi $x_0 = a$ e $x_{n+1} = b$ sono identificati l'uno con l'altro. Per questo in (30) è sufficiente includere uno tra $U_0$ e $U_{n+1}$ nel vettore delle incognite ma non possiamo metterli entrambi.




[3] La definizione delle matrici a predominanza diagonale (stretta o meno) non è davvero necessaria per dimostrare la Proposizione 4.22: il teorema di Gershgorin è sufficiente a mostrare che la matrice è definita positiva e quindi invertibile. La proprietà di predominanza diagonale è comunque una proprietà molto utile da conoscere: essendo facile da verificare, soprattutto per matrici tridiagonali o sparse, offre un comodo criterio per controllare l'invertibilità di molte matrici.

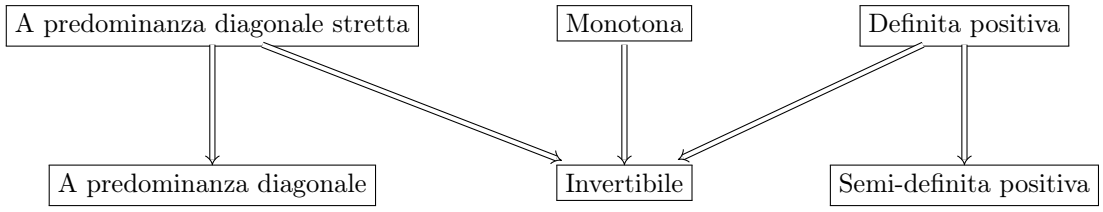



*Figura 10: Abbiamo visto diverse nozioni di "positività" per matrici quadrate, questo schema le riassume. Ogni freccia significa "implica". Per ciascuna delle 25 frecce non disegnate si può trovare una matrice $2 \times 2$ che funge da controesempio, cioè che mostra che l'implicazione corrispondente non è valida.*
*È anche possibile trovare una matrice reale $2 \times 2$ che sia invertibile, a predominanza diagonale, semi-definita positiva ma né a predominanza diagonale stretta né definita positiva (e nemmeno monotona).*

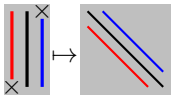
## 4.5 Implementazione

Il metodo delle differenze finite si riduce alla costruzione e alla soluzione di un sistema lineare. La principale caratteristica delle matrici ottenute è che sono **sparse**: la maggioranza degli elementi è zero. Più in particolare, la matrice è **tridiagonale**: gli unici termini diversi da zero si trovano sulla diagonale principale e le due adiacenti. Metodi alle differenze finite di ordine più alto possono dare matrici a banda con bande più larghe.

Per implementare il metodo in modo efficiente in Matlab, o in qualsiasi altro linguaggio, dobbiamo tener conto della sparsità della matrice. Una matrice di tipo sparso come quelle considerate finora occupa una quantità di memoria proporzionale a $n$, la stessa matrice salvata come matrice densa occupa una quantità di memoria proporzionale a $n^2$. Ad esempio una matrice tridiagonale $n \times n$ per $n = 10^6$ salvata come sparsa occupa circa 50 MB e 8 TB se salvata come densa (ricordiamo che un numero floating point in double precision occupa 8 bytes). Equivalentemente, 1 GB di memoria può memorizzare una matrice densa di dimensione circa $n = 11\,000$ o una sparsa tridiagonale di dimensione $n = 1.8 \cdot 10^7$ (18 milioni).

Il comando Matlab più utile per assemblare le matrici del metodo alle differenze finite è `spdiags`. Provare ad esempio

```
n = 10
C = [(1:n)', (1:n)'+n, (1:n)'+2*n]
M = spdiags( C, [-1:1], n, n )
```



Per convertire la matrice sparsa `M` in una densa e leggere più comodamente i suoi elementi si può usare `full(M)`. Notare cosa è successo ai valori 10 e 21. Altri comandi utili per maneggiare matrici sparse sono `spalloc` e `sparse`.

Ricordiamo che un codice Matlab è più veloce se invece di usare cicli `for` usa operazioni vettorizzate. Ad esempio per calcolare il vettore $\vec{\mathbf{B}}$ in (24), si può usare

```
xNodes = linspace(a,b,n+2)';
xInnerNodes = xNodes(2:end-1);
B = f_fun(xInnerNodes);
B([1,end]) = B([1,end]) + [Alpha; Beta] / h^2;
```

dove `f_fun` è una funzione vettorizzata (cioè che accetta vettori come input e restituisce come output vettori della stessa dimensione) che rappresenta il termine di sorgente $f$. Tutti i vettori e le matrici usati in questa sezione possono essere assemblati senza cicli lungo i nodi della mesh.

**Nota 4.28** (Numero di condizionamento)**.** Cosa possiamo dire del **numero di condizionamento** di $\underline{\underline{\mathbf{A}}}$? Abbiamo visto in §4.3 che (per il problema di Dirichlet) $\left\|\underline{\underline{\mathbf{A}}}^{-1}\right\|_\infty \leq \frac{1}{8}(b-a)^2$. Dalla definizione della norma infinito vediamo anche che $\left\|\underline{\underline{\mathbf{A}}}\right\|_\infty \leq \frac{4}{h^2} + \|q\|_{L^\infty(a,b)}$. Ne segue che il numero di condizionamento, almeno in norma infinito, cresce al massimo come $h^{-2}$: $\kappa_\infty(\underline{\underline{\mathbf{A}}}) = \left\|\underline{\underline{\mathbf{A}}}\right\|_\infty \left\|\underline{\underline{\mathbf{A}}}^{-1}\right\|_\infty \leq \frac{1}{8}(b-a)^2(\frac{4}{h^2} + \|q\|_{L^\infty(a,b)})$. Poiché $\underline{\underline{\mathbf{A}}}$ è simmetrica, $\left\|\underline{\underline{\mathbf{A}}}\right\|_\infty = \left\|\underline{\underline{\mathbf{A}}}\right\|_1$ e lo stesso vale per $\underline{\underline{\mathbf{A}}}^{-1}$, quindi $\kappa_\infty(\underline{\underline{\mathbf{A}}}) = \kappa_1(\underline{\underline{\mathbf{A}}})$. Stimeremo $\kappa_2(\underline{\underline{\mathbf{A}}})$ più avanti.

**Esercizio 💻 4.29.** Verificare la dipendenza da $h$ del numero di condizionamento di $\underline{\underline{\mathbf{A}}}$ in norma 2.

La funzione Matlab `cond` richiede matrici dense, mentre `condest`, che dà solo una stima del numero di condizionamento, accetta matrici sparse.

**Nota 4.30** (Buone norme di scrittura in Matlab)**.** Scrivere bene un codice numerico non è facile, anche per esercizi semplici. È fondamentale essere ordinati e disciplinati. Anche se alcuni dei suggerimenti qui riportati sembrano banali, l'esperienza insegna che confrontarli con il proprio codice è utile per tutti.

- Prima di iniziare a scrivere in Matlab pianifichiamo in dettaglio con *carta e penna* tutto il codice che vogliamo scrivere. Decidiamo quante e quali funzioni vogliamo scrivere, i loro input e output, i cicli, le variabili principali, i grafici...
- Diamo *nomi* sensati alle *variabili*. Un buon nome dovrebbe ricordarci il suo significato ed evitare ambiguità.

  Se chiamiamo `v` il primo vettore incontrato, `v1` il secondo e `vett` il terzo, quando ci saranno decine di vettori come faremo a capire cosa rappresenta ciascuno di essi?

  Se abbiamo una funzione $f$ e il vettore dei suoi valori in un insieme di punti, non chiamiamoli `f` ma, ad esempio, `f_fun` e `f_val` (ad esempio `f_fun=@(x) cos(x); f_val=f_fun(linspace(0,pi,n+2));`).
- Diamo un nome a tutti i parametri usati più di una volta.

  Ad esempio, se nell'Esercizio 3.6 vogliamo calcolare l'errore per $52$ diversi esperimenti, definiremo una variabile `NumEsperimenti = 52;` e non scriveremo mai più il numero $52$. Così, se ci accorgiamo che $50$ esperimenti sono sufficienti, cambiamo il valore in un solo punto del codice ed evitiamo errori.
- Usiamo le *parentesi*. Se siamo in dubbio se in una certa riga del codice servono oppure no, allora servono (almeno per essere sicuri di sapere cosa farà il codice e di capirlo quando lo rileggeremo).

  Ad esempio, se scriviamo `10/2*3`, siamo sicuri di sapere quant'è il risultato? E per `4^3/2`? E `1:10/2` è uguale a `1:5` o a `0.5:0.5:5`? E `2>4/3` restituirà `1` o `0`?
- Scriviamo codice *ordinato* e leggibile: indentiamo i cicli, mandiamo a capo le linee troppo lunghe con "`...`", usiamo spaziature sensate.

  Inseriamo tutti i *commenti* (con il simbolo `%`) che possono essere utili per capire cosa fa il nostro programma. Vorremmo essere in grado di capire quello che abbiamo scritto anche quando riguarderemo il codice tra qualche giorno/mese/anno.
- Una causa frequente di errori difficili da individuare è data dal fatto che uno *script* (cioè un file .m che non contiene una `function`) "vede" le variabili già presenti nel workspace. Facendo girare uno script più volte, come succede quando stiamo scrivendolo e testandolo, capita che dimentichiamo di sovrascrivere alcune variabili e il codice usi quelle calcolate in precedenza senza che ce ne accorgiamo. Un modo per evitare questi errori è far sì che ogni file .m sia una `function`, anche se non ha input e output.

  Un'altra tipica causa di confusione è la seguente. Stiamo scrivendo un programma che genera una figura. Facciamo un errore e la figura è chiaramente sbagliata. Correggiamo l'errore, riplottiamo la figura ma vediamo ancora quella sbagliata e non capiamo perché. Se il programma usa il comando `hold on` e non chiudiamo la figura prima di aggiornarla, quella nuova corretta viene plottata insieme a quella vecchia sbagliata. Per evitare questo problema basta usare `close all` all'inizio di ogni programma oppure `figure` prima di una nuova figura.
- Le scorciatoie da tastiera e le combinazioni di tasti velocizzano la scrittura. Ad esempio `F5` fa girare la funzione aperta nell'editor, `ctrl+I` permette di indentare automaticamente il codice selezionato, `ctrl+R` e `ctrl+T` commentano e de-commentano le righe selezionate... (le combinazioni dipendono dalle impostazioni del computer).
- Matlab ci aiuta a trovare e correggere gli errori. Se un comando "built-in" non dà il risultato desiderato usiamo `help` oppure `doc`. Se otteniamo un messaggio di errore o un warning leggiamolo attentamente: ci aiuta a individuare cosa abbiamo sbagliato. Se parte del codice appare sottolineato in rosso potrebbe esserci un problema: avviciniamo il mouse e vediamo cosa ci consiglia Matlab. Usiamo il debugger.
- Attenzione al significato di ciascuna variabile. Seguire la notazione delle dispense aiuta a non confondersi.

  Ad esempio, in questo capitolo $n$ denota il numero di nodi interni nell'intervallo $(a, b)$. Quindi i sotto-intervalli in cui è diviso l'intervallo sono $n+1$, e il totale dei nodi inclusi gli estremi è $n+2$. Risolveremo un sistema di dimensione $n$ per il problema di Dirichlet e uno di dimensione $n+2$ per quello di Neumann. Se nel codice chiamiamo `n` la variabile che denota il numero dei nodi interni allora possiamo usare le formule delle dispense così come sono. Se invece vogliamo chiamare `n` la dimensione del sistema lineare, nel caso di Neumann tutte le formule cambiano ed è più facile fare pasticci.
- Se il file principale che risolve un esercizio è una funzione che richiede diversi input, ricordiamoci di salvare anche il comando che lo lancia.

```
function [U,X]=DF(f,q,a,b,Alpha,Beta,n)
h = (b-a)/(n+1);
...    % calcolo X, A, B
U = A\B;
end
```

```
function EsempiDF(SceltaBVP)
n = 40;
switch SceltaBVP
  case 1
    a = -1;
    b = 1;
    ...    % inizializzazione parametri
  case 2
    ...    % altri problemi al bordo
end
[U,X] = DF(f,q,a,b,Alpha,Beta,n);
plot([a;X;b],[Alpha;U;Beta],'-o');
end
```

```
function ConvergenzaDF
a = -1;
b = 1;
...    % inizializzazione parametri
NN = 2.^(1:14);
NumEsperimenti = length(NN);
Errori = zeros(NN,1);
h = zeros(NN,1);
for j = 1:NumEsperimenti
  [U,X]= DF(f,q,a,b,Alpha,Beta,NN(j));
  h(j) = X(2)-X(1);
  ...    % calcolo errore
end
loglog(h,Errori,'-o')
end
```

*Figura 11: Un esempio di come è possibile strutturare la soluzione degli esercizi 4.1 e 4.13 in tre functions seguendo l'ultimo punto della Nota 4.30.*

Ad esempio se `function [U,X]=DF(f_fun,q_fun,a,b,Alpha,Beta,n)` applica il metodo delle differenze finite, invece di riscrivere o incollare questa espressione ogni volta, possiamo scrivere un altro file .m che genera tutti gli input per uno o più problemi al bordo e chiama il comando `DF`.

Questo suggerisce che molti degli esercizi di questo corso possono essere strutturati in modo simile. Avremo:
*(i)* una prima function che implementa il metodo numerico (ad esempio differenze finite) prendendo in input tutti i parametri necessari;
*(ii)* una seconda che genera i parametri per un problema al bordo (magari scelto tra varie opzioni usando i comandi `switch` e `case`), chiama la prima, raccoglie l'output e lo plotta;
*(iii)* una terza che chiama la prima un certo numero di volte per diversi valori di un parametro (ad esempio $n$), calcola gli errori dall'output e plotta un grafico di convergenza.
Si veda Figura 11 per un esempio di questa struttura.[4]

Ricordiamo che questo è un corso in cui impariamo dei metodi numerici, quindi non usiamo mai il toolbox di calcolo simbolico.

### 4.5.1 Risoluzione di sistemi tridiagonali

I sistemi lineari $\underline{\underline{\mathbf{A}}}\mathbf{x} = \mathbf{y}$ in cui la matrice $\underline{\underline{\mathbf{A}}}$ è tridiagonale possono essere risolti in modo estremamente efficiente. In generale se $\underline{\underline{\mathbf{A}}}$ è una matrice a bande che ammette una decomposizione LU $\underline{\underline{\mathbf{A}}} = \underline{\underline{\mathbf{LU}}}$, allora le matrici $\underline{\underline{\mathbf{L}}}$ e $\underline{\underline{\mathbf{U}}}$ hanno la stessa larghezza di banda di $\underline{\underline{\mathbf{A}}}$ [QSSG14, Proprietà 3.5]. (Ricordiamo anche [QSSG14, Proprietà 3.2]: $\underline{\underline{\mathbf{A}}}$ ammette una decomposizione LU se e solo se le sottomatrici principali di ordine $1, 2, \ldots, n-1$ sono invertibili; questo è sempre vero per matrici a predominanza diagonale stretta.)

[4] Per chi vuole essere ancora più efficiente, un modo comodo per passare i dati tra funzioni è quello di usare le "structures" di Matlab. Queste sono variabili che raccolgono altre variabili di qualsiasi tipo; vedere il comando `struct` per come usarle. Nell'esempio precedente, le functions *(ii)* e *(iii)* possono generare una structure che rappresenta il problema al bordo e la sua discretizzazione e contiene le variabili $a, b, \alpha, \beta, f, q, n, \ldots$ (notare che alcune saranno di tipo double, altre sono funzioni). Poi *(ii)* e *(iii)* passeranno quest'unica variabile alla function *(i)* che restituirà un'altra structure contenente i dati della soluzione discreta $\vec{\mathbf{U}}, \{x_j\}, \ldots$. Il vantaggio è che se più avanti decideremo di passare altre variabili (come il tipo di condizioni al bordo se vogliamo che *(i)* tratti tipi diversi, oppure il numero di condizionamento) basterà aggiungerle alla structure senza cambiare la sintassi delle functions.

Nel caso di una matrice tridiagonale questo significa che si può scrivere

$$\underbrace{\begin{bmatrix} a_1 & c_1 & & & \\ b_1 & a_2 & c_2 & & \\ & b_2 & a_3 & \ddots & \\ & & \ddots & \ddots & c_{n-1} \\ & & & b_{n-1} & a_n \end{bmatrix}}_{=\underline{\underline{\mathbf{A}}}} = \underbrace{\begin{bmatrix} 1 & & & & \\ \ell_1 & 1 & & & \\ & \ell_2 & 1 & & \\ & & \ddots & \ddots & \\ & & & \ell_{n-1} & 1 \end{bmatrix}}_{=\underline{\underline{\mathbf{L}}}} \underbrace{\begin{bmatrix} u_1 & r_1 & & & \\ & u_2 & r_2 & & \\ & & u_3 & \ddots & \\ & & & \ddots & r_{n-1} \\ & & & & u_n \end{bmatrix}}_{=\underline{\underline{\mathbf{U}}}} \tag{31}$$

Si vede facilmente che $r_j = c_j$. I valori di $\ell_1, \dots, \ell_{n-1}$, $u_1, \dots, u_n$, possono essere calcolati come segue:

$u_1 = a_1$
`For` $i = 2$ `To` $n$ `Do`
  $\ell_{i-1} = b_{i-1}/u_{i-1}$
  $u_i = a_i - \ell_{i-1} c_{i-1}$
`Next` $i$

La soluzione del sistema $\underline{\underline{\mathbf{A}}}\mathbf{x} = \underline{\underline{\mathbf{L}}}\underline{\underline{\mathbf{U}}}\mathbf{x} = \mathbf{y}$ può poi essere calcolata velocemente con la sostituzione in avanti e indietro. L'intera operazione richiede meno di $8n$ operazioni: il solutore ha complessità lineare.

**Esercizio ✎+💻 4.31.**

- Verificare che l'algoritmo proposto effettivamente corrisponde alla decomposizione LU di $\underline{\underline{\mathbf{A}}}$.
- Scrivere esplicitamente il metodo di sostituzione in avanti e indietro per le matrici $\underline{\underline{\mathbf{L}}}$ e $\underline{\underline{\mathbf{U}}}$ ottenute.
- Implementare una funzione `x = TridiagSolver(a, b, c, y)` che dati i quattro vettori $\vec{\mathbf{a}} \in \mathbb{R}^n, \vec{\mathbf{b}} \in \mathbb{R}^{n-1}, \vec{\mathbf{c}} \in \mathbb{R}^{n-1}, \vec{\mathbf{y}} \in \mathbb{R}^n$ restituisce il vettore $\vec{\mathbf{x}}$ soluzione del sistema lineare $\underline{\underline{\mathbf{A}}}\vec{\mathbf{x}} = \vec{\mathbf{y}}$.
  Testare il codice con `a,b,c,y` casuali e verificare che `x` sia effettivamente soluzione del sistema desiderato.
- Confrontare il tempo impiegato dal proprio codice con quello del comando backslash in Matlab. (Per sistemi con elementi random e dimensione $n \approx 10^4 \sim 10^6$ si dovrebbe guadagnare almeno un ordine di grandezza.)
  Per misurare i tempi computazionali si possono usare i comandi `tic` e `toc`.
- Usare questa funzione per risolvere uno dei problemi in Esercizio 4.1 con il metodo delle differenze finite.

### 4.5.2 Il caso periodico

Tutti i metodi alle differenze finite incontrati finora portano a sistemi lineari tridiagonali, eccetto il caso (30) delle condizioni al bordo periodiche. In questo caso gli elementi $A_{1,n+1}$ e $A_{n+1,1}$ sono diversi da zero e la matrice è sparsa ma non a bande. La matrice $\underline{\underline{\mathbf{A}}}$ si può comunque scrivere come una **perturbazione di rango uno** di una matrice tridiagonale:

$$\underline{\underline{\mathbf{A}}} = \underline{\underline{\mathbf{M}}} + \vec{\mathbf{u}}\vec{\mathbf{w}}^\top$$

per $\underline{\underline{\mathbf{M}}}$ tridiagonale e $\vec{\mathbf{u}}, \vec{\mathbf{w}}$ vettori.

**Esercizio ✎ 4.32.** Scrivere esplicitamente la matrice $\underline{\underline{\mathbf{M}}}$ e i vettori $\vec{\mathbf{u}}, \vec{\mathbf{w}}$ per la matrice in (30). Questa decomposizione è unica?

Per ogni $\underline{\underline{\mathbf{M}}}$ invertibile, la soluzione del sistema lineare $\underline{\underline{\mathbf{A}}}\vec{\mathbf{x}} = (\underline{\underline{\mathbf{M}}} + \vec{\mathbf{u}}\vec{\mathbf{w}}^\top)\vec{\mathbf{x}} = \vec{\mathbf{y}}$, se anche $\underline{\underline{\mathbf{A}}}$ è invertibile, si può scrivere a partire dalla soluzione del sistema lineare non perturbato come

$$\vec{\mathbf{x}} = \left( \underline{\underline{\mathbf{I}}} - \frac{(\underline{\underline{\mathbf{M}}}^{-1}\vec{\mathbf{u}})\vec{\mathbf{w}}^\top}{1 + \vec{\mathbf{w}}^\top(\underline{\underline{\mathbf{M}}}^{-1}\vec{\mathbf{u}})} \right) (\underline{\underline{\mathbf{M}}}^{-1}\vec{\mathbf{y}}). \tag{32}$$

Questa formula permette di calcolare la soluzione di $\underline{\underline{\mathbf{A}}}\vec{\mathbf{x}} = \vec{\mathbf{y}}$ risolvendo due sistemi con matrice $\underline{\underline{\mathbf{M}}}$. Ovviamente non calcoliamo mai esplicitamente l'inversa $\underline{\underline{\mathbf{M}}}^{-1}$ ma calcoliamo i due vettori $\underline{\underline{\mathbf{M}}}^{-1}\vec{\mathbf{u}}$ e $\underline{\underline{\mathbf{M}}}^{-1}\vec{\mathbf{y}}$ come soluzione dei corrispondenti sistemi lineari. La complessità dell'algoritmo è pari a due volte quella della soluzione di un sistema per $\underline{\underline{\mathbf{M}}}$ (più due prodotti scalari). In particolare se $\underline{\underline{\mathbf{M}}}\vec{\mathbf{z}} = \vec{\mathbf{y}}$ può essere risolto in $\mathcal{O}(n)$ operazioni, come nel caso di $\underline{\underline{\mathbf{M}}}$ tridiagonale, risolvere $\underline{\underline{\mathbf{A}}}\vec{\mathbf{x}} = \vec{\mathbf{y}}$ con (32) ha la stessa complessità.

La formula (32) è utile in molte altre situazioni, ad esempio per calcolare le soluzioni per una famiglia di sistemi lineari al variare di un singolo elemento delle matrici (Esercizio 4.35).

**Esercizio ✎ 4.33.** Verificare la formula (32).
Suggerimento: definire i vettori ausiliari $\vec{\mathbf{p}}$ e $\vec{\mathbf{q}}$ come le soluzioni dei sistemi lineari $\underline{\underline{\mathbf{M}}}\vec{\mathbf{p}} = \vec{\mathbf{y}}$ e $\underline{\underline{\mathbf{M}}}\vec{\mathbf{q}} = \vec{\mathbf{u}}$.

Il prossimo esercizio mostra che l'implementazione efficiente della formula (32) richiede attenzione.

**Esercizio 💻 4.34.** Sia $\underline{\underline{\mathbf{M}}} \in \mathbb{R}^{n\times n}$ una matrice invertibile data. Immaginiamo di avere a disposizione una funzione Matlab `a = FastSolve(b)` che risolve il sistema lineare $\underline{\underline{\mathbf{M}}}\vec{\mathbf{a}} = \vec{\mathbf{b}}$ con complessità computazionale $\mathcal{O}(n)$, dato un qualsiasi vettore $\vec{\mathbf{b}} \in \mathbb{R}^n$.

Vogliamo risolvere il sistema perturbato $(\underline{\underline{\mathbf{M}}}+\vec{\mathbf{u}}\vec{\mathbf{w}}^\top)\vec{\mathbf{x}} = \vec{\mathbf{y}}$ dove $\vec{\mathbf{u}}, \vec{\mathbf{w}}, \vec{\mathbf{y}} \in \mathbb{R}^n$ sono vettori dati. Applichiamo la formula (32) usando il seguente comando Matlab

```
x = (eye(n) - FastSolve(u)*w' / (1 + w'*FastSolve(u))) * FastSolve(y);
```

dove `y,u,w` sono vettori colonna di lunghezza `n`. (Ricordiamo che `eye(n)` è la matrice identità $n \times n$.)

(i) Questo codice fornisce il vettore `x` corretto ma non è soddisfacente per $n$ grande: perché?

Qual è la complessità computazionale in $n$ di questa funzione? E la quantità di memoria necessaria (come potenza di $n$)?

(ii) Scrivere una breve funzione Matlab che calcoli la soluzione $\vec{\mathbf{x}}$ di $(\underline{\underline{\mathbf{M}}} + \vec{\mathbf{u}}\vec{\mathbf{w}}^\top)\vec{\mathbf{x}} = \vec{\mathbf{y}}$ con la complessità desiderata, sfruttando `FastSolve` e usando la formula (32).

Suggerimento: manipolare (32) in modo da ottenere un'espressione più utile. Usare le proprietà dei prodotti matriciali in modo furbo.

(iii) Scrivere una funzione Matlab che dati $\vec{\mathbf{u}}, \vec{\mathbf{w}}, \vec{\mathbf{a}}, \vec{\mathbf{b}}, \vec{\mathbf{c}}, \vec{\mathbf{y}}$ risolve in $\mathcal{O}(n)$ operazioni il sistema $(\underline{\underline{\mathbf{M}}}+\vec{\mathbf{u}}\vec{\mathbf{w}}^\top)\vec{\mathbf{x}} = \vec{\mathbf{y}}$, per $\underline{\underline{\mathbf{M}}}$ tridiagonale come in (31).

Questa funzione può chiamare due volte la `TridiagSolver` implementata nell'Esercizio 4.31.

(iv) Confrontare il tempo impiegato dal proprio codice con quello usato dal comando backslash in Matlab.

Suggerimento: per implementare la matrice $\underline{\underline{\mathbf{M}}} + \vec{\mathbf{u}}\vec{\mathbf{w}}^\top$ in modo sparso (necessario per risolvere il sistema con backslash), imporre che almeno uno tra $\vec{\mathbf{u}}$ e $\vec{\mathbf{w}}$ abbia pochi elementi diversi da zero.

(v) Usare la funzione implementata per risolvere il problema alle differenze finite dell'Esercizio 4.27.

**Esercizio ✎ 4.35.** Sia $\underline{\underline{\mathbf{M}}} \in \mathbb{R}^{n\times n}$ una matrice invertibile, $\vec{\mathbf{y}} \in \mathbb{R}^n$, e siano $j, k \in \{1, \dots, n\}$ fissati. Per $t \in \mathbb{R}$ sia $\underline{\underline{\mathbf{M}}}^{(t)}$ la matrice invertibile che coincide con $\underline{\underline{\mathbf{M}}}$ in tutti gli elementi tranne che in quello $j, k$ per cui vale $\underline{\underline{\mathbf{M}}}^{(t)}_{j,k} = \underline{\underline{\mathbf{M}}}_{j,k} + t$. Mostrare che, per un opportuna scelta di due vettori $\vec{\mathbf{f}}$ e $\vec{\mathbf{g}}$ (indipendenti da $t$), il vettore $\vec{\mathbf{x}}^{(t)} := \vec{\mathbf{f}} - \frac{tf_k}{1+tg_k}\vec{\mathbf{g}}$ è soluzione del sistema lineare $\underline{\underline{\mathbf{M}}}^{(t)}\vec{\mathbf{x}}^{(t)} = \vec{\mathbf{y}}$, se il denominatore della frazione non è zero.

Quando $p$ e $q$ sono costanti, la matrice $\underline{\underline{\mathbf{A}}}$ di (30) è una matrice "circolante", vedremo più avanti un altro metodo estremamente veloce per risolvere i sistemi lineari corrispondenti.

### 4.5.3 Altri usi delle differenze finite

**Nota 4.36** (Metodo aggiunto)**.** Supponiamo di avere il problema al bordo (23) (con $\alpha = \beta = 0$ per semplicità) e di voler calcolare un funzionale lineare della soluzione $J(u) = \int_a^b u(x)\phi(x)\,\mathrm{d}x$, dove $\phi$ è una funzione data (ad esempio $\phi = \frac{1}{b-a}$ calcola la media di $u$). Immaginiamo di voler calcolare $J(u)$ per molti valori diversi di $f$, cioè al variare del termine di sorgente del problema al bordo. Come possiamo approssimare $J$ in modo efficiente? Possiamo approssimare $u$ con un vettore $\vec{\mathbf{U}}$ attraverso il metodo delle differenze finite, e approssimare $J$ con il metodo del trapezio $J(u) \approx \frac{1}{h}\sum_{j=1}^n U_j\phi(x_j) = \frac{1}{h}\vec{\mathbf{U}}\cdot\vec{\mathbf{P}}$, per $P_j := \phi(x_j)$. Però così dovremmo risolvere un sistema lineare $\underline{\underline{\mathbf{A}}}\vec{\mathbf{U}} = \vec{\mathbf{B}}$ per ogni dato $f$ (con $B_j = f(x_j)$ come in (24)). Vediamo un metodo più efficiente.

- Partiamo dal caso di un sistema lineare algebrico. Sia $\underline{\underline{\mathbf{M}}} \in \mathbb{R}^{d\times d}$ una matrice invertibile e $\vec{\mathbf{u}} \in \mathbb{R}^d$ la soluzione del sistema lineare $\underline{\underline{\mathbf{M}}}\vec{\mathbf{u}} = \vec{\mathbf{f}}$. Se ci interessa calcolare il funzionale lineare $J(\vec{\mathbf{u}}) := \vec{\mathbf{p}}\cdot\vec{\mathbf{u}}$ per un certo $\vec{\mathbf{p}} \in \mathbb{R}^d$ fissato al variare di $\vec{\mathbf{f}}$, abbiamo

$$J(\vec{\mathbf{u}}) = \vec{\mathbf{p}}\cdot\vec{\mathbf{u}} = \vec{\mathbf{p}}\cdot(\underline{\underline{\mathbf{M}}}^{-1}\vec{\mathbf{f}}) = (\underline{\underline{\mathbf{M}}}^{-\top}\vec{\mathbf{p}})\cdot\vec{\mathbf{f}}$$

(qui $^\top$ e $^{-\top}$ denotano matrice trasposta e inversa della trasposta). Se chiamiamo $\vec{\mathbf{z}} = \underline{\underline{\mathbf{M}}}^{-\top}\vec{\mathbf{p}}$ la soluzione del "sistema lineare aggiunto" $\underline{\underline{\mathbf{M}}}^\top\vec{\mathbf{z}} = \vec{\mathbf{p}}$, allora basta risolvere questo sistema per calcolare il funzionale come $J(\vec{\mathbf{u}}) = \vec{\mathbf{z}}\cdot\vec{\mathbf{f}}$. Per calcolare $J$ per $m$ diverse scelte di $\vec{\mathbf{f}}$ bastano un sistema lineare e $m$ prodotti scalari.

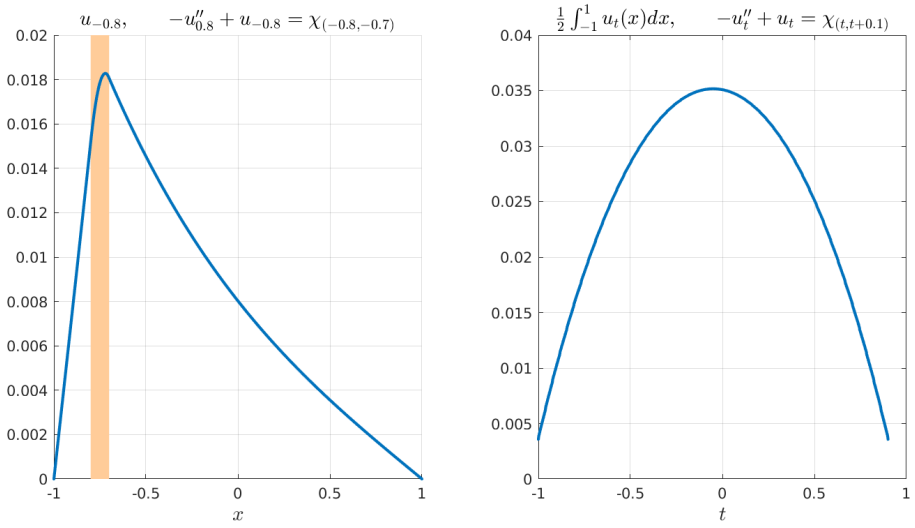



*Figura 12: Sinistra: la soluzione $u_t$ del problema al bordo descritto nella Nota 4.36 per $t = -0.8$, calcolato con il metodo delle differenze finite. La barra colorata indica la parte del dominio $(a, b)$ dove $f > 0$. Destra: la media integrale della soluzione $u_t$ al variare di $t$, calcolata con il metodo aggiunto. Un solo sistema lineare è stato risolto per calcolare tutti i valori.*

- Torniamo al problema al bordo (23). Definiamo il "problema aggiunto":

$$-z'' + qz = \phi \qquad \text{in } (a,b), \qquad z(a) = z(b) = 0.$$

  Integrando per parti e notando che i termini al bordo valgono zero (ricordiamo che $\alpha = \beta = 0$), otteniamo

$$\begin{aligned} J(u) &= \int_a^b u\phi \,\mathrm{d}x = \int_a^b u\big(-z'' + qz\big)\,\mathrm{d}x = \int_a^b \big(u'z' + qz\big)\,\mathrm{d}x - u(b)z'(b) + u(a)z'(a) \\ &= \int_a^b \big(-u'' + qu\big)z\,\mathrm{d}x + u'(b)z(b) - u'(a)z(a) = \int_a^b fz\,\mathrm{d}x. \end{aligned}$$

  Questo significa che il funzionale $J(u)$ si può calcolare senza bisogno di $u$, ma solo conoscendo $f$ e la soluzione $z$ del problema aggiunto, che non dipende da $f$.

- Se $\underline{\underline{\mathbf{A}}}$ è la matrice del metodo delle differenze finite e $P_j := \phi(x_j)$, possiamo risolvere il sistema aggiunto $\underline{\underline{\mathbf{A}}}^\top \vec{\mathbf{Z}} = \vec{\mathbf{P}}$, la cui soluzione $\vec{\mathbf{Z}}$ approssima $z$ nei nodi, e approssimare $J(u)$ con $\frac{1}{h}\vec{\mathbf{Z}} \cdot \vec{\mathbf{B}}$.

  Quindi siamo in grado di approssimare il funzionale $J$ per $m$ diversi sorgenti $f$ risolvendo un solo sistema lineare e calcolando $m$ prodotti scalari tra vettori.

**Esercizio**: plottare la media $\frac{1}{2}\int_{-1}^{1} u_t(x)\,\mathrm{d}x$, dove $u_t$ è soluzione di $-u''_t + u_t = \chi_{(t,t+0.1)}$ in $(-1, 1)$, $u(\pm 1) = 0$, al variare di $t \in [-1, 0.9]$. Qui $\chi_{(t,t+0.1)}(x) = 1$ se $x \in (t, t + 0.1)$ e $\chi_{(t,t+0.1)}(x) = 0$ altrimenti. Combinare il metodo delle differenze finite con il metodo aggiunto. Confrontare il risultato con Figura 12.

È facile immaginare come può essere usato il metodo aggiunto (*adjoint method*), ad esempio per scegliere il termine di sorgente che ottimizza una proprietà della soluzione del problema al bordo corrispondente.

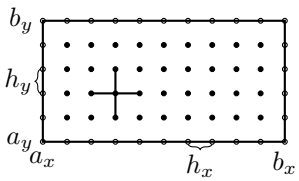
Il metodo aggiunto non è legato in particolare alle differenze finite ma può essere combinato con qualsiasi metodo per equazioni (differenziali o meno) lineari, come il metodo di collocazione o quello degli elementi finiti che vedremo in seguito.

**Esercizio** ✎+💻 **4.37** (Differenze finite in due dimensioni)**.** In §2.1 abbiamo introdotto equazioni a derivate parziali in dimensione arbitraria, mentre in questo capitolo e nei prossimi ci occupiamo solo di problemi al bordo uno-dimensionali. In questo esercizio vediamo un semplice esempio di come le tecniche studiate fin qui si estendono a problemi in dimensione maggiore. In particolare, consideriamo l'approssimazione di un problema due-dimensionale con il metodo delle differenze finite.

Sia $\Omega = (a_x, b_x) \times (a_y, b_y) \subset \mathbb{R}^2$ un rettangolo. Vogliamo approssimare il problema di diffusione–reazione stazionario (11) (con $\mathbf{p} = \mathbf{0}$) con condizioni di Dirichlet

$$-\Delta u + qu = f \quad \text{in } \Omega, \qquad u = g \quad \text{su } \partial\Omega,$$

dove $q \geq 0$, $f$, $g$ sono dati, Scegliamo dei nodi equispaziati sui lati del rettangolo:




$$x_j = a_x + jh_x, \quad j = 0, \dots, n_x + 1, \quad n_x \in \mathbb{N}, \quad h_x = (b_x - a_x)/n_x,$$

$$y_{j'} = a_y + j'h_y, \quad j' = 0, \dots, n_y + 1, \quad n_y \in \mathbb{N}, \quad h_y = (b_y - a_y)/n_y.$$

Assumiamo per semplicità che $h_x = h_y = h$. Cerchiamo dei valori $U_{j,j'}$ che approssimino $u(x_j, y_{j'})$, cioè approssimiamo la soluzione solo in un numero finito di punti nel rettangolo. Approssimando il Laplaciano $\Delta u(x_j, y_{j'}) = \frac{\partial^2 u}{\partial x^2}(x_j, y_{j'}) + \frac{\partial^2 u}{\partial y^2}(x_j, y_{j'})$ con le differenze finite centrate sia in $x$ che in $y$ otteniamo

$$\frac{4U_{j,j'} - U_{j-1,j'} - U_{j+1,j'} - U_{j,j'-1} - U_{j,j'+1}}{h^2} + q(x_j, y_{j'})U_{j,j'} = f(x_j, y_{j'}) \qquad \begin{matrix} j = 1, \dots, n_x, \\ j' = 1, \dots, n_y. \end{matrix} \tag{33}$$

Ogni nodo della griglia "comunica" con i quattro nodi più vicini. Le condizioni al bordo sono imposte fissando

$$U_{j,0} = g(x_j, a_y), \quad U_{j,n_y+1} = g(x_j, b_y), \quad U_{0,j'} = g(a_x, y_{j'}), \quad U_{n_x+1,j'} = g(b_x, y_{j'}).$$

Scegliendo un ordinamento dei nodi $(x_j, y_{j'})$ e raccogliendo i valori $U_{j,j'}$ in un vettore di $\mathbb{R}^{n_x n_y}$, si ottiene un sistema lineare quadrato in $n_x n_y$ dimensioni.

- Verificare la derivazione di (33) usando la definizione di $D_h^{2C}$ e studiare l'errore di troncamento.
- Com'è la matrice del sistema lineare ottenuto? Simmetrica? Sparsa? A bande? Quanti elementi sono diversi da zero? Quanti per ogni riga?
- Dimostrare che la matrice è invertibile. (Facile per $q > 0$, più difficile per $q = 0$.)
- Implementare il metodo per il problema

$$-\Delta u + u = \mathrm{e}^y \cosh x \quad \text{in } (-1, 1) \times (0, 1),$$

$$u(x, 0) = -\cosh x, \quad u(x, 1) = -\frac{1}{2}(\mathrm{e}^{x+1} + \mathrm{e}^{1-x}), \quad u(\pm 1, y) = -\mathrm{e}^y \cosh 1.$$

  Plottare la soluzione discreta e confrontare il risultato con la soluzione analitica.
- Come cambia il pattern di sparsità se i nodi sono ordinati prima lungo le ascisse o prima lungo le ordinate?

Per domini con i lati non paralleli agli assi Cartesiani oppure curvilinei, il metodo delle differenze finite richiede accorgimenti più complicati.

## 4.6 Problemi di diffusione–trasporto e metodo upwind

### 4.6.1 Un problema modello di diffusione–trasporto

Finora abbiamo considerato equazioni differenziali con termini di diffusione ($-u''$) e reazione ($u$), e abbiamo trascurato i termini di trasporto ($u'$). Per capire come la presenza di un termine di trasporto condiziona la soluzione, partiamo da un semplice problema omogeneo di diffusione–trasporto con coefficienti costanti e condizioni di Dirichlet:

$$\begin{cases} -\epsilon u''(x) + pu'(x) = 0 & \text{in } (0, 1), \\ u(0) = 0, \\ u(1) = 1, \end{cases} \tag{34}$$

dove $\epsilon > 0$ e $p$ sono costanti. Dall'equazione caratteristica $-\epsilon\lambda^2 + p\lambda = 0$ ricaviamo che due soluzioni linearmente indipendenti dell'equazione sono $u_1(x) = 1$ e $u_2(x) = \mathrm{e}^{\frac{p}{\epsilon}x}$. Le condizioni al bordo danno

$$u(x) = \frac{\mathrm{e}^{\frac{p}{\epsilon}x} - 1}{\mathrm{e}^{\frac{p}{\epsilon}} - 1}.$$

Se $0 < p \ll \epsilon$, cioè se il termine dominante è quello diffusivo, espandendo gli esponenziali attorno a $\frac{p}{\epsilon} = 0$ troviamo

$$u(x) = \frac{1 + \frac{p}{\epsilon}x + \frac{p^2x^2}{2\epsilon^2} + \dots - 1}{1 + \frac{p}{\epsilon} + \frac{p^2}{2\epsilon^2} + \dots - 1} = \frac{x + \frac{px^2}{2\epsilon} + \dots}{1 + \frac{p}{2\epsilon} + \dots} \approx x.$$

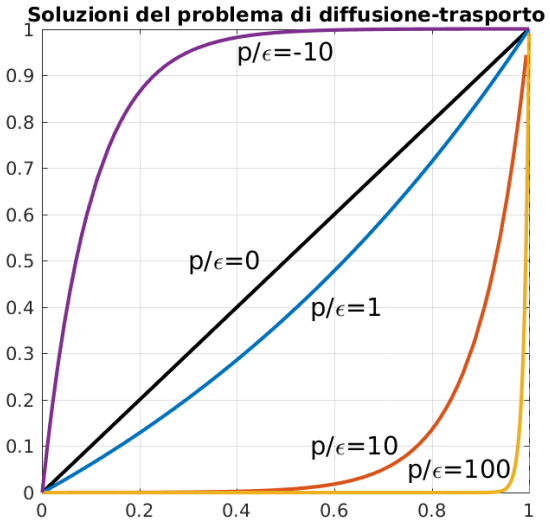



*Figura 13: Soluzioni di* $-\epsilon u'' + pu' = 0, u(0) = 0, u(1) = 1$ *per diversi valori di* $p/\epsilon$.

La soluzione del problema è vicina a quella del problema di pura diffusione $-\epsilon u''(x) = 0, u(0) = 0, u(1) = 1 \Rightarrow u(x) = x$. Questo è un esempio di **perturbazione regolare**: la soluzione del problema dipendente da un parametro piccolo $0 < \frac{p}{\epsilon} \ll 1$ converge alla soluzione del problema con $\frac{p}{\epsilon} = 0$ al diminuire di questo parametro.

Nel caso in cui il termine dominante è quello di trasporto $0 < \epsilon \ll p$ abbiamo

$$u(x) \approx \mathrm{e}^{\frac{p}{\epsilon}(x-1)}.$$

Questa soluzione è vicina a zero in tutto l'intervallo $(0, 1)$ tranne che molto vicino a 1. Vicino a 1, c'è uno "**strato limite**" (*boundary layer*): la derivata $u'$ è molto grande e $u$ "salta" improvvisamente da un valore molto piccolo a 1, per soddisfare la condizione al bordo. La risoluzione numerica di uno strato limite può essere problematica.

Fisicamente possiamo pensare a $u(x)$ come la temperatura nel punto $x$ di un fluido che si muove in un condotto isolato $(a, b)$ con velocità $p$ e con coefficiente di diffusione termica $\epsilon$. Il movimento del fluido è verso destra poiché $p$ è positivo. Le condizioni al bordo impongono due temperature diverse ai due estremi del tubo. Se la velocità è bassa e la diffusione domina sul trasporto ($p \ll \epsilon$), allora la condizione al bordo all'estremo di uscita $b$ ha effetto all'interno del condotto. Se la velocità è alta ($p \gg \epsilon$) allora possiamo aspettarci che il valore della temperatura all'interno del tubo sia uguale a quella dell'estremo di entrata, tranne che in un piccolo intorno dell'estremo opposto. Vedere [LeVeque07, §2.17] per una discussione più approfondita.

Un importante parametro adimensionale è il **numero di Péclet globale**:

$$\mathrm{Pe} := \frac{|p|(b-a)}{2\epsilon}. \tag{35}$$

Questo misura il rapporto tra il termine convettivo $pu'$ e quello diffusivo $-\epsilon u$.

### 4.6.2 Il metodo delle differenze finite per problemi con termine di trasporto

Consideriamo ora il problema generale di diffusione–trasporto con condizioni di Dirichlet:

$$\begin{cases} -\epsilon u''(x) + p(x)u'(x) = f(x) & \text{in } (a, b), \\ u(a) = \alpha, \\ u(b) = \beta. \end{cases} \tag{36}$$

Ricordiamo che per il Teorema 2.13, se $p, f \in C^0([a, b])$, il problema ammette un'unica soluzione. Vogliamo approssimare $u'$ nei nodi con delle differenze finite. Per preservare l'ordine quadratico di convergenza una scelta naturale è quella di prendere le differenze centrate $u'(x_j) \approx D_h^C u(x_j) = \frac{u(x_{j+1}) - u(x_{j-1})}{2h}$. Con la

notazione usata nelle sezioni precedenti e denotando $p_j = p(x_j)$, ricaviamo che il metodo delle differenze finite corrisponde al sistema lineare $\underline{\underline{\mathbf{A}}}\vec{\mathbf{U}} = \vec{\mathbf{B}}$ con

$$\underline{\underline{\mathbf{A}}} = \frac{\epsilon}{h^2}\begin{pmatrix} 2 & -1+\frac{hp_1}{2\epsilon} & & & \\ -1-\frac{hp_2}{2\epsilon} & 2 & -1+\frac{hp_2}{2\epsilon} & & \\ & -1-\frac{hp_3}{2\epsilon} & 2 & -1+\frac{hp_3}{2\epsilon} & \\ & & \ddots & \ddots & \ddots \\ & & & -1-\frac{hp_n}{2\epsilon} & 2 \end{pmatrix}, \quad \vec{\mathbf{B}} = \begin{pmatrix} f_1 + \frac{\alpha\epsilon}{h^2} + \frac{p_1\alpha}{2h} \\ f_2 \\ f_3 \\ \vdots \\ f_n + \frac{\beta\epsilon}{h^2} - \frac{p_n\beta}{2h} \end{pmatrix}. \tag{37}$$

Notiamo che rispetto al caso di pura diffusione (cioè (24) con $q = 0$) solo i termini di $\underline{\underline{\mathbf{A}}}$ fuori dalla diagonale principale sono cambiati. A differenza degli esempi precedenti la matrice $\underline{\underline{\mathbf{A}}}$ non è simmetrica.

**Esercizio 💻 4.38.** Implementare il metodo definito da (37) per il problema (34) con diversi valori (costanti in $x$) di $p/\epsilon$, ad esempio 1 e 1000. Plottare le soluzioni numeriche ottenute. Cosa si osserva?

Attenzione! La formula $u(x) = \frac{e^{\frac{p}{\epsilon}x}-1}{e^{\frac{p}{\epsilon}}-1}$ non è adatta per essere implementata: se $p/\epsilon \gg 1$ gli esponenziali fanno overflow e Matlab restituisce `inf`. Come si può riscrivere $u$ per poter mostrare il grafico in modo stabile?

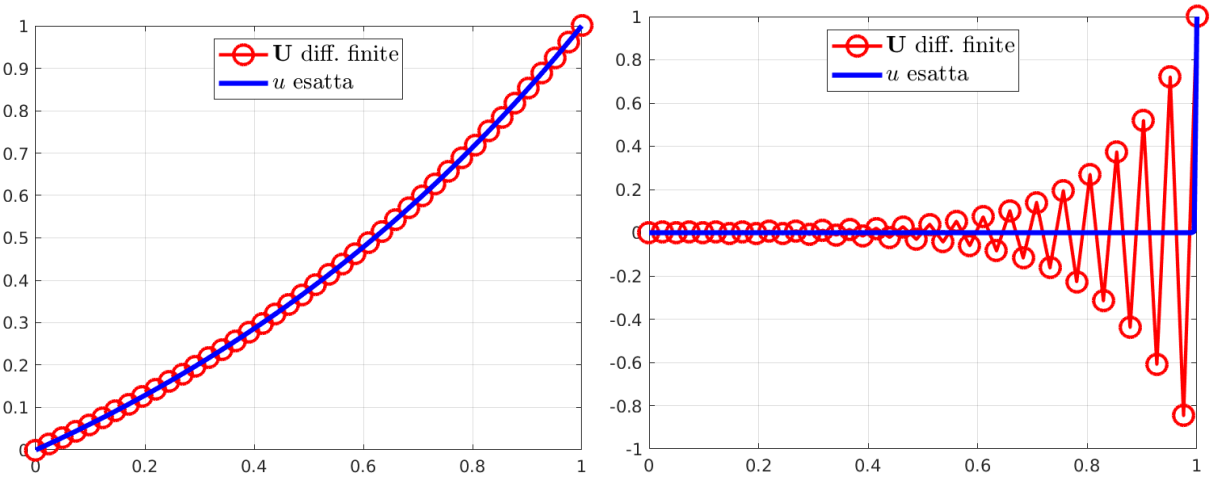



*Figura 14: La soluzione di $-u'' + pu' = 0$, $u(0) = 0$, $u(1) = 1$ per $p = 1$ (sinistra) e $p = 1000$ (destra), e la soluzione del metodo delle differenze finite per $n = 40$.*

Notiamo dalla Figura 14 che la soluzione discreta è accurata se $p/\epsilon$ è piccolo ma ha ampie **oscillazioni spurie** (cioè non presenti nella soluzione esatta) se $p/\epsilon$ è più grande. Ciò significa che il metodo non è stabile. Infatti la matrice $\underline{\underline{\mathbf{A}}}$ in (37) è a predominanza diagonale solo se $2 \geq |-1-\frac{hp_j}{2\epsilon}| + |-1+\frac{hp_j}{2\epsilon}|$ cioè se vale la disuguaglianza

$$\mathrm{Pe}_h := \frac{h|p_j|}{2\epsilon} \leq 1 \tag{38}$$

dove $\mathrm{Pe}_h$ è detto **numero di Péclet locale**. Si verifica numericamente che le oscillazioni spurie sono presenti solo per $\mathrm{Pe}_h > 1$. Questo ci vincola a scegliere $h$ molto piccolo in dipendenza da $p$ per garantire la stabilità del metodo, che può diventare computazionalmente troppo dispendioso.

L'esercizio seguente mostra che le oscillazioni spurie sono legate alla soluzione del "problema limite" con $\epsilon = 0$. Poiché questo è un problema del primo ordine, ammette una sola condizione al bordo. L'imposizione di due condizioni in (36) per un problema "quasi del primo ordine" genera le oscillazioni spurie. I problemi dipendenti da un parametro piccolo $\epsilon$ che cambiano natura per $\epsilon = 0$ vengono detti di **perturbazione singolare**.

**Esercizio ✎ 4.39.** Scrivere $\underline{\underline{\mathbf{A}}}$ e $\vec{\mathbf{B}}$ per il metodo definito da (37) per il problema $-\epsilon u'' + pu' = 0$, $u(0) = \alpha$, $u(1) = \beta$ per $p$ costante, leggermente più generale di (34), nel caso limite $\epsilon \to 0$. Mostrare che il sistema lineare ottenuto non è invertibile per $n$ dispari. Nel caso in cui $n$ è pari, calcolare a mano la soluzione $\vec{\mathbf{U}}$ e mostrare che per $j = 1, \ldots, n/2$, $U_{2j}$ dipende solo dalla condizione al bordo in $a$ e $U_{2j-1}$ da quella in $b$.

**Esercizio ✎ 4.40.** • Dimostrare il seguente "principio del massimo discreto pesato":

$$\text{Dati } n \in \mathbb{N}, \quad V_0, \ldots, V_{n+1} \in \mathbb{R}, \quad a_1, \ldots, a_n \geq 0, \quad b_1, \ldots, b_n > 0, \quad c_1, \ldots, c_n \geq 0, \quad a_j + c_j \leq b_j,$$
$$-a_jV_{j-1} + b_jV_j - c_jV_j \leq 0 \text{ per } 1 \leq j \leq n, \quad V_0 \leq 0, \ V_{n+1} \leq 0 \implies V_j \leq 0 \text{ per } 1 \leq j \leq n.$$

- Dimostrare che per $\mathrm{Pe}_h \leq 1$ la matrice $\underline{\underline{\mathbf{A}}}$ in (37) è monotona (e in particolare invertibile).
- Mostrare un esempio in cui $\mathrm{Pe}_h > 1$ e la matrice $\underline{\underline{\mathbf{A}}}$ non è monotona.

### 4.6.3 Il metodo upwind

Per ovviare a questo fenomeno possiamo scegliere differenze finite del primo ordine:

$$\text{se } p_j \geq 0 \;\Rightarrow\; u'(x_j) \approx \frac{U_j - U_{j-1}}{h} \quad \text{(d.f. all'indietro)},$$
$$\text{se } p_j \leq 0 \;\Rightarrow\; u'(x_j) \approx \frac{U_{j+1} - U_j}{h} \quad \text{(d.f. in avanti)}.$$

Questo è il metodo "*upwind*" (letteralmente "controvento" o "sopravento").

Poiché abbiamo usato differenze finite in avanti e all'indietro il metodo potrà convergere al più linearmente in $h$, quindi più lentamente dei metodi studiati in precedenza. Abbiamo sacrificato dell'accuratezza per migliorare la stabilità del metodo.

Da dove viene questa scelta? Interpretiamo il termine di trasporto $p$ come la velocità del fluido con densità $u$. Se nel nodo $x_j$ il fluido scorre verso destra abbiamo $p_j > 0$: in questo caso possiamo immaginare che il valore di $u(x_j)$ sarà maggiormente influenzato dal valore di $u(x_{j-1})$, a sinistra o sopravento rispetto a $x_j$. Quindi usando $\frac{U_j - U_{j-1}}{h}$ includiamo questa informazione e ignoriamo l'informazione proveniente da $U_{j+1}$, sottovento. Vediamo in Figura 13 che la soluzione con $p/\epsilon = 100$ si comporta in questo modo: in tutto il dominio (tranne lo strato limite) il suo valore è molto vicino a quello della condizione al bordo sopravento, cioè $u(0) = 0$.

Il metodo upwind per il problema (36) con $p \geq 0$ si può quindi scrivere in forma matriciale come $\underline{\underline{\mathbf{A}}}\vec{\mathbf{U}} = \vec{\mathbf{B}}$ per

$$\underline{\underline{\mathbf{A}}} = \frac{\epsilon}{h^2}\begin{pmatrix} 2+\frac{p_1 h}{\epsilon} & -1 & & & \\ -1-\frac{p_2 h}{\epsilon} & 2+\frac{p_2 h}{\epsilon} & -1 & & \\ & -1-\frac{p_3 h}{\epsilon} & 2+\frac{p_3 h}{\epsilon} & -1 & \\ & & \ddots & \ddots & \ddots \\ & & & -1-\frac{p_n h}{\epsilon} & 2+\frac{p_n h}{\epsilon} \end{pmatrix}, \quad \vec{\mathbf{B}} = \begin{pmatrix} f_1 + \frac{\alpha\epsilon}{h^2} + \frac{p_1\alpha}{h} \\ f_2 \\ f_3 \\ \vdots \\ f_n + \frac{\beta\epsilon}{h^2} \end{pmatrix}. \tag{39}$$

Le differenze finite all'indietro (o in avanti) si possono scrivere come somma di differenze centrate del primo e del secondo ordine (pesate con $-h/2$):

$$D_h^- u(x_j) = \frac{u_j - u_{j-1}}{h} = \frac{u_{j+1} - u_{j-1}}{2h} - \frac{h}{2}\cdot\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} = D_h^C u(x_j) - \frac{h}{2} D_h^{2C} u(x_j).$$

Lo schema upwind si può quindi pensare come uno schema alle differenze finite centrate in cui è stata aggiunta una **diffusione artificiale** proporzionale ad $h$: per $p_j \geq 0$

$$-\epsilon\frac{U_{j+1} - 2U_j + U_{j-1}}{h^2} + p_j\frac{U_j - U_{j-1}}{h} = -\epsilon_h\frac{U_{j+1} - 2U_j + U_{j-1}}{h^2} + p_j\frac{U_{j+1} - U_{j-1}}{2h} = f_j$$

dove $\epsilon_h := \epsilon + \frac{h}{2}p_j$ è la "viscosità numerica". Questa è una discretizzazione del problema perturbato

$$-\epsilon_h\widetilde{u}''(x) + p(x)\widetilde{u}'(x) = f$$

che ha numero di Péclet locale $\mathrm{Pe}_h = \frac{p_j h}{2\epsilon_h} = \frac{p_j h}{2\epsilon + p_j h} < 1$.

**Esercizio ✎ 4.41.** Scrivere il metodo upwind in forma matriciale per $p$ con segno arbitrario.

**Esercizio ✎ 4.42.** Mostrare che $\underline{\underline{\mathbf{A}}}$ in (39) è invertibile per $p \geq 0$.

Suggerimento: considerare un vettore $\vec{\mathbf{v}}$ tale che $\underline{\underline{\mathbf{A}}}\vec{\mathbf{v}} = \vec{\mathbf{0}}$. Mostrare (i) che se $v_1 = 0$ allora $\vec{\mathbf{v}} = \vec{\mathbf{0}}$ e (ii) che se $v_1 > 0$ allora $v_j \leq v_{j+1}$ per $j = 1, \ldots, n-1$. Dedurre l'invertibilità. Alternativamente si può dimostrare che $\underline{\underline{\mathbf{A}}}$ è monotona.

**Esercizio 💻 4.43.** Implementare il metodo upwind (39) e confrontare i risultati con quelli dell'Esercizio 4.38.

I risultati numerici dell'esercizio precedente mostrano che il metodo upwind è più preciso quando $n \lesssim \frac{p}{\epsilon}$, cioè nel regime in cui $\mathrm{Pe}_h > 1$ e il metodo (37) è affetto da oscillazioni spurie. Al contrario, per $n > \frac{p}{\epsilon}$, i

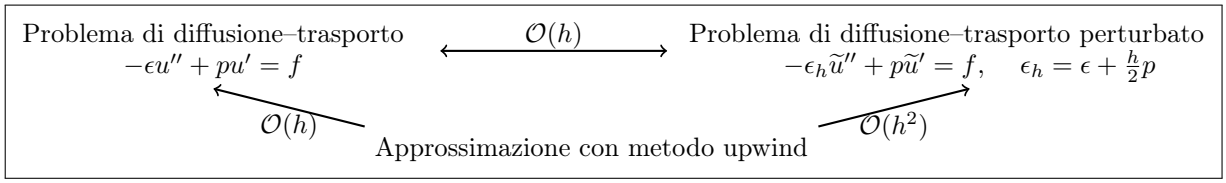



*Figura 15: Il metodo upwind coincide con quello delle differenze centrate per un problema perturbato. Quindi la soluzione discreta $\vec{\mathbf{U}}$ ottenuta con il metodo upwind "converge" con velocità quadratica in $h$ alla soluzione $\widetilde{u}$ del problema perturbato. Ma $\widetilde{u}$ dipende da $h$ e differisce dalla soluzione $u$ del problema originale di un ordine $\mathcal{O}(h)$. Quindi anche la convergenza del metodo upwind alla soluzione esatta ha un errore lineare in $h$.*

punti $x_j$ sono sufficienti a descrivere lo strato limite e il metodo (37) converge più velocemente grazie alla più accurata discretizzazione di $u'$.

Una descrizione approfondita ma accessibile dei problemi di diffusione e trasporto, dei metodi numerici per approssimarli (incluso quello di upwind) e della loro analisi si trova nel libro [M. Stynes, D. Stynes, *Convection-diffusion problems: An introduction to their analysis and numerical solution*, AMS, 2018]. In particolare, il capitolo 3 di questo libro si occupa di metodi alle differenze finite per il problema che abbiamo studiato in questa sezione.

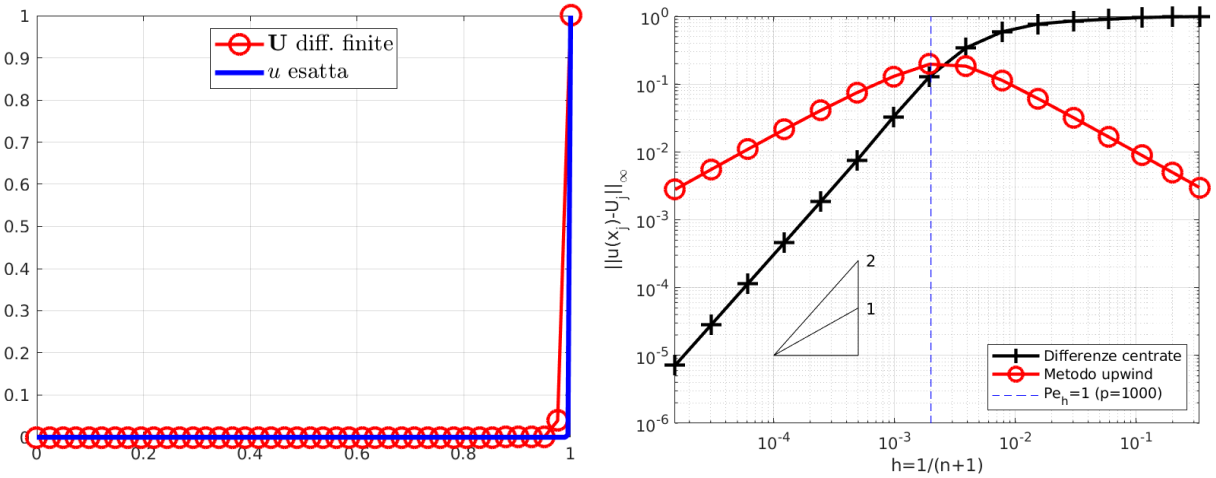



*Figura 16: Sinistra. La soluzione del problema modello $-u'' + pu' = 0$, $u(0) = 0$, $u(1) = 1$ per $p = 1000$ e la soluzione del metodo upwind per $n = 40$. Nonostante l'errore commesso vicino allo strato limite non sia trascurabile, non sono presenti oscillazioni spurie, il risultato è qualitativamente corretto e la parte "piatta" della soluzione è approssimata accuratamente.*
*Destra. Gli errori in norma infinito commessi dal metodo delle differenze centrate (37) e dal metodo upwind (39) per lo stesso problema modello con $n = 2^1, 2^2, \ldots, 2^{16}$. Si osserva che per $\mathrm{Pe}_h > 1$ il metodo upwind è più accurato; al contrario per $\mathrm{Pe}_h < 1$ quello delle differenze centrate è più accurato e gode di un ordine di convergenza più alto ($\mathcal{O}(h^2)$ invece di $\mathcal{O}(h)$).*

**Esercizio 💻 4.44.** Nel grafico a destra in Figura 16 sembra che l'errore commesso dal metodo upwind peggiori raffinando la griglia, almeno finché non si raggiunge un valore di $h$ sufficientemente fine. Questo è dovuto al fatto che stiamo misurando l'errore in norma $\|\cdot\|_\infty$. Spiegare questo fatto. Cercare una norma ragionevole per cui l'errore decresce monotonicamente e generare il plot corrispondente (una norma di questo tipo è già stata usata in queste note).

**Esercizio ✎ 4.45.** (Continuazione dell'Esercizio 4.39.) Scrivere $\underline{\underline{\mathbf{A}}}$ e $\vec{\mathbf{B}}$ per il metodo definito da (39) per il problema $-\epsilon u'' + pu' = 0$, $u(0) = \alpha$, $u(1) = \beta$, $p > 0$ costante, nel caso limite $\epsilon \to 0$.

Mostrare che $\underline{\underline{\mathbf{A}}}$ è invertibile e calcolare $\vec{\mathbf{U}}$.

Sia $\underline{\underline{\mathbf{M}}}$ la matrice $n \times n$ con elementi $M_{j,k} = h/p$ se $k \le j$ e $M_{j,k} = 0$ altrimenti. Mostrare che $\underline{\underline{\mathbf{M}}}$ è l'inversa di $\underline{\underline{\mathbf{A}}}$ per $\epsilon = 0$ (◹ ◣ = ◹).

Mostrare che $\left\|\underline{\underline{\mathbf{A}}}^{-1}\right\|_\infty \le 1/p$ e che quindi il metodo upwind è stabile per questo limite.

**Esercizio ✎+💻 4.46.** Considerare il problema di diffusione–trasporto

$$-\epsilon u'' + u' = w_1 \chi_{(0.1,0.2)} - w_2 \chi_{(0.5,0.7)} \quad \text{in } (0,1), \qquad u(0) = u(1) = 0,$$

con $\epsilon \ll 1$, per due numeri $w_1, w_2 > 0$. (Qui $\chi_{(x_1,x_2)}(x) = 1$ se $x_1 < x < x_2$ e $\chi_{(x_1,x_2)}(x) = 0$ altrimenti.)

Provare a indovinare il grafico della soluzione. Risolvere numericamente con il metodo upwind e interpretare il risultato come la temperatura di un fluido. Per quali valori di $w_1, w_2$ compare uno strato limite?

**Esercizio ✎+💻 4.47.** Anche i problemi di diffusione–reazione (senza trasporto) possono mostrare fenomeni di perturbazione singolare. Date due costanti $\epsilon, q > 0$, consideriamo il problema modello

$$-\epsilon u'' + qu = 0 \quad \text{in } (-1,1), \qquad u(-1) = u(1) = 1.$$

- Studiare i limiti $q/\epsilon \to 0$ e $\epsilon/q \to 0$.
- Calcolare e plottare la soluzione esatta e quella ottenuta con il metodo delle differenze finite centrate per diversi valori di $q/\epsilon \gg 1$ e di $n$. Compaiono strati limiti? Compaiono oscillazioni spurie? In che parte del dominio la soluzione numerica è accurata?
- Plottare il grafico di convergenza del metodo delle differenze finite al variare di $n$ per diversi valori di $q/\epsilon$. Cosa si osserva? Come si spiega?
- Confrontare con il caso di diffusione–trasporto. In quale caso il metodo delle differenze finite centrate dà risultati più affidabili?

## 4.7 Problemi agli autovalori

Gli operatori differenziali e i loro inversi, cioè gli operatori "soluzione di un problema al bordo", sono mappe lineari tra spazi vettoriali. In quanto tali ammettono autovalori e autovettori, che in questo caso sono chiamati autofunzioni. Ad esempio, si verifica facilmente che il problema

$$\begin{cases} -u'' = \lambda u, \\ u(0) = u(b) = 0 \end{cases} \tag{40}$$

per $b > 0$ ammette gli (infiniti) autovalori $\lambda_\ell$ e le autofunzioni $u_\ell$ definite come

$$k_\ell := \frac{\pi \ell}{b}, \qquad \lambda_\ell := k_\ell^2, \qquad u_\ell(x) := \sin(k_\ell x) \qquad \ell \in \mathbb{N}. \tag{41}$$

Abbiamo discretizzato i problemi al bordo (lineari) trasformandoli in sistemi lineari algebrici. Similmente i problemi come (40) si possono discretizzare con problemi algebrici agli autovalori. Gli autovalori $\lambda_\ell$ e le autofunzioni $u_\ell$ (meglio: i loro valori $u_\ell(x_j)$ nei nodi $x_j$) di (40) vengono approssimati dagli autovalori e dagli autovettori della matrice $\underline{\underline{\mathbf{A}}}$ del metodo delle differenze finite. L'equazione $j$esima del sistema algebrico agli autovalori $\underline{\underline{\mathbf{A}}}\vec{\mathbf{U}} = \lambda \vec{\mathbf{U}}$ è $-\frac{U_{j+1} - 2U_j + U_{j-1}}{h^2} = \lambda U_j$ che chiaramente approssima $-u''(x_j) = \lambda u(x_j)$. Usando una matrice $n \times n$ solo $n$ degli infiniti autovalori possono essere approssimati.

**Esercizio 💻 4.48.** Confrontare numericamente i primi $n$ autovalori di (40) con gli autovalori della matrice del metodo delle differenze finite con $n$ nodi. Confrontare le corrispondenti autofunzioni come in Figura 17.

Suggerimento: ricordare che le autofunzioni sono definite a meno di un fattore moltiplicativo. Per confrontare quelle ottenute numericamente con le $u_\ell$ definite analiticamente è necessaria una normalizzazione. Ad esempio se `ul` è l'autovettore $\ell$esimo della matrice $\underline{\underline{\mathbf{A}}}$ calcolato con `eig` o `eigs` e $(\vec{\mathbf{u}}_\ell)_j = u_\ell(x_j) = \sin(k_\ell x_j)$, allora si può moltiplicare `ul` stesso per $\frac{\|\vec{\mathbf{u}}_\ell\| \operatorname{sign}(u_\ell(x_1))}{\|\texttt{ul}\| \operatorname{sign}(\texttt{ul}_1)}$.

Come si vede dalla Figura 17 (sinistra) solo i gli autovalori più bassi sono approssimati con una certa accuratezza dalla matrice delle differenze finite. Le figure sulla destra confrontano le autofunzioni esatte (in blu) e quelle ottenute dal metodo delle differenze finite (in rosso). Per indici $\ell$ più alti, $u_\ell$ oscilla fortemente nell'intervallo $(0, b)$ e le differenze finite non sono in grado di approssimarle accuratamente.

Dalla figura sembra che il valore delle autofunzioni discrete in $x_j$ coincide con quello delle autofunzioni esatte. Verifichiamo che questo è vero. Fissato $n \in \mathbb{N}$, $b > 0$ e $h = b/(n+1)$, definiamo

$$u_\ell^j := u_\ell(x^j) = \sin \frac{\pi \ell}{b} \frac{jb}{n+1} = \sin \frac{\pi \ell j}{n+1}.$$

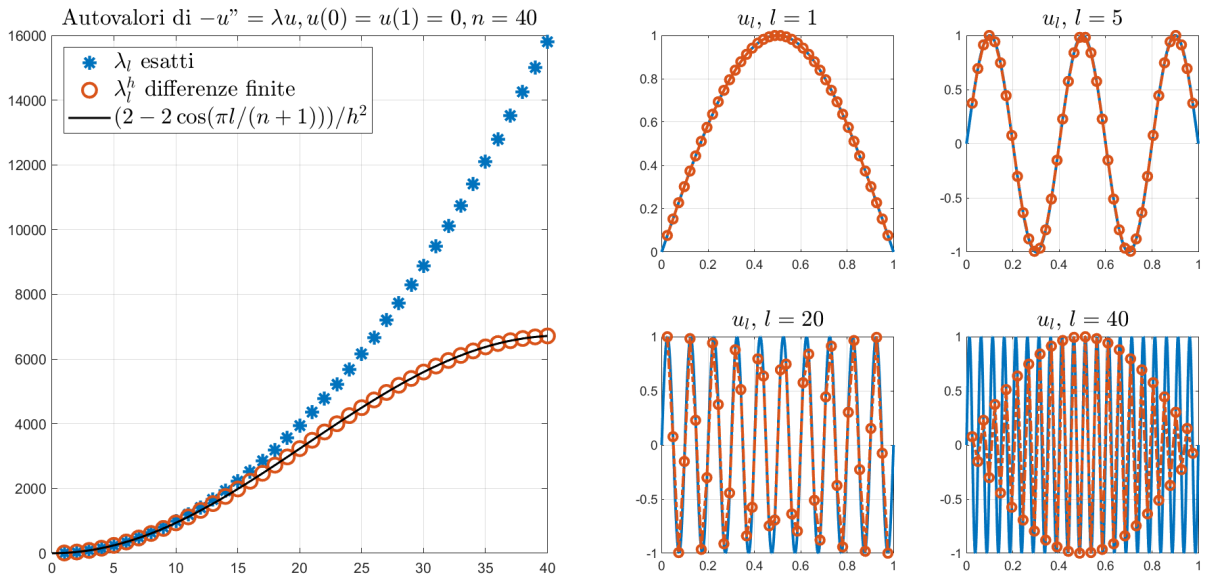



*Figura 17: Sinistra: i primi 40 autovalori di $-u'' = \lambda u$ con condizioni di Dirichlet in $(0,1)$, e gli autovalori della matrice delle differenze finite corrispondente. Destra: le autofunzioni $u_1$, $u_5$, $u_{20}$, $u_{40}$ e le approssimazioni corrispondenti calcolate con il metodo delle differenze finite.*

Per $\vec{\mathbf{u}}_\ell = (u_\ell^1, \dots, u_\ell^n)^\top$ e $2 \le j \le n-1$ abbiamo

$$\begin{aligned}
(\underline{\underline{\mathbf{A}}}\vec{\mathbf{u}}_\ell)_j &= \frac{-u_\ell^{j-1} + 2u_\ell^j - u_\ell^{j+1}}{h^2} \\
&= \frac{-\sin\frac{\pi\ell(j-1)}{n+1} + 2\sin\frac{\pi\ell j}{n+1} - \sin\frac{\pi\ell(j+1)}{n+1}}{h^2} \\
&= \frac{-\sin\frac{\pi\ell j}{n+1}\cos\frac{\pi\ell}{n+1} + \cos\frac{\pi\ell j}{n+1}\sin\frac{\pi\ell}{n+1} + 2\sin\frac{\pi\ell j}{n+1} - \sin\frac{\pi\ell j}{n+1}\cos\frac{\pi\ell}{n+1} - \cos\frac{\pi\ell j}{n+1}\sin\frac{\pi\ell}{n+1}}{h^2} \\
&= \frac{(2 - 2\cos\frac{\pi\ell}{n+1})\sin\frac{\pi\ell j}{n+1}}{h^2} = \frac{(2 - 2\cos\frac{\pi\ell}{n+1})}{h^2}(\vec{\mathbf{u}}_\ell)_j
\end{aligned}$$

dove abbiamo usato $\sin(x+y) = \sin x\cos y + \cos x \sin y$.

**Esercizio ✎ 4.49.** Verificare che $(\underline{\underline{\mathbf{A}}}\vec{\mathbf{u}}_\ell)_1 = \frac{(2-2\cos\frac{\pi\ell}{n+1})}{h^2}(\vec{\mathbf{u}}_\ell)_1$ e $(\underline{\underline{\mathbf{A}}}\vec{\mathbf{u}}_\ell)_n = \frac{(2-2\cos\frac{\pi\ell}{n+1})}{h^2}(\vec{\mathbf{u}}_\ell)_n$.

Abbiamo dimostrato che gli autovalori di $\underline{\underline{\mathbf{A}}}$, in Figura 17, sono

$$\lambda_\ell^h := \frac{(2 - 2\cos\frac{\pi\ell}{n+1})}{h^2}, \quad \ell = 1, \dots, n.$$

Per $\ell \approx n$ questi sono chiaramente molto diversi dai $\lambda_j$ in (41). Per $\ell \ll n$ invece

$$\begin{aligned}
\lambda_\ell - \lambda_\ell^h &= \left(\frac{\pi\ell}{b}\right)^2 - \frac{(2 - 2\cos\frac{\pi\ell}{n+1})}{h^2} \\
&= \left(\frac{\pi\ell}{b}\right)^2 - \frac{\left(2 - 2 + (\frac{\pi\ell}{n+1})^2 - \frac{1}{12}(\frac{\pi\ell}{n+1})^4 + \mathcal{O}\big((\frac{\pi\ell}{n+1})^6\big)\right)}{h^2} \quad = \frac{\pi^4}{12b^4}\ell^4 h^2 + \mathcal{O}\big(\ell^6 h^4\big)
\end{aligned}$$

dove abbiamo usato l'espansione di Taylor $\cos x = \sum_{j=0}^{\infty} \frac{(-1)^j x^{2j}}{(2j)!} = 1 - \frac{1}{2}x^2 + \frac{1}{24}x^4 + \mathcal{O}(x^6)$ e $\frac{1}{n+1} = \frac{h}{b}$. Questa relazione ci garantisce che per $\ell$ fissato l'$\ell$simo autovalore discreto $\lambda_\ell^h$ converge all'$\ell$esimo autovalore esatto $\lambda_\ell$ quadraticamente in $h$. D'altro canto l'errore cresce in $\ell$ come $\ell^4$: ad esempio il decimo autovalore avrà un errore circa $10^4$ volte più grande del primo.

**Nota 4.50.** Il calcolo degli autovalori di $\underline{\underline{\mathbf{A}}}$ mostra che questa matrice è invertibile, senza usare la proprietà di monotonia. Inoltre ci fornisce immediatamente una stima delle norme $\left\|\underline{\underline{\mathbf{A}}}\right\|_2$, $\left\|\underline{\underline{\mathbf{A}}}^{-1}\right\|_2$ e del numero di condizionamento $\kappa_2(\underline{\underline{\mathbf{A}}}) = \frac{\lambda_n^h}{\lambda_1^h} = \frac{1-\cos\frac{\pi n}{n+1}}{1-\cos\frac{\pi}{n+1}} = \frac{1+\cos\frac{\pi}{n+1}}{1-\cos\frac{\pi}{n+1}} \approx (\frac{2(n+1)}{\pi})^2$.

**Esercizio ✎+💻 4.51.** Ripetere quanto fatto in questa sezione per il problema con condizioni di Neumann.

Il problema al bordo lineare più generale

$$\begin{cases} -u'' + pu' + qu = \lambda u, \\ u(0) = u(b) = 0 \end{cases}$$

ammette una discussione simile, anche se in generale gli autovalori esatti e discreti non sono calcolabili esplicitamente. Se $p \neq 0$ gli autovalori (esatti e discreti) possono essere complessi.

Più informazioni si possono trovare in [LeVeque07, §2.10] e in [SM03, §13.6]. Una discussione dei problemi agli autovalori e della loro importanza nelle applicazioni si trova su [TBD18, §6]. In una delle prossime lezioni vedremo un'applicazione importante di quanto visto qui: il metodo di Fourier per risolvere l'equazione del calore.

### 4.7.1 Problemi di Sturm–Liouville

Una classe importante di problemi differenziali agli autovalori sono i cosiddetti problemi di Sturm–Liouville: per $0 < K \in C^1([a,b])$, $0 \leq q \in C^0([a,b])$

$$\begin{cases} -\big(K(x)u'(x)\big)' + q(x)u(x) = \lambda u(x), \\ u(a) = u(b) = 0. \end{cases} \tag{42}$$

Definiamo l'operatore differenziale $\mathcal{L} : C^2([a,b]) \to C^0([a,b])$ come $\mathcal{L}u := -(Ku')' + qu$ e lo spazio vettoriale a valori complessi $C^2_0([a,b]) := \{w \in C^2([a,b]), w(a) = w(b) = 0\}$. Dall'integrazione per parti vediamo che $\mathcal{L}$ è un "**operatore autoaggiunto**" [5] rispetto al prodotto scalare di $L^2(a,b)$, cioè per ogni $u, w \in C^2_0([a,b])$

$$\begin{aligned} \int_a^b (\mathcal{L}u)\overline{w}\,\mathrm{d}x &= \int_a^b (Ku'\overline{w}' + qu\overline{w})\,\mathrm{d}x - K(b)u'(b)\underbrace{\overline{w}(b)}_{=0} + K(a)u'(a)\underbrace{\overline{w}(a)}_{0} \\ &= \int_a^b u\overline{\mathcal{L}w}\,\mathrm{d}x + K(b)\underbrace{u(b)}_{=0}\overline{w}'(b) - K(a)\underbrace{u(a)}_{=0}\overline{w}'(a) \qquad = \int_a^b u\overline{\mathcal{L}w}\,\mathrm{d}x. \end{aligned}$$

Sappiamo che le matrici autoaggiunte hanno autovalori reali, lo stesso vale per gli operatori: se $\mathcal{L}u = \lambda u$ per $u \in C^2_0([a,b])$ allora

$$\lambda \|u\|^2_{L^2(a,b)} = \int_a^b \lambda u\overline{u}\,\mathrm{d}x = \int_a^b \mathcal{L}u\overline{u}\,\mathrm{d}x = \int_a^b u\overline{\mathcal{L}u}\,\mathrm{d}x = \int_a^b u\overline{\lambda}\overline{u}\,\mathrm{d}x = \overline{\lambda}\|u\|^2_{L^2(a,b)} \qquad \Rightarrow \lambda \in \mathbb{R}.$$

Inoltre gli autovalori sono positivi:

$$\lambda \|u\|^2_{L^2(a,b)} = \int_a^b \mathcal{L}u\overline{u}\,\mathrm{d}x = \int_a^b (Ku'\overline{u}' + qu\overline{u})\,\mathrm{d}x > 0.$$

Infine se $\mathcal{L}u_1 = \lambda_1 u_1$ e $\mathcal{L}u_2 = \lambda_2 u_2$ per $u_1, u_2 \in C^2_0([a,b])$

$$\lambda_1 \int_a^b u_1\overline{u}_2\,\mathrm{d}x = \int_a^b \mathcal{L}u_1\overline{u}_2\,\mathrm{d}x = \int_a^b u_1\overline{\mathcal{L}u_2}\,\mathrm{d}x = \overline{\lambda}_2 \int_a^b u_1\overline{u}_2\,\mathrm{d}x.$$

Da questo segue che le autofunzioni corrispondenti a due autovalori distinti sono ortogonali in $L^2(a,b)$.

Abbiamo dimostrato parte del seguente risultato, tralasciamo il resto della dimostrazione che richiede alcuni strumenti di analisi funzionale (si veda ad esempio il Teorema 2.4 di [J.D. Pryce, *Numerical solution of Sturm–Liouville problems*, OUP, 1993]).

**Proposizione 4.52.** Per $0 < K \in C^1([a,b])$ e $0 \leq q \in C^0([a,b])$, il problema (42) ammette una sequenza $0 < \lambda_1 < \lambda_2 < \dots$ tendente a infinito di autovalori, le cui corrispondenti autofunzioni $u_1, u_2, \dots$ sono ortogonali in $L^2(a,b)$. Ciascun autospazio ha dimensione 1. Inoltre $u_\ell$ ha esattamente $\ell - 1$ zeri nell'intervallo aperto $(a,b)$ e ogni $w \in L^2(a,b)$ può essere rappresentata dalla sua serie di Fourier

$$w(x) = \sum_{\ell=1}^{\infty} \left( \frac{\int_a^b w\overline{u}_\ell\,\mathrm{d}x}{\int_a^b u_\ell\overline{u}_\ell\,\mathrm{d}x} \right) u_\ell(x).$$

[5] Ricordiamo che $\underline{\underline{\mathbf{M}}} \in \mathbb{C}^{n\times n}$ è autoaggiunta se $\underline{\underline{\mathbf{M}}} = \underline{\underline{\mathbf{M}}}^H$, dove $^H$ denota la trasposta coniugata.
Esercizio: dimostrare che $\underline{\underline{\mathbf{M}}}$ è autoaggiunta se e solo se per ogni $\vec{\mathbf{u}}, \vec{\mathbf{w}} \in \mathbb{C}^n$ vale $\vec{\mathbf{w}}^H \underline{\underline{\mathbf{M}}}\vec{\mathbf{u}} = (\underline{\underline{\mathbf{M}}}\vec{\mathbf{w}})^H \vec{\mathbf{u}}$.
Per analogia, diciamo che l'operatore $\mathcal{L}$ è autoaggiunto se $\int_a^b (\mathcal{L}u)\overline{w}\,\mathrm{d}x = \int_a^b u\overline{\mathcal{L}w}\,\mathrm{d}x$ per ogni $u, w \in C^2_0([a,b])$.

Possiamo discretizzare il problema (42) con il metodo delle differenze finite, che ben conosciamo nel caso $K(x) = 1$, calcolando autovalori e autovettori della matrice $\underline{\underline{\mathbf{A}}}$ che si ottiene.

**Nota 4.53** (Operatori definiti positivi)**.** Sappiamo che una matrice quadrata è definita positiva se $\mathbf{v}^\top \underline{\underline{\mathbf{M}}}\mathbf{v} > 0$ per ogni vettore $\mathbf{v} \neq \mathbf{0}$ (nel caso complesso, se $\Re\{\overline{\mathbf{v}}^\top \underline{\underline{\mathbf{M}}}\mathbf{v}\} > 0$). Inoltre, se la matrice è reale simmetrica (oppure complessa autoaggiunta), è definita positiva se e solo se tutti i suoi autovalori sono positivi.

La proprietà analoga per un operatore $\mathcal{L}$ che agisce su funzioni in $(a, b)$ è che $\int_a^b (\mathcal{L}w)\overline{w} > 0$ per ogni funzione $w \neq 0$. Abbiamo visto che per l'operatore $\mathcal{L} : u \mapsto -(Ku')' + qu$ vale $\int_a^b (\mathcal{L}w)\overline{w} > 0$ per ogni funzione $w \neq 0$ in $C_0^2([a, b])$. Inoltre è autoaggiunto e ha autovalori positivi. Quindi diciamo che l'operatore è "definito positivo". Questo motiva l'uso del segno meno davanti alla derivata seconda nelle equazioni differenziali incontrate finora.

**Nota 4.54.** La teoria di Sturm–Liouville spiega cosa succede ai problemi al bordo come (23) quando il termine di reazione $q$ è negativo, caso che abbiamo escluso dall'analisi nelle sezioni precedenti. Assumiamo di avere condizioni al bordo di Dirichlet omogenee $u(a) = u(b) = 0$. L'equazione differenziale $-u'' + qu = f$ con $q \in C^0([a, b])$ generico si può scrivere come $-u'' + (\tilde{q} - \lambda)u = f$ per qualche funzione $\tilde{q} \geq 0$ e una costante $\lambda \geq 0$. Per la Proposizione 4.52 esistono infiniti valori di $\lambda$ per cui il problema omogeneo (con $f = 0$) ammette soluzioni non nulle, quindi per cui il problema è mal posto. Per tutti gli altri valori di $\lambda$ invece il problema al bordo sarà ben posto (ricordare il ragionamento fatto in §2.2.3). Il metodo delle differenze finite applicato a questo problema è in generale instabile e la sua matrice può essere singolare.

Possiamo osservare questo fatto dal risultato dell'Esercizio 4.16, mostrato in Figura 18. La norma dell'inversa della matrice delle differenze finite sull'intervallo $(a, b) = (0, 1)$ per $q < 0$ costante diventa molto grande (quindi il metodo è poco stabile) per $q = -(\ell\pi)^2$, cioè in corrispondenza degli autovalori del problema. L'instabilità del problema discreto segnala che siamo "nelle vicinanze" di un problema al bordo mal posto.

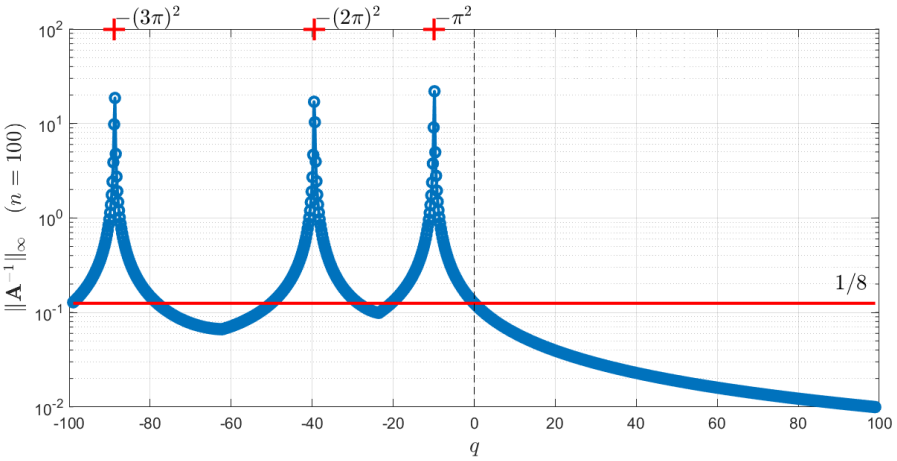



*Figura 18: Il grafico richiesto nell'Esercizio 4.16: la norma $\|\underline{\underline{\mathbf{A}}}^{-1}\|_\infty$ per la matrice delle differenze finite a $n = 100$ nodi per il problema di Dirichlet sull'intervallo $(0, 1)$ con $q$ costante, al variare di $q$ da $-100$ a 100. La norma esplode in vicinanza degli autovalori (segni + rossi), cioè $q = -(\pi\ell)^2$, mentre è maggiorata uniformemente da $1/8$ (linea orizzontale rossa) per $q \geq 0$. Questo conferma quanto detto nella Nota 4.54. Notare che il grafico è in scala semilogaritmica.*

**Esercizio ✎ 4.55** (Potenziali simmetrici)**.** Assumiamo che il problema di Sturm–Liouville (42) sia simmetrico rispetto all'origine, cioè sia $a = -b$, $K(-x) = K(x)$, $q(-x) = q(x)$. Usare la Proposizione 4.52 per dimostrare che ciascuna autofunzione $u_\ell$ o è una funzione pari o una dispari. Più precisamente le $u_\ell$ di indice $\ell$ dispari sono funzioni pari e viceversa.

**Esercizio 💻 4.56** (Potenziali a pozzo e localizzazione delle autofunzioni)**.**

- Calcolare numericamente le prime autofunzioni di (42) con $a = -2, b = 2$, $K(x) = 1$ e $q(x) = q_* > 0$ se $|x| > 1$ e $q(x) = 0$ altrimenti. Scegliere ad esempio $q_* = 10^3$.

Le autofunzioni saranno localizzate nella parte del dominio in cui $q$ è zero; vedere Figura 19. [6]

- Disegnare il grafico del "potenziale a doppio pozzo" $q(x) = (x^2 - 1)^2$. Considerate le autofunzioni dell'esempio precedente, provare a prevedere le proprietà qualitative delle prime autofunzioni $-u'' + cq(x)u = \lambda u$ nell'intervallo $(a, b) = (-2, 2)$, al variare di un parametro $c > 0$, e verificarle con il metodo delle differenze finite.
- Come si comportano le autofunzioni di un problema il cui potenziale $q$ ha pozzi di diversa profondità?

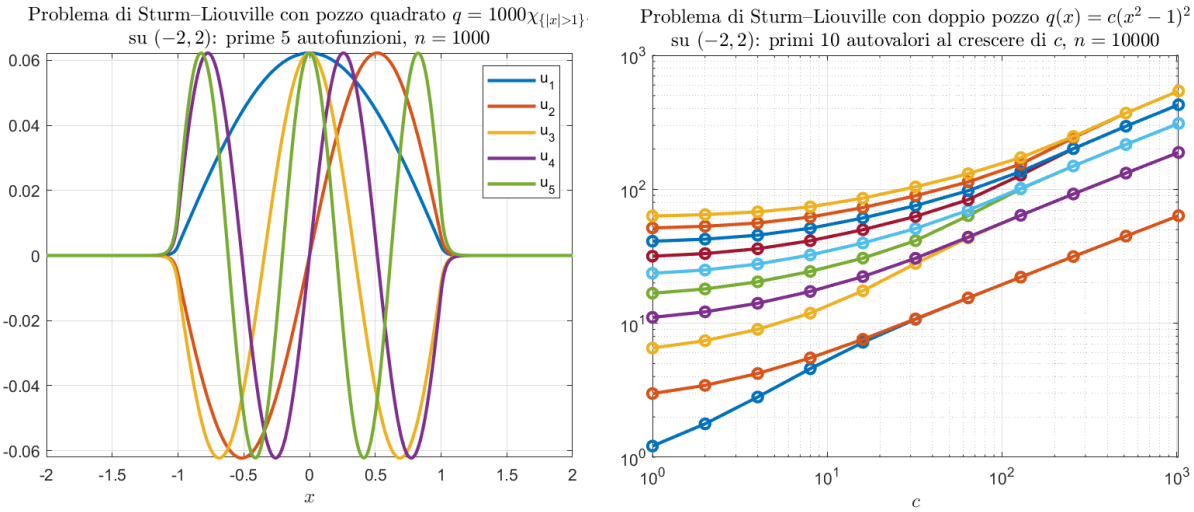



*Figura 19: Grafici relativi all'Esercizio 4.56.*
*Sinistra: Le prime 5 autofunzioni di $-u'' + qu = \lambda u$ su $(-2, 2)$ con $q = 0$ in $(-1, 1)$ e $q = 10^3$ altrimenti. Destra: i primi autovalori per il potenziale $q(x) = c(x^2 - 1)^2$ al variare di $c = 2^0, 2^1, \ldots, 2^{10}$ (in scala logaritmica). Si nota che, nonostante per ogni valore di $c$ gli autovalori siano tutti distinti, per valori grandi di $c$ tendono a coincidere a 2 a 2: $\lambda_{2\ell-1} \approx \lambda_{2\ell}$. Una delle due autofunzioni corrispondenti è pari e l'altra dispari. Per $c$ grande, il metodo delle differenze finite non è in grado di distinguere i due autospazi associati ad autovalori molto vicini e spesso calcola autofunzioni localizzate in un solo pozzo.*

**Esercizio 💻 4.57** (Per i più motivati — Localizzazione di Anderson)**.** Negli esempi dell'Esercizio 4.56 le prime autofunzioni di $-u'' + qu = \lambda u$ si concentrano nel "pozzo di potenziale", cioè dove il termine di reazione (o potenziale) $q$ è più piccolo. In questo caso basta osservare $q$ per prevedere dove si localizzeranno le autofunzioni, cioè dove saranno significativamente diverse da zero. Se il potenziale è più complicato non è facile prevedere la regione di localizzazione osservando $q$.

[6] È possibile verificare la localizzazione delle prime autofunzioni nel pozzo in modo analitico. Sia $u$ un'autofunzione per questo problema e $\lambda$ il suo autovalore. L'Esercizio 4.55 garantisce che $u$ è una funzione pari o una dispari: assumiamo sia pari, il caso dispari si studia in modo analogo.

Chiamiamo $I_{in} = (-1, 1)$ e $I_{out} = (-2, 2) \setminus [-1, 1]$ le due regioni in cui $q$ è costante. Su $I_{in}$ avremo $-u'' = \lambda u$, quindi $u(x) = \cos(kx)$ per qualche $k > 0$ e $\lambda = k^2$. In $I_{out}$ avremo $-u'' = (k^2 - q_*)u$. Per autovalori grandi, cioè $\lambda = k^2 > q_*$, avremo soluzioni oscillanti con frequenze più alte in $I_{in}$ e più basse in $I_{out}$. Se $q_*$ è abbastanza grande, gli autovalori più piccoli per cui $0 < \lambda = k^2 < q_*$ sono quelli che danno autofunzioni localizzate nel pozzo; scegliamo uno di questi $\lambda$.

In questo caso chiamiamo $\mu = \sqrt{q_* - k^2} > 0$, così che $-u'' + \mu^2 u = 0$ in $I_{out}$ e $u$ è combinazione lineare di $e^{\mu x}$ e $e^{-\mu x}$ in $(1, 2)$. La condizione al bordo $u(2) = 0$ impone $u(x) = C \sinh(\mu(2 - x))$ in $(1, 2)$, per qualche $C \in \mathbb{R}$. Stiamo considerando la soluzione di un problema al bordo (agli autovalori) con un coefficiente discontinuo: non sappiamo ancora dare un significato preciso a questa equazione, per ora ci basti sapere che la soluzione deve essere di classe $C^1$. La continuità di $u$ e $u'$ in $x = 1$ ci permette di eliminare $C$ e di scrivere una relazione che lega $k$ e $\mu$:

$$\cos k = C \sinh \mu, \quad -k \sin k = -\mu C \cosh \mu, \qquad \Rightarrow \qquad k \tan k \tanh \mu = \mu.$$

Le intersezioni di questa curva (composta da infinite componenti) nel piano $k\mu$ con la circonferenza $\mu^2 + k^2 = q_*$ (e con $k > 0$, $\mu > 0$) forniscono gli autovalori $\lambda = k^2$ a cui è associata $u(x) = \begin{cases} \cos(kx) & x \in I_{in}, \\ \frac{\cos k}{\sinh \mu} \sinh(\mu(2 - |x|)) & x \in I_{out}. \end{cases}$ Questa è una funzione che oscilla in $I_{in}$ e decade esponenzialmente in $I_{out}$, come desiderato. Studiando la relazione tra $k$ e $\mu$, si vede che il numero di autofunzioni di questo tipo è $\mathcal{O}(\sqrt{q_*})$.

In meccanica quantistica queste autofunzioni sono dette *bound states* e corrispondono a particelle a bassa energia ($E \sim \lambda < q_*$) confinate in una buca di potenziale; le code esponenziali dell'autofunzione rappresentano la probabilità della particella di sfuggire dalla buca per effetto tunnel.

Per esercizio verificare i calcoli fatti, plottare le curve nel piano $k\mu$ (ad esempio per $q_* = 1000$) e individuare gli autovalori, estendere il ragionamento alle autofunzioni dispari, e verificare se i risultati ottenuti con il metodo delle differenze finite sono coerenti con questi fatti.

- Usando il metodo delle differenze finite, calcolare e plottare alcune autofunzioni $-u'' + qu = \lambda u$ sull'intervallo $(0,1)$ dove $q$ è ottenuto generando un numero casuale in ciascun nodo $x_j$. Ad esempio, fissare il vettore dei valori di $q$ come `q = n*n*rand(n,1);`, cioè valori distribuiti uniformemente in $[0, n^2]$, dove $n$ è il numero dei nodi. Scegliere ad esempio $n = 300$.

Si vede che le autofunzioni sono localizzate, cioè valgono quasi zero in gran parte del dominio, mentre il potenziale $q$ non ha pozzi ben definiti. Come prevedere in quale parte del dominio si concentrano? Per rispondere fissiamo un'autofunzione $u$, cioè $-u'' + qu = \lambda u$, $u(0) = u(1) = 0$ e definiamo $w$ soluzione del problema al bordo con sorgente costante $-w'' + qw = 1$, $w(0) = w(1) = 0$. Allora $g := u - \lambda \|u\|_{L^\infty(0,1)} w$ soddisfa $g(0) = g(1) = 0$ e $-g'' + qg \le 0$, quindi per il principio del massimo (Corollario 2.10) vale $g \le 0$ in $(0,1)$. (Verificare questi passaggi.) Manipolando questa disuguaglianza otteniamo il seguente fatto: dato $0 \le q \in C^0(0,1)$, $u, w \in C^2_0([0,1])$ con

$$\begin{aligned} -u'' + qu &= \lambda u, & u(0) = u(1) = 0, \\ -w'' + qw &= 1, & w(0) = w(1) = 0, \end{aligned} \quad \Rightarrow \quad \frac{|u(x)|}{\lambda \|u\|_{L^\infty(0,1)}} \le w(x).$$

A parole: ciascuna autofunzione $u$, normalizzata in modo tale da avere norma $L^\infty(0,1)$ pari a $1/\lambda$ è maggiorata in ogni punto dalla soluzione $w$ del problema al bordo con sorgente costante $f = 1$. La funzione $w$ è detta "*landscape function*".

- Calcolare $w$ soluzione di $-w'' + qw = 1$, per lo stesso $q$ come sopra, con il metodo delle differenze finite. Mostrare che le prime autofunzioni sono effettivamente maggiorate da $w$ come descritto. Vedere Figura 20.

Si nota anche che i massimi locali più alti di $w$ individuano i supporti delle prime autofunzioni, quindi $w$ permette di stimare la posizione delle prime autofunzioni.

La localizzazione delle autofunzioni con un potenziale "disordinato" è un importante fenomeno fisico detto "localizzazione di Anderson". La funzione di landscape è stata scoperta nel 2012 da Filoche e Mayboroda (https://doi.org/10.1073/pnas.1120432109) ed è già stata usata ad esempio nella progettazione di LED.

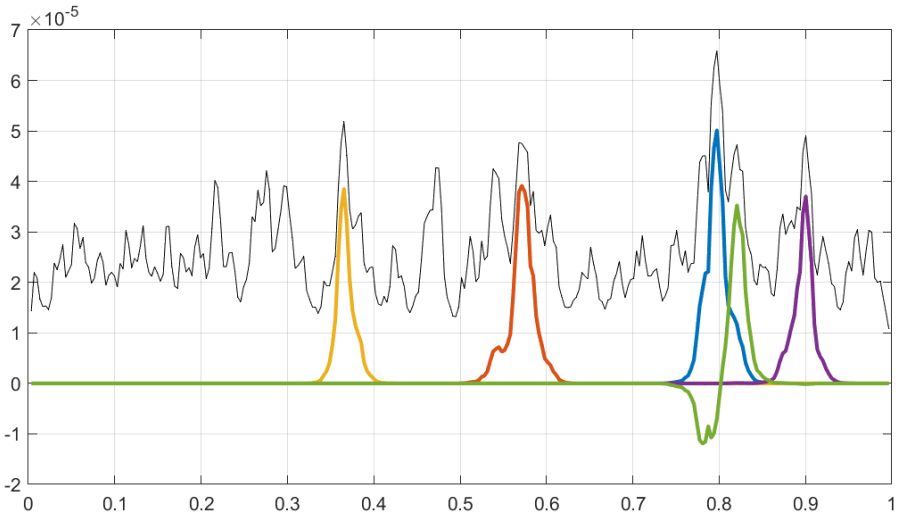



*Figura 20: (Esercizio 4.57.) Le prime 5 autofunzioni di $-u'' + qu = \lambda u$, dove $q$ prende valori casuali uniformemente distribuiti in $[0, n^2]$ in ciascuno degli $n$ nodi. Qui $n = 300$. In nero la funzione di landscape $w$. Le autofunzioni sono normalizzata in modo che $\|u\|_{L^\infty(0,1)} = 1/\lambda$ così che sono maggiorate da $w$.*

**Esercizio ✎ 4.58.** I problemi agli autovalori sono strettamente legati ai problemi al bordo visti nelle sezioni precedenti. Consideriamo il problema al bordo con condizioni di Dirichlet omogenee

$$\begin{cases} -u'' + qu = f & \text{in } (a,b), \\ u(a) = u(b) = 0. \end{cases}$$

Immaginiamo di conoscere gli autovalori $\lambda_\ell$ e le autofunzioni $u_\ell$ di $\mathcal{L} = -\frac{\partial^2}{\partial x^2} + q$. Proposizione 4.52 permette di espandere il dato $f$ come $f(x) = \sum_{\ell=1}^\infty \hat{f}_\ell u_\ell(x)$ per dei coefficienti $\hat{f}_\ell$.

- Verificare che la soluzione $u$ del problema al bordo si può scrivere come $u = \sum_{\ell=1}^\infty \frac{\hat{f}_\ell}{\lambda_\ell} u_\ell$.

  Qual è l'analogo finito-dimensionale, cioè in termini di algebra lineare?

- Definiamo $G(x,y) := \sum_{\ell=1}^{\infty} \frac{1}{\lambda_\ell} \frac{u_\ell(x)\overline{u}_\ell(y)}{\int_a^b |u_\ell|^2}$ per $x, y \in (a,b)$.

  Mostrare (solo formalmente!) che $u(x) = \int_a^b G(x,y) f(y)\,\mathrm{d}y$ è soluzione del problema al bordo.

  (Ignorare il fatto che non è ovvio che le condizioni al bordo siano soddisfatte. L'espressione in Proposizione 4.52 permette di scrivere qualsiasi funzione di $L^2(a,b)$ come somma di funzioni che valgono $0$ agli estremi; $G$ invece vale zero perché il fattore $\frac{1}{\lambda_\ell} \xrightarrow{\ell\to\infty} 0$ garantisce una convergenza più forte.)

  Da questo deduciamo che $G(x,y)$ è la funzione di Green del problema ed estende quanto detto in §2.2.4 per il caso $q = 0$.
- Plottare la funzione $G$ troncando la serie, nel caso $q = 0$, $(a,b) = (0,1)$, e confrontare con Figura 6.

**Nota 4.59.** Perché è rilevante studiare i problemi agli autovalori per operatori differenziali (e quindi i metodi numerici per approssimarli)? Diamo alcune motivazioni molto generali (e vaghe!), provenienti dalla matematica, dalla fisica e dall'ingegneria.

- Come visto nella Proposizione 4.52, è possibile *approssimare qualsiasi funzione* con una somma di autofunzioni. Questo è utile ad esempio nella teoria dell'approssimazione e nella tecnica della "*separazione delle variabili*": vedremo un esempio importante nella sezione §7.2. Abbiamo già sfruttato la possibilità di espandere funzioni in basi di autofunzioni nell'Esercizio 4.58. L'approssimazione di funzioni con serie di funzioni trigonometriche, cioè autofunzioni di $-\frac{\partial^2}{\partial x^2}$, è alla base della teoria delle serie di Fourier.
- Gli autovalori ci danno informazioni sulla *stabilità* dei problemi di evoluzione. Ricordiamo il semplice problema di Cauchy per un sistema di equazioni differenziali lineari omogenee:

$$\vec{\mathbf{y}}'(t) = \underline{\underline{\mathbf{A}}}\vec{\mathbf{y}}(t), \quad \vec{\mathbf{y}}(0) = \vec{\mathbf{y}}_0, \tag{43}$$

  per una matrice $\underline{\underline{\mathbf{A}}} \in \mathbb{R}^{m\times m}$ invertibile e un dato iniziale $\vec{\mathbf{y}}_0 \in \mathbb{R}^m$. Sappiamo che l'unica soluzione stazionaria del sistema è $\vec{\mathbf{y}}(t) = \mathbf{0}$, e che questo è un equilibrio stabile o meno a seconda del segno della parte reale degli autovalori di $\underline{\underline{\mathbf{A}}}$. Per problemi non lineari $\vec{\mathbf{y}}(t) = \mathbf{G}(\vec{\mathbf{y}}(t))$ la stabilità delle soluzioni stazionarie $\vec{\mathbf{y}}^*$ con $\mathbf{G}(\vec{\mathbf{y}}^*) = \mathbf{0}$ viene studiata attraverso gli autovalori della Jacobiana di $\mathbf{G}$. In modo simile, per un'equazione alle derivate parziali $\frac{\partial u}{\partial t}(\mathbf{x},t) = \mathcal{L}u(\mathbf{x},t)$, dove $\mathcal{L}$ è un operatore differenziale nella variabile spaziale $\mathbf{x}$ come quelli visti in §2.1 o in §4.7.1, la stabilità di una soluzione stazionaria (indipendente da $t$) può essere analizzata a partire dagli autovalori di $\mathcal{L}$ o di una sua linearizzazione.
- Nella *meccanica quantistica*, le autofunzioni di opportuni operatori differenziali (detti Hamiltoniani, spesso nella forma che già conosciamo $u \mapsto -\Delta u + qu$, dove $q$ è chiamato potenziale) rappresentano gli stati stazionari delle particelle associate a questi operatori; si veda [TBD18, pp. 70–73]. La *spettroscopia* studia le transizioni tra stati di energia differenti: la lunghezza d'onda di una radiazione emessa o assorbita è inversamente proporzionale alla differenza tra due livelli energetici, corrispondenti a due autovalori di un operatore differenziale.
- Nelle applicazioni ingegneristiche gli autovalori di un operatore hanno spesso a che vedere con il concetto di *risonanza*, cioè la risposta amplificata di un sistema a certi input. Introduciamo questo concetto a partire da un esempio noto. Consideriamo il sistema di equazioni differenziali (43). Se la matrice $\underline{\underline{\mathbf{A}}} \in \mathbb{C}^{m\times m}$ è diagonalizzabile, avremo una base $\{\vec{\mathbf{w}}_1, \dots, \vec{\mathbf{w}}_m\}$ di autovettori, corrispondenti agli autovalori $\lambda_1, \dots, \lambda_m$. Espandendo il dato iniziale come $\vec{\mathbf{y}}_0 = \sum_{j=1}^m \hat{y}_j \vec{\mathbf{w}}_j$, la soluzione del problema di Cauchy è

$$\vec{\mathbf{y}}(t) = \sum_{j=1}^m \hat{y}_j \vec{\mathbf{w}}_j \mathrm{e}^{\lambda_j t} = \sum_{j=1}^m \hat{y}_j \vec{\mathbf{w}}_j \mathrm{e}^{\Re\lambda_j t}\big(\cos(\Im\lambda_j t) + \mathrm{i}\sin(\Im\lambda_j t)\big), \qquad \lambda_j = \Re\lambda_j + \mathrm{i}\Im\lambda_j.$$

  Se tutti gli autovalori hanno parte reale $\Re\lambda_j \le 0$ e solo uno ha $\Re\lambda_{j_*} = 0$, dopo un certo intervallo di tempo tutti contributi per $j \neq j_*$ saranno trascurabili, per via del decadimento esponenziale, e avremo $\vec{\mathbf{y}}(t) \approx \hat{y}_{j_*}\vec{\mathbf{w}}_{j_*}(\cos(\Im\lambda_j t) + \mathrm{i}\sin(\Im\lambda_j t))$. Cioè l'unico contributo del dato iniziale $\vec{\mathbf{y}}_0$ che permane nella soluzione per tempi lunghi è quello dato da $\hat{y}_{j_*}\vec{\mathbf{w}}_{j_*}$, oscillante con frequenza angolare $\omega = \Im\lambda_{j_*}$. Lo stesso succede per equazioni a derivate parziali del tipo $\frac{\partial u}{\partial t}(\mathbf{x},t) = \mathcal{L}u(\mathbf{x},t)$, per cui le soluzioni sono dominate dalle autofunzioni dell'operatore differenziale in spazio $\mathcal{L}$ corrispondenti agli autovalori di parte reale più grande. (Più spesso le risonanze sono definite a partire dal termine noto $\vec{\mathbf{f}}$ del sistema lineare $\vec{\mathbf{y}}' = \underline{\underline{\mathbf{A}}}\vec{\mathbf{y}} + \vec{\mathbf{f}}$, invece che dalle condizioni iniziali $\vec{\mathbf{y}}_0$.)

  In alcune applicazioni vogliamo sistemi che oscillano a frequenze precise come nel caso di strumenti musicali o di antenne. (Si veda [TBD18, Ex. 6.3] per l'applicazione a flauti e clarinetti.) In altri casi invece vogliamo prevenire la presenza di risonanze: ad esempio vogliamo che tutte le vibrazioni elastiche in un edificio, un ponte o una struttura simile siano attenuate il più velocemente possibile.

## 4.8 Differenze finite per problemi non lineari

Finora abbiamo applicato il metodo delle differenze finite a equazioni differenziali lineari. Vediamo come estendere quello che abbiamo imparato a un caso semplice di equazione non lineare, l'equazione del pendolo vista nella Sezione 1.2.2. Consideriamo il problema già visto:

$$\begin{cases} u'' = -\sin u, \\ u(a) = \alpha, \qquad u(b) = \beta. \end{cases}$$

Sostituendo la derivata con le differenze finite centrate e usando la solita notazione abbiamo $n$ equazioni

$$G_j(\vec{\mathbf{U}}) := \frac{U_{j-1} - 2U_j + U_{j+1}}{h^2} + \sin U_j = 0, \quad j = 1, \dots, n, \qquad U_0 = \alpha, \quad U_{n+1} = \beta. \tag{44}$$

Questo è un sistema non lineare di $n$ incognite, che possiamo scrivere in breve come $\vec{\mathbf{G}}(\vec{\mathbf{U}}) = \vec{\mathbf{0}}$ per una $\vec{\mathbf{G}} : \mathbb{R}^n \to \mathbb{R}^n$. Non possiamo usare metodi come quello di bisezione per calcolarne gli zeri poiché siamo in dimensione maggiore di 1, quindi usiamo il metodo di Newton. Dato un vettore iniziale $\vec{\mathbf{U}}^0 \in \mathbb{R}^n$, le iterate di Newton sono

$$\vec{\mathbf{U}}^{k+1} = \vec{\mathbf{U}}^k - \left(\underline{\underline{\mathbf{J}}}\vec{\mathbf{G}}(\vec{\mathbf{U}}^k)\right)^{-1} \vec{\mathbf{G}}(\vec{\mathbf{U}}^k) \qquad k = 0, 1, \dots$$

dove $\underline{\underline{\mathbf{J}}}\vec{\mathbf{G}}(\vec{\mathbf{w}})$ è la matrice Jacobiana di $\vec{\mathbf{G}}$ calcolata in $\vec{\mathbf{w}}$, cioè la matrice con $(\underline{\underline{\mathbf{J}}}\vec{\mathbf{G}})_{i,j}(\vec{\mathbf{w}}) = \frac{\partial}{\partial w_j} G_i(\vec{\mathbf{w}})$ (la riga $i$esima di $\underline{\underline{\mathbf{J}}}\vec{\mathbf{G}}$ è il gradiente di $G_i$). Poiché $\vec{\mathbf{G}}$ è data da (44), vediamo che

$$\underline{\underline{\mathbf{J}}}\vec{\mathbf{G}}(\vec{\mathbf{U}}) = \begin{pmatrix} -\frac{2}{h^2} + \cos U_1 & \frac{1}{h^2} & & & & \\ \frac{1}{h^2} & -\frac{2}{h^2} + \cos U_2 & \frac{1}{h^2} & & & \\ & \frac{1}{h^2} & -\frac{2}{h^2} + \cos U_3 & \frac{1}{h^2} & & \\ & & \ddots & \ddots & \ddots & \\ & & & \frac{1}{h^2} & -\frac{2}{h^2} + \cos U_{n-1} & \frac{1}{h^2} \\ & & & & \frac{1}{h^2} & -\frac{2}{h^2} + \cos U_n \end{pmatrix}.$$

Ogni iterazione del metodo consiste nella valutazione di $\vec{\mathbf{G}}$ e $\underline{\underline{\mathbf{J}}}\vec{\mathbf{G}}$ in $\vec{\mathbf{U}}^k$ e nella soluzione di un sistema lineare con matrice (tridiagonale) $\underline{\underline{\mathbf{J}}}\vec{\mathbf{G}}(\vec{\mathbf{U}}^k)$ e termine noto $\vec{\mathbf{G}}(\vec{\mathbf{U}}^k)$.

Il metodo proposto può ovviamente essere esteso a equazioni più generali $-u''(x) = f(x, u(x), u'(x))$, approssimando $f(x_j, u(x_j), u'(x_j))$ con $f(x_j, U_j^k, \frac{U_{j+1} - U_{j-1}}{2h})$.

**Esercizio 💻 4.60.** Implementare il metodo di Newton per questo problema con i dati scelti in Sezione 1.2.2 ($u(0) = u(2\pi) = \pi/3$).

La funzione $\vec{\mathbf{G}} : \mathbb{R}^n \to \mathbb{R}^n$ e la funzione $\underline{\underline{\mathbf{J}}}\vec{\mathbf{G}} : \mathbb{R}^n \to \mathbb{R}^{n \times n}$ possono essere implementate come `function` oppure come funzioni "anonime", (`G = @(U) ...`). Ancora una volta il comando `spdiags` è d'aiuto.

Scegliendo come dato iniziale $\vec{\mathbf{U}}^0 = \vec{\mathbf{0}}$, il metodo converge velocemente a una delle soluzioni. Individuare due scelte iniziali che portano a convergere a due soluzioni distinte, come quelle in Figura 5.

Altri semplici problemi al bordo che ammettono più soluzioni distinte si possono trovare in [TBD18, §16], tra cui quelli relativi alle traiettorie che permettono di inviare un'astronave in un certo punto nello spazio in presenza di uno o più corpi celesti.

**Nota 4.61.** Dall'iterazione di Newton e dall'espressione di $\vec{\mathbf{G}}$ e $\underline{\underline{\mathbf{J}}}\vec{\mathbf{G}}$ può non essere evidente come i valori al bordo $\alpha$ e $\beta$ entrano nell'algoritmo. Ricordando che $U_0 = \alpha$, il primo elemento del vettore $\vec{\mathbf{G}}(\vec{\mathbf{U}})$ è $(\vec{\mathbf{G}}(\vec{\mathbf{U}}))_1 = \frac{\alpha - 2U_1 + U_2}{h^2} + \sin U_1$; similmente $\beta$ compare in $G_n$.

**Esercizio ✎+💻 4.62.** Vogliamo usare il metodo delle differenze finite, combinato con quello di Newton, per approssimare la soluzione del problema ai limiti dell'Esercizio 1.6 ($u'' = u^3 - uu'$, $u(1) = \frac{1}{2}$, $u(3) = \frac{1}{4}$). Discretizzeremo la derivata prima di $u$ con differenze finite centrate.

- Scrivere (a mano) la funzione $\vec{\mathbf{G}} : \mathbb{R}^n \to \mathbb{R}^n$ di cui cercheremo gli zeri e la sua Jacobiana $\underline{\underline{\mathbf{J}}}\vec{\mathbf{G}}$.
- Implementare il metodo e plottare le soluzioni numeriche.

  ($\vec{\mathbf{G}}$ e $\underline{\underline{\mathbf{J}}}\vec{\mathbf{G}}$ possono essere scritti in Matlab come funzioni anonime ma richiedono molta attenzione.)

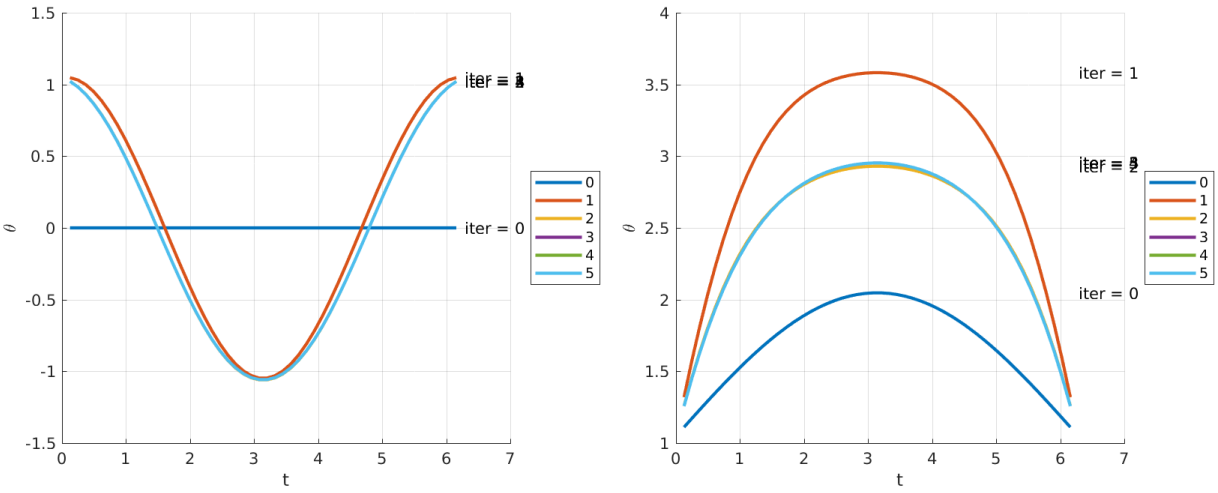



*Figura 21: Le prime 5 iterazioni di Newton per il problema del pendolo $u'' = -\sin u$ con $u(0) = u(2\pi) = \pi/3$. L'intervallo è stato discretizzato con $n = 50$ punti. Per diverse scelte iniziali $(\vec{\mathbf{U}}^0)_j = 0$ e $(\vec{\mathbf{U}}^0)_j = \sin(x_j/2) + \pi/3$ il metodo converge a diverse soluzioni isolate dello stesso problema ai limiti. (La legenda si può generare con* `legend(num2str((0:MaxIt)'))`*, dove* `MaxIt` *è il numero delle iterazioni.)*

- Calcolare l'errore nei nodi, ricordando che $u(x) = 1/(1+x)$. Plottare i valori del residuo $\|\vec{\mathbf{G}}(\vec{\mathbf{U}}^k)\|_\infty$, dell'incremento $\|\vec{\mathbf{U}}^k - \vec{\mathbf{U}}^{k-1}\|_\infty$ e dell'errore $\|\vec{\mathbf{U}}^k - \vec{\mathbf{u}}\|_\infty$ in dipendenza da $k$.

Scegliendo $\vec{\mathbf{U}}^0 = \vec{\mathbf{0}}$ come vettore iniziale e $n = 50$ nodi, bastano 5 iterazioni di Newton per ottenere $\|\vec{\mathbf{G}}(\vec{\mathbf{U}}^5)\|_\infty \approx 6 \cdot 10^{-14}$ e $\|\vec{\mathbf{U}}^5 - \vec{\mathbf{U}}^4\|_\infty \approx 3 \cdot 10^{-15}$. Tuttavia, dopo la quarta iterazione l'errore $\|\vec{\mathbf{U}}^k - \vec{\mathbf{u}}\|_\infty$ (con $u_j = u(x_j) = 1/(1+x_j)$, $j = 1, \ldots, n$) non decresce ma si stabilizza intorno a $4 \cdot 10^{-6}$.

Come possiamo misurare l'accuratezza della soluzione ottenuta? Dei buoni criteri di arresto per il metodo di Newton sono la lunghezza $\|\vec{\mathbf{U}}^{k+1} - \vec{\mathbf{U}}^k\|$ del passo $k$-simo e il residuo $\|\vec{\mathbf{G}}(\vec{\mathbf{U}}^k)\|$. Nell'esempio del pendolo (con $n = 50$ e $\vec{\mathbf{U}}^0 = \vec{\mathbf{0}}$) la quinta e la sesta iterata di Newton differiscono in norma infinito di un fattore $\approx 2 \cdot 10^{-15}$ e $\|\vec{\mathbf{G}}(\vec{\mathbf{U}}^5)\|_\infty \approx 1 \cdot 10^{-14}$. Questo è una buona stima dell'errore commesso? No: questo misura quanto accuratamente abbiamo risolto il sistema non lineare $\vec{\mathbf{G}}(\vec{\mathbf{U}}) = \vec{\mathbf{0}}$ ma non dice niente sull'errore commesso dalla discretizzazione alle differenze finite. (L'equivalente nel caso lineare sarebbe l'errore commesso nella soluzione del sistema lineare, che abbiamo sempre assunto essere trascurabile.) L'errore $\vec{\mathbf{e}} = \vec{\mathbf{u}} - \vec{\mathbf{U}}$ non ha questa precisione (ricordiamo che nella notazione di §4.3 $\vec{\mathbf{u}}$ e $\vec{\mathbf{U}}$ sono i vettori dei valori nei nodi della soluzione esatta e di quella discreta, rispettivamente).

Nel caso lineare abbiamo studiato il troncamento, rappresentato dal vettore $\vec{\mathbf{T}} = \underline{\underline{\mathbf{A}}}\vec{\mathbf{u}} - \vec{\mathbf{B}} = \vec{\mathbf{A}}\vec{\mathbf{e}}$. Nel caso non lineare considerato, definiamo il troncamento come $\vec{\mathbf{T}} := \vec{\mathbf{G}}(\vec{\mathbf{u}})$, cioè il valore in (44) calcolato usando i valori $u_j = u(x_j)$ della soluzione esatta nei nodi. Gli elementi di $\vec{\mathbf{T}}$ sono

$$T_j = \frac{u_{j-1} - 2u_j + u_{j+1}}{h^2} + \sin u_j = u''(x_j) + \frac{h^2}{12}u^{(iv)}(\xi) + \sin u(x_j) = \frac{h^2}{12}u^{(iv)}(\xi), \qquad \xi \in (x_{j-1}, x_{j+1}),$$

per $j = 1, \ldots, n$, dove abbiamo usato l'errore di troncamento (22) e l'equazione differenziale $u'' = -\sin u$. Per $u$ sufficientemente liscia, il metodo ha un **errore di troncamento quadratico in** $h$.

Ci aspettiamo che l'errore $\vec{\mathbf{e}}$ converga quadraticamente se il metodo è "stabile". Assumiamo che $\vec{\mathbf{U}}$ sia la soluzione esatta dell'equazione $\vec{\mathbf{G}}(\vec{\mathbf{U}}) = \vec{\mathbf{0}}$, cioè il metodo di Newton (o un altro metodo) ha raggiunto la convergenza; nell'esempio precedente questo è vero a meno di precisione macchina. Quindi $\vec{\mathbf{T}} = \vec{\mathbf{G}}(\vec{\mathbf{u}}) - \vec{\mathbf{G}}(\vec{\mathbf{U}})$. Se $\vec{\mathbf{G}}$ fosse lineare potremmo controllare $\vec{\mathbf{e}}$ come $\vec{\mathbf{e}} = \vec{\mathbf{G}}^{-1}(\vec{\mathbf{T}})$, e infatti nelle sezioni precedenti abbiamo studiato la stabilità analizzando la norma dell'inversa della matrice del metodo. Nel problema non lineare linearizziamo con il polinomio di Taylor del primo grado:

$$\vec{\mathbf{G}}(\vec{\mathbf{U}}) = \vec{\mathbf{G}}(\vec{\mathbf{u}}) + \underline{\underline{\mathbf{J}}}\vec{\mathbf{G}}(\vec{\mathbf{u}})(\vec{\mathbf{U}} - \vec{\mathbf{u}}) + \mathcal{O}\Big(\Big\|\vec{\mathbf{U}} - \vec{\mathbf{u}}\Big\|^2\Big) \quad \Rightarrow \quad \underline{\underline{\mathbf{J}}}\vec{\mathbf{G}}(\vec{\mathbf{u}})\vec{\mathbf{e}} = \vec{\mathbf{T}} + \mathcal{O}(\|\vec{\mathbf{e}}\|^2).$$

Diciamo quindi che il metodo è **stabile** in una certa norma se le matrici $(\underline{\underline{\mathbf{J}}}\vec{\mathbf{G}}(\vec{\mathbf{u}}))^{-1}$ sono maggiorate uniformemente per $h \to 0$, cioè esistono $h_0$ e $C$ positivi tali che

$$\Big\|\big(\underline{\underline{\mathbf{J}}}\vec{\mathbf{G}}(\vec{\mathbf{u}})\big)^{-1}\Big\| \le C \quad \text{per } h \le h_0.$$

Ricordiamo che per $h \to 0$ la dimensione $n$ di $\underline{\underline{\mathbf{J}}}\vec{\mathbf{G}}$ cresce proporzionalmente a $1/h$. Per completare l'analisi dovremmo studiare il termine $\mathcal{O}(\|\vec{\mathbf{e}}\|^2)$, non ce ne occuperemo qui.

**Nota 4.63** (Significati diversi di "convergenza quadratica")**.** Attenzione ad una fonte di confusione frequente. Abbiamo visto che il metodo delle differenze finite converge quadraticamente e ricordiamo dai corsi precedenti che anche il metodo di Newton converge quadraticamente. L'uso dell'espressione "convergenza quadratica" ha significato diverso nei due casi.

Per metodi come quello delle differenze finite, in cui si studia la convergenza al variare di un parametro $h$ che decresce a 0 (nel nostro caso la distanza tra i nodi), convergenza quadratica significa che l'errore è proporzionale alla seconda potenza del parametro $h$. Dimezzando $h$ (o raddoppiando $n$) l'errore diminuisce di un fattore 4.

Il metodo di Newton invece è un metodo iterativo: la sua accuratezza non migliora scegliendo il valore di un parametro, ma calcolando più iterazioni. Ogni iterazione richiede il calcolo di quelle precedenti. Si dice che un metodo iterativo converge quadraticamente se l'errore dell'iterazione $k+1$-sima è proporzionale al quadrato dell'errore dell'iterazione $k$-sima. Il numero di cifre decimali esatte raddoppia ad ogni iterazione

## 4.9 Quantificazione dell'incertezza (uncertainty quantification)

> *As far as the laws of mathematics refer to reality, they are not certain, and as far as they are certain, they do not refer to reality.* A. Einstein

Nelle applicazioni che coinvolgono la soluzione di un problema al bordo per un'equazione differenziale, molto spesso conosciamo il modello (cioè la forma dell'equazione differenziale) ma non conosciamo tutti i dati che definiscono il problema. Per "dati" qui intendiamo il dominio, i coefficienti, il termine noto, le condizioni al bordo... (nei problemi visti fin qui: $(a,b)$, $q$, $p$, $f$, $\alpha$, $\beta$). Questa incertezza può essere dovuta alla scarsa conoscenza del problema specifico: se modelliamo il flusso di una sostanza (acqua, petrolio, gas, inquinante...) nel sottosuolo non c'è modo di conoscere la permeabilità della roccia in ogni punto; se consideriamo un oggetto prodotto da un macchinario non sappiamo a priori le imprecisioni nella sua forma e nel materiale; se vogliamo prevedere l'evoluzione del meteo abbiamo a disposizione i valori passati di temperatura, pressione, vento solo nei punti in cui ci sono degli strumenti di misura. Oppure l'incertezza può essere dovuta all'intrinseca aleatorietà del fenomeno: se vogliamo prevedere come si propagherà un possibile terremoto non avremo a disposizione la posizione precisa dell'epicentro.

Per questo vogliamo considerare i problemi al bordo da un punto di vista stocastico: assumendo che i dati siano variabili aleatorie distribuite secondo una certa legge, vogliamo ricavare qualche informazione sulla distribuzione statistica delle soluzioni. Ad esempio vogliamo calcolare quali saranno il valore atteso e la varianza di certe quantità d'interesse che dipendono dalla soluzione del problema al bordo, o la probabilità che queste quantità cadano in un certo intervallo. Questo studio si chiama "**quantificazione dell'incertezza**" ("*uncertainty quantification*", UQ).

Per semplicità consideriamo un problema di diffusione e reazione molto semplice:

$$-u'' + qu = 0 \quad \text{in } (0,1), \qquad u(0) = u(1) = 1. \tag{45}$$

In questo caso la soluzione $u$ è determinata da un solo parametro aleatorio: la funzione $q \geq 0$. Al contrario, $f = 0, a = 0, b = 1, \alpha = \beta = 1$ sono stati fissati. Nel modello di diffusione–reazione, $q(x)$ può rappresentare il tasso di degradazione locale della sostanza con densità $u$. Da queste ipotesi segue che $0 \leq u(x) \leq 1$.

Sia $\boxed{(\Omega, \mathcal{F}, \mathbb{P})}$ uno **spazio di probabilità** ($\Omega$ è lo spazio degli eventi, $\mathcal{F}$ la sigma-algebra e $\mathbb{P}$ la misura di probabilità). Sia $\boxed{\vec{\mathbf{y}} : \Omega \to Y}$ una **variabile aleatoria** vettoriale (*random variable*), cioè una funzione misurabile, dove $\boxed{Y = (0,1)^d}$ è un ipercubo $d$-dimensionale. Assumiamo che la misura di probabilità $\boxed{\rho}$ indotta da $\vec{\mathbf{y}}$ in $Y$ sia quella uniforme, cioè coincida con la misura di Lebesgue (il cubo $Y$ ha misura unitaria). Stiamo dicendo che descriveremo ogni evento che può realizzarsi usando $d$ parametri $y_1, \ldots, y_d$, ognuno nell'intervallo $(0,1)$, e che ciascuno di questi valori si verifica con la stessa probabilità ed in modo indipendente. Pensiamo a $\vec{\mathbf{y}}$ semplicemente come a un vettore in $Y$ che non conosciamo. (Ovviamente molte altre scelte per $Y$ e $\rho$ sono possibili, anzi necessarie nei casi applicativi.)

Ora fissiamo $d$ funzioni non-negative e limitate $\phi_1, \ldots, \phi_d \geq 0$ definite sull'intervallo (fisico) $(0,1)$. Nel contesto dell'UQ, vogliamo modellare il coefficiente $q$ come una combinazione lineare casuale di queste funzioni: $\boxed{q(x, \vec{\mathbf{y}}) = \sum_{i=1}^{d} y_i \phi_i(x)}$. Per ogni realizzazione del parametro stocastico $\vec{\mathbf{y}} \in Y$, $x \mapsto q(x, \vec{\mathbf{y}})$ è una funzione limitata e non-negativa definita sull'intervallo $(0,1)$. In altre parole, $q$ è un "campo aleatorio", una variabile aleatoria a valori in uno spazio di funzioni. Le funzioni $\phi_i$ verranno scelte a seconda delle informazioni disponibili sul modello fisico, ad esempio a seconda di quanto ci aspettiamo che $q$ sia liscio.

Se le $\phi_i$ sono sufficientemente regolari (ad esempio continue), sappiamo che per ogni realizzazione di $q$ esiste un'unica soluzione $u$ del problema al bordo (45). Poiché $u$ dipende da $q$, dipende anche da $\vec{\mathbf{y}}$, cioè è una variabile aleatoria: scriveremo $\boxed{u(x, \vec{\mathbf{y}})}$ per esplicitare la sua dipendenza dalla posizione $x$ e dal vettore $\vec{\mathbf{y}}$ dei parametri stocastici. Ripetiamo: $u$ è la soluzione unicamente determinata di un problema al bordo, come quelli che abbiamo visto nelle sezioni precedenti, i cui dati sono aleatori, quindi anche $u$ stessa è una variabile aleatoria (a valori in uno spazio di funzioni come $C^0([0,1]))$[7].

Poiché $u$ non è deterministica, non siamo interessati a calcolarla esattamente, come funzione del parametro stocastico. Vogliamo invece calcolare (in modo approssimato) le proprietà statistiche di alcune "**quantità di interesse**" $\boxed{\mathcal{Q}(u)}$ (*quantity of interest*, QoI). Per semplicità assumiamo che $\mathcal{Q}(u)$ sia un numero reale. Esempi di quantità d'interesse sono la media integrale (in spazio) di $u$, cioè $\mathcal{Q}(u) = \int_0^1 u(x, \vec{\mathbf{y}})\,\mathrm{d}x$, il suo minimo $\mathcal{Q}(u) = \min_{x\in[0,1]} u(x, \vec{\mathbf{y}})$, la posizione del punto di minimo $\mathcal{Q}(u) = \arg\min_{x\in[0,1]} u(x, \vec{\mathbf{y}})$, il massimo della sua derivata $\mathcal{Q}(u) = \max_{x\in[0,1]} |\frac{\partial u(x,\vec{\mathbf{y}})}{\partial x}|\ldots$ Tutte queste quantità dipendono da $\vec{\mathbf{y}}$, quindi sono variabili aleatorie reali. Vorremmo approssimare le proprietà statistiche di queste quantità: valore atteso $\mathbb{E}[\mathcal{Q}(u)]$, varianza $\mathrm{Var}[\mathcal{Q}(u)]$, probabilità di eccedere una soglia $\mathbb{P}(\mathcal{Q}(u) > c)$

$$\mathbb{E}[\mathcal{Q}(u)] = \int_Y \mathcal{Q}\big(u(\cdot,\mathbf{y})\big)\,\mathrm{d}\rho(\vec{\mathbf{y}}), \quad \mathrm{Var}[\mathcal{Q}(u)] = \mathbb{E}\big(|\mathcal{Q}(u) - \mathbb{E}[\mathcal{Q}(u)]|^2\big), \quad \mathbb{P}\big(\mathcal{Q}(u) > c\big) = \int_{\{\vec{\mathbf{y}}:\mathcal{Q}(u(\vec{\mathbf{y}}))>c\}} 1\,\mathrm{d}\rho(\vec{\mathbf{y}}).$$

Nel seguito ci concentreremo solo sulla stima del valore atteso $\mathbb{E}[\mathcal{Q}(u)]$.

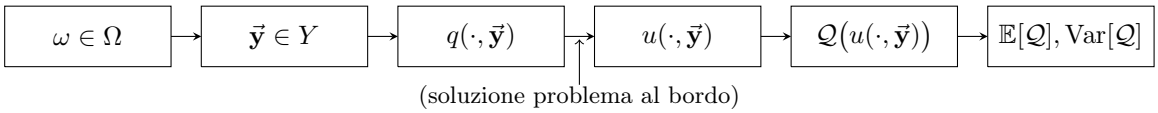



*Figura 22: Ad ogni evento elementare $\omega \in \Omega$ corrisponde una realizzazione del vettore stocastico $\vec{\mathbf{y}} \in Y = (0,1)^2$. Questo descrive un coefficiente $q(\cdot, \vec{\mathbf{y}}) = \sum_{i=1}^d y_i \phi_i$. A sua volta, a questo corrisponde la soluzione $u$ del problema al bordo. La quantità $\mathcal{Q}$ a cui siamo interessati dipende da $u$, quindi in ultima analisi da $\vec{\mathbf{y}}$ (e $\omega$) ed è quindi una variabile aleatoria. Di questa ci interessano le sue proprietà statistiche, come valore atteso e varianza.*

### 4.9.1 Metodo Monte Carlo

Come possiamo approssimare il valore atteso $\mathbb{E}[\mathcal{Q}(u)]$? C'è un modo molto semplice: generiamo una realizzazione casuale $\vec{\mathbf{y}}^{(1)}$ del vettore $\vec{\mathbf{y}}$, calcoliamo la $q$ corrispondente, risolviamo numericamente il problema al bordo associato (ad esempio con il metodo delle differenze finite), e abbiamo un'approssimazione $u_h^{(1)}$ della realizzazione $u(\cdot, \vec{\mathbf{y}}^{(1)})$ di $u$. (Usiamo il pedice $_h$ con il significato di "discreto", per ricordarci che $u_h^{(1)}$ è calcolata con il metodo delle differenze finite con passo $h$.) Calcoliamo $\mathcal{Q}(u_h^{(1)})$, ripetiamo l'operazione $M \in \mathbb{N}$ volte, generando dei vettori casuali $\vec{\mathbf{y}}^{(2)}, \ldots, \vec{\mathbf{y}}^{(M)}$, e calcoliamo la media empirica (o media campionaria) dei valori ottenuti:

$$\boxed{\mathcal{Q}^M_{\mathrm{MC}} := \frac{1}{M}\sum_{m=1}^{M} \mathcal{Q}(u_h^{(m)})} \approx \frac{1}{M}\sum_{m=1}^{M} \mathcal{Q}\big(u(\cdot, \vec{\mathbf{y}}^{(m)})\big) \approx \int_Y \mathcal{Q}\big(u(\cdot,\mathbf{y})\big)\,\mathrm{d}\rho(\vec{\mathbf{y}}) = \mathbb{E}[\mathcal{Q}(u)]. \tag{46}$$

Questo si chiama **metodo Monte Carlo**: simulare un fenomeno aleatorio generando tante realizzazioni e calcolando la media empirica della quantità d'interesse. Nella formula, il primo segno "$\approx$" indica l'approssimazione dovuta al metodo delle differenze finite; il secondo segno "$\approx$" indica l'approssimazione del valore atteso attraverso la media empirica.

La quantità $\mathcal{Q}^M_{\mathrm{MC}}$ è detta "**stimatore Monte Carlo**". Poiché prende valori diversi a seconda delle realizzazioni $\vec{\mathbf{y}}^{(1)}, \ldots, \vec{\mathbf{y}}^{(M)}$ anche $\mathcal{Q}^M_{\mathrm{MC}}$ è una variabile aleatoria.

Vediamo un esempio in Figura 23. Fissiamo la dimensione del parametro stocastico $d = 4$, il numero di realizzazioni stocastiche $M = 7$, la dimensione del sistema lineare per le differenze finite $n = 500$. Per ogni realizzazione $m = 1, \ldots, M$ generiamo $d$ valori casuali in $(0,1)$ con il comando `rand`. Costruiamo

[7] In questo caso vale anche che, per ogni $x \in (0,1)$, il valore puntuale $\vec{\mathbf{y}} \mapsto u(x, \vec{\mathbf{y}})$ è una variabile aleatoria reale. Questo non sarebbe vero, ad esempio, se $\vec{\mathbf{y}} \mapsto u(\cdot, \vec{\mathbf{y}})$ prendesse valori solo in $L^2(0,1)$ invece che in $C^0([0,1])$.

delle funzioni $q$ secondo le espressioni nella didascalia della figura[8], assembliamo la matrice del metodo delle differenze finite e risolviamo. Due classi di funzioni $q$ (trigonometriche e costanti a tratti) sono rappresentati nei riquadri in basso della Figura 23; le corrispondenti soluzioni $u$ sono nei riquadri in alto. Se poi vogliamo calcolare il valore atteso di una quantità d'interesse, ad esempio il valore minimo preso da $u$, lo calcoliamo per ognuna delle 7 realizzazioni e ne facciamo la media.

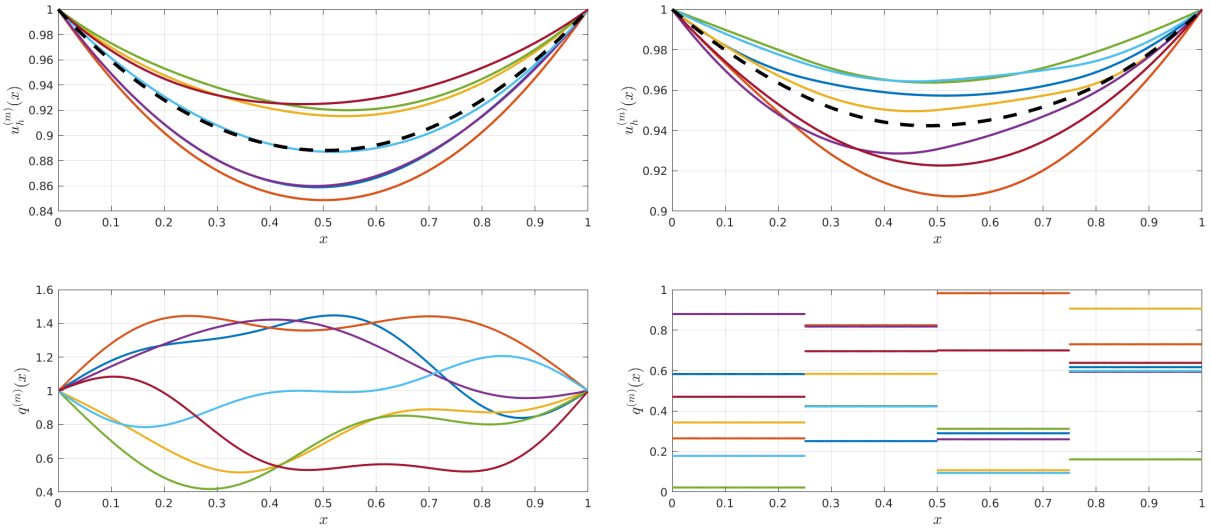



*Figura 23: Il metodo Monte Carlo applicato al problema al bordo (45).*
*In basso a sinistra vediamo $M = 7$ realizzazioni stocastiche $q^{(1)}, \dots, q^{(7)}$ di $q(x, \vec{\mathbf{y}}) = 1 + \sum_{i=1}^{4}(y_i - \frac{1}{2})\frac{\sin(\pi i x)}{i}$ (espressione leggermente più generale di quella nel testo), dove $y_1, \dots, y_4$ sono uniformemente distribuite in $(0, 1)$. In alto vediamo le 7 soluzioni corrispondenti $u_h^{(1)}, \dots, u_h^{(7)}$ del problema al bordo, calcolate con il metodo delle differenze finite con $n = 500$. A colori uguali corrispondono le stesse realizzazioni di $\vec{\mathbf{y}} \in (0, 1)^4$. La linea nera tratteggiata rappresenta la media $\frac{1}{7}(u_h^{(1)} + \cdots + u_h^{(7)})$ delle 7 realizzazioni.*
*A destra, la stessa cosa con $q(x, \vec{\mathbf{y}}) = \sum_{i=1}^{4} y_i \chi_{(\frac{i-1}{4}, \frac{i}{4})}(x)$ costante a tratti.*
*In entrambi i casi stiamo usando $d = 4$, $M = 7$, $n = 500$.*

**Esercizio 💻 4.64.** Ricostruire Figura 23. Usare il metodo Monte Carlo per approssimare il valore atteso del minimo di $u$ per ciascuna parametrizzazione di $q$.

Se abbiamo a disposizione un codice per approssimare il problema al bordo (45) (con differenze finite o qualsiasi altro metodo non importa), è molto facile implementare il metodo Monte Carlo: basta generare i vettori casuali $\vec{\mathbf{y}}^{(1)}, \dots, \vec{\mathbf{y}}^{(M)}$, costruire i coefficienti $q^{(1)}, \dots, q^{(M)}$ corrispondenti, risolvere il problema al bordo con il codice esistente, e calcolare la media empirica dei risultati ottenuti. A volte si dice che il metodo Monte Carlo è "**non-intrusivo**" perché può combinarsi facilmente con qualsiasi metodo e codice che discretizza l'equazione differenziale. Inoltre è "*embarassingly parallel*", cioè può essere implementato in parallelo senza nessuna difficoltà: se abbiamo a disposizione $M$ processori possiamo assegnare a ciascuno di essi il calcolo di uno degli $u^{(m)}$; il calcolo della media e la generazione dei valori casuali hanno costo trascurabile. La qualità dei risultati ottenuti dipende da quella del generatore di numeri casuali usato.

Quanto sarà accurato il metodo Monte Carlo? Ci aspettiamo che per $M \to \infty$ e $n \to \infty$ lo stimatore Monte Carlo $\mathcal{Q}_{\text{MC}}^{M}$ converga al valore atteso $\mathbb{E}[\mathcal{Q}(u)]$. La **legge dei grandi numeri**[9] garantisce che

[8] Qui stiamo violando alcune delle regole che ci siamo dati. Nel primo esempio (plots a sinistra), l'espressione di $q$ non è precisamente la somma degli $y_j\phi_j(x)$ ma è ottenuta sottraendo $\frac{1}{2}$ ad ogni coefficiente e aggiungendo 1 alla somma (per includere questo caso basterebbe considerare coefficienti nella forma leggermente più generale $q(x, \vec{\mathbf{y}}) = q_0(x) + \sum_{i=1}^{d} y_i\phi_i(x)$ per un $q_0 > 0$ dato, in questo caso $q_0(x) = 1 - \sum_{i=1}^{4} \frac{\sin(\pi i x)}{2i}$).
Nel secondo esempio stiamo risolvendo dei problemi al bordo con coefficienti discontinui, cosa significa? Come può l'equazione differenziale essere verificata nei punti di salto di $q$, cioè $\{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$? Per ora non ce ne preoccupiamo, tratteremo problemi al bordo con coefficienti discontinui in seguito.

[9] La legge forte dei grandi numeri afferma che, data una successione $(X_m)_{m\in\mathbb{N}}$ di variabili aleatorie indipendenti e identicamente distribuite (i.i.d.) con $\mathbb{E}[|X_m|] < \infty$, la media empirica $\overline{X}_M = \frac{1}{M}\sum_{m=1}^{M} X_m$ "converge quasi certamente" al valore atteso $\mu = \mathbb{E}[X_m]$. "Convergenza quasi certa" significa che $\mathbb{P}(\{\lim_{M\to\infty} \overline{X}_M = \mu\}) = 1$. Le realizzazioni della quantità d'interesse $\mathcal{Q}(u^{(m)})$ possono essere considerate indipendenti e identicamente distribuite (se abbiamo a disposizione un buon generatore di numeri casuali) e lo stimatore Monte Carlo non è altro che la loro media empirica.

$\mathcal{Q}_{\mathrm{MC}}^M$ converge al valore atteso $\mathbb{E}[\mathcal{Q}(u_h)]$ per $M \to \infty$, dove $u_h(\cdot, \vec{\mathbf{y}})$ è la soluzione calcolata con il metodo delle differenze finite per la realizzazione $q(\cdot, \vec{\mathbf{y}})$. Se inoltre raffiniamo $n$ simultaneamente a $M$ in modo appropriato, allora lo stimatore Monte Carlo converge a $\mathbb{E}[\mathcal{Q}(u)]$.

Quanto sarà veloce la convergenza? Per indagare la convergenza del metodo Monte Carlo, pensiamo a cosa succede quando lo applichiamo. Stiamo approssimando il valore atteso di una quantità con la media di $M$ realizzazioni. Il valore atteso non è altro che un integrale sullo spazio dei parametri $\vec{\mathbf{y}}$, cioè sul cubo $Y$: $\mathbb{E}[\mathcal{Q}(u)] = \int_Y \mathcal{Q}(u(\cdot, \vec{\mathbf{y}}))\, \mathrm{d}\rho(\vec{\mathbf{y}})$. Poiché abbiamo assunto che le variabili $y_1, \ldots, y_d$ sono uniformemente distribuite, questo è un semplice integrale $d$-dimensionale sull'ipercubo rispetto alla misura di Lebesgue. Quindi **il metodo Monte Carlo non è altro che una formula di quadratura**! È una formula di quadratura a $M$ nodi per l'integrale rispetto alla misura di probabilità, in cui i nodi sono scelti in modo casuale e i pesi sono uniformi $(1/M)$.

Visto che non possiamo testare l'accuratezza del metodo Monte Carlo per il problema di UQ in esame (non abbiamo una "soluzione esatta" contro cui confrontarlo), lo testiamo in un caso più semplice. In Figura 24 usiamo il metodo Monte Carlo per approssimare l'integrale di una funzione liscia $F$ su un intervallo. Vediamo che con varie migliaia di realizzazioni, quindi di valutazioni di $F$, abbiamo un errore dell'ordine $10^{-2}/10^{-3}$. (Per confronto, la quadratura Gaussiana con solo 10 punti commette un errore dell'ordine $10^{-8}$.) La convergenza del metodo Monte Carlo è estremamente lenta, come $M^{-1/2}$. In compenso potremmo verificare che il metodo Monte Carlo converge allo stesso modo anche se la funzione da approssimare o il suo dominio di definizione sono molto irregolari. (Figura 25 mostra un secondo modo di usare il metodo Monte Carlo per approssimare un integrale.)

Usando il teorema centrale del limite[10] si può dimostrare che anche nel caso dell'UQ il metodo Monte Carlo converge con velocità $M^{-1/2}$.

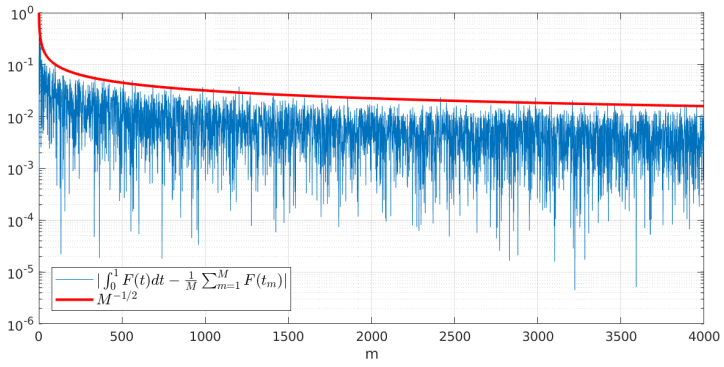



*Figura 24: In blu, l'errore commesso dal metodo Monte Carlo nella quadratura di una funzione liscia di una variabile reale. Qui abbiamo scelto* $F(t) = \sin(\pi t) - \frac{1}{10}\sin(5\pi t)$. *Per* $M = 1, \ldots, 4000$ *calcoliamo* $M$ *valori casuali* $t_1, \ldots, t_M$ *uniformemente distribuiti in* $(0, 1)$ *con il comando* `rand`, *e calcoliamo la media* $\frac{1}{M}\sum_{m=1}^M F(t_m)$. *Vediamo che l'errore decresce solo come* $M^{-1/2}$ *(in rosso).*

**Esercizio** 💻 **4.65.** Ricostruire le Figure 24 e 25: due diversi modi di usare il metodo Monte Carlo per approssimare l'integrale di una funzione reale.

### 4.9.2 Metodo della quadratura Gaussiana

Nel caso dell'integrale di una funzione reale il metodo Monte Carlo è estremamente lento: per questo motivo normalmente preferiamo usare formule di quadratura più efficienti (Gauss, trapezi, ...). Nel

[10] Il teorema centrale del limite afferma che, data una sequenza di variabili aleatorie $(X_m)_{m\in\mathbb{N}}$ i.i.d. con media $\mu$ e varianza $\sigma^2 < \infty$, vale $Z_M := \frac{\sqrt{M}}{\sigma}(\frac{1}{M}\sum_{m=1}^M X_m - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$. Il simbolo $\xrightarrow{\mathcal{D}}$ significa che la successione di variabili aleatorie $(Z_M)_{M\in\mathbb{Z}}$ converge in distribuzione a una variabile normale di media nulla e varianza 1. In questo caso convergenza "in distribuzione" significa che le funzioni cumulative (o funzioni di ripartizione) $F_{Z_M}(x) := \mathbb{P}(Z_M \le x)$ convergono puntualmente in $\mathbb{R}$ a quella della variabile normale $\Phi(x) := \frac{1}{\sqrt{2\pi}}\int_{-\infty}^x \mathrm{e}^{-\frac{1}{2}t^2}\, \mathrm{d}t$. Questo implica che per ogni $c > 0$ valgono $\lim_{M\to\infty}\mathbb{P}(Z_M \le c) = \Phi(c)$ e quindi $\lim_{M\to\infty}\mathbb{P}(|Z_M| \le c) = \Phi(c) - \Phi(-c) = 2\Phi(c) - 1$ Quindi la probabilità che lo stimatore Monte Carlo (la media empirica) $\overline{X}_M := \frac{1}{M}\sum_{m=1}^M X_m$ differisca dal valore desiderato $\mu$ di un valore minore di $\frac{c\sigma}{\sqrt{M}}$, nel limite $M \to \infty$ dipende solo da $c$. Scelto ad esempio $c = 3$, la probabilità che lo stimatore Monte Carlo $\overline{X}_M$ commetta un errore minore di $\frac{3\sigma}{\sqrt{M}}$ è circa $2\Phi(3) - 1 \approx 0.9973$. Questo è il significato dell'affermazione che "il metodo Monte Carlo converge con velocità $\frac{1}{\sqrt{M}}$", a volte scritta informalmente come $|\overline{X}_M - \mu| \sim \frac{\sigma}{\sqrt{M}}$. Questo si applica anche al caso dell'UQ per problemi al bordo considerati in questa sezione.

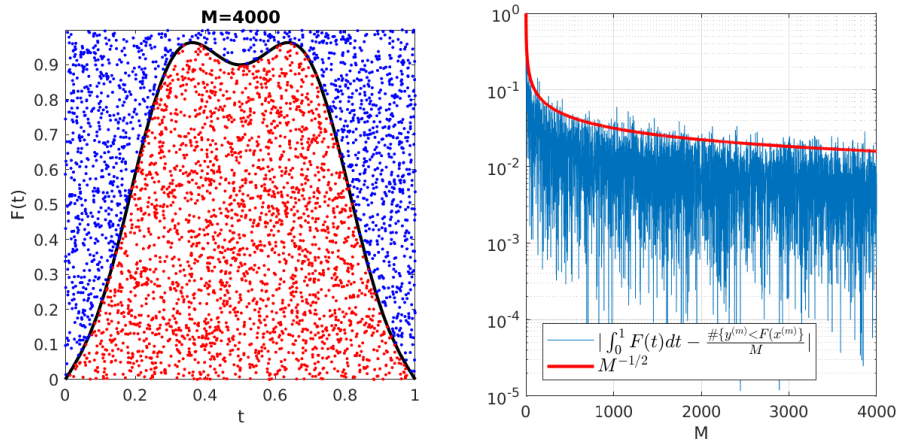



*Figura 25: Un secondo modo di applicare il metodo Monte Carlo per approssimare* $\int_0^1 F(t)\,\mathrm{d}t$ *per una* $F : (0,1) \to (0,1)$. *Qui scegliamo* $F(t) = \sin(\pi t) - \frac{1}{10}\sin(5\pi t)$. *Per* $M \in \mathbb{N}$ *generiamo* $M$ *punti* $\{(x^{(m)}, y^{(m)})\}_{m=1,\dots,M}$ *uniformemente distribuiti nel quadrato* $(0,1)^2$ *e contiamo per quanti di questi* $y^{(m)} < F(x^{(m)})$. *Questo numero diviso per* $M$ *indica la probabilità che un punto a caso del quadrato si trovi sotto il grafico di* $F$, *quindi è un'approssimazione dell'integrale desiderato. Nel riquadro a sinistra vediamo i punti generati per* $M = 4000$, *i colori distinguono quelli al di sotto e quelli al di sopra del grafico di* $F$. *Nel riquadro a destra vediamo che l'errore di questo metodo decresce circa come* $M^{-1/2}$, *cioè molto lentamente. (A voler esser più precisi, qui stiamo usando il metodo di Monte Carlo come formula di quadratura per la funzione caratteristica* $f(x,y) = \chi_{\{y<F(x)\}}(x,y)$ *definita su* $(x,y) \in (0,1)^2$.*)*

caso dell'UQ per il problema al bordo (45) abbiamo usato Monte Carlo perché il problema di partenza è naturalmente di tipo stocastico. Come migliorare la velocità di convergenza? Ricordiamo che, per ogni realizzazione di $\vec{\mathbf{y}}$, il calcolo di $\mathcal{Q}(u_h)$ richiede la soluzione del sistema lineare del metodo delle differenze finite: nel semplice caso in esame è economica, ma nelle applicazioni ogni risoluzione può essere costosa, quindi un metodo che converga più velocemente rispetto al numero di "*solve*" è necessario.

L'interpretazione del metodo come formula di quadratura ci suggerisce un'idea: usare una formula di quadratura migliore, ad esempio quella di Gauss. Ricordiamo che vogliamo approssimare il valore atteso di $\mathcal{Q}(u)$, quindi un integrale sull'ipercubo $Y$.

Fissiamo un nuovo parametro $g \in \mathbb{N}$. Denotiamo $t_1, \dots, t_g$ i nodi della quadratura di Gauss–Legendre (a $g$ nodi) per l'intervallo $(0,1)$ e $w_1, \dots, w_g$ i pesi corrispondenti (vedere [QSSG14, §9.4])[11]. Dobbiamo "tensorizzare" questi nodi e pesi, cioè estenderli al caso $d$-dimensionale. Vediamo una rappresentazione di questi nodi in Figura 26. Formalmente, per $M = g^d$ e $m = 1, \dots, M$, chiamiamo $\vec{\mathbf{y}}^{(m)}_{\text{Gauss}}$ gli $M$ punti distinti in $(0,1)^d$ le cui coordinate appartengono all'insieme $\{t_1, \dots, t_g\}$. Chiamiamo $w^{(m)}_{\text{Gauss}}$ i valori dei pesi corrispondenti (cioè se $\vec{\mathbf{y}}^{(m)}_{\text{Gauss}} = (t_{j_1}, \dots, t_{j_d})$, allora $w^{(m)}_{\text{Gauss}}$ è il prodotto $w_{j_1} \cdots w_{j_d}$).

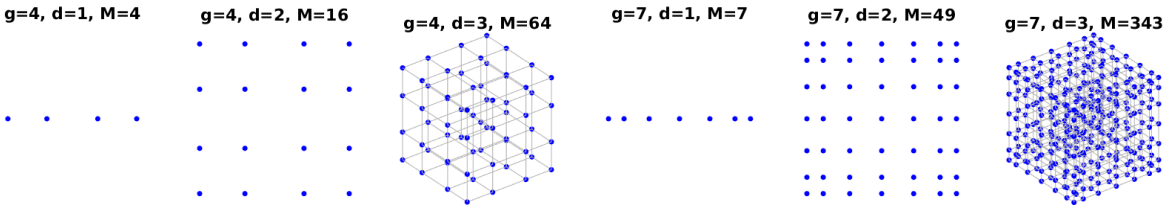



*Figura 26: I nodi di quadratura Gaussiana* $\{\vec{\mathbf{y}}^{m}_{\text{Gauss}}\}_{m=1,\dots,M=g^d}$ *su* $Y = (0,1)^d$ *in dimensione* $d = 1, 2, 3$ *per* $g = 4$ *e* $g = 7$. *Si nota l'aumento esponenziale del numero dei nodi* $M = g^d$ *al crescere della dimensione* $d$ *dello spazio dei parametri.*

A questo punto siamo pronti per usare la quadratura Gaussiana nell'UQ. A ciascuno degli $M$ nodi $\vec{\mathbf{y}}^{(m)}_{\text{Gauss}}$ corrisponde una funzione $q^{(m)}_{\text{Gauss}}(x) = q(x, \vec{\mathbf{y}}^{(m)}_{\text{Gauss}}) = \sum_{i=1}^{d} y^{(m)}_{\text{Gauss},i}\phi_i(x)$, risolviamo gli $M$ sistemi lineari corrispondenti ed otteniamo $M$ soluzioni discrete $u^{(m)}_{h,\text{Gauss}}$. Il valore atteso della quantità d'interesse

[11] Se avessimo fatto scelte diverse per $Y$ e $\rho$ ora dovremmo scegliere una formula di quadratura diversa. Ad esempio per $Y = \mathbb{R}^d$ e $\rho$ la misura Gaussiana sceglieremo nodi e pesi della quadratura di Gauss–Hermite. Se $Y = (0,\infty)^d$ e ogni $y_i$ è distribuito secondo una distribuzione esponenziale allora useremo la quadratura di Gauss–Laguerre.

sarà approssimato come

$$\boxed{\mathcal{Q}^M_{\text{Gauss}} := \sum_{m=1}^{M} w^{(m)}_{\text{Gauss}} \mathcal{Q}\big(u^{(m)}_{h,\text{Gauss}}\big)} \approx \sum_{m=1}^{M} w^{(m)}_{\text{Gauss}} \mathcal{Q}\big(u(\cdot, \vec{\mathbf{y}}^{(m)}_{\text{Gauss}})\big) \approx \int_Y \mathcal{Q}\big(u(\cdot,\vec{\mathbf{y}})\big)\,\mathrm{d}\rho(\vec{\mathbf{y}}) = \mathbb{E}[\mathcal{Q}(u)].$$

(Confrontare con la formula analoga per il metodo Monte Carlo, cioè (46).)

Se la funzione $\vec{\mathbf{y}} \mapsto \mathcal{Q}(u(\cdot,\vec{\mathbf{y}}))$ è liscia[12] ci aspettiamo una convergenza molto veloce. Se la funzione è analitica abbiamo addirittura convergenza esponenziale in $g$. Come Monte Carlo, anche il metodo della quadratura Gaussiana è non-intrusivo e *embarassingly parallel*.

Anche questo metodo ha un problema: pur convergendo con un errore proporzionale a $\mathrm{e}^{-cg}$ per qualche $c > 0$, il costo computazionale, misurato come il numero di problemi discreti (sistemi lineari $n \times n$) da risolvere, è pari a $M = g^d$. Cioè il costo cresce esponenzialmente nel numero di parametri $d$ necessario a descrivere il dominio stocastico. (Al contrario il metodo Monte Carlo è robusto rispetto a $d$: abbiamo un errore $\sim M^{-1/2}$ invece di $\sim \mathrm{e}^{-bM^{1/d}}$.) Questo problema è tanto importante che viene chiamato "*curse of dimensionality*". Esistono tecniche raffinate per affrontarlo, come le "griglie sparse" ("*sparse grids*") che permettono di ottenere la stessa convergenza ad un costo computazionale molto più basso, al prezzo di maggiori complicazioni.

Un altro difetto del metodo è che i nodi della quadratura Gaussiana per un certo $g$ non sono nodi di quadratura per $g^* > g$. Quindi non possiamo decidere di fermare in modo adattivo il codice quando otteniamo un risultato soddisfacente, ma scelto un certo $g$ dobbiamo risolvere tutti i $g^d$ problemi. Se il risultato non è sufficientemente accurato dobbiamo aumentare $g$ e ripartire da capo. Altre formule deterministiche di quadratura (trapezi…) possono ovviare a questo problema ma non al *curse of dimensionality*.

### 4.9.3 Confronto tra metodi Monte Carlo e di quadratura Gaussiana: un esempio concreto

Vogliamo confrontare il metodo di Monte Carlo e quello della quadratura Gaussiana per l'approssimazione del valore atteso di una quantità d'interesse.

Consideriamo il problema al bordo (45), con $q(x,y) = y_1\chi_{(0,\frac12)}(x) + y_2\chi_{(\frac12,1)}(x)$ costante a tratti su due intervalli di lunghezza $\frac12$ (la dimensione stocastica è $d = 2$). Vogliamo approssimare il valore atteso del minimo di $u$:

$$\begin{aligned}\mathbb{E}[\mathcal{Q}(u)] = \mathbb{E}[\min_{x\in(0,1)} u(x,\cdot)] &= \int_{(0,1)^2} \min_{x\in(0,1)} u(x,\vec{\mathbf{y}})\,\mathrm{d}\vec{\mathbf{y}} \\ &\approx \begin{cases} \displaystyle\sum_{m=1}^{M} w^{(m)}_{\text{Gauss}} \min_{x\in(0,1)} u_h(x,\vec{\mathbf{y}}^{(m)}_{\text{Gauss}}) \approx \sum_{m=1}^{M} w^{(m)}_{\text{Gauss}} \min_{j=1,\dots,n} U^{(m)}_{j,\text{Gauss}} = \mathcal{Q}^M_{\text{Gauss}}, \\ \displaystyle\sum_{m=1}^{M} \frac{1}{M} \min_{x\in(0,1)} u_h(x,\vec{\mathbf{y}}^{(m)}_{\text{MC}}) \quad \approx \sum_{m=1}^{M} \frac{1}{M} \min_{j=1,\dots,n} U^{(m)}_{j,\text{MC}} \quad = \mathcal{Q}^M_{\text{MC}}. \end{cases}\end{aligned}$$

dove $\vec{\mathbf{U}}^{(m)}_{\text{Gauss}} = (U^{(m)}_{1,\text{Gauss}},\dots,U^{(m)}_{n,\text{Gauss}})^\top$ e $\vec{\mathbf{U}}^{(m)}_{\text{MC}} = (U^{(m)}_{1,\text{MC}},\dots,U^{(m)}_{n,\text{MC}})^\top \in \mathbb{R}^n$ sono i vettori soluzione ottenuti dal metodo delle differenze finite.

Approssimiamo questo integrale con (i) la quadratura di Gauss–Legendre con $g = 1,\dots,6$ nodi per direzione, cioè $M = 1, 2^2, 3^2, \dots, 6^2$ nodi in totale, e con (ii) il metodo Monte Carlo con lo stesso numero di valutazioni di $\mathcal{Q}$, cioè con $M = 1, 2^2, 3^2, \dots, 6^2$ realizzazioni. Ciascuna di queste valutazioni di $\mathcal{Q}$ richiede la soluzione di un sistema lineare $n \times n$ del metodo delle differenze finite (fissiamo $n = 500$). Quindi risolveremo $2(1 + 2^2 + 3^2 + \ldots + 6^2) = 182$ sistemi lineari e i due metodi avranno lo stesso costo.

Nella Figura 27 gli asterischi blu rappresentano i valori di $\mathcal{Q}(u_h(x,\vec{\mathbf{y}}^{(m)}_{\text{Gauss}}))$ per ogni nodo del metodo della quadratura Gaussiana, al variare di $g = 1,\dots,6$ (sulle ascisse) e di $m = 1,\dots,g^d$ (i diversi punti

[12] Se una funzione $F : [0,1] \to \mathbb{R}$ è di classe $C^k$ per $k \in \mathbb{N}$, la quadratura di Gauss–Legendre garantisce convergenza algebrica (all'aumentare del numero $g$ dei nodi) di ordine $k$: $\left|\int_0^1 F(x)\,\mathrm{d}x - \sum_{j=1}^g w_j F(t_j)\right| \sim g^{-k}$, [QSSG14, eq. (9.45)]. Se $F$ ammette un'estensione analitica in un intorno dell'intervallo nel piano complesso si ha convergenza esponenziale $\sim \mathrm{e}^{-cg}$, si veda il Teorema 19.3 in [L.N. Trefethen, *Approximation theory and approximation practice*, SIAM, 2013], http://www.chebfun.org/ATAP/ Lo stesso vale per l'estensione all'ipercubo $d$-dimensionale $Y$.

La funzione che ci interessa mappa $Y \subset \mathbb{R}^d$ in $\mathbb{R}$ ed è composizione di operatori che mappano tra spazi di funzioni, attraverso la soluzione di un problema al bordo: $\vec{\mathbf{y}} \mapsto q(\cdot,\vec{\mathbf{y}}) \mapsto u(\cdot,\vec{\mathbf{y}}) \mapsto \mathcal{Q}(u(\cdot,\vec{\mathbf{y}}))$. È un oggetto molto complicato! Ci basta sapere che questa funzione è liscia in molti casi rilevanti, perfino quando le funzioni $\phi_i$ che descrivono $q$ sono irregolari, come le costanti a tratti viste negli esempi. L'interpretazione di questa regolarità è che a piccole perturbazioni dei parametri $\vec{\mathbf{y}}$ corrispondono piccole perturbazioni di $\mathcal{Q}(u)$.

per una stessa ascissa). I dischi rossi rappresentano i valori analoghi $\mathcal{Q}(u_h(x, \vec{\mathbf{y}}_{\mathrm{MC}}^{(m)}))$ per ogni realizzazione usata per il metodo Monte Carlo. Le linee continue rappresentano gli stimatori $\mathcal{Q}_{\mathrm{Gauss}}^M$ e $\mathcal{Q}_{\mathrm{MC}}^M$ per $M = 1, 2^2, \ldots, 6^2$, cioè le medie (pesate con i pesi della quadratura appropriata) dei valori di asterischi e dischi.

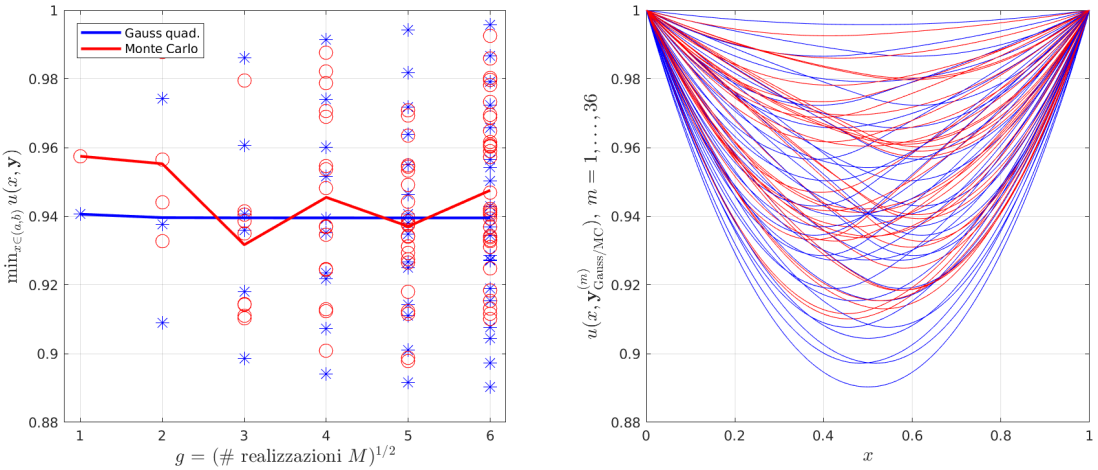



*Figura 27: Confronto tra il metodo di Monte Carlo e quello della quadratura Gaussiana per l'approssimazione del valore atteso di una quantità d'interesse come descritto in §4.9.3.*
*Qui consideriamo il problema al bordo (45), con $q(x, y) = y_1\chi_{(0,\frac{1}{2})}(x) + y_2\chi_{(\frac{1}{2},1)}(x)$ costante a tratti su due intervalli di lunghezza $\frac{1}{2}$ (la dimensione stocastica è $d = 2$).*
*A sinistra: le stime del valore atteso della quantità d'interesse $\mathbb{E}[\mathcal{Q}(u)] = \mathbb{E}[\min_{x\in(0,1)} u(x, \cdot)]$ ottenute con il metodo della quadratura Gaussiana e il metodo Monte Carlo.*
*A destra: le soluzioni $u_h(x, \vec{\mathbf{y}}_{\mathrm{Gauss}}^{(1)}), \ldots, u_h(x, \vec{\mathbf{y}}_{\mathrm{Gauss}}^{(36)})$ (in blu) e $u_h(x, \vec{\mathbf{y}}_{\mathrm{MC}}^{(1)}), \ldots, u_h(x, \vec{\mathbf{y}}_{\mathrm{MC}}^{(36)})$ (in rosso) del problema al bordo usate per calcolare $\mathcal{Q}_{\mathrm{Gauss}}^6$ e $\mathcal{Q}_{\mathrm{MC}}^6$.*

Si nota che lo stimatore del metodo di quadratura Gaussiana già per poche realizzazioni è praticamente costante a $\mathcal{Q}_{\mathrm{Gauss}}^M = 0.93952$, mentre quello Monte Carlo oscilla molto: la stima del metodo di quadratura Gaussiana è molto più preciso e affidabile.

**Esercizio 💻 4.66.** Implementare la simulazione descritta e ricostruire Figura 27.

Il metodo della quadratura Gaussiana è deterministico: ad ogni realizzazione fornisce gli stessi valori. Al contrario, il metodo Monte Carlo ogni volta dà valori diversi a seconda dei numeri casuali generati.

In questo esempio i nodi e i pesi di Gauss–Legendre sono calcolati con l'algoritmo di Golub–Welsch:

```
% Nodi e pesi di Gauss-Legendre su intervallo (0,1).
function [x,w] = gaussquad(g)
b      = (1:(g-1)) ./ sqrt(4*(1:(g-1)).^2-1) ;
J      = diag(b,-1) + diag(b,1);
[V,D]  = eig(J);
x      = (diag(D)+1)/2;
w      = ((V(1,:).*V(1,:)))';
```

Per $d = 2, 3$ i nodi "tensorizzati" in $Y$ si possono calcolare con il comando `meshgrid`.

# 5 Il metodo di collocazione spettrale

## 5.1 Il metodo di collocazione in generale

Consideriamo ancora il problema al bordo lineare con dati $p, q, f$ lisci e condizione al bordo omogenee, ad esempio di Dirichlet:

$$\begin{cases} \mathcal{L}u(x) := -u''(x) + p(x)u'(x) + q(x)u(x) = f(x) & \text{in } (a, b) \\ u(a) = 0, \\ u(b) = 0. \end{cases} \tag{47}$$

Il metodo delle differenze finite richiede la risoluzione di un sistema lineare (sparso) $n \times n$ e ha un errore proporzionale a $n^{-2}$. Ora vogliamo descrivere un metodo numerico che converge molto più velocemente e siamo disposti a "pagare" la maggiore velocità accettando di risolvere un sistema lineare denso.

Inoltre, il metodo delle differenze finite permette di approssimare il valore della soluzione $u$ in alcuni punti predefiniti, i nodi $x_j = a + hj$. La soluzione esatta $u$ però è una funzione definita su tutto l'intervallo $(a, b)$: vogliamo un metodo il cui output sia una funzione $u_h$ definita sullo stesso intervallo e che approssimi $u$.

Il metodo di collocazione (*collocation*) consiste nel

- scegliere uno spazio vettoriale $V_h$ di dimensione finita $n$ di funzioni di classe $C^2$ su $(a, b)$ che soddisfano le condizioni al bordo omogenee,
- scegliere una base $\{\varphi_k\}_{k=1,\dots,n}$ di $V_h$,
- fissare dei nodi $x_1, x_2, \dots, x_n$ nell'intervallo $(a, b)$,
- "collocare" l'equazione differenziale in questi nodi, cioè cercare un elemento $u_h \in V_h$ che soddisfi l'equazione in ogni nodo.

(Il simbolo $_h$ in $u_h$ e $V_h$ denota semplicemente gli oggetti discreti, qui non c'è nessuna mesh di passo $h$.) Per avere lo stesso numero di gradi di libertà e di condizioni da soddisfare scegliamo il numero dei nodi uguale alla dimensione di $V_h$. Data una base $\varphi_1, \dots, \varphi_n$ di $V_h$, un elemento $u_h \in V_h$ può essere scritto $u_h(x) = \sum_{k=1}^n U_k \varphi_k(x)$, dove $\vec{\mathbf{U}} = (U_1, \dots, U_n)^\top$ è un vettore di $\mathbb{R}^n$. Collocare l'equazione nei nodi significa imporre

$$(\mathcal{L}u_h)(x_j) = \mathcal{L}\Big(\sum_{k=1}^n U_k \varphi_k\Big)(x_j) = \sum_{k=1}^n U_k (\mathcal{L}\varphi_k)(x_j) = f(x_j)$$

per $j = 1, \dots, n$. In formato vettoriale questo si scrive

$$\boxed{\underline{\underline{\mathbf{A}}}\vec{\mathbf{U}} = \vec{\mathbf{f}}, \qquad \text{dove} \quad A_{j,k} := (\mathcal{L}\varphi_k)(x_j) = -\varphi_k''(x_j) + p(x_j)\varphi_k'(x_j) + q(x_j)\varphi_k(x_j), \quad f_j := f(x_j).} \tag{48}$$

Quindi dato l'operatore differenziale lineare $\mathcal{L}$ (determinato dall'equazione), e scelti i nodi $x_j$ e la base $\{\varphi_k\}$, tutto quello che dobbiamo fare è costruire la matrice $\underline{\underline{\mathbf{A}}}$ definita in (48) e il vettore $\vec{\mathbf{f}}$ e risolvere un sistema lineare. Questo è il metodo di collocazione.

**Nota 5.1.** Cosa possiamo fare se le condizioni al bordo non sono omogenee ma $u(a) = \alpha$ e $u(b) = \beta$? Scegliendo una funzione $\tilde{u}$ che soddisfa entrambe le condizioni, ad esempio $\tilde{u}(x) = (x-a)\frac{\beta-\alpha}{b-a} + \alpha$, la funzione $u_0 := u - \tilde{u}$ soddisfa un nuovo problema ai limiti nella forma (47) (con $\mathcal{L}u_0 = \tilde{f} := f - \mathcal{L}\tilde{u}$) e il metodo può essere applicato a $u_0$.

Notiamo che per poter applicare l'operatore $\mathcal{L}$, che include una derivata seconda, gli elementi dello spazio discreto $V_h$ devono essere funzioni di classe $C^2$.

Quando le funzioni di base $\varphi_k$ hanno un supporto "piccolo" in $(a, b)$, allora, dato un indice $k$, $(\mathcal{L}\varphi_k)(x_j)$ sarà zero per molti $j$, quindi la matrice $\underline{\underline{\mathbf{A}}}$ sarà sparsa e a bande. Un esempio di funzioni di base con supporto piccolo è dato dalle B-splines; si veda il capitolo 8 di [P.M. Prenter, *Splines and variational methods*, Wiley, 1989] per l'analisi di questa versione del metodo. Al contrario, se si scelgono funzioni di base con supporto su tutto $(a, b)$, **la matrice $\underline{\underline{\mathbf{A}}}$ sarà densa**.

Se $V_h$ è scelto in modo da garantire che la convergenza di $u_h$ a $u$ in qualche norma è "superalgebrica", cioè più veloce di $\mathcal{O}(n^{-c})$ per ogni $c > 0$, allora il metodo si dice **metodo di collocazione spettrale** o **metodo pseudospettrale**. (A volte l'uso della parola "spettrale" è limitato a scelte specifiche di $V_h$.)

## 5.2 Il metodo di collocazione spettrale polinomiale

La scelta più semplice per lo spazio discreto $V_h$ nel metodo di collocazione per il problema di Dirichlet (47) è

$$V_h := \{P(x) \text{ polinomio di grado } \le n+1 \text{ tale che } P(a) = P(b) = 0\}.$$

Fissiamo per semplicità l'intervallo $(a, b) = (-1, 1)$.

Una base molto semplice di $V_h$ è data dalle funzioni

$$(1-x^2), \quad x(1-x^2), \quad x^2(1-x^2), \quad \dots, \quad x^{n-1}(1-x^2).$$

Tuttavia non conviene usare questa base poiché conduce a un metodo spettrale estremamente malcondizionato. Una base migliore e usata spesso è quella dei **polinomi di Legendre integrati**:

$$M_k(x) := \int_{-1}^x P_k(t)\,\mathrm{d}t = \frac{P_{k+1}(x) - P_{k-1}(x)}{2k+1} \qquad k = 1, \dots, n, \tag{49}$$

dove $P_k$ sono i **polinomi di Legendre**, definiti da

$$P_k(x) := \frac{1}{2^k k!}\frac{\partial^k}{\partial x^k}[(x^2-1)^k].$$

In pratica i polinomi di Legendre vengono calcolati con una semplice formula ricorsiva a tre termini:

$$P_{k+1}(x) = \frac{(2k+1)xP_k(x) - kP_{k-1}(x)}{k+1}, \qquad P_0 = 1, \quad P_1(x) = x.$$

**Esercizio ✎ 5.2.** Mostrare che $P_k$ è pari per $k$ pari e dispari altrimenti.
In particolare $P_k(1) = 1$ e $P_k(-1) = (-1)^k$.
Dedurre che i polinomi di Legendre integrati $M_k$ soddisfano le condizioni di Dirichlet omogenee su $(-1, 1)$.

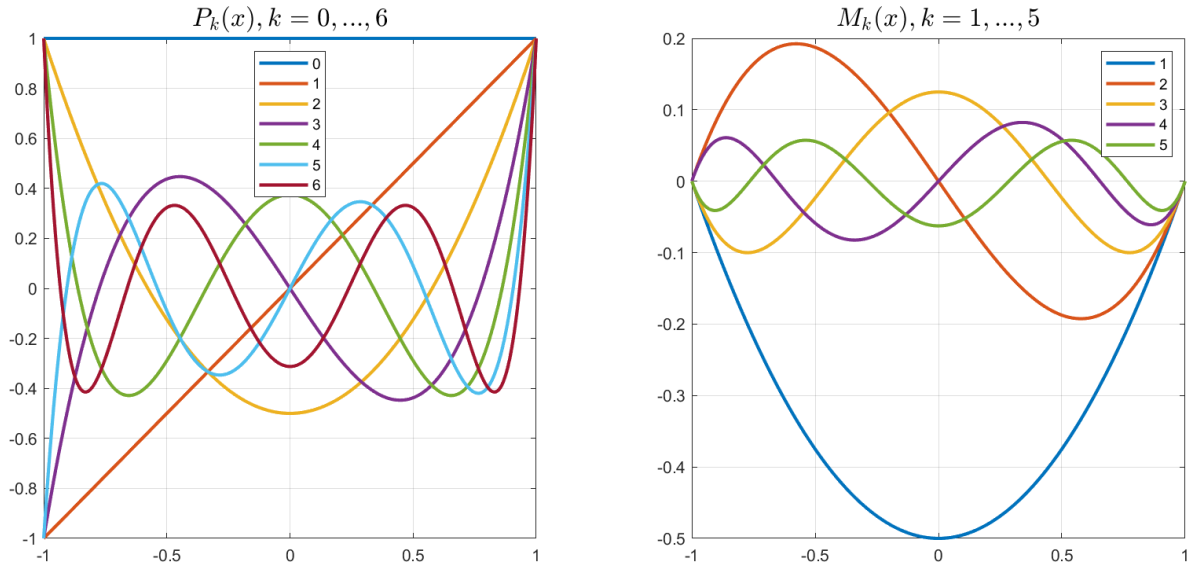



*Figura 28: I polinomi di Legendre $P_0, \ldots, P_6$ (sinistra) e i polinomi di Legendre integrati $M_1, \ldots, M_5$.*

Poiché $M_k' = P_k$, gli elementi della matrice $\underline{\underline{\mathbf{A}}}$ del metodo di collocazione sono

$$A_{j,k} = (\mathcal{L}M_k)(x_j) = -P_k'(x_j) + p(x_j)P_k(x_j) + q(x_j)M_k(x_j).$$

**Esercizio ✎ 5.3.** L'espressione di $A_{j,k}$ richiede la valutazione della derivata $P_k'$ dei polinomi di Legendre nei nodi. Derivare da (49) una semplice formula ricorsiva che permette di calcolare $P_{k+1}'(x_j)$ a partire dai valori di $P_\ell(x_j)$ e $P_\ell'(x_j)$ per $\ell \le k$.

**Nota 5.4.** I polinomi di Legendre sono autofunzioni di un operatore differenziale in $(-1, 1)$:

$$-\big((1-x^2)P_k'(x)\big)' = \lambda_k P_k(x), \qquad \lambda_k = k(k+1).$$

Nonostante i $P_k$ non soddisfino condizioni al bordo omogenee e la funzione $K(x) = (1-x^2)$ si annulli agli estremi di $(-1, 1)$, procedendo come in §4.7.1 da questa equazione si può facilmente derivare l'ortogonalità dei polinomi di Legendre:

$$\int_{-1}^{1} P_k(x)P_j(x)\,\mathrm{d}x = \frac{2}{2k+1}\delta_{k,j}.$$

La seconda scelta importante per definire un metodo di collocazione è quella dei nodi. L'esempio di Runge ci dice che i nodi equispaziati generano interpolanti poco accurati. La scelta più comune è quella dei **nodi di Chebyshev**:

$$x_j = \cos\frac{2j-1}{2n}\pi, \qquad j = 1, \ldots, n.$$

**Esercizio 💻 5.5** (Per i più coraggiosi)**.** Implementare il metodo di collocazione spettrale polinomiale con base $M_k$, $k = 1, \ldots, n$, e nodi di Chebyshev:

- nel caso $(a,b)=(-1,1)$ e condizioni al bordo omogenee $\alpha=\beta=0$
  (ad esempio per l'equazione $-u''+25u=1$ [soluzione $u(x)=\frac{1}{25}(1-\frac{\cosh 5x}{\cosh 5})$],
  oppure per $-u''+4x^2u=2\mathrm{e}^{-x^2}-\frac{4}{\mathrm{e}}x^2$ [soluzione $u(x)=\mathrm{e}^{-x^2}-\mathrm{e}^{-1}$]);
- nel caso $(a,b)=(-1,1)$ e condizioni al bordo non omogenee $\alpha,\beta\neq 0$, ricordando la Nota 5.1;
  (ad esempio per $-u''+u=0$, $\alpha=\beta=1$, come nell'Esercizio 4.1 [soluzione $u(x)=\frac{\cosh x}{\cosh 1}$]);
- nel caso $(a,b)$ un intervallo limitato generico. Fare attenzione a come vengono scalati su $(a,b)$ i nodi di Chebyshev, le funzioni di base e le loro derivate.

Suggerimento: implementare tre matrici contenenti i valori nei nodi dei polinomi di Legendre integrati, le loro derivate prime e seconde, rispettivamente. Usare la formula ricorsiva per $P_k$, formula (49) e l'Esercizio 5.3. Fare attenzione che i $P_k$ sono definiti a partire da $k=0$ e gli $M_k$ da $k=1$.

**Nota 5.6** (Altre basi e altri spazi discreti)**.** La base dei polinomi di Legendre integrati (49) permette di calcolare facilmente il loro valore e le loro derivate, e quindi di assemblare la matrice $\underline{\underline{\mathbf{A}}}$. Altre basi dello stesso spazio polinomiale possono essere scelte per il metodo spettrale, si veda ad esempio la Nota 5.23. È anche possibile usare spazi discreti $V_h$ non polinomiali. Ad esempio, si possono usare le prime $n$ autofunzioni di un operatore di Sturm–Liouville $\mathcal{L}$: infatti Proposizione 4.52 mostra che uno spazio $V_h$ costruito in questo modo può approssimare qualsiasi funzione di $L^2(a,b)$.

**Nota 5.7** (Collocazione spettrale vs differenze finite)**.** Sia il metodo di collocazione spettrale polinomiale che quello delle differenze finite permettono di approssimare la soluzione del problema al bordo (47). In quali casi è conveniente preferire l'uno o l'altro?

Il metodo delle differenze finite richiede la soluzione di un sistema lineare *sparso* $n\times n$ e, se $u\in C^4(a,b)$, garantisce convergenza *quadratica* in $n$ (cioè $\mathcal{O}(n^{-2})$). Se $u$ è più regolare, l'ordine di convergenza non migliora, poiché questo è dettato dall'ordine di troncamento delle differenze finite centrate.

Il metodo di collocazione spettrale richiede la soluzione di un sistema lineare *denso* $n\times n$ quindi sarà conveniente solo quando l'ordine di convergenza è più che quadratico. Si può verificare che l'ordine di convergenza dipende dalla regolarità di $u$ (che a sua volta dipende da quella di $f,p,q$). Ad esempio se $u$ è analitica, l'errore converge a zero con velocità *esponenziale* in $n$ (cioè $\mathcal{O}(\mathrm{e}^{-cn})$ per qualche $c>0$).

In sintesi se la soluzione $u$ è molto regolare allora conviene il metodo di collocazione spettrale perché a parità di $n$ ottiene un'accuratezza molto più elevata; se $u$ ha regolarità più bassa la sparsità rende il metodo delle differenze finite conveniente. Vedremo più in dettaglio una situazione simile in §5.3.

**Nota 5.8.** Nell'Esercizio 3.5 abbiamo interpretato le differenze finite come derivate di un polinomio che interpola la funzione da differenziare. L'ordine di convergenza e l'accuratezza di uno schema di differenze finite aumentano aumentando il grado del polinomio interpolante. L'interpolazione con polinomi di grado alto richiede l'uso di più nodi: i polinomi di grado 2 e 4 nell'Esercizio 3.5 richiedono 3 e 5 nodi, rispettivamente. Se tutti i nodi a disposizione vengono coinvolti nell'interpolazione si ottiene la derivata pseudospettrale, ricordare la Nota 3.10.

Se le differenze finite vengono usate per costruire uno schema numerico per risolvere un problema al bordo, il numero di nodi corrispondente ad ogni differenza finita determina l'ampiezza di banda della matrice corrispondente al metodo. Per le differenze finite centrate del secondo ordine abbiamo infatti ottenuto matrici tridiagonali. Portando il ragionamento all'estremo, il metodo delle differenze finite basato sull'interpolazione con un polinomio globale corrisponde al metodo di collocazione spettrale e conduce a una matrice densa.

## 5.3 Il metodo di collocazione spettrale trigonometrico

Le funzioni trigonometriche sono particolarmente efficaci per approssimare soluzioni periodiche. Consideriamo il **problema periodico**

$$\begin{cases} \mathcal{L}u(x) := -u''(x)+p(x)u'(x)+q(x)u(x)=f(x) & \text{in } (0,2\pi),\\ u(0)=u(2\pi),\\ u'(0)=u'(2\pi). \end{cases} \tag{50}$$

dove per semplicità abbiamo fissato $a=0$ e $b=2\pi$. Dato $n\in\mathbb{N}$, uno spazio $n$-dimensionale di funzioni discrete tipicamente usato in questo caso è

$$V_h := \operatorname{span}\left\{\varphi_k,\ k=-\left\lfloor\frac{n}{2}\right\rfloor,\dots,\left\lfloor\frac{n-1}{2}\right\rfloor\right\},\qquad \begin{aligned}\varphi_k(x)&:=\cos(kx) \quad k=0,\dots,\left\lfloor\frac{n-1}{2}\right\rfloor,\\ \varphi_k(x)&:=\sin(kx) \quad k=-\left\lfloor\frac{n}{2}\right\rfloor,\dots,-1.\end{aligned}$$

Usiamo i nodi equispaziati[13] $x_j = \frac{2\pi}{n}(j-1)$ per $j = 1, \dots, n$. Il sistema lineare del metodo di collocazione diventa

$$\underline{\underline{\mathbf{A}}}\vec{\mathbf{U}} = \vec{\mathbf{f}}, \qquad \text{dove} \tag{51}$$

$$A_{j,k} = \big(k^2 + q(x_j)\big)\cos\frac{2\pi k(j-1)}{n} - kp(x_j)\sin\frac{2\pi k(j-1)}{n}, \qquad f_j = f(x_j), \qquad \begin{cases} j = 1, \dots, n, \\ k = 0, \dots, \lfloor \frac{n-1}{2} \rfloor, \end{cases}$$

$$A_{j,k} = \big(k^2 + q(x_j)\big)\sin\frac{2\pi k(j-1)}{n} + kp(x_j)\cos\frac{2\pi k(j-1)}{n}, \qquad f_j = f(x_j), \qquad \begin{cases} j = 1, \dots, n, \\ k = -\lfloor \frac{n}{2} \rfloor, \dots, -1. \end{cases}$$

L'implementazione del metodo di collocazione e la manipolazione delle basi è più semplice se consideriamo spazi di funzioni periodiche a valori complessi con base

$$\psi_k(x) := \mathrm{e}^{\mathrm{i}(k - \lfloor \frac{n}{2} \rfloor - 1)x}, \qquad 1 \le k \le n.$$

Ricordiamo che le basi $\{\psi_k\}$ e $\{\varphi_k\}$ sono legate tra loro dalla relazione di Eulero e dall'espressione delle funzioni trigonometriche in termini di esponenziali complessi:

$$\mathrm{e}^{\mathrm{i}t} = \cos t + \mathrm{i}\sin t, \qquad \sin t = \frac{\mathrm{e}^{\mathrm{i}t} - \mathrm{e}^{-\mathrm{i}t}}{2\mathrm{i}}, \qquad \cos t = \frac{\mathrm{e}^{\mathrm{i}t} + \mathrm{e}^{-\mathrm{i}t}}{2}$$

**Esercizio ✎ 5.9.** Mostrare che, se $n = 2N + 1$ è dispari, le due basi $\{\varphi_k\}_{k=-N,\dots,N}$ e $\{\psi_k\}_{k=1,\dots,2N+1}$ generano lo stesso spazio vettoriale complesso $V_h$.

Mostrare che entrambe le basi sono ortogonali in $L^2(0, 2\pi)$.

Gli elementi di $V_h$, scritti come combinazione lineare di funzioni trigonometriche o di esponenziali complessi, sono detti "polinomi trigonometrici"[14].

**Esercizio 💻 5.10.**

- Scrivere in forma matriciale il metodo di collocazione spettrale scegliendo come funzioni di base gli esponenziali complessi $\{\psi_k\}_{k=1,\dots,n}$ e i nodi $x_j = \frac{2\pi}{n}(j-1)$ per $j = 1, \dots, n$.
- Implementare[15] questo metodo per l'equazione

  $$-u''(x) + (\cos^2 x)u(x) = \mathrm{e}^{\sin x}\sin x$$

  con condizioni al bordo periodiche su $(0, 2\pi)$, la cui soluzione esatta è $u(x) = \mathrm{e}^{\sin x}$.
- Plottare la soluzione discreta $u_h$ per qualche valore di $n$.

  Il valore di $u_h$ in un vettore di $m$ punti `xPlot` può essere calcolato come il prodotto $\underline{\underline{\mathbf{M}}}\vec{\mathbf{U}}$ tra un'opportuna matrice $\underline{\underline{\mathbf{M}}} \in \mathbb{R}^{m\times n}$ e il vettore dei coefficienti ottenuti dal metodo di collocazione. Normalmente $m$ è indipendente da $n$.
- Plottare le norme $L^\infty(0, 2\pi)$ e $L^2(0, 2\pi)$ dell'errore commesso dal metodo in dipendenza da $n$, per $n$ da 1 a 40. Con che velocità convergono? Come si possono stimare numericamente le norma richieste? Confrontare l'errore commesso con quello commesso dal metodo delle differenze finite.

**Esercizio 💻 5.11.** Ripetere l'Esercizio 5.10 per l'equazione

$$-u''(x) + \frac{1}{4}u(x) = (x - \pi)^2$$

con condizioni al bordo periodiche su $(0, 2\pi)$, la cui soluzione esatta è $u(x) = 4(x-\pi)^2 + 32 - \frac{16\pi}{\sinh\frac{\pi}{2}}\cosh\frac{x-\pi}{2}$. Cosa si nota nella convergenza del metodo? Come si può spiegare questo fatto?

[13] Questa è l'unica parte di queste note in cui, usando nodi equispaziati, dividiamo l'intervallo su cui è definito il problema al bordo in $n$ parti invece che in $n + 1$.

[14] Infatti i polinomi di variabile complessa $z^k$ e $\overline{z}^k$ ristretti alla circonferenza unitaria $\{z \in \mathbb{C}, |z| = 1\}$ hanno valore $\mathrm{e}^{\mathrm{i}k\theta}$ e $\mathrm{e}^{-\mathrm{i}k\theta}$, detto $\theta$ l'argomento complesso della variabile $z$ (cioè $z = |z|\mathrm{e}^{\mathrm{i}k\theta} = |z|\cos\theta + \mathrm{i}|z|\sin\theta$). Questo equivale a identificare l'intervallo "periodico" $[0, 2\pi)$ con la circonferenza unitaria.

[15] Attenzione: qui matrici e vettori hanno valori complessi. In Matlab l'apostrofo " `'` " indica il trasposto coniugato; per trasporre una matrice o un vettore senza coniugio si deve usare " `.'` ". L'unità immaginaria si può scrivere come `1i`.

**Nota 5.12** (Calcolare la norma $L^2(0,2\pi)$)**.** Negli esercizi 5.10–5.11 dobbiamo calcolare le norme dell'errore del metodo di collocazione. Mentre la norma $L^\infty(0,2\pi)$ è facile da approssimare, quella $L^2(0,2\pi)$ richiede più attenzione. Partizioniamo l'intervallo $(0,2\pi)$ in $m$ sottointervalli di uguale lunghezza $\frac{2\pi}{m}$, e chiamiamo $y_j = \frac{2\pi j}{m}$, $j = 0,\dots,m$ i loro estremi. Ad esempio questi possono essere i punti usati per plottare la soluzione, normalmente $m \gg n$. Chiamiamo $e := u - u_h$ la funzione errore ed $\vec{\mathbf{e}} \in \mathbb{R}^m$ il vettore con $e_j := e(y_j)$. Poiché $e$ è periodica, $e_0 = e_m$ e non è necessario ripetere questo elemento nel vettore. Usando la formula di quadratura composita dei trapezi e la periodicità di $e$ abbiamo

$$\|e\|_{L^2(0,2\pi)} = \sqrt{\int_0^{2\pi} |e(y)|^2 \,\mathrm{d}y} \approx \sqrt{\frac{2\pi}{m}} \sqrt{\frac{1}{2}|e(y_0)|^2 + \sum_{j=1}^{m-1} |e(y_j)|^2 + \frac{1}{2}|e(y_m)|^2} = \sqrt{\frac{2\pi}{m}} \sqrt{\sum_{j=1}^{m} |e(y_j)|^2}$$
$$= \sqrt{\frac{2\pi}{m}}\, \|\vec{\mathbf{e}}\|_2 .$$

Questo significa che possiamo calcolare (approssimativamente) la norma $L^2(0,2\pi)$ dell'errore (o di qualsiasi altra funzione periodica) come la norma $\|\cdot\|_2$ (vettoriale) della sua valutazione nei punti equispaziati. È importante non dimenticarsi del coefficiente $\sqrt{\frac{2\pi}{m}}$.

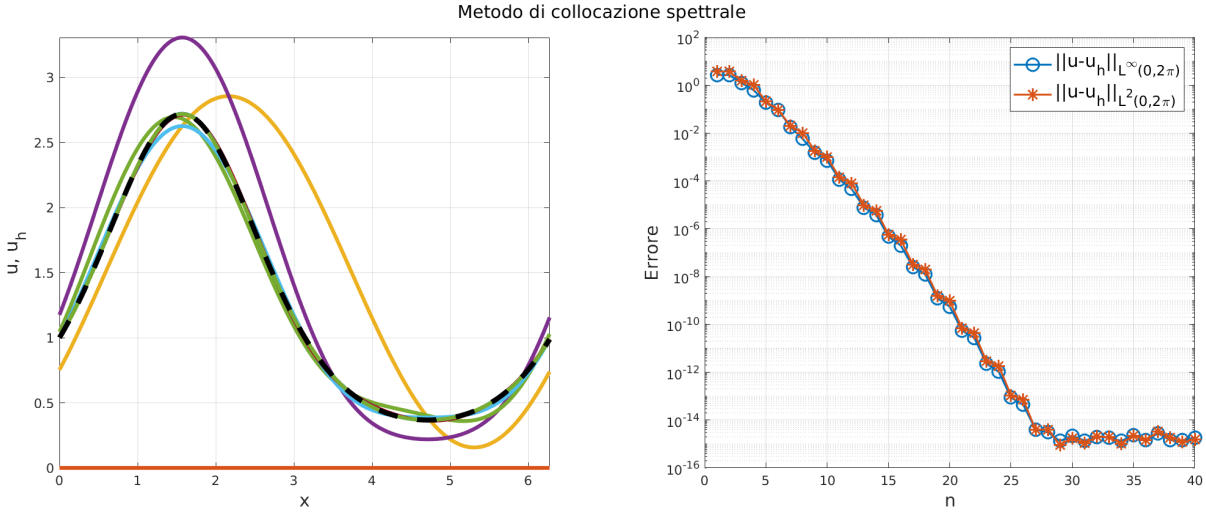



*Figura 29: Sinistra: le soluzioni $u_h$ ottenute con il metodo di collocazione spettrale per il problema periodico dell'Esercizio 5.10 per $n = 1,\dots,40$; la soluzione esatta è tratteggiata in nero. Destra: l'errore in norma $L^\infty(0,2\pi)$ e $L^2(0,2\pi)$. Il metodo raggiunge precisione macchina con meno di 30 gradi di libertà. Il metodo delle differenze finite applicato allo stesso problema sfiora un errore (misurato solo sui nodi) dell'ordine di 1e-9 con oltre $n = 50\,000$ gradi di libertà; aumentando ulteriormente $n$ l'errore di roundoff prende il sopravvento.*

L'Esercizio 5.10 e Figura 29 mostrano che, per questo problema al bordo, il metodo converge con velocità esponenziale (o spettrale): $\|u - u_h\|_{L^\infty(0,2\pi)} = \mathcal{O}(\mathrm{e}^{-bn})$ per $b > 0$. Al contrario, l'Esercizio 5.11 e Figura 30 mostrano che per un altro problema al bordo molto simile la convergenza è solo algebrica. Qual è la differenza tra i due problemi, che porta a velocità di convergenza tanto diverse? Le funzioni trigonometriche (o equivalentemente gli esponenziali complessi $\mathrm{e}^{\mathrm{i}kx}$) approssimano con velocità superalgebrica tutte le funzioni lisce e periodiche (e velocità esponenziale quelle analitiche e periodiche). La soluzione $u(x) = \mathrm{e}^{\sin x}$ del primo problema al bordo soddisfa queste condizioni. Invece la soluzione del secondo problema, quando viene estesa periodicamente oltre l'intervallo $(0,2\pi)$, non è liscia ma solo di classe $C^2(\mathbb{R})$.

### 5.3.1 La trasformata di Fourier discreta

Vediamo ora un modo per rendere più efficiente la soluzione del sistema lineare denso generato dal metodo di collocazione spettrale trigonometrico. Consideriamo ancora il problema al bordo periodico (50), assumendo che i coefficienti $p$ e $q > 0$ siano costanti in $x$. Fissiamo $n \in \mathbb{N}$ pari. Come prima, fissiamo gli elementi di base e i nodi

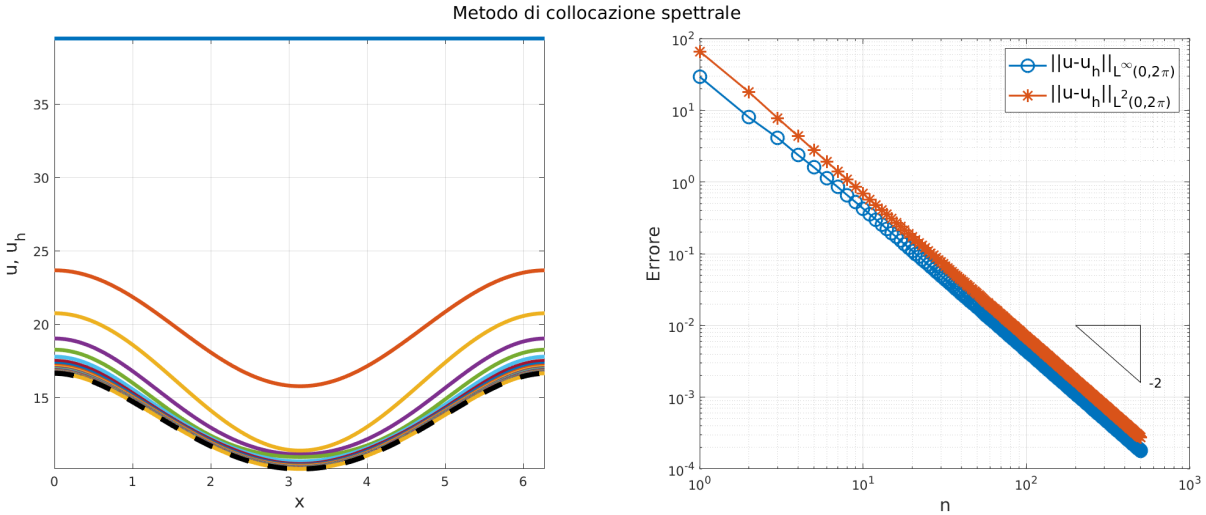



*Figura 30: Come Figura 29 per l'equazione dell'Esercizio 5.11 e $n$ fino a 500. In questo caso la convergenza è solamente algebrica e con $n = 500$ gradi di libertà raggiunge solamente un errore dell'ordine di 1e-4.*

$$\psi_k(x) := \mathrm{e}^{\mathrm{i}(k-\frac{n}{2}-1)x}, \qquad x_j := \frac{2\pi}{n}(j-1), \qquad k,j = 1,\dots,n.$$


Notiamo che valgono

$$\frac{\partial^m}{\partial x^m}\psi_k(x) = \mathrm{i}^m\Big(k-\frac{n}{2}-1\Big)^m\psi_k(x), \quad \psi_k(x_j) = \mathrm{e}^{\frac{2\pi\mathrm{i}}{n}(k-\frac{n}{2}-1)(j-1)} = (-1)^{j-1}\omega_n^{(k-1)(j-1)} \text{ dove } \boxed{\omega_n := \mathrm{e}^{\frac{2\pi\mathrm{i}}{n}}}.$$

Notiamo che $\omega_n$ è una radice $n$-sima dell'unità nel piano complesso, cioè $\omega_n^n = 1$. In particolare $\omega_n^j = \omega_n^{j+kn}$ per ogni $j,k \in \mathbb{Z}$. Il metodo di collocazione spettrale produce il vettore $\vec{\mathbf{f}}$ con $f_j = f(x_j)$ e la matrice $\underline{\underline{\mathbf{A}}}$ con elementi

$$A_{j,k} = \underbrace{\Big[\Big(k-\frac{n}{2}-1\Big)^2 + \mathrm{i}p\Big(k-\frac{n}{2}-1\Big)+q\Big]}_{=:D_k}\psi_k(x_j) = (-1)^{j-1}\omega_n^{(k-1)(j-1)}D_k, \qquad 1 \le j,k \le n.$$

Definendo

$$f_j^* := (-1)^{j-1}f(x_j), \qquad U_k^* := D_kU_k, \qquad W_{j,k} := \omega_n^{(j-1)(k-1)},$$

possiamo riscrivere il sistema lineare del metodo di collocazione:

$$\underline{\underline{\mathbf{A}}}\vec{\mathbf{U}} = \vec{\mathbf{f}} \iff \sum_{k=1}^n A_{j,k}U_k = f_j \iff \sum_{k=1}^n (-1)^{j-1}W_{j,k}D_kU_k = f_j \iff \sum_{k=1}^n W_{j,k}U_k^* = f_j^* \iff \underline{\underline{\mathbf{W}}}\vec{\mathbf{U}}^* = \vec{\mathbf{f}}^*.$$

Chiaramente $\vec{\mathbf{f}}^*$ e $\vec{\mathbf{U}}$ si possono calcolare con $\mathcal{O}(n)$ operazioni da $\vec{\mathbf{f}}$ e $\vec{\mathbf{U}}^*$, rispettivamente. Invece di $\underline{\underline{\mathbf{A}}}\vec{\mathbf{U}} = \vec{\mathbf{f}}$ punteremo a risolvere $\underline{\underline{\mathbf{W}}}\vec{\mathbf{U}}^* = \vec{\mathbf{f}}^*$: vedremo infatti che questo sistema si può risolvere più economicamente. La matrice

$$\underline{\underline{\mathbf{W}}} = \begin{pmatrix} 1 & 1 & 1 & 1 & \dots & 1 \\ 1 & \omega_n & \omega_n^2 & \omega_n^3 & \dots & \omega_n^{n-1} \\ 1 & \omega_n^2 & \omega_n^4 & \omega_n^6 & \dots & \omega_n^{2(n-1)} \\ \vdots & & & & \ddots & \vdots \\ 1 & \omega_n^{n-1} & \omega_n^{2(n-1)} & \omega_n^{3(n-1)} & \dots & \omega_n^{(n-1)^2} \end{pmatrix}$$

è detta **matrice di Fourier**. Due prime importanti proprietà di $\underline{\underline{\mathbf{W}}}$ sono le seguenti:

$$\underline{\underline{\mathbf{W}}} = \underline{\underline{\mathbf{W}}}^\top, \qquad \underline{\underline{\mathbf{W}}}^{-1} = \frac{1}{n}\underline{\underline{\mathbf{W}}}^H = \frac{1}{n}\overline{\underline{\underline{\mathbf{W}}}}, \tag{52}$$

dove $\overline{\phantom{x}}$ denota il coniugio complesso e $^H$ indica la matrice coniugata trasposta. In particolare $\frac{1}{\sqrt{n}}\underline{\underline{\mathbf{W}}}$ è **unitaria** (cioè le sue colonne sono vettori ortonormali, o $(\frac{1}{\sqrt{n}}\underline{\underline{\mathbf{W}}})^{-1} = \frac{1}{\sqrt{n}}\underline{\underline{\mathbf{W}}}^H$).

**Esercizio ✎ 5.13.** Dimostrare l'uguaglianza (52).
Dimostrare che $|\det(\underline{\underline{\mathbf{W}}})| = n^{n/2}$.
Suggerimento: usare la somma parziale della serie geometrica $\sum_{\ell=0}^{n-1} z^\ell = \frac{1-z^n}{1-z}$.

Il vantaggio di avere una matrice unitaria consiste nel fatto che il corrispondente sistema lineare si può risolvere con un semplice prodotto matrice vettore: $\vec{\mathbf{U}}^* = \underline{\underline{\mathbf{W}}}^{-1}\vec{\mathbf{f}}^* = \frac{1}{n}\overline{\underline{\underline{\mathbf{W}}}}\vec{\mathbf{f}}^*$. Questo ha complessità computazionale pari a $\mathcal{O}(n^2)$, invece di $\mathcal{O}(n^3)$ tipico della risoluzione di un sistema lineare denso. Quindi invece di risolvere direttamente $\underline{\underline{\mathbf{A}}}\vec{\mathbf{u}} = \vec{\mathbf{f}}$ procediamo così:

$$\vec{\mathbf{f}} \overset{\text{moltiplicando per } (-1)^{j-1}}{\rightsquigarrow} \vec{\mathbf{f}}^* \overset{\text{prodotto matrice–vettore}}{\rightsquigarrow} \vec{\mathbf{U}}^* = \frac{1}{n}\overline{\underline{\underline{\mathbf{W}}}}\vec{\mathbf{f}}^* \overset{\text{dividendo per } D_k}{\rightsquigarrow} \vec{\mathbf{U}}.$$

Da questo segue anche che, sotto le ipotesi fatte in questa sezione, il metodo di collocazione trigonometrico è ben posto.

La moltiplicazione di un vettore $\vec{\mathbf{v}}$ per la matrice $\overline{\underline{\underline{\mathbf{W}}}}$, cioè la mappa lineare (da $\mathbb{C}^n$ a $\mathbb{C}^n$)

$$\vec{\mathbf{v}} \mapsto \overline{\underline{\underline{\mathbf{W}}}}\vec{\mathbf{v}}, \qquad (\overline{\underline{\underline{\mathbf{W}}}}\vec{\mathbf{v}})_j = \sum_{k=1}^{n} \mathrm{e}^{-\frac{2\pi \mathrm{i}}{n}(j-1)(k-1)} v_k, \qquad 1 \le j \le n,$$

è detta **trasformata di Fourier discreta** (*discrete Fourier transform*, DFT). Useremo la notazione $\boxed{DFT_n(\vec{\mathbf{v}}) = \overline{\underline{\underline{\mathbf{W}}}}\vec{\mathbf{v}}}$ e denoteremo l'operazione inversa come $\boxed{IFT_n(\vec{\mathbf{z}}) = \frac{1}{n}\underline{\underline{\mathbf{W}}}\vec{\mathbf{z}}}$ (dove la lettera "I" sta per *inverse*).

**Nota 5.14.** Attenzione: molto spesso la trasformata di Fourier discreta è definita con diversi indici (da 0 a $n-1$, da $-\lfloor \frac{n}{2} \rfloor$ a $\lfloor \frac{n-1}{2} \rfloor$, ...), segni ($\mathrm{e}^{+\frac{2\pi \mathrm{i}}{n}}$) e normalizzazioni (fattore $\frac{1}{\sqrt{n}}$). La scelta fatta qui coincide esattamente con l'azione del comando Matlab "`fft`" (e "`ifft`" per la trasformata inversa).

**Nota 5.15.** Esistono diverse relazioni tra la trasformata di Fourier discreta (DFT) e le serie di Fourier (e la trasformata di Fourier continua). In generale, la DFT mette in relazione i valori in nodi equispaziati di una funzione periodica con i coefficienti della serie di Fourier corrispondente.

Fissiamo $n \in \mathbb{N}$ pari. Fissiamo un vettore $\vec{\mathbf{f}} \in \mathbb{C}^n$ e denotiamo $\vec{\mathbf{F}} = DFT_n(\vec{\mathbf{f}}) \in \mathbb{C}^n$. Scriviamo il seguente polinomio trigonometrico con coefficienti $F_k$:

$$f(x) = \sum_{k=1}^{n} F_k \psi_k(x) = \sum_{k=1}^{n} \left(DFT_n(\vec{\mathbf{f}})\right)_k \mathrm{e}^{\mathrm{i}(k-\frac{n}{2}-1)x} = \sum_{k=1}^{n}\sum_{\ell=1}^{n} \mathrm{e}^{-\frac{2\pi \mathrm{i}}{n}(\ell-1)(k-1)} f_\ell \, \mathrm{e}^{\mathrm{i}(k-\frac{n}{2}-1)x}.$$

Se valutiamo $f$ nei nodi $x_j = \frac{2\pi}{n}(j-1)$ per $j = 1, \dots, n$

$$\begin{aligned} f(x_j) &= \sum_{k=1}^{n}\sum_{\ell=1}^{n} \mathrm{e}^{-\frac{2\pi \mathrm{i}}{n}(\ell-1)(k-1)} f_\ell \, \mathrm{e}^{\frac{2\pi \mathrm{i}}{n}(k-\frac{n}{2}-1)(j-1)} \\ &= (-1)^{j-1} \sum_{\ell=1}^{n} f_\ell \sum_{k=1}^{n} \mathrm{e}^{\frac{2\pi \mathrm{i}}{n}(k-1)(j-\ell)} = (-1)^{j-1} \sum_{\ell=1}^{n} f_\ell \, n \, \delta_{j,\ell} = (-1)^{j-1} n f_j \end{aligned}$$

ritroviamo gli elementi di $f$, a meno di un fattore $(-1)^{j-1}n$. A parole: se interpretiamo gli elementi di un vettore $\vec{\mathbf{f}}$ come i valori in nodi equispaziati di un polinomio trigonometrico, la DFT ci fornisce i coefficienti di quel polinomio, cioè la sua serie di Fourier (che ha un numero finito di termini diversi da zero):

$$f(x) = \sum_{k=1}^{n} \mathrm{e}^{\mathrm{i}(k-\frac{n}{2})x} F_k, \qquad F_k = \frac{1}{2\pi}\int_0^{2\pi} f(x)\mathrm{e}^{-\mathrm{i}(k-\frac{n}{2})x}\,\mathrm{d}x.$$

Se invece abbiamo una funzione $f$ periodica di periodo $2\pi$ e sufficientemente liscia, la DFT del *sampling* $\vec{\mathbf{f}} = (f(x_1), \dots, f(x_n))^\top$ di $f$ (pesato con $(-1)^j$) sui nodi equispaziati $x_j$ corrisponde all'approssimazione dei coefficienti $F_k = \frac{1}{2\pi}\int_0^{2\pi} f(x)\mathrm{e}^{-\mathrm{i}kx}\,\mathrm{d}x$ della sua serie di Fourier $f(x) = \sum_{k\in\mathbb{Z}} \mathrm{e}^{\mathrm{i}kx} F_k$ attraverso la regola del trapezio:

$$F_k = \frac{1}{2\pi}\int_0^{2\pi} f(x)\mathrm{e}^{-\mathrm{i}(k-\frac{n}{2}-1)x}\,\mathrm{d}x \approx \frac{1}{2\pi}\sum_{j=1}^{n} f(x_j)\mathrm{e}^{-\mathrm{i}(k-\frac{n}{2}-1)x_j}(x_j - x_{j-1})$$

$$= \frac{1}{2\pi}\sum_{j=1}^{n} f(x_j)\mathrm{e}^{-\frac{2\pi\mathrm{i}}{n}(k-\frac{n}{2}-1)(j-1)}\frac{2\pi}{n}$$
$$= \frac{1}{n}\sum_{j=1}^{n}(-1)^j f(x_j)\mathrm{e}^{-\frac{2\pi\mathrm{i}}{n}(k-1)(j-1)} = \frac{1}{n}DFT_n\big([(-1)^j f(x_j)]_{j=1,\dots,n}\big).$$

### 5.3.2 La FFT: la trasformata di Fourier veloce

Se plottiamo il tempo impiegato da Matlab per risolvere il sistema lineare $\underline{\underline{\mathbf{A}}}\vec{\mathbf{U}} = \vec{\mathbf{f}}$ usando (1) il comando backslash `U=A\f`, (2) la moltiplicazione per $\underline{\underline{\mathbf{W}}}$ e (3) il comando `fft` per calcolare l'azione di $\underline{\underline{\mathbf{W}}}$ otteniamo un grafico come quello in Figura 31. (Ovviamente i tempi dipendono dai dettagli dell'implementazione, dal computer usato, dalla versione di Matlab, ...)

**Nota 5.16.** I vettori soluzione del metodo di collocazione spettrale usato per Figura 31 sono stati calcolati con il seguente codice Matlab. Notare che in tutti e tre i casi non è necessario salvare in memoria nessuna matrice ma ogni vettore può essere calcolato in una sola riga.

```
% DATI: n (naturale pari), p (reale), q (positivo), f_fun (funzione)
xnodes = 2*pi/n * (0:n-1)';     % Vettore colonna dei nodi
f = f_fun(xnodes);              % Vettore colonna valori di f nei nodi
kk = -n/2 : (n/2-1);            % Vettore riga indici funzioni di base
U_bcks  = ((ones(n,1)*kk.^2 + 1i*p*kk + q) .* exp(1i*xnodes * kk)) \ f;
U_multW = ((exp(1i*2*pi/n)).^(-(0:n-1)'*(0:n-1))*((-1).^(0:n-1)'.*f ))
   ./ (n*(kk'.^2 + 1i*p*kk' + q));
U_FFT   = (fft( (-1).^(0:n-1)' .* f)) ./ (n*(kk'.^2 + 1i*p*kk' + q));
```

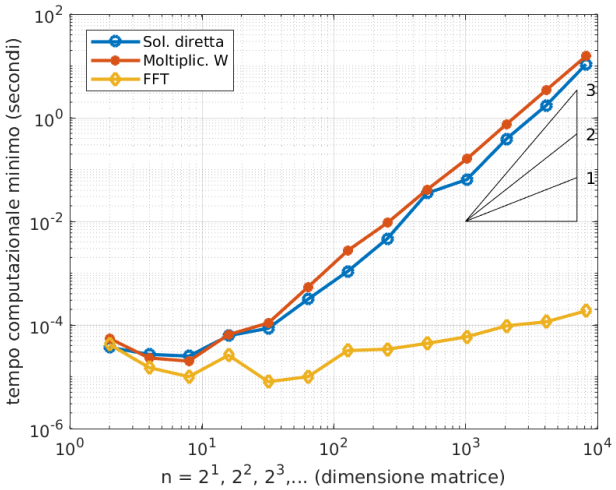



*Figura 31: I tempi impiegato da Matlab per risolvere il sistema lineare prodotto dal metodo spettrale per il problema al bordo dell'Esercizio 5.11 con $n = 2^1, \dots, 2^{13}$. Il sistema lineare è risolto (1) direttamente con il comando backslash, (2) moltiplicando per $\underline{\underline{\mathbf{W}}}$, (3) con il comando `fft`.*

**Esercizio 💻 5.17.** Ricostruire il grafico di Figura 31: confrontare e plottare i tempi di calcolo necessari per approssimare il problema al bordo dell'Esercizio 5.11 con il metodo di collocazione spettrale al crescere di $n$ risolvendo il sistema lineare nei tre modi descritti nella Nota 5.16.

Per calcolare i tempi computazionali in Matlab si possono usare i comandi `tic` e `toc`. (I comandi `clock` e `etime` possono svolgere lo stesso compito ma sono molto meno precisi, meglio usare `tic` e `toc`.) Ogni esperimento è stato ripetuto 5 volte e il tempo minore dei 5 è stato plottato.

È importante assicurarsi che le soluzioni ottenute siano quelle corrette. Ad esempio si possono misurare le norme delle differenze tra i vettori `U_bcks`, `U_multW`, `U_FFT` (con i nomi usati nel codice nella Nota 5.16).

Notiamo prima di tutto che la risoluzione diretta del sistema lineare ha un tempo di esecuzione simile a quello della moltiplicazione con $\underline{\underline{\mathbf{W}}}$ e cresce con ordine poco più che quadratico in $n$. Questo avviene

perché Matlab riconosce la struttura della matrice e risolve il sistema lineare in modo efficiente. Inoltre $n$ è abbastanza moderato quindi siamo in un regime "preasintotico".

Ma c'è un secondo fatto più interessante: la moltiplicazione per $\underline{\underline{\mathbf{W}}}$ calcolata con il comando `fft` è estremamente più economica della risoluzione diretta: per $n = 2^{13} = 8192$ il prodotto matrice–vettore richiede circa 10 secondi mentre la stessa operazione con il comando `fft` ci mette meno di 0.2 millisecondi, un risparmio di un fattore 50 000! Come fa Matlab a calcolare questo prodotto matrice vettore così velocemente? Il trucco sta nell'uso della **trasformata di Fourier veloce** (*fast Fourier transform*, FFT). Questo algoritmo, pubblicato nel 1965 da Cooley e Tukey (ma già scoperto da Gauss) permette di calcolare il prodotto della matrice $\overline{\underline{\mathbf{W}}}$ di dimensione $n \times n$ per un vettore di lunghezza $n$ con costo $\mathcal{O}(n \log_2 n)$. La FFT è un algoritmo estremamente efficiente per calcolare la DFT.

L'FFT è stata scelta come uno dei "10 algoritmi del ventesimo secolo". La sua importanza non è dovuta solo all'uso nei metodi spettrali: la DFT (calcolata efficientemente attraverso la FFT) è uno strumento fondamentale nell'analisi, sintesi, elaborazione e compressione di segnali. In particolare è usata per rappresentare segnali "in tempo" in termini delle loro componenti "in frequenza", e viceversa. Ad esempio i formati .mp3 e .jpeg per la compressione di segnali audio e di immagini, rispettivamente, si basano sulla DFT.

Descriviamo ora l'algoritmo della FFT di Cooley–Tukey nel caso più semplice. Vogliamo calcolare efficientemente la DFT di $\vec{\mathbf{x}} \in \mathbb{C}^n$, cioè il prodotto

$$\vec{\mathbf{y}} = DFT_n(\vec{\mathbf{x}}) = \overline{\underline{\mathbf{W}}}\vec{\mathbf{x}}, \quad \text{dove} \quad (\overline{\underline{\mathbf{W}}})_{j,k} = \overline{\omega}_n^{(j-1)(k-1)} = \mathrm{e}^{-\mathrm{i}\frac{2\pi}{n}(j-1)(k-1)}.$$

Notiamo che, se $n = 2m$ è pari, possiamo separare il contributo degli elementi di $\vec{\mathbf{x}}$ con indici pari e dispari ($k = 2\ell - 1$ e $k = 2\ell$ per $1 \le \ell \le m$):

$$\begin{aligned}
y_j = \big(DFT_n(\vec{\mathbf{x}})\big)_j &= \sum_{k=1}^{n} x_k \mathrm{e}^{-\mathrm{i}\frac{2\pi}{n}(j-1)(k-1)} \\
&= \sum_{\ell=1}^{m} x_{2\ell-1} \mathrm{e}^{-\mathrm{i}\frac{2\pi}{n}(j-1)2(\ell-1)} + \mathrm{e}^{-\mathrm{i}\frac{2\pi}{n}(j-1)} \sum_{\ell=1}^{m} x_{2\ell} \mathrm{e}^{-\mathrm{i}\frac{2\pi}{n}(j-1)2(\ell-1)} \\
&= \sum_{\ell=1}^{m} x_{2\ell-1} \mathrm{e}^{-\mathrm{i}\frac{2\pi}{m}(j-1)(\ell-1)} + \mathrm{e}^{-\mathrm{i}\frac{2\pi}{n}(j-1)} \sum_{\ell=1}^{m} x_{2\ell} \mathrm{e}^{-\mathrm{i}\frac{2\pi}{m}(j-1)(\ell-1)} \\
&= \big(DFT_m(\vec{\mathbf{x}}^{\text{dispari}})\big)_j + \mathrm{e}^{-\mathrm{i}\frac{2\pi}{n}(j-1)} \big(DFT_m(\vec{\mathbf{x}}^{\text{pari}})\big)_j, \qquad 1 \le j \le n,
\end{aligned}$$

dove abbiamo usato $\frac{n}{2} = m$, abbiamo definito

$$\vec{\mathbf{x}}^{\text{dispari}} := (x_1, x_3, \cdots, x_{n-1})^\top \in \mathbb{C}^m, \qquad \vec{\mathbf{x}}^{\text{pari}} := (x_2, x_4, \cdots, x_n)^\top \in \mathbb{C}^m,$$

e abbiamo assunto per convenienza la **periodicità** $(DFT_m(\vec{\mathbf{v}}))_{j+m} := (DFT_m(\vec{\mathbf{v}}))_j$, $j = 1, \ldots, m$. Questa identità ci dice che **la DFT di lunghezza $2m$ si può calcolare come due DFT di lunghezza $m$ più $2m$ addizioni e moltiplicazioni**.

Se $n$ è una potenza di 2, cioè $n = 2^N$, allora possiamo iterare il processo in modo ricorsivo: $DFT_{2^N}(\vec{\mathbf{x}})$ si può calcolare come due $DFT_{2^{N-1}}$, cioè come quattro $DFT_{2^{N-2}}$, come otto $DFT_{2^{N-3}}$, …, come $2^N$ $DFT_1$ (la DFT di lunghezza 1 è semplicemente l'identità, poiché $W_1 = 1$). Questo processo, detto "*divide and conquer*", è l'idea alla base della FFT. La complessità computazionale è pari a $\mathcal{O}(2^N N) = \boxed{\mathcal{O}(n \log_2 n)}$ (in Matlab il costo è circa di $5n \log_2 n$). Al contrario, il calcolo del prodotto matrice–vettore con la somma sugli indici richiede $\mathcal{O}(n^2)$ operazioni.

Per avere un'idea della velocità ottenuta, l'FFT di un vettore di lunghezza $n = 10^7$ (10 milioni) con Matlab sul mio desktop richiede meno di 0.2 secondi. La matrice $\underline{\underline{\mathbf{W}}}$ in formato double richiederebbe 800 TB di memoria, 100 000 volte più della RAM installata sullo stesso computer.

L'operazione inversa della DFT è la moltiplicazione per la matrice $\overline{\underline{\mathbf{W}}}^{-1} = \frac{1}{n}\underline{\underline{\mathbf{W}}}$, che può essere calcolato con un algoritmo identico e con la stessa complessità sostituendo $\overline{\omega}_n = \mathrm{e}^{-\frac{2\pi\mathrm{i}}{n}}$ con $\omega_n = \mathrm{e}^{\frac{2\pi\mathrm{i}}{n}}$.

**Nota 5.18.** Abbiamo assunto che $n = 2^N$, cosa succede per altri valori di $n$? Se $n$ può essere scomposto in fattori primi piccoli si può ottenere un'algoritmo simile e con costo computazionale equivalente. Se la fattorizzazione di $n$ contiene primi molto grandi allora l'algoritmo della FFT diventa più complicato e costoso.

D'altro canto per molte applicazioni il valore di $n$ non è specificato a priori ma può essere scelto come la più vicina potenza di due. Ad esempio per il metodo di collocazione spettrale $n$ è un parametro scelto da chi usa il metodo.

Il codice seguente è un'implementazione in Matlab dell'algoritmo di Cooley–Tukey per $n = 2^N$. Notare la struttura ricorsiva dell'algoritmo.

```
function Y = fft_rec(y)     % y e Y sono vettori colonna
n = length (y);
if n == 1
    Y = y;
else
    Ydisp = fft_rec(y(1:2:n));
    Ypari = fft_rec(y(2:2:n));
    Y = [Ydisp; Ydisp] + (exp(-2* pi*1i/n).^((0:n-1)')).*[Ypari;Ypari];
end
end
```

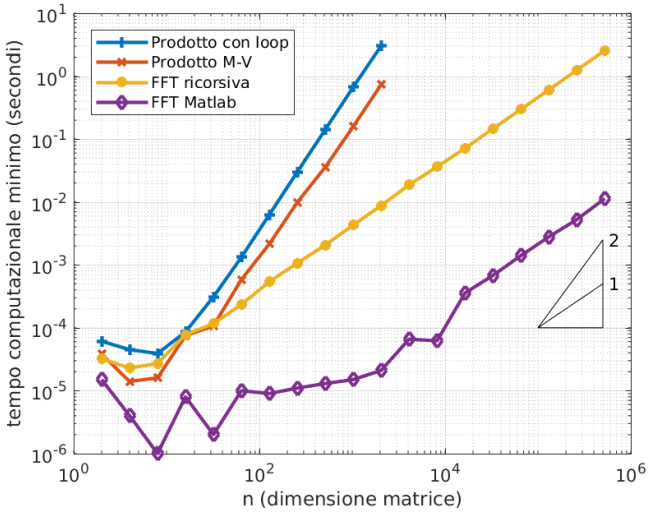



*Figura 32: I tempi impiegato da Matlab per calcolare la DFT di un vettore random usando: (1) il prodotto matrice–vettore implementato con un doppio ciclo for, (2) il prodotto matrice–vettore implementato con un unico comando, (3) l'implementazione ricorsiva della FFT mostrata nel riquadro, (4) il comando* `fft`*. Gli ordini di convergenza numerici dedotti dai valori più alti di $n$ sono:* 2.223, 2.088, 1.033, 0.994.

**Esercizio 💻 5.19.** Ricostruire il grafico di Figura 32: confrontare i tempi di calcolo della DFT, implementata in quattro modi diversi, al crescere di $n$.

**Esempio 5.20.** Un'applicazione della FFT è l'accelerazione del prodotto matrice–vettore per matrici circolanti.

In questo esempio usiamo la notazione periodica per gli indici dei vettori $\vec{\mathbf{v}} \in \mathbb{C}^n$, cioè $v_{j+n} = v_j$, $\forall j \in \mathbb{Z}$.

Sia $\underline{\underline{\mathbf{C}}} \in \mathbb{C}^{n\times n}$ una matrice circolante, cioè con componenti $C_{j,k} = c_{j-k+1} = c_{n+j-k+1}$, dove il vettore $\vec{\mathbf{c}} = (c_1, \ldots, c_n)^\top$ è la sua prima colonna:

$$\underline{\underline{\mathbf{C}}} = \begin{pmatrix} c_1 & c_n & c_{n-1} & c_{n-2} & \cdots & c_2 \\ c_2 & c_1 & c_n & c_{n-1} & \cdots & c_3 \\ c_3 & c_2 & c_1 & c_n & \cdots & c_4 \\ c_4 & c_3 & c_2 & c_1 & \cdots & c_5 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ c_n & c_{n-1} & c_{n-2} & c_{n-3} & \cdots & c_1 \end{pmatrix}.$$

A parole: una matrice è circolante se ogni colonna è la traslazione (periodica) verso il basso della colonna precedente. Sia $\vec{\mathbf{v}}^{(\ell)}$ la colonna $\ell$esima della matrice di Fourier $\underline{\underline{\mathbf{W}}}$, cioè $v_k^{(\ell)} = \omega_n^{(k-1)(\ell-1)}$. Moltiplicando $\underline{\underline{\mathbf{C}}}$ a $\vec{\mathbf{v}}^{(\ell)}$ otteniamo

$$\begin{aligned}(\underline{\underline{\mathbf{C}}}\vec{\mathbf{v}}^{(\ell)})_j &= \sum_{k=1}^{n} c_{j-k+1}\omega_n^{(k-1)(\ell-1)} \overset{(m=j-k+1)}{=} \sum_{m=1}^{n} c_m \omega_n^{(j-m)(\ell-1)} = \omega_n^{(j-1)(\ell-1)} \sum_{m=1}^{n} c_m \omega_n^{-(m-1)(\ell-1)} \\ &= v_j^{(\ell)}\ (\underline{\underline{\mathbf{W}}}\vec{\mathbf{c}})_\ell.\end{aligned}$$

Questo significa che $\vec{\mathbf{v}}^{(\ell)}$ è un autovettore di $\underline{\underline{\mathbf{C}}}$ con autovalore $(\overline{\underline{\mathbf{W}}}\vec{\mathbf{c}})_\ell$. (In particolare, tutte le matrici circolanti hanno gli stessi autovettori!) Usando ancora le proprietà (52) di $\overline{\underline{\mathbf{W}}}$, possiamo diagonalizzare $\underline{\underline{\mathbf{C}}}$:

$$\underline{\underline{\mathbf{C}}} = \underline{\underline{\mathbf{W}}}\,\mathrm{diag}(\overline{\underline{\mathbf{W}}}\vec{\mathbf{c}})\,\underline{\underline{\mathbf{W}}}^{-1} = \frac{1}{n}\underline{\underline{\mathbf{W}}}\,\mathrm{diag}(\overline{\underline{\mathbf{W}}}\vec{\mathbf{c}})\,\overline{\underline{\mathbf{W}}} \qquad \Rightarrow \qquad \underline{\underline{\mathbf{C}}}\vec{\mathbf{v}} = DFT_n^{-1}\big(DFT_n(\vec{\mathbf{c}}) : DFT_n(\vec{\mathbf{v}})\big),$$

dove il segno ":" sta per il prodotto termine a termine tra due vettori, $(\vec{\mathbf{p}} : \vec{\mathbf{q}})_j = p_j q_j$ (`.*` in Matlab). Il prodotto matrice–vettore tra una matrice circolante e un vettore si può quindi calcolare con due FFT e una FFT inversa, con costo computazionale $\mathcal{O}(n\log_2 n)$, invece di $\mathcal{O}(n^2)$. Similmente anche un sistema lineare con matrice circolante si può risolvere con lo stesso costo.

In Matlab il prodotto $\underline{\underline{\mathbf{C}}}\vec{\mathbf{v}}$ si può calcolare semplicemente con il comando `ifft( fft(c) .* fft(v) )`.

Il prodotto tra $\underline{\underline{\mathbf{C}}}$ e un vettore $\vec{\mathbf{v}} = (v_1, \dots, v_n)^\top$ è la convoluzione discreta periodica tra i vettori $\vec{\mathbf{c}}$ e $\vec{\mathbf{v}}$: $(\underline{\underline{\mathbf{C}}}\vec{\mathbf{v}})_j = (\vec{\mathbf{c}} \star \vec{\mathbf{v}})_j = \sum_{k=1}^n c_{j-k+1} v_k$. Quindi anche la convoluzione discreta periodica può essere calcolata in $\mathcal{O}(n\log_2 n)$ operazioni usando la FFT.

**Esercizio** ✎+💻 **5.21.** Usare la FFT per risolvere in $\mathcal{O}(n\log_2 n)$ operazioni il sistema lineare $\underline{\underline{\mathbf{C}}}\vec{\mathbf{x}} = \vec{\mathbf{y}}$ per $\vec{\mathbf{y}} \in \mathbb{C}^n$ e $\underline{\underline{\mathbf{C}}}$ circolante. Implementare e testare l'algoritmo in Matlab.

È possibile calcolare $\vec{\mathbf{x}}$ a partire da $\vec{\mathbf{y}}$ e $\vec{\mathbf{c}}$ con un brevissimo comando di una sola riga.

Per testare la correttezza della soluzione provare a costruire la matrice $\underline{\underline{\mathbf{C}}}$ da $\vec{\mathbf{c}}$ usando `repmat` e `mod`.

**Esercizio** 💻 **5.22.** Dato $n \in \mathbb{N}$,calcolare numericamente il vettore $\vec{\mathbf{x}} \in \mathbb{R}^n$ tale che

$$\sum_{k=1}^{n}(j-k)x_k + \sum_{k=j+1}^{n} n x_k = j \qquad \text{per ogni } j = 1, \dots, n.$$

Ad esempio, per $n = 100\,000$, si ottiene $x_1 = 1+\delta$, $x_2 = \cdots = x_n = \delta$, per $\delta = 2.000020000200002 \cdot 10^{-10}$.

Suggerimento: scrivere il problema come un sistema lineare (scriverlo a mano esplicitamente per $n$ piccolo) e risolverlo con la FFT come nell'Esempio 5.20. Dato $n$, è possibile calcolare $\vec{\mathbf{x}}$ con un unico comando Matlab di meno di 30 caratteri!

(È anche possibile verificare la correttezza dei risultati ottenuti calcolando a mano il valore di $\vec{\mathbf{x}}$, sapendo che per ogni $n$ la soluzione è nella forma $x_1 = 1+\delta_n$, $x_2 = \cdots = x_n = \delta_n$, per qualche $\delta_n > 0$.)

**Nota 5.23** (FFT e collocazione polinomiale)**.** Abbiamo visto che la FFT si può usare per rendere molto efficiente il metodo di collocazione per il problema periodico, sfruttando la relazione tra le funzioni di base periodiche e le radici dell'unità. La FFT può essere usata anche per il metodo di collocazione con basi polinomiali. Invece dei polinomi di Legendre considerati in §5.2 è conveniente usare i polinomi di Chebyshev, definiti come $T_k(x) := \cos(k \arccos x)$, per $k \in \mathbb{N}_0$ e $-1 \le x \le 1$. Questa relazione tra i polinomi e le funzioni trigonometriche permette di adattare la FFT a questa versione del metodo di collocazione.

# 6 Problemi variazionali e metodo di Galerkin

## 6.1 Formulazione debole di un problema al contorno

Finora abbiamo scritto i problemi ai limiti come equazioni differenziali del secondo ordine completate da condizioni al bordo. Questa forma richiede i valori puntuali delle derivate prime e seconde della soluzione $u$. A volte è importante considerare **problemi la cui soluzione non è differenziabile due volte con continuità**. Ad esempio, se $u$ rappresenta la temperatura di una barra metallica estesa dal punto $a$ al punto $b$ di cui solo una parte $(c,d) \subsetneq (a,b)$ viene riscaldata, possiamo modellare il problema con un termine di sorgente $f$ discontinuo: la soluzione $u$ non potrà essere di classe $C^2$. In questo caso è necessario scrivere il problema in una forma più generale.

Consideriamo ancora una volta il problema di Dirichlet per l'equazione di diffusione–reazione con condizioni al bordo omogenee (e $q \ge 0$):

$$\begin{cases} -u''(x) + q(x)u(x) = f(x) & \text{in } (a,b) \\ u(a) = 0, \\ u(b) = 0. \end{cases} \tag{53}$$

Moltiplicando l'equazione differenziale per una qualsiasi funzione $w$ sufficientemente liscia con $w(a) = w(b) = 0$ e integrando per parti troviamo

$$\int_a^b (-u'' + qu)w\,\mathrm{d}x = \int_a^b (u'w' + quw)\,\mathrm{d}x - u'(b)w(b) + u'(a)w(a)$$

$$\Rightarrow \qquad \int_a^b fw\,\mathrm{d}x = \int_a^b (u'w' + quw)\,\mathrm{d}x. \tag{54}$$

Questa equazione contiene $u'$ ma non la derivata seconda $u''$. Vogliamo usare questa equazione "variazionale", cioè che deve essere valida al variare di $w$ in uno spazio funzionale adeguato, come alternativa a (53). Per questo introduciamo alcuni importanti spazi di funzioni.[16]

**Definizione 6.1.** Denotiamo con $L^2(a,b)$ lo spazio vettoriale delle funzioni definite sull'intervallo $(a,b)$ a quadrato integrabile, cioè con norma e prodotto scalare

$$\|u\|_{L^2(a,b)} := \Big(\int_a^b u^2\,\mathrm{d}x\Big)^{1/2} < \infty, \qquad (u,w)_{L^2(a,b)} := \int_a^b uw\,\mathrm{d}x.$$

Dato $k \in \mathbb{N}$, chiamiamo $H^k(a,b)$ lo spazio delle funzioni $u$ su $[a,b]$ tali che $u$ e tutte le derivate fino a $u^{(k-1)}$ sono assolutamente continue e tali che $u^{(k)} \in L^2(a,b)$. Usiamo come norma e prodotto scalare su $H^k(a,b)$

$$\|u\|_{H^k(a,b)} := \Big(\sum_{m=0}^{k} \big\|u^{(m)}\big\|^2_{L^2(a,b)}\Big)^{1/2}, \qquad (u,w)_{H^k(a,b)} := \sum_{m=0}^{k} \int_a^b u^{(m)} w^{(m)}\,\mathrm{d}x.$$

(Qui intendiamo $u^{(0)} = u$.) Definiamo inoltre il sottospazio

$$H^1_0(a,b) = \{u \in H^1(a,b),\ u(a) = u(b) = 0\}.$$

Useremo anche la seminorma $|u|_{H^1(a,b)} := \|u'\|_{L^2(a,b)}$, definita per $u \in H^1(a,b)$.
Gli spazi $H^k(a,b)$ sono detti **spazi di Sobolev**.

Possiamo pensare $H^k(a,b)$ semplicemente come il sottospazio vettoriale di $L^2(a,b)$ delle **funzioni le cui derivate fino all'ordine $k$ sono in $L^2(a,b)$**.[17]

Chiaramente, se $u \in H^k(a,b)$ allora $u^{(j)} \in H^{k-j}(a,b)$ per $1 \le j < k$ e $u^{(k)} \in L^2(a,b)$. Inoltre, per ogni $k \in \mathbb{N}$, $H^k(a,b) \subset C^{k-1}([a,b])$ e, se $(a,b)$ è un intervallo limitato, $C^k([a,b]) \subset H^k(a,b)$.

Gli spazi $H^k(a,b)$ sono spazi di Hilbert: spazi vettoriali forniti di un prodotto scalare che induce una norma e quindi una distanza rispetto alla quale sono completi (le successioni di Cauchy sono convergenti).

Tra questi spazi useremo solo $L^2(a,b), H^1(a,b), H^2(a,b)$ e $H^1_0(a,b)$.

**Esercizio ✎ 6.2.** Fissato $(a,b) = (-1,1)$, dire a quali degli spazi introdotti nella Definizione 6.1 appartengono le seguenti funzioni:

$$\sin(\pi x), \quad \mathrm{sign}(x), \quad \max\Big\{0, \frac{1}{2} - |x|\Big\}, \quad \max\{0, x^{13}\}, \quad |x|^\gamma \text{ al variare di } \gamma \in \mathbb{R}.$$

La disuguaglianza di Cauchy–Schwarz in $L^2(a,b)$ (o di Hölder) garantisce che, date due funzioni $\varphi, \psi$ su $(a,b)$, il loro prodotto è integrabile se entrambe sono in $L^2(a,b)$:

$$\int_a^b \varphi\psi\,\mathrm{d}x \le \|\varphi\|_{L^2(a,b)}\,\|\psi\|_{L^2(a,b)}\,. \tag{55}$$

Da questo segue che tutti i termini nell'uguaglianza (54) sono limitati se $u, w \in H^1(a,b)$, $f \in L^2(a,b)$ e $q \in L^\infty(a,b)$. Questo ci suggerisce la seguente definizione.

[16] Ricordiamo che $u : [a,b] \to \mathbb{R}$ è detta "assolutamente continua" se per ogni $\epsilon > 0$ esiste $\delta > 0$ tale che per ogni successione di sottointervalli $(x_k, y_k) \subset (a,b)$ a due a due disgiunti e di lunghezza totale $\sum_K (y_k - x_k) < \delta$, vale $\sum_k |u(y_k) - u(x_k)| < \epsilon$. Se $u$ è assolutamente continua allora la sua derivata $u'$ è definita quasi ovunque in $[a,b]$, $u'$ è integrabile secondo Lebesgue e vale il "teorema fondamentale del calcolo integrale di Lebesgue", cioè $u(x) = u(a) + \int_a^x u'(s)\,\mathrm{d}s$. Viceversa, se $v : [a,b] \to \mathbb{R}$ non è assolutamente continua, allora non esiste nessuna funzione $g$ integrabile su $[a,b]$ tale che $v(x) = v(a) + \int_a^x g(s)\,\mathrm{d}s$.

[17] Gli spazi di Sobolev $H^k(a,b)$ sono spesso definiti in modo equivalente usando "derivate deboli" e "distribuzioni", ad esempio in [QSSG14, §11.3.2]. Questa formulazione è necessaria per estendere la definizione a spazi di funzioni definite su aperti di $\mathbb{R}^d$ per $d > 1$: in questo caso la definizione con l'assoluta continuità non sarebbe quella corretta. Poiché qui ci accontentiamo del caso unidimensionale possiamo usare tranquillamente Definizione 6.1, come in [SM03, §14.1], senza dover introdurre il concetto di derivata distribuzionale. Ricordiamo che se $u$ è assolutamente continua allora è differenziabile quasi ovunque, quindi possiamo chiederci se $u'$ appartiene o meno a $L^2(a,b)$ senza dover introdurre nuovi concetti di derivata.

**Definizione 6.3.** Data $f \in L^2(a,b)$, $0 \le q \in L^\infty(a,b)$, la **formulazione debole** del problema (53) è:

$$\text{trovare } u \in H^1_0(a,b) \text{ tale che} \qquad \int_a^b (u'w' + quw)\,\mathrm{d}x = \int_a^b fw\,\mathrm{d}x \quad \text{per ogni } w \in H^1_0(a,b). \tag{56}$$

Una soluzione di (56) è detta **soluzione debole** del problema (53).

Le funzioni $w \in H^1_0(a,b)$ in (56) sono dette **funzioni test**. Se $u \in C^2(a,b) \cap C^0([a,b])$ è soluzione di (53) "puntualmente", con $f, q \in C^0(a,b)$, allora è detta **soluzione classica**.

La derivazione in (54) mostra che una soluzione classica di (53) è anche soluzione debole.

Supponiamo ora che $u$ e $\tilde{u}$ siano entrambe soluzioni deboli dello stesso problema. Avremo

$$\int_a^b \big((u'w' + quw) - (\tilde{u}'w' + q\tilde{u}w)\big)\,\mathrm{d}x = \int_a^b (fw - fw)\,\mathrm{d}x = 0.$$

Scegliendo $w = u - \tilde{u}$ (che è in $H^1_0(a,b)$, quindi è una scelta ammissibile) e usando $q \ge 0$ abbiamo

$$\|u' - \tilde{u}'\|^2_{L^2(a,b)} \le \int_a^b \big((u' - \tilde{u}')^2 + q(u - \tilde{u})^2\big)\,\mathrm{d}x = 0.$$

Questo garantisce che $u' = \tilde{u}'$, cioè $u - \tilde{u}$ è costante in $(a,b)$. Poiché $u(a) = 0 = \tilde{u}(a)$, abbiamo $u = \tilde{u}$. Abbiamo dimostrato l'**unicità della soluzione debole** del problema (53).

**Nota 6.4.** Immaginiamo di avere due vettori $\vec{\mathbf{u}}, \vec{\mathbf{v}} \in \mathbb{R}^n$ e di non avere accesso alle singole componenti $u_j, v_j$ ma di essere in grado di calcolare i loro prodotti scalari con altri vettori. Vogliamo verificare se $\vec{\mathbf{u}}$ e $\vec{\mathbf{v}}$ sono uguali tra loro. Un modo ovvio è verificare se $\vec{\mathbf{u}} \cdot \vec{\mathbf{w}} = \vec{\mathbf{v}} \cdot \vec{\mathbf{w}}$ per ogni $\vec{\mathbf{w}} \in \mathbb{R}^n$, cioè "testare" l'identità desiderata sfruttando il prodotto scalare di $\mathbb{R}^n$. Si verifica immediatamente che testare contro tutti gli infiniti $\vec{\mathbf{w}}$ di $\mathbb{R}^n$ è equivalente a testare contro gli $n$ elementi di una base dello stesso spazio. La formulazione debole (56) (e quelle variazionali più generali che vedremo in seguito) estendono questa idea a spazi infinito-dimensionali, ad esempio spazi di funzioni come $H^1_0(a,b)$.

Il seguente risultato permette di interpretare il problema debole come un problema di minimo in uno spazio funzionale: la soluzione debole minimizza una "energia" $J$.

**Proposizione 6.5** (Principio di Ritz)**.** Definiamo il funzionale quadratico

$$J : H^1_0(a,b) \to \mathbb{R}, \qquad J(w) := \frac{1}{2}\int_a^b \big((w')^2 + qw^2\big)\,\mathrm{d}x - \int_a^b fw\,\mathrm{d}x.$$

Il problema di minimo

$$\text{trovare } u \in H^1_0(a,b) \text{ tale che } J(u) = \min_{w \in H^1_0(a,b)} J(w) \tag{57}$$

è equivalente alla formulazione debole (56).

In questa proposizione l'equivalenza tra i due problemi (56) e (57) significa che per ogni scelta di dati $(f, q, a, b)$ una soluzione di un problema è soluzione anche dell'altro.

*Dimostrazione.* (I) Assumiamo che $u$ sia soluzione di (57) e mostriamo che è anche soluzione di (56). Fissiamo $w \in H^1_0(a,b)$ e $\epsilon \in \mathbb{R}$ arbitrari. Allora

$$\Psi(\epsilon) := J(u + \epsilon w) \ge J(u).$$

Espandiamo la funzione (reale di variabile reale) $\Psi$ e la sua derivata:

$$\begin{aligned}
\Psi(\epsilon) &= \frac{1}{2}\int_a^b \big((u' + \epsilon w')^2 + q(u + \epsilon w)^2\big)\,\mathrm{d}x - \int_a^b f(u + \epsilon w)\,\mathrm{d}x \\
&= \frac{1}{2}\int_a^b \big((u')^2 + \epsilon^2(w')^2 + 2\epsilon u'w' + qu^2 + q\epsilon^2 w^2 + 2q\epsilon uw\big)\,\mathrm{d}x - \int_a^b (fu + \epsilon fw)\,\mathrm{d}x \\
&= J(u) + \epsilon\int_a^b \big(u'w' + quw - fw\big)\,\mathrm{d}x + \frac{1}{2}\epsilon^2\int_a^b \big((w')^2 + qw^2\big)\,\mathrm{d}x, \\
\frac{\partial\Psi}{\partial\epsilon}(\epsilon) &= \int_a^b \big(u'w' + quw - fw\big)\,\mathrm{d}x + \epsilon\int_a^b \big((w')^2 + qw^2\big)\,\mathrm{d}x.
\end{aligned}$$

La funzione $\Psi$ ha un minimo in zero, quindi

$$0 = \Psi'(0) = \int_a^b \big(u'w' + quw - fw\big)\,\mathrm{d}x;$$

poiché $w \in H^1_0(a,b)$ è arbitraria, questo significa che $u$ è soluzione debole del problema.

(II) Assumiamo ora che $u$ sia soluzione di (56) e calcoliamo $J(u+w)$ per una $w \in H^1_0(a,b)$ arbitraria:

$$\begin{aligned} J(u+w) &= \frac{1}{2}\int_a^b \big((u'+w')^2 + q(u+w)^2\big)\,\mathrm{d}x - \int_a^b f(u+w)\,\mathrm{d}x \\ &= \underbrace{\frac{1}{2}\int_a^b \big((u')^2 + qu^2\big)\,\mathrm{d}x - \int_a^b fu\,\mathrm{d}x}_{=J(u)} + \underbrace{\int_a^b \big(u'w' + quw - fw\big)\,\mathrm{d}x}_{=0} + \underbrace{\frac{1}{2}\int_a^b \big((w')^2 + qw^2\big)\,\mathrm{d}x}_{\geq 0} \\ &\geq J(u), \end{aligned}$$

quindi $u$ è punto di minimo per $J$. (Poiché l'ultimo termine è strettamente positivo per ogni $w \neq 0$, abbiamo che la soluzione di (56) è l'*unico* punto di minimo per $J$.) □

## 6.2 Problemi variazionali astratti

Il problema debole (56) è scritto come uguaglianza tra una forma bilineare e un funzionale lineare al variare di una funzione test. Possiamo quindi formularlo più in astratto e in generale.

**Definizione 6.6.** Sia $V$ uno spazio di Hilbert con norma $\|\cdot\|_V$, $\mathcal{A} : V \times V \to \mathbb{R}$ una forma bilineare su $V$ e $\mathcal{F} : V \to \mathbb{R}$ un funzionale lineare su $V$. Il **problema variazionale** relativo a $V, \mathcal{A}, \mathcal{F}$ è

$$\text{trovare } u \in V \text{ tale che} \quad \mathcal{A}(u,w) = \mathcal{F}(w) \quad \text{per ogni } w \in V. \tag{58}$$

Un risultato fondamentale dell'analisi funzionale è il Teorema di Lax–Milgram.

**Teorema 6.7** (Teorema di Lax–Milgram)**.** Sia dato il problema (58) e siano soddisfatte le seguenti ipotesi:

- $\mathcal{A}$ è **continua**, cioè esiste $C_{\mathcal{A}} > 0$ tale che $|\mathcal{A}(u,w)| \leq C_{\mathcal{A}} \|u\|_V \|w\|_V \quad \forall u, w \in V$;
- $\mathcal{A}$ è **coerciva** (o $V$-**ellittica**), cioè esiste $\gamma_{\mathcal{A}} > 0$ tale che $\mathcal{A}(w,w) \geq \gamma_{\mathcal{A}} \|w\|_V^2 \quad \forall w \in V$;
- $\mathcal{F}$ è **continuo** (o **limitato**), cioè esiste $C_{\mathcal{F}} > 0$ tale che $|\mathcal{F}(w)| \leq C_{\mathcal{F}} \|w\|_V \quad \forall w \in V$.

Allora esiste un'unica soluzione $u \in V$ del problema variazionale (58).

**Corollario 6.8.** Sotto le ipotesi del Teorema di Lax–Milgram 6.7, la soluzione $u$ del problema variazionale (58) dipende con continuità da $\mathcal{F}$, cioè

$$\|u\|_V \leq \frac{C_{\mathcal{F}}}{\gamma_{\mathcal{A}}}.$$

*Dimostrazione.* La coercività di $\mathcal{A}$, il problema (58) con la scelta $w = u$, e la continuità di $\mathcal{F}$ danno

$$\gamma_{\mathcal{A}} \|u\|_V^2 \leq \mathcal{A}(u,u) = \mathcal{F}(u) \leq C_{\mathcal{F}} \|u\|_V \qquad \Rightarrow \qquad \gamma_{\mathcal{A}} \|u\|_V \leq C_{\mathcal{F}}.$$

□

**Esercizio ✎ 6.9** (Principio di Ritz in astratto)**.** Assumiamo oltre alle ipotesi del Teorema di Lax–Milgram che $\mathcal{A}$ sia **simmetrica**, cioè $\mathcal{A}(w,\tilde{w}) = \mathcal{A}(\tilde{w},w)$ per ogni $w, \tilde{w} \in V$. Seguendo la dimostrazione della Proposizione 6.5, dimostrare che il problema variazionale (58) è equivalente al problema di minimo

$$\text{trovare } u \in V \text{ tale che } J(u) = \min_{w \in V} J(w), \qquad \text{dove } J(w) := \frac{1}{2}\mathcal{A}(w,w) - \mathcal{F}(w).$$

(Notiamo che, nel caso simmetrico, $\mathcal{A}(\cdot,\cdot)$ è un prodotto scalare.)

**Nota 6.10** (Problemi variazionali simmetrici e annullamento della derivata del funzionale $J$)**.** L'esercizio 6.9 mostra che la soluzione del problema variazionale astratto (58) simmetrico è soluzione di un problema di minimo per un funzionale quadratico $J$. Possiamo pensare al problema variazionale come alla condizione

di annullamento della derivata prima di $J$. Per capire meglio questa affermazione consideriamo un esempio finito-dimensionale.

Sia $J : \mathbb{R}^d \to \mathbb{R}$ una funzione convessa e liscia. Allora $\vec{\mathbf{u}} \in \mathbb{R}^d$ è punto di minimo per $J$ se e solo se il gradiente si annulla, o equivalentemente se ogni derivata direzionale si annulla:

$$J(\vec{\mathbf{u}}) \le J(\vec{\mathbf{v}}) \quad \forall \vec{\mathbf{v}} \in \mathbb{R}^d \qquad \iff \qquad \nabla J(\vec{\mathbf{u}}) = \vec{\mathbf{0}} \qquad \iff \qquad \vec{\mathbf{w}} \cdot \nabla J(\vec{\mathbf{u}}) = 0 \quad \forall \vec{\mathbf{w}} \in \mathbb{R}^d.$$

Se inoltre $J$ è una funzione quadratica, cioè un polinomio di secondo grado in $d$ variabili reali, allora si può scrivere

$$J(\vec{\mathbf{v}}) = \frac{1}{2} \sum_{j=1}^{d} \sum_{k=1}^{d} A_{j,k} v_j v_k - \sum_{j=1}^{d} f_j v_j + c = \frac{1}{2} \vec{\mathbf{v}}^\top \underline{\underline{\mathbf{A}}} \vec{\mathbf{v}} - \vec{\mathbf{f}}^\top \vec{\mathbf{v}} + c$$

per una matrice $\underline{\underline{\mathbf{A}}} \in \mathbb{R}^{d \times d}$, un vettore $\vec{\mathbf{f}} \in \mathbb{R}^d$ e uno scalare $c \in \mathbb{R}$. Si calcola facilmente che il suo gradiente è $(\nabla J)(\vec{\mathbf{v}}) = \frac{1}{2}(\underline{\underline{\mathbf{A}}} + \underline{\underline{\mathbf{A}}}^\top)\vec{\mathbf{v}} - \vec{\mathbf{f}}$ (verificare questo conto per esercizio). Denotiamo con $\underline{\underline{\mathbf{A}}}^S := \frac{1}{2}(\underline{\underline{\mathbf{A}}} + \underline{\underline{\mathbf{A}}}^\top)$ la parte simmetrica della matrice $\underline{\underline{\mathbf{A}}}$. Notiamo anche che $\underline{\underline{\mathbf{A}}}^S$ è la matrice Hessiana di $J$ (e ricordiamo che l'Hessiana è sempre simmetrica). Allora un punto di minimo $\vec{\mathbf{u}}$ per $J$ (quadratica e convessa) può essere caratterizzato come

$$J(\vec{\mathbf{u}}) \le J(\vec{\mathbf{v}}) \quad \forall \vec{\mathbf{v}} \in \mathbb{R}^d \qquad \iff \qquad \underline{\underline{\mathbf{A}}}^S \vec{\mathbf{u}} = \vec{\mathbf{f}} \qquad \iff \qquad \vec{\mathbf{w}}^\top \underline{\underline{\mathbf{A}}}^S \vec{\mathbf{u}} = \vec{\mathbf{w}}^\top \vec{\mathbf{f}} \quad \forall \vec{\mathbf{w}} \in \mathbb{R}^d.$$

Il punto di minimo $\vec{\mathbf{u}}$ è soluzione di un sistema lineare algebrico e di un problema variazionale simmetrico su $V = \mathbb{R}^d$ con $\mathcal{A}(\vec{\mathbf{u}}, \vec{\mathbf{w}}) := \vec{\mathbf{w}}^\top \underline{\underline{\mathbf{A}}}^S \vec{\mathbf{u}}$ e $\mathcal{F}(\vec{\mathbf{w}}) := \vec{\mathbf{w}}^\top \vec{\mathbf{f}}$.

Per analogia interpretiamo il problema variazionale astratto (58) (con $\mathcal{A}(\cdot,\cdot)$ simmetrica) come l'annullamento delle derivate direzionali del funzionale quadratico $J$ definito nell'Esercizio 6.9. Questo può essere reso preciso con un'opportuna definizione delle derivate dei funzionali definiti su spazi di Hilbert.

**Esercizio ✎ 6.11** (Problemi variazionali in dimensione finita)**.** I problemi variazionali più semplici sono quelli posti su uno spazio $V$ finito-dimensionale. Fissiamo $V = \mathbb{R}^n$, scegliamo $\|\cdot\|_V$ come la norma Euclidea $\|\cdot\|_2$, e denotiamo $\vec{\mathbf{e}}_1, \dots, \vec{\mathbf{e}}_n$ gli elementi di una base ortonormale. Allora, espandendo in coefficienti i vettori $\vec{\mathbf{u}}, \vec{\mathbf{w}} \in \mathbb{R}^n$ come $\vec{\mathbf{u}} = \sum_{j=1}^n u_j \vec{\mathbf{e}}_j$ e $\vec{\mathbf{w}} = \sum_{j=1}^n w_j \vec{\mathbf{e}}_j$, usando la bilinearità di $\mathcal{A}$ e la linearità di $\mathcal{F}$ otteniamo

$$\mathcal{A}(\vec{\mathbf{u}}, \vec{\mathbf{w}}) = \sum_{j,k=1}^{n} u_k w_j \mathcal{A}(\vec{\mathbf{e}}_k, \vec{\mathbf{e}}_j), \qquad \mathcal{F}(\vec{\mathbf{w}}) = \sum_{j=1}^{n} w_j \mathcal{F}(\vec{\mathbf{e}}_j).$$

Quindi la forma bilineare e il funzionale lineare sono univocamente definiti dalla loro azione sugli elementi della base. In altre parole basta conoscere la matrice $\underline{\underline{\mathbf{A}}} \in \mathbb{R}^{n \times n}$ e il vettore $\vec{\mathbf{F}} \in \mathbb{R}^n$ definiti come $\underline{\underline{\mathbf{A}}}_{j,k} := \mathcal{A}(\vec{\mathbf{e}}_k, \vec{\mathbf{e}}_j)$ e $F_j := \mathcal{F}(\vec{\mathbf{e}}_j)$ per identificare il problema variazionale.

1. Mostrare che le ipotesi di continuità del Teorema di Lax–Milgram valgono con $C_{\mathcal{A}} = \left\|\underline{\underline{\mathbf{A}}}\right\|_2$ e $C_{\mathcal{F}} = \left\|\vec{\mathbf{F}}\right\|_2$.
2. Dare un esempio di problema variazionale con $V = \mathbb{R}^n$ che soddisfa l'ipotesi di coercività e un esempio di uno che non la soddisfa.
3. Quali ipotesi sulla matrice $\underline{\underline{\mathbf{A}}}$ garantiscono la coercività di $\mathcal{A}$?

## 6.3 Formulazione variazionale astratta di problemi al bordo

Per verificare che la versione debole (56) del problema al bordo soddisfa le ipotesi del Teorema di Lax–Milgram, premettiamo un risultato importante.

**Proposizione 6.12** (Disuguaglianza di Poincaré)**.** Per ogni $w \in H^1_0(a,b)$ vale la disuguaglianza di Poincaré (o di Friedrichs):

$$\|w\|_{L^2(a,b)} \le C_P \, |w|_{H^1(a,b)}, \qquad C_P := \frac{b-a}{\sqrt{2}}. \tag{59}$$

In particolare, la seminorma $|\cdot|_{H^1(a,b)}$ è una norma in $H^1_0(a,b)$ ed è equivalente alla norma $\|\cdot\|_{H^1(a,b)}$.

*Dimostrazione.* Ogni $w \in H^1(a,b)$ è assolutamente continua, quindi vale il teorema fondamentale del calcolo (di Lebesgue): $w(x) = w(a) + \int_a^x w'(t)\,\mathrm{d}t$. La condizione $w \in H^1_0(a,b)$ garantisce inoltre che $w(a) = 0$. Con alcune manipolazioni abbiamo la disuguaglianza (59):

$$\|w\|^2_{L^2(a,b)} = \int_a^b w^2(x)\,\mathrm{d}x = \int_a^b \Big( \int_a^x w'(t)\,\mathrm{d}t \Big)^2 \mathrm{d}x \le \int_a^b (x-a) \int_a^x \big(w'(t)\big)^2 \,\mathrm{d}t\,\mathrm{d}x$$

$$\leq \int_a^b (x-a)\,\mathrm{d}x \int_a^b \big(w'(t)\big)^2\,\mathrm{d}t = \frac{(b-a)^2}{2}|w|^2_{H^1(a,b)}.$$

Nella prima disuguaglianza abbiamo usato quella di Cauchy–Schwarz (55) su $(a,x)$ con $\varphi = w'$ e $\psi = 1$ o, equivalentemente, la disuguaglianza di Jensen $(\frac{1}{|\Omega|}\int_\Omega f\,\mathrm{d}x)^2 \leq \frac{1}{|\Omega|}\int_\Omega f^2\,\mathrm{d}x$.

Dalla definizione delle norme segue che $|\cdot|_{H^1(a,b)}$ è una norma equivalente a $\|\cdot\|_{H^1(a,b)}$:

$$|w|^2_{H^1(a,b)} \leq \|w\|^2_{H^1(a,b)} = \|w\|^2_{L^2(a,b)} + |w|^2_{H^1(a,b)} \leq (C_P^2+1)|w|^2_{H^1(a,b)}.$$

□

Abbiamo dimostrato che $|\cdot|_{H^1(a,b)}$ è una norma su $H^1_0(a,b)$; questo non è vero sullo spazio più grande $H^1(a,b)$: infatti $|c|_{H^1(a,b)} = 0$ per ogni funzione costante $c$. L'unica funzione costante in $H^1_0(a,b)$ è quella costantemente nulla.

Il problema debole (56) è un caso molto speciale di problema variazionale per

$$V = H^1_0(a,b), \quad \|\cdot\|_V = \|\cdot\|_{H^1(a,b)}, \quad \mathcal{A}(u,w) = \int_a^b (u'w' + quw)\,\mathrm{d}x, \quad \mathcal{F}(w) = \int_a^b fw\,\mathrm{d}x. \tag{60}$$

**Esercizio ✎ 6.13.** Usare (59) per dimostrare che la forma debole (56) del problema al contorno con $f \in L^2(a,b)$ e $q \in L^\infty(a,b)$, scritta nella forma astratta (58) con le scelte (60) soddisfa le tre ipotesi del Teorema di Lax–Milgram con

$$C_{\mathcal{A}} = \max\{1, \|q\|_{L^\infty(a,b)}\}, \qquad \gamma_{\mathcal{A}} = \frac{1}{1+C_P^2} = \frac{1}{1+\frac{(b-a)^2}{2}}, \qquad C_{\mathcal{F}} = \|f\|_{L^2(a,b)}, \tag{61}$$

dove $C_P$ è la costante di Poincaré (59).

Suggerimento: per dimostrare la continuità di $\mathcal{A}$, usare sia la disuguaglianza di Cauchy–Schwarz (55) in $L^2(a,b)$ che quella in $\mathbb{R}^2$ ($|x_1y_1 + x_2y_2| \leq \sqrt{x_1^2+x_2^2}\sqrt{y_1^2+y_2^2}$).

Questo esercizio, insieme al Teorema di Lax–Milgram e al Corollario 6.8, garantisce che **esiste un'unica soluzione $u$ del problema debole** (56) e che questa soddisfa la **stima di stabilità** (confrontare con (18))

$$\|u\|_{H^1(a,b)} \leq (1+C_P^2)\,\|f\|_{L^2(a,b)}. \tag{62}$$

**Esercizio ✎ 6.14.** Ripetere l'Esercizio 6.13 per la scelta $\|\cdot\|_V = |\cdot|_{H^1(a,b)}$ e dedurre la stima di stabilità

$$|u|_{H^1(a,b)} \leq C_P\,\|f\|_{L^2(a,b)}.$$

Se $f, q \in C^0([a,b])$, allora sappiamo dal Teorema 2.13 che la soluzione classica del problema (53) appartiene a $C^2(a,b)$. Dalla derivazione in (54) (cioè dall'integrazione per parti) sappiamo che $u$ è anche soluzione debole, e dal Teorema di Lax–Milgram sappiamo che è l'unica soluzione debole del problema (56). Quindi abbiamo un risultato di regolarità: se i dati $f$ e $q$ sono continui, allora la soluzione debole $u$ è di classe $C^2$ e coincide con la soluzione classica.

Riassumiamo nella prossima proposizione le proprietà dimostrate per la soluzione del problema variazionale.

**Proposizione 6.15.** Siano dati $f \in L^2(a,b)$ e $q \in L^\infty(a,b)$ con $q \geq 0$.
Allora il problema al bordo in forma debole (56) ammette un'unica soluzione $u \in H^1_0(a,b)$.
Inoltre vale la stima di stabilità (62) per $u$.
Se $f, q \in C^0([a,b])$ allora $u \in C^2(a,b)$ e $u$ è soluzione classica del problema al bordo (53).

**Esercizio ✎ 6.16.** Sia $u \in H^1_0(a,b)$ la soluzione del problema (56). Sia $u_\epsilon \in H^1_0(a,b)$ la soluzione dello stesso problema con dato perturbato $f_\epsilon = f + \epsilon$. Maggiorare la differenza $\|u - u_\epsilon\|_{H^1(a,b)}$ in dipendenza da $\epsilon$.

### 6.3.1 Esempi di problemi deboli che non ammettono soluzioni classiche

Abbiamo motivato la formulazione debole dicendo che è più generale di quella classica. Esistono problemi che ammettono soluzioni deboli ma non classiche? Sì, vediamo alcuni esempi.

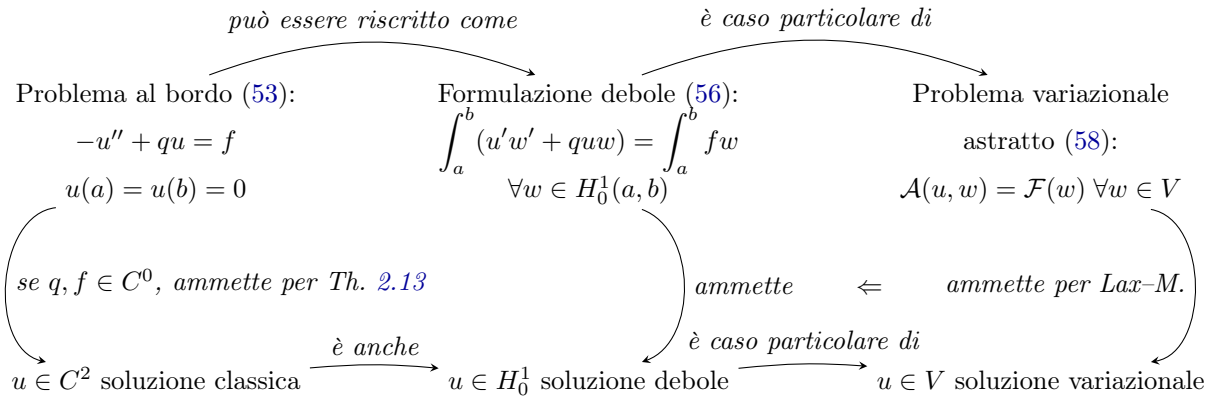



*Figura 33: Schema delle relazioni tra i problemi (53), (56), (58) e le loro soluzioni.*

**Primo esempio** Siano $(a,b) = (-1,1)$, $q(x) = 0$, $f(x) = \chi_{(-\frac{1}{2},\frac{1}{2})}(x)$, la funzione caratteristica dell'intervallo $(-\frac{1}{2},\frac{1}{2})$. È facile verificare a mano che

$$u(x) = \begin{cases} \frac{1+x}{2} & x < -\frac{1}{2}, \\ \frac{3}{8} - \frac{1}{2}x^2 & -\frac{1}{2} \le x \le \frac{1}{2}, \\ \frac{1-x}{2} & x > \frac{1}{2} \end{cases}$$

soddisfa $u(a) = u(b) = 0$ ed è soluzione puntuale di $-u'' = f$ in tutti i punti di $(a,b)$ tranne $\pm\frac{1}{2}$. Poiché $f \notin C^0(-1,1)$ e $u \in C^1(-1,1) \setminus C^2(-1,1)$, $u$ non è soluzione classica. Tuttavia, per ogni $w \in H_0^1(a,b)$, integrando per parti otteniamo

$$\begin{aligned} \int_{-1}^{1} (u'w' + quw)\,\mathrm{d}x &= \int_{-1}^{-\frac{1}{2}} \frac{1}{2}w'\,\mathrm{d}x + \int_{-\frac{1}{2}}^{\frac{1}{2}} -xw'\,\mathrm{d}x + \int_{\frac{1}{2}}^{1} -\frac{1}{2}w'\,\mathrm{d}x \\ &= \frac{w(-1/2) - w(-1)}{2} + \Big( \int_{-\frac{1}{2}}^{\frac{1}{2}} w\,\mathrm{d}x - \frac{1}{2}w(1/2) - \frac{1}{2}w(-1/2) \Big) - \frac{w(1) - w(1/2)}{2} \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} w\,\mathrm{d}x = \int_{-1}^{1} fw\,\mathrm{d}x. \end{aligned}$$

cioè $u$ è soluzione debole della formulazione variazionale del problema al bordo. (Notiamo che $u \in H^2(-1,1)$, nonostante $u \notin C^2(-1,1)$.)

**Secondo esempio** Generalizziamo leggermente l'esempio precedente e consideriamo la sequenza di problemi al bordo

$$-u_n''(x) = f_n(x) := \frac{n}{2}\chi_{(-\frac{1}{n},\frac{1}{n})}(x), \qquad u_n(-1) = u_n(1) = 0, \qquad n \in \mathbb{N}, \tag{63}$$

dove $\chi_{(-\frac{1}{n},\frac{1}{n})}$ è la funzione caratteristica dell'intervallo $(-\frac{1}{n},\frac{1}{n})$. Come prima, si verifica che le soluzioni sono

$$u_n(x) = \begin{cases} \frac{1+x}{2} & x < -\frac{1}{n}, \\ \frac{1}{2} - \frac{1}{4n} - \frac{n}{4}x^2 & -\frac{1}{n} \le x \le \frac{1}{n}, \\ \frac{1-x}{2} & x > \frac{1}{n}. \end{cases}$$

Alcune soluzioni sono mostrate in Figura 34. Possiamo interpretare $u_n$ come la temperatura di una barra che viene scaldata in un breve tratto centrale di lunghezza $2/n$ e i cui estremi sono mantenuti a temperatura costante. Poiché la quantità

$$\int_{-1}^{1} f_n(x)\,\mathrm{d}x = \int_{-1}^{1} \frac{n}{2}\chi_{(-\frac{1}{n},\frac{1}{n})}(x)\,\mathrm{d}x = 1$$

è indipendente da $n$, il calore introdotto è lo stesso per tutti i problemi. Chiaramente la sequenza di soluzioni $u_n$ tende uniformemente a

$$u_\infty(x) := \begin{cases} \frac{1+x}{2} & x < 0, \\ \frac{1-x}{2} & x > 0. \end{cases}$$

Questa $u_\infty$ corrisponde alla soluzione del problema in cui la barra è scaldata in un unico punto. Possiamo scrivere un "problema al bordo limite"? Vediamo che $f_n(x) \to 0$ per $x \in (-1,0)\cup(0,1)$ e $f_n(0) \to \infty$, quindi non possiamo scrivere un'equazione differenziale limite. Inoltre $u_\infty$ non è differenziabile in 0 neppure una volta: è impossibile scrivere un'equazione differenziale in $(-1,1)$ con soluzione $u_\infty$.

Possiamo però scrivere un problema variazionale limite. Per ogni $n \in \mathbb{N}$, (63) si scrive

$$\mathcal{A}(u_n, w) = \int_{-1}^{1} u_n' w' \,\mathrm{d}x = \frac{n}{2} \int_{-\frac{1}{n}}^{\frac{1}{n}} w \,\mathrm{d}x = \fint_{-\frac{1}{n}}^{\frac{1}{n}} w \,\mathrm{d}x =: \mathcal{F}_n(w) \qquad \forall w \in H_0^1(-1,1).$$

La forma bilineare $\mathcal{A}$ è indipendente da $n$, mentre per il teorema della media integrale e la continuità di $w \in H_0^1(a,b)$, $\mathcal{F}_n(w)$ tende a $\mathcal{F}_\infty(w) := w(0)$ per $n \to \infty$ e per ogni $w \in H_0^1(a,b)$. Scriviamo il problema variazionale con il funzionale "esotico" $\mathcal{F}_\infty(w) = w(0)$:

$$\text{cerchiamo } u \in H_0^1(a,b) \text{ tale che} \qquad \int_{-1}^{1} u' w' \,\mathrm{d}x = w(0) \qquad \forall w \in H_0^1(-1,1). \tag{64}$$

Si può dimostrare che il funzionale $\mathcal{F}_\infty$ è continuo (Esercizio 6.17). Questo problema variazionale ammette un'unica soluzione, grazie al Teorema di Lax–Milgram. Mostriamo che questa soluzione è la $u_\infty$ definita sopra:

$$\mathcal{A}(u_\infty, w) = \int_{-1}^{1} u_\infty' w' \,\mathrm{d}x = \frac{1}{2}\int_{-1}^{0} w' \,\mathrm{d}x - \frac{1}{2}\int_{0}^{1} w' \,\mathrm{d}x = w(0) = \mathcal{F}_\infty(w) \qquad \forall w \in H_0^1(-1,1).$$

In sintesi, (64) è un problema ben posto scritto in forma variazionale che ammette l'unica soluzione variazionale $u_\infty$, ma che non corrisponde a nessuna equazione differenziale. Non è neppure un problema debole come in Definizione 6.3 perché non possiamo scrivere $\mathcal{F}_\infty(w)$ nella forma $\int_{-1}^{1} fw \,\mathrm{d}x$ per una $f \in L^2(-1,1)$. Tuttavia è problema "limite" di una sequenza di problemi al contorno.

A volte questo problema è scritto come un problema al bordo per l'equazione $-u'' = \delta$, dove $\delta$ non è una funzione ma un oggetto matematico più generale: è uguale a zero nei due intervalli $(-1,0)$ e $(0,1)$ ma allo stesso tempo è limite di funzioni con integrale 1, quindi ha "massa" 1 concentrata nell'origine. La "funzione generalizzata" $\delta$ è detta "delta di Dirac" ed è descritta rigorosamente attraverso la "teoria delle distribuzioni".

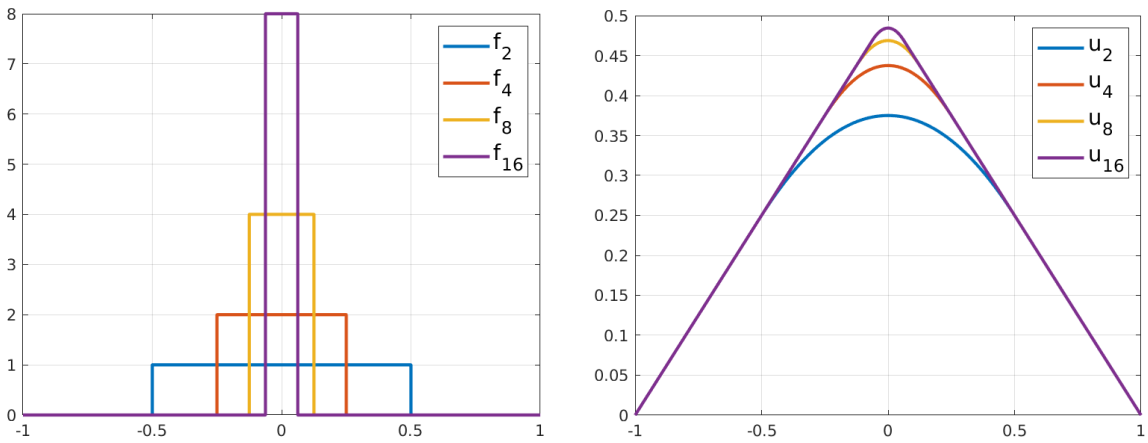



*Figura 34: I termini di sorgente $f_n$ e le soluzioni $u_n$ per $n = 2, 4, 8, 16$ per i problemi al bordo dell'esempio di Sezione 6.3.1.*

**Esercizio ✎ 6.17.** Dimostrare che il funzionale $\mathcal{F}_\infty(w) := w(0)$ è continuo in $H_0^1(-1,1)$ (nel senso usato nel Teorema di Lax–Milgram).

Si può usare il teorema fondamentale del calcolo e procedere in modo simile alla dimostrazione della disuguaglianza di Poincaré.

## 6.4 Il metodo di Galerkin

Il metodo di Galerkin è una tecnica molto generale per l'approssimazione di problemi variazionali. Assumiamo siano dati $\mathcal{A}, \mathcal{F}$ e $V$ come in Definizione 6.6 e consideriamo il problema variazionale astratto (58).

Consideriamo un **sottospazio vettoriale finito-dimensionale** $V_h \subset V$ che chiameremo "spazio discreto". Il metodo di Galerkin consiste nel

$$\boxed{\text{cercare} \quad u_h \in V_h \quad \text{tale che} \quad \mathcal{A}(u_h, w_h) = \mathcal{F}(w_h) \qquad \forall w_h \in V_h.} \tag{65}$$

Il metodo di Galerkin non è che la "restrizione" del problema variazionale a un sottospazio di dimensione finita. Questa restrizione agisce in due modi: (1) cerchiamo una soluzione discreta solo in $V_h$, (2) richiediamo che $\mathcal{A}(u_h, w_h) = \mathcal{F}(w_h)$ solo per funzioni test $w_h$ in $V_h$.

Questo metodo sembra molto astratto, come può essere scritto più in concreto? Scegliamo uno spazio $V_h \subset V$ di dimensione $n$ con base $\{\varphi_1, \dots, \varphi_n\}$. Gli elementi della base sono a volte chiamati "funzioni di forma" (*shape functions*). Ogni elemento $w_h \in V_h$ può essere espanso come $w_h = \sum_{j=1}^n W_j \varphi_j$, per $\vec{\mathbf{W}} \in \mathbb{R}^n$. L'identificazione di una funzione discreta con i suoi coefficienti $w_h \mapsto \vec{\mathbf{W}}$ è un isomorfismo $V_h \to \mathbb{R}^n$. Le componenti $W_j$ del vettore $\vec{\mathbf{W}}$ sono dette "gradi di libertà" (*degrees of freedom*) corrispondenti a $w_h$. L'equazione (65) è vera per ogni $w_h \in V_h$ se è verificata per le $n$ funzioni di base $\varphi_1, \dots, \varphi_n$:

$$\begin{aligned}
\mathcal{A}(u_h, w_h) = \mathcal{F}(w_h) \quad \forall w_h \in V_h \quad &\iff \quad \mathcal{A}\Big(u_h, \sum_{j=1}^n W_j \varphi_j\Big) = \mathcal{F}\Big(\sum_{j=1}^n W_j \varphi_j\Big) \quad \forall \vec{\mathbf{W}} \in \mathbb{R}^n \\
&\iff \quad \sum_{j=1}^n W_j \mathcal{A}(u_h, \varphi_j) = \sum_{j=1}^n W_j \mathcal{F}(\varphi_j) \quad \forall \vec{\mathbf{W}} \in \mathbb{R}^n \\
&\iff \quad \mathcal{A}(u_h, \varphi_j) = \mathcal{F}(\varphi_j) \quad j = 1, \dots, n.
\end{aligned}$$

In altre parole l'equazione (65) che chiediamo per ogni elemento di $V_h$ si riduce a $n$ equazioni lineari. Espandendo $u_h = \sum_{k=1}^n U_k \varphi_k$, queste $n$ equazioni diventano un sistema lineare quadrato per il vettore $\vec{\mathbf{U}}$ dei gradi di libertà di $u_h$:

$$\begin{aligned}
\mathcal{A}(u_h, w_h) = \mathcal{F}(w_h) \quad \forall w_h \in V_h \quad &\iff \quad \sum_{k=1}^n \mathcal{A}(\varphi_k, \varphi_j) U_k = \mathcal{F}(\varphi_j) \quad j = 1, \dots, n \\
&\iff \quad \boxed{\underline{\underline{\mathbf{A}}} \vec{\mathbf{U}} = \vec{\mathbf{F}}, \qquad A_{j,k} := \mathcal{A}(\varphi_k, \varphi_j), \quad F_j := \mathcal{F}(\varphi_j).}
\end{aligned} \tag{66}$$

Per implementare (66) basta:

- scegliere una base $\{\varphi_k\}$ di $V_h$,
- calcolare la matrice $\underline{\underline{\mathbf{A}}}$ a partire dalla forma bilineare $\mathcal{A}$ e dalla base,
- calcolare il vettore $\vec{\mathbf{F}}$ dal funzionale lineare $\mathcal{F}$ e dalla base, e
- risolvere il sistema lineare $\underline{\underline{\mathbf{A}}} \vec{\mathbf{U}} = \vec{\mathbf{F}}$.

La soluzione ottenuta è un vettore $\vec{\mathbf{U}} \in \mathbb{R}^n$, che corrisponde a un elemento $u_h \in V_h \subset V$, cioè (normalmente) a una funzione.

Rimandiamo a §6.5–6.6 la scelta di uno spazio $V_h$ concreto e nel resto di §6.4 osserviamo alcune proprietà generali del metodo di Galerkin in astratto.

### 6.4.1 Proprietà del metodo di Galerkin

Consideriamo il metodo di Galerkin applicato a un problema che soddisfa le ipotesi del Teorema di Lax–Milgram 6.7.

Ricordiamo che la forma bilineare $\mathcal{A}$ e la matrice $\underline{\underline{\mathbf{A}}}$ sono legate dalla seguente relazione: per ogni $u_h, w_h \in V_h$, se $u_h = \sum_j U_j \varphi_j$ e $w_h = \sum_j W_j \varphi_j$, vale $\mathcal{A}(u_h, w_h) = \vec{\mathbf{W}}^\top \underline{\underline{\mathbf{A}}} \vec{\mathbf{U}}$.

Il metodo è **ben posto**: se $\vec{\mathbf{W}} \in \mathbb{R}^n$ è diverso da $\vec{\mathbf{0}}$, allora per la coercività di $\mathcal{A}$ abbiamo

$$\vec{\mathbf{W}}^\top \underline{\underline{\mathbf{A}}} \vec{\mathbf{W}} = \mathcal{A}(w_h, w_h) \geq \gamma_{\mathcal{A}} \|w_h\|_V^2 > 0,$$

quindi $\underline{\underline{\mathbf{A}}}$ è definita positiva e invertibile. Equivalentemente, il Teorema di Lax–Milgram applicato direttamente allo spazio finito-dimensionale $V_h$ fornisce la buona posizione del metodo.

La soluzione discreta $u_h$ dipende con continuità dai dati $\mathcal{F}$ e gode della stessa stabilità della soluzione $u$ del problema originario (ricordare il Corollario 6.8)

$$\|u_h\|_V^2 \leq \frac{1}{\gamma_{\mathcal{A}}} \mathcal{A}(u_h, u_h) = \frac{1}{\gamma_{\mathcal{A}}} \mathcal{F}(u_h) \leq \frac{C_{\mathcal{F}}}{\gamma_{\mathcal{A}}} \|u_h\|_V \qquad \Rightarrow \qquad \|u_h\|_V \leq \frac{C_{\mathcal{F}}}{\gamma_{\mathcal{A}}}.$$

Dalla formulazione del problema variazionale (58) e da quella del metodo di Galerkin (65), ricordando che $V_h \subset V$, deriviamo la seguente utile proprietà, detta **ortogonalità di Galerkin**:

$$\mathcal{A}(u - u_h, w_h) = \mathcal{A}(u, w_h) - \mathcal{A}(u_h, w_h) = \mathcal{F}(w_h) - \mathcal{F}(w_h) = 0 \qquad \forall w_h \in V_h. \tag{67}$$

L'ortogonalità di Galerkin, la coercività e la continuità di $\mathcal{A}$ ci permettono di maggiorare l'errore commesso dal metodo di Galerkin:

$$\begin{aligned} \|u - u_h\|_V^2 &\le \frac{1}{\gamma_{\mathcal{A}}} \mathcal{A}(u - u_h, u - u_h) \\ &= \frac{1}{\gamma_{\mathcal{A}}} \mathcal{A}(u - u_h, u - w_h) \qquad \text{per qualsiasi } w_h \in V_h, \text{grazie a (67)}, \\ &\le \frac{C_{\mathcal{A}}}{\gamma_{\mathcal{A}}} \|u - u_h\|_V \|u - w_h\|_V . \end{aligned}$$

Dividendo per $\|u - u_h\|_V$ e scegliendo la $w_h$ discreta che meglio approssima $u$ otteniamo il seguente importante risultato.

**Lemma 6.18** (Lemma di Céa)**.** Sotto le ipotesi del Teorema di Lax–Milgram, siano $u$ la soluzione di (58) e $u_h$ la soluzione di (65). Allora vale

$$\|u - u_h\|_V \le \frac{C_{\mathcal{A}}}{\gamma_{\mathcal{A}}} \inf_{w_h \in V_h} \|u - w_h\|_V . \tag{68}$$

La disuguaglianza (68) è detta stima di **quasi-ottimalità** e ci dice un fatto molto rilevante: sotto le ipotesi fatte, l'errore commesso dal metodo di Galerkin dipende da due termini. Il primo termine, $C_{\mathcal{A}}/\gamma_{\mathcal{A}}$, è un termine di stabilità; dipende solo dal problema continuo e non dallo spazio discreto $V_h$. Il secondo termine, $\inf_{w_h \in V_h} \|u - w_h\|_V$, è un termine di approssimazione: misura quanto bene lo spazio discreto $V_h$ è in grado di approssimare la soluzione $u$. Ancora una volta abbiamo separato i contributi di stabilità e approssimazione nella stima dell'errore (ricordare ad esempio (27) per il metodo delle differenze finite).

**Esercizio ✎ 6.19.** Dimostrare che la soluzione discreta è controllata da quella continua: $\|u_h\|_V \le \frac{C_{\mathcal{A}}}{\gamma_{\mathcal{A}}} \|u\|_V$.

**Nota 6.20** (Il caso simmetrico e il principio di Ritz discreto)**.** Se la forma bilineare $\mathcal{A}$ è simmetrica, cioè se $\mathcal{A}(w, \tilde{w}) = \mathcal{A}(\tilde{w}, w)$ per ogni $w, \tilde{w} \in V$, allora $\mathcal{A}(\cdot, \cdot)$ costituisce un prodotto scalare su $V$. Seguendo ancora una volta la dimostrazione della Proposizione 6.5 si verifica che vale il principio di Ritz discreto: **la soluzione $u_h$ del metodo di Galerkin è l'elemento di $V_h$ che minimizza il funzionale** $J$ definito nell'Esercizio 6.9. L'uguaglianza $\mathcal{A}(u_h, w_h) = \mathcal{A}(u, w_h)$ per ogni $w_h \in V_h$ implica che $u_h$ è la proiezione ortogonale rispetto al prodotto scalare $\mathcal{A}(\cdot, \cdot)$ di $u$ su $V_h$. Da questi fatti si può ricavare una stima di quasi-ottimalità migliore di (68), con costante di quasi-ottimalità $\sqrt{C_{\mathcal{A}}/\gamma_A}$. Il metodo di Galerkin per problemi simmetrici è talvolta chiamato metodo di Ritz o di Rayleigh–Ritz.

### 6.4.2 Il metodo di Galerkin per problemi al bordo

Ricordiamo che il caso che ci interessa maggiormente è il metodo di Galerkin (65) applicato alla formulazione debole (56) per il problema di Dirichlet (53), con le scelte (60). In questo caso la matrice $\underline{\underline{\mathbf{A}}}$ e il vettore $\vec{\mathbf{F}}$ da "assemblare" sono

$$A_{j,k} = \mathcal{A}(\varphi_k, \varphi_j) = \int_a^b (\varphi_k' \varphi_j' + q \varphi_k \varphi_j)\, \mathrm{d}x, \qquad F_j = \mathcal{F}(\varphi_j) = \int_a^b f \varphi_j \, \mathrm{d}x.$$

Abbiamo già stimato i valori di $C_{\mathcal{A}}, \gamma_{\mathcal{A}}$ e $C_{\mathcal{F}}$ in (61). In particolare, dato qualsiasi spazio discreto $V_h \subset H_0^1(a, b)$ la corrispondente soluzione $u_h$ del metodo di Galerkin soddisfa

$$\|u - u_h\|_{H^1(a,b)} \le C_{qo} \inf_{w_h \in V_h} \|u - w_h\|_{H^1(a,b)} \tag{69}$$

$$\text{dove } C_{qo} = \frac{C_{\mathcal{A}}}{\gamma_{\mathcal{A}}} = (1 + C_P^2) \max\{1, \|q\|_{L^\infty(a,b)}\} = \left(1 + \frac{(b-a)^2}{2}\right) \max\{1, \|q\|_{L^\infty(a,b)}\}.$$

(In realtà, grazie a quanto detto nella Nota 6.20, questa stima si può migliorare prendendo la radice quadrata di $C_{qo}$.)

**Nota 6.21** (Il caso con condizioni al bordo non omogenee)**.** In questa sezione abbiamo considerato il problema (53) con condizioni al bordo $u(a) = u(b) = 0$ e quindi il problema variazionale in $H_0^1(a,b)$, i cui elementi valgono zero in $a$ e $b$. Come possiamo estendere il metodo di Galerkin al caso con condizioni non-omogenee $u(a) = \alpha$ e $u(b) = \beta$?

In questo caso prima costruiamo una funzione (detta sollevamento, o *lifting*) $\tilde{u}$ che soddisfa entrambe le condizioni al bordo. Poi notiamo che $u_0 := u - \tilde{u}$ soddisfa l'equazione differenziale $-u_0'' + qu_0 = f + \tilde{u}'' - q\tilde{u}$ e le condizioni $u_0(a) = u_0(b) = 0$, quindi possiamo approssimarla con una $u_{0,h} \in V_h \subset H_0^1(a,b)$ con il metodo di Galerkin come descritto in precedenza. A questo punto possiamo ricostruire $u_h = \tilde{u} + u_{0,h}$, che approssima $u$ e soddisfa le condizioni al bordo. (Ovviamente $\tilde{u}$ può essere scelta come una funzione lineare.)

**Nota 6.22** (Il caso di Neumann)**.** Finora abbiamo considerato il caso del problema di Dirichlet, cosa succede nel caso del problema di Neumann? Consideriamo il problema al bordo (19).

Applichiamo la solita forma bilineare $\mathcal{A}$ alla soluzione $u$ e a una qualsiasi $w \in H^1(a,b)$ (attenzione: qui useremo $H^1(a,b)$ e non più $H_0^1(a,b)$!), integrando per parti e usando il problema (19) otteniamo

$$\begin{aligned}\mathcal{A}(u,w) &= \int_a^b (u'w' + quw)\,\mathrm{d}x = \int_a^b (-u''w + quw)\,\mathrm{d}x + u'(b)w(b) - u'(a)w(a)\\ &= \int_a^b fw\,\mathrm{d}x + \beta w(b) - \alpha w(a) =: \mathcal{F}_N(w).\end{aligned}$$

Il problema variazionale $\mathcal{A}(u,w) = \mathcal{F}_N(w)$ per ogni $w \in H^1(a,b)$ è la formulazione debole del problema di Neumann (19).

Notiamo che rispetto al problema di Dirichlet, la forma bilineare è la stessa ma lo spazio $V$ e il funzionale lineare sono diversi. L'analisi in Sezione 6.3 si basa sulla disuguaglianza di Poincaré, che non vale in $V = H^1(a,b)$. Tuttavia se $q > 0$ è possibile dimostrare che questo problema variazionale soddisfa le ipotesi del Teorema di Lax–Milgram (provare a dimostrarlo per esercizio). Il metodo di Galerkin si scrive allo stesso modo (ora con $V_h \subset H^1(a,b)$), e il Lemma di Céa segue immediatamente.

Un'importante differenza tra problemi di Dirichlet e di Neumann è che nel primo caso le condizioni al bordo (omogenee) sono incorporate nella scelta dello spazio funzionale (cioè usiamo $H_0^1(a,b)$), mentre nel secondo caso non le imponiamo su tutti gli elementi dello spazio funzionale ma vengono imposte dal termine noto $\mathcal{F}_N$. Si dice che le condizioni di Dirichlet sono "essenziali" e quelle di Neumann sono "naturali".

**Nota 6.23** (Il caso del problema di diffusione, trasporto e reazione)**.** Se l'equazione differenziale che vogliamo approssimare è un'equazione di diffusione–trasporto–reazione $-\epsilon u'' + pu' + qu = f$, il problema variazionale e il metodo di Galerkin si scrivono in modo simile al caso considerato finora. Il termine di trasporto dà luogo a un termine $\int_a^b pu'w\,\mathrm{d}x$ nella forma bilineare $\mathcal{A}$. Questo rende la forma bilineare **non-simmetrica**: in generale $\mathcal{A}(u,w) \neq \mathcal{A}(w,u)$, quindi la soluzione debole del problema non è necessariamente un punto di minimo di un funzionale $J$. La costante di coercività $\gamma_{\mathcal{A}}$ nel Teorema di Lax–Milgram è proporzionale a $\epsilon$, quindi se $\epsilon \ll 1$ la stima sull'errore fornita dal Lemma di Céa è debole: l'errore $u - u_h$ commesso dal metodo può essere molto grande nonostante lo spazio discreto contiene una buona approssimazione di $u$. Come nel caso delle differenze finite, i problemi di trasporto dominante richiedono un trattamento particolare per curarne la mancanza di stabilità, ad esempio l'introduzione di una "viscosità artificiale".

Al contrario, l'equazione differenziale $-(Ku')' + qu = f$ (già incontrata in §4.7.1) con coefficiente $K \in C^0([a,b])$ strettamente positivo dà sempre origine a un problema variazionale coercivo e simmetrico (provare a scriverlo e ad analizzarlo per esercizio).

**Nota 6.24** (Galerkin vs collocazione)**.** Il metodo di Galerkin ricorda quello di collocazione, quali sono i vantaggi di ciascun metodo?

Un primo vantaggio del metodo di Galerkin è che si basa sulla formulazione debole del problema al bordo. Questo permette di approssimare soluzioni poco regolari.

L'uso della formulazione debole permette di usare spazi discreti $V_h \subset H_0^1(a,b) \setminus C^2(a,b)$: le funzioni di base possono avere derivate discontinue, ad esempio. Costruire sottospazi discreti di $C^2(\Omega)$ "locali" (cioè con elementi di base con supporto piccolo e quindi matrici sparse) è molto difficile in dimensione maggiore di 1 ed è uno dei motivi per cui il popolare metodo degli elementi finiti, che vedremo tra poco, viene formulato come un caso particolare di quello di Galerkin e non di quello di collocazione.

Inoltre il metodo di Galerkin offre un'analisi teorica molto semplice e potente, al contrario quello di collocazione.

Infine, la matrice del metodo di Galerkin è simmetrica ogni volta che lo è la forma bilineare corrispondente.

Un vantaggio del metodo di collocazione è che, a parità di spazio discreto $V_h$, la sua implementazione può essere più semplice e non richiede formule di quadratura, che sono invece necessarie per calcolare gli integrali $A_{j,k}$ e $F_j$ in (66) nel metodo di Galerkin.

Per descrivere una versione del metodo di Galerkin per i problemi di Dirichlet dobbiamo solo definire e studiare degli spazi discreti $V_h$. Accenneremo solo brevemente al metodo spettrale, in cui $V_h$ è costituito da polinomi globali, e studieremo più in dettaglio il metodo degli elementi finiti, in cui $V_h$ è costituito da polinomi a tratti.

## 6.5 Il metodo spettrale

Un semplice sottospazio $n$-dimensionale di $H^1_0(a,b)$ è quello dei polinomi di grado non maggiore di $n+1$ che si annullano in $a$ e $b$ introdotto nella sezione 5.2. Il metodo di Galerkin con questa scelta dello spazio discreto $V_h$ è chiamato **metodo spettrale**.

Fissiamo $(a,b) = (-1,1)$ e assumiamo di usare come funzioni di base i polinomi di Legendre integrati $M_k$ definiti in (49). Se $q = 0$, la matrice $\underline{\underline{\mathbf{A}}}$ è diagonale e i suoi elementi si possono calcolare esattamente usando la relazione $M_k'(x) = P_k(x)$ e l'ortogonalità dei polinomi di Legendre $P_k$:

$$A_{j,k} = \int_{-1}^{1} M_k'(x) M_j'(x)\,\mathrm{d}x = \frac{2}{2k+1}\delta_{j,k}.$$

Se $q$ è costante, i valori degli elementi della matrice $\underline{\underline{\mathbf{A}}}$ si possono calcolare esattamente usando (49). Se $q$ non è costante è necessario usare una formula di quadratura. Il termine noto $\vec{\mathbf{F}}$ è sempre calcolato con una formula di quadratura.

Se la funzione $u$ è liscia, sappiamo dalla teoria dell'approssimazione che i polinomi l'approssimano con velocità superalgebrica in $n$. Se $u$ è analitica la convergenza è addirittura esponenziale. (Le stime di approssimazione polinomiale viste nel corso di analisi numerica riguardano la norma $L^\infty(a,b)$, è possibile dimostrare stime simili in norma $H^1(a,b)$.) Grazie alla quasi-ottimalità che segue dal Lemma di Céa, anche l'errore commesso dal metodo spettrale decade con la stessa velocità. È necessario però che tutte le quadrature utilizzate abbiano lo stesso ordine di accuratezza.

## 6.6 Il metodo degli elementi finiti

Il metodo degli elementi finiti (*finite element method*, spesso abbreviato in FEM o FE) non è altro che il metodo di Galerkin con uno spazio discreto $V_h$ composto da funzioni polinomiali a tratti.

Scelti dei nodi $a = x_0 < x_1 < x_2 < \cdots < x_n < x_{n+1} = b$, chiamiamo "elementi" gli intervalli $K_j := [x_{j-1}, x_j]$ tra due nodi consecutivi. La collezione $\mathcal{T}_h := \{K_j, j = 1,\dots,n+1\}$ degli elementi è detta "griglia computazionale" (o *mesh*). Chiamiamo $h_j := x_j - x_{j-1}$ la lunghezza dell'elemento $j$esimo e $h := \max_{j=1,\dots,n+1} h_j$ l'ampiezza della griglia ("*mesh width*" o "*mesh size*").

Lo spazio dei polinomi a tratti di grado $p \in \mathbb{N}$ definiti sulla mesh $\mathcal{T}_h$ è

$$S^p(\mathcal{T}_h) := \{w \in C^0(a,b),\ w|_{K_j} \in \mathbb{P}^p(K_j),\ j = 1,\dots,n+1\},$$

dove $\mathbb{P}^p(I)$ denota i polinomi di grado massimo $p$ nell'intervallo $I$.

Il metodo di Galerkin (65) applicato al problema di Dirichlet (53) richiede che gli elementi di $V_h$ siano in $H^1_0(a,b)$. Gli elementi di $S^p(\mathcal{T}_h)$ sono sicuramente in $H^1(a,b)$, poiché continui e con derivate polinomiali (di grado $p-1$) a tratti, quindi in $L^2(a,b)$. Bisogna però imporre che le funzioni dello spazio discreto valgano zero agli estremi. Possiamo quindi scegliere lo spazio discreto $V_h = S^p_0(\mathcal{T}_h)$ con

$$S^p_0(\mathcal{T}_h) := S^p(\mathcal{T}_h) \cap H^1_0(a,b) = \{w \in S^p(\mathcal{T}_h),\ w(a) = w(b) = 0\}.$$

**Esercizio ✎ 6.25.** Mostrare che $\dim(S^p(\mathcal{T}_h)) = (p+1)(n+1) - n$ e che $\dim(S^p_0(\mathcal{T}_h)) = pn + p - 1$.

Gli elementi di $S^p_0(\mathcal{T}_h)$ sono funzioni continue con derivate prime discontinue. Possiamo usarli per approssimare le soluzioni di equazioni differenziali di secondo grado poiché di queste consideriamo la forma debole.

Una scelta alternativa è quella di usare spazi discreti di splines, cioè $S^p(\mathcal{T}_h) \cap C^k(a,b) \cap H^1_0(a,b)$ per qualche $1 \le k \le p-1$; si veda [SF08, §1.7].

Nel resto della sezione consideriamo solo il caso dei problemi di Dirichlet. Il caso del problema di Neumann (con $q > 0$) si tratta in modo simile, come descritto nella Nota 6.22 si può scegliere $V_h = S^p(\mathcal{T}_h)$.

Un'introduzione al metodo degli elementi finiti per problemi in una dimensione, come quelli considerati qui, si può trovare in [SF08], [SM03, §14] e in [QSSG14, §11.3.5].

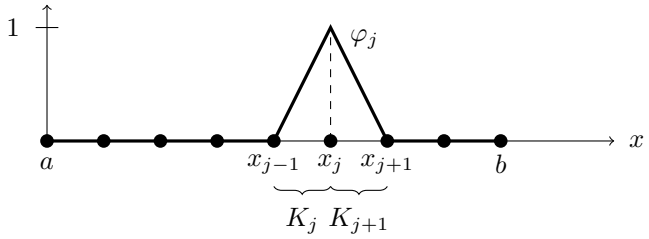
In dimensioni più alte, i problemi al bordo per le equazioni alle derivate parziali di tipo ellittico come (11) (in particolare l'equazione di Poisson $-\Delta u = f$) si possono scrivere come problemi variazionali

(usando il teorema della divergenza al posto dell'integrazione per parti) a cui si può applicare il metodo di Galerkin. Se gli spazi discreti sono composti da polinomi a tratti su griglie scelte opportunamente (ad esempio decomposizioni del dominio in triangoli o quadrilateri in 2D, tetraedri o parallelepipedi in 3D) si parla di elementi finiti. Tutti i risultati qui presentati si estendono a questa situazione più generale, con qualche (interessante!) complicazione.

### 6.6.1 Elementi finiti lineari ($p = 1$)

Il caso più semplice, e allo stesso tempo il più importante, è quello degli elementi finiti lineari, cioè di grado $p = 1$: $V_h = S_0^1(\mathcal{T}_h)$. La dimensione di questo spazio è $\dim(S_0^1(\mathcal{T}_h)) = n$, il numero dei nodi interni. La base più semplice di questo spazio è costituita dalle funzioni "a tenda", cioè le $\varphi_j \in V_h = S_0^1(\mathcal{T}_h)$ tali che $\varphi_j(x_k) = \delta_{j,k}$. Le $\varphi_j$ sono dette anche "funzioni nodali", poiché ciascuna è associata a un nodo della griglia. Esplicitamente:

$$\varphi_j(x) = \begin{cases} \dfrac{x - x_{j-1}}{x_j - x_{j-1}} = \dfrac{x - x_{j-1}}{h_j} & \text{in } K_j, \\ \dfrac{x_{j+1} - x}{x_{j+1} - x_j} = \dfrac{x_{j+1} - x}{h_{j+1}} & \text{in } K_{j+1}, \\ 0 & \text{altrimenti} \end{cases} \qquad j = 1, \dots, n.$$




Per la soluzione discreta $u_h(x) = \sum_{k=1}^n U_k \varphi_k(x)$ (come per ogni altra funzione in $V_h$), i valori $U_k$ rappresentano sia i coefficienti nell'espansione rispetto alla base scelta, sia il valore nei nodi: $u_h(x_j) = U_j$.

Un'importante proprietà di queste funzioni di base è che il loro **supporto** è il più piccolo possibile: solo due elementi. Questo garantisce che i supporti di $\varphi_j$ e $\varphi_k$ si sovrappongono solo se $|k - j| \le 1$. Quindi l'elemento $A_{j,k}$ della matrice di Galerkin è uguale a zero se $|j - k| > 1$, in altre parole la matrice è **tridiagonale** e il sistema lineare (66) può essere risolto molto efficientemente come in §4.5.1.

Per implementare il metodo è necessario innanzitutto assemblare la matrice $\underline{\underline{\mathbf{A}}}$, i cui elementi sono

$$A_{j,k} = \int_a^b (\varphi_k' \varphi_j' + q \varphi_k \varphi_j) \, \mathrm{d}x.$$

L'integrale di $\varphi_k' \varphi_j'$ si può calcolare esattamente usando $\varphi_j'|_{K_j} = 1/h_j$ e $\varphi_j'|_{K_{j+1}} = -1/h_{j+1}$. Al contrario, se $q$ è una funzione generica per calcolare $\int_a^b q \varphi_k \varphi_j \, \mathrm{d}x$ è necessario usare una formula di **quadratura** su ogni elemento. Possiamo scegliere di approssimare $q$ con una funzione costante su ciascun elemento: $q|_{K_j} \approx q_j$, dove $q_j$ è scelto come il valore nel punto medio dell'elemento, cioè $q_j = q(\frac{x_{j-1}+x_j}{2})$. (In alternativa si può prendere $q_j$ come la media di $q$ sullo stesso elemento, cioè $q_j = \frac{1}{h_j} \int_{x_{j-1}}^{x_j} q(x) \, \mathrm{d}x$). In questo caso ($q|_{K_j} \approx q_j$), integrando analiticamente il prodotto $\varphi_k \varphi_j$ otteniamo

$$\underline{\underline{\mathbf{A}}} = \begin{pmatrix} \boxed{\frac{1}{h_1} + \frac{1}{h_2} + \frac{q_1 h_1 + q_2 h_2}{3}} & -\frac{1}{h_2} + \frac{q_2 h_2}{6} & & \\ -\frac{1}{h_2} + \frac{q_2 h_2}{6} & \boxed{\frac{1}{h_2} + \frac{1}{h_3} + \frac{q_2 h_2 + q_3 h_3}{3}} & -\frac{1}{h_3} + \frac{q_3 h_3}{6} & \\ & \ddots & \ddots & \ddots \\ & & -\frac{1}{h_n} + \frac{q_h h_n}{6} & \boxed{\frac{1}{h_n} + \frac{1}{h_{n+1}} + \frac{q_n h_n + q_{n+1} h_{n+1}}{3}} \end{pmatrix}. \tag{70}$$

Se $q$ è zero e tutti gli elementi hanno la stessa lunghezza $h_j = h$, allora $\frac{1}{h} \underline{\underline{\mathbf{A}}}$ coincide con la matrice del metodo delle differenze finite (24) per lo stesso problema. (Qui stiamo commettendo un piccolo abuso di notazione: la matrice in (70) è l'approssimazione della matrice $\underline{\underline{\mathbf{A}}}$ del metodo di Galerkin definita in (66), da cui differisce per l'errore commesso nell'integrazione di $q \varphi_k \varphi_j$; si veda l'Esercizio 6.31.)

**Esercizio ✎ 6.26.** Verificare l'espressione degli elementi di $\underline{\underline{\mathbf{A}}}$ in (70).

Allo stesso modo, gli elementi $F_j = \int_a^b \varphi_j f \,\mathrm{d}x$ del termine noto $\vec{\mathbf{F}}$ si calcolano usando una formula di quadratura. Ad esempio, approssimando $f$ con una costante a tratti $f|_{K_j} \approx f_j := f(\frac{x_{j-1}+x_j}{2})$, otteniamo

$$F_j = \int_{K_j} f_j \varphi_j(x)\,\mathrm{d}x + \int_{K_{j+1}} f_{j+1}\varphi_j(x)\,\mathrm{d}x = \frac{h_j f_j + h_{j+1} f_{j+1}}{2}. \tag{71}$$

Ricordando la definizione delle $\varphi_j$, questo coincide con la formula dei rettangoli composita sugli intervalli $K_j$, cioè $F_j = \sum_{k=1}^{n+1} h_k f(\frac{x_{k-1}+x_k}{2})\varphi_j(\frac{x_{k-1}+x_k}{2}) \approx \int_a^b f\varphi_j \,\mathrm{d}x$.

**Esercizio ✎ 6.27.** Per $f, q \in C^0(a,b)$ e per griglie uniformi, cioè con $h_j = h$ per ogni $j$, mostrare che se $\int_a^b q\varphi_k\varphi_j \,\mathrm{d}x$ e $F_j = \int_a^b \varphi_j f \,\mathrm{d}x$ sono approssimati usando la regola dei trapezi composita sui nodi $x_j$ (cioè $\int_a^b g(x)\,\mathrm{d}x \approx \sum_{j=1}^{n+1} h_j \frac{g(x_{j-1})+g(x_j)}{2}$), allora il sistema lineare (66) del metodo di Galerkin coincide con quello del metodo delle differenze finite (24) per lo stesso problema. (Ricordare il valore di $\varphi_j(x_k)$.)

**Esercizio 💻 6.28.**

- Implementare il metodo agli elementi finiti per una griglia uniforme ($h_j = h$ per ogni $j$) per il problema $-u'' + u = (1+\pi^2)\sin(\pi x)$ in $(a,b) = (0,1)$ con $u(a) = u(b) = 0$.

  Plottare la soluzione discreta $u_h$ e quella esatta $u$.

  Notare che $q$ costante semplifica l'implementazione. Implementare $\underline{\underline{\mathbf{A}}}$ come matrice sparsa.
- Estendere il metodo al caso con condizioni al bordo non omogenee e $q$ variabile. Considerare gli esempi già visti nell'Esercizio 4.1.
- Studiare la convergenza in $h$ dell'errore in norma $L^2(a,b)$ e $H^1(a,b)$.

  Per calcolare queste norme integrali (ricordarsi Definizione 6.1) è necessario usare una formula di quadratura su ciascun elemento per integrare $(u-u_h)^2$ e $(u'-u_h')^2$, ad esempio quella di Cavalieri–Simpson composita:

$$\int_a^b g(x)\,\mathrm{d}x \approx \sum_{j=1}^{n+1} \frac{h_j}{6}\Big(g(x_{j-1}) + 4g(\tfrac{x_j+x_{j-1}}{2}) + g(x_j)\Big).$$

  Ricordare che $u_h$ è una funzione lineare a tratti e $u_h'$ è costante a tratti. In particolare l'integrando $g(x) = (u'(x) - u_h'(x))^2$ usato per calcolare la seminorma $H^1(a,b)$ è discontinuo in ogni nodo $x_j$: il termine $g(x_{j-1}) + 4g(\frac{x_j+x_{j-1}}{2}) + g(x_j)$ nella formula di quadratura deve approssimare $\int_{x_{j-1}}^{x_j} g(x)\,\mathrm{d}x$, quindi $g(x_{j-1})$ e $g(x_j)$ vanno interpretati come opportuni limiti da destra e da sinistra, rispettivamente. Può essere utile precalcolare un vettore $n+1$-dimensionale `Uder` la cui componente $j$sima è il valore della derivata prima di $u_h$ nell'elemento $j$simo ( `Uder(j)`$= u_h'|_{K_j}$).
- Studiare la dipendenza da $h$ del numero di condizionamento della matrice.

Confrontare i grafici ottenuti con Figura 35.

**Esercizio 💻 6.29** (Elementi finiti per soluzioni singolari)**.** Una delle motivazioni per l'introduzione del metodo di Galerkin (e quindi degli elementi finiti) è il trattamento di problemi con soluzioni deboli e non classiche.

- Implementare il metodo degli elementi finiti per il primo esempio di §6.3.1, cioè $-u'' = \chi_{(-1/2,1/2)}$ in $H_0^1(-1,1)$, la cui soluzione appartiene a $H^2(-1,1) \setminus C^2(-1,1)$.

  Attenzione: per avere ordini di convergenza ottimali in norma $L^2$ è necessario che l'errore di quadratura nel calcolo del vettore $\vec{\mathbf{F}}$ sia $\mathcal{O}(h^2)$ (cosa che è sempre garantita nei casi precedenti in cui $f$ è liscia). In questo caso $f$ è discontinua, quindi è necessario trattare con cura gli $F_j$ corrispondenti a $\varphi_j$ il cui supporto interseca le discontinuità di $f$. Ad esempio l'approssimazione di $f$ con una costante a tratti (cioè $F_j = \frac{h}{2}[f(\frac{x_{j-1}+x_j}{2}) + f(\frac{x_j+x_{j+1}}{2})]$) calcola il valore esatto di $\int_{-1}^1 \varphi_j f \,\mathrm{d}x$ se il numero di elementi $n+1$ è multiplo di 4.
- Approssimare con il metodo degli elementi finiti la soluzione del problema variazionale $\int_{-1}^1 u'w' \,\mathrm{d}x = w(0)$ in $H_0^1(-1,1)$ (cioè $-u'' = \delta$ con $u(\pm 1) = 0$) descritto in §6.3.1.

  Cosa si osserva quando $n$ è pari? E quando è dispari?

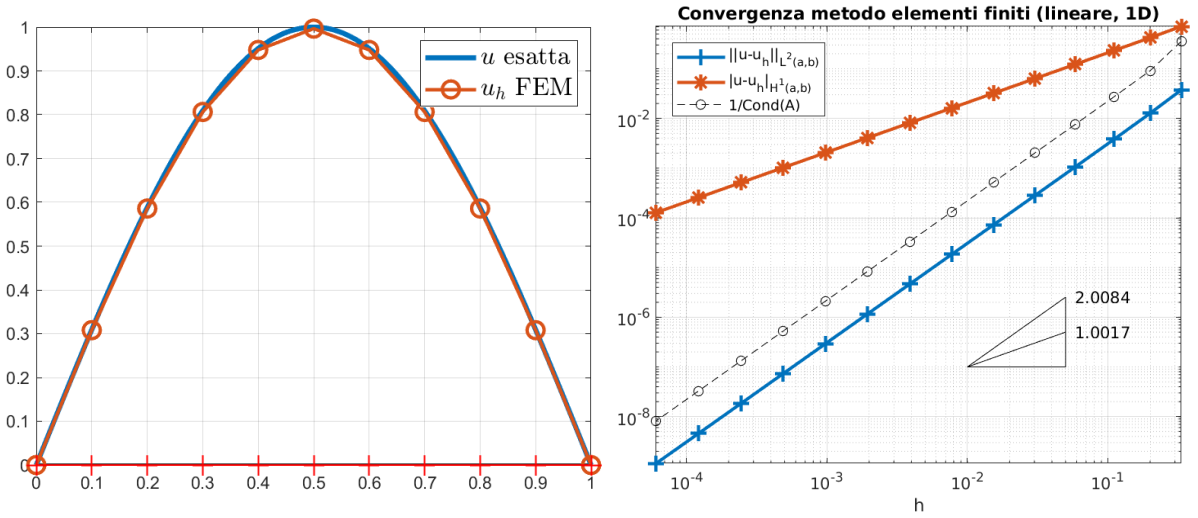



*Figura 35: Sinistra: la soluzione del primo problema al bordo nell'Esercizio 6.28 e la soluzione del metodo agli elementi finiti con $n = 9$. (Si veda anche Figura 36.)*
*Destra: l'errore commesso dal metodo degli elementi finiti misurato in norma $L^2(0,1)$ e seminorma $H^1(0,1)$ per lo stesso problema e $n = 2^1, \dots, 2^{14}$.*

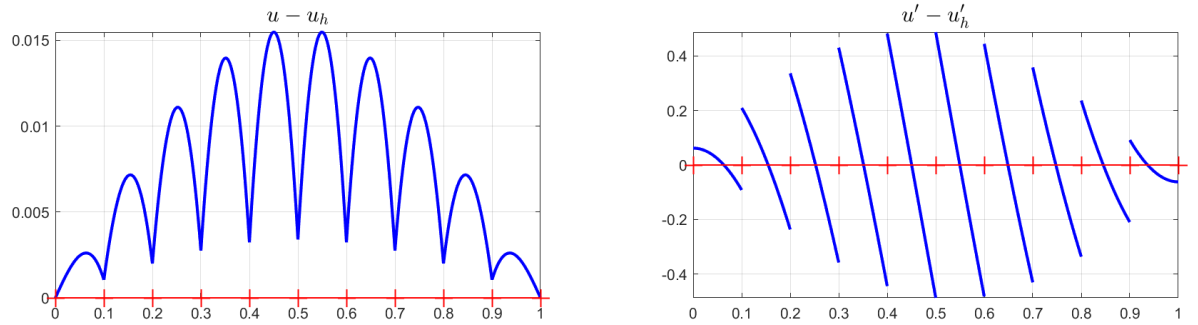



*Figura 36: L'errore $u - u_h$ commesso dal metodo degli elementi finiti e la sua derivata prima $u' - u_h'$ per il problema in Figura 35 ($n = 9$). Notare le scale diverse. Dal primo grafico si nota che $u_h$ non è l'interpolante di $u$, come Figura 35 potrebbe suggerire, ma il valore nei nodi è leggermente sottostimato.*

**Esercizio 💻 6.30** ("Per solutori più che abili"). Finora abbiamo considerato il metodo degli elementi finiti su griglie uniformi, cioè con $h_j = h$ per ogni $j$. Quando la soluzione $u$ presenta una singolarità può essere conveniente usare una griglia "graduata", cioè con tanti elementi più piccoli vicino alla singolarità per approssimarla accuratamente e pochi elementi più ampi nella parte dove la soluzione è più liscia e facilmente approssimabile.

Supponiamo di voler approssimare il problema al bordo

$$-u'' = -\mu(\mu-1)x^{\mu-2} \quad \text{in } (0,1), \qquad u(0) = 0,\ u(1) = 1,$$

per un parametro $\mu > 1/2$, $\mu \notin \mathbb{N}$. La soluzione è $u(x) = x^\mu$ che ha una singolarità in $x = 0$. Per fissare le idee, scegliamo $\mu = 1.2$, per cui $u'(x) = 1.2x^{0.2}$ è una funzione di Hölder con esponente $0.2$ ($u' \in C^{0,0.2}([0,1])$, $|u'(x) - u'(y)| \le C|x-y|^{0.2}$). Verificare che $u \in H^1(0,1)$, per cui vale la teoria astratta per il metodo di Galerkin, e che $u \notin H^2(0,1)$ (per cui non valgono le stime di approssimazione che vedremo tra poco).

Mostrare numericamente che l'errore del metodo degli elementi finiti lineari sulla griglia uniforme di $n+1$ elementi (nodi $x_j = j/(n+1)$ per $j = 1, \dots, n$) converge con ordine $\mathcal{O}(n^{-1.2})$ in norma $L^2(0,1)$ e $\mathcal{O}(n^{-0.7})$ in norma $H^1(0,1)$.

Implementare il metodo degli elementi finiti con la griglia "graduata" con $n$ nodi $x_j = (j/(n+1))^\zeta$ per $j = 1, \dots, n$, dove $\zeta > 0$ è un parametro. Quando $\zeta > 1$ gli elementi si concentrano vicino alla singolarità in $x = 0$. Verificare numericamente che con $\zeta = 2$ si ottiene convergenza con ordine $\mathcal{O}(n^{-2})$ in norma $L^2(0,1)$ e $\mathcal{O}(n^{-1})$ in norma $H^1(0,1)$. Attenzione: l'implementazione di $\underline{\underline{\mathbf{A}}}$, $\vec{\mathbf{F}}$ e il calcolo delle norme dell'errore ora richiede più attenzione poiché $h_j$ ha un valore diverso per ogni elemento $K_j$. Studiare come variano gli ordini di convergenza al variare di $\mu$ e $\zeta$.

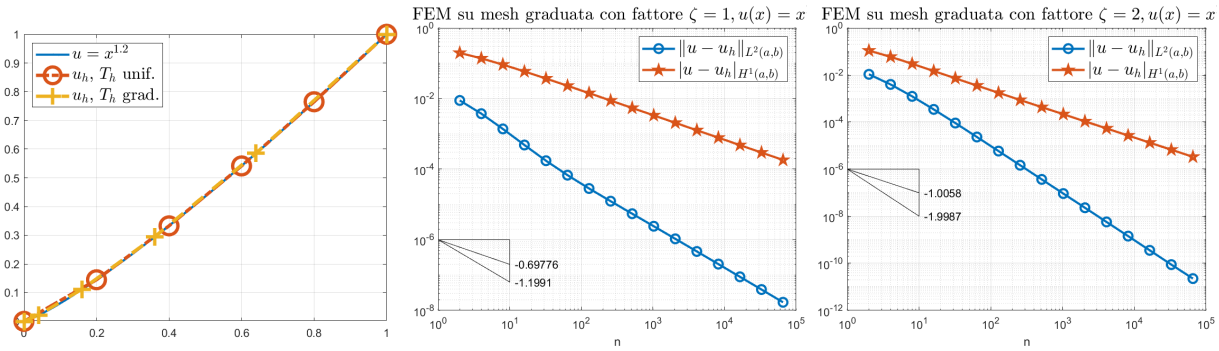



*Figura 37: Sinistra: la soluzione del problema al bordo descritto nell'Esercizio 6.30 con $\mu = 1.2$ e le soluzioni del metodo degli elementi finiti con $n = 4$ usando una griglia uniforme (cerchi blu) e una griglia graduata (croci rosse). Nonostante $u(x) = x^{1.2}$ abbia un aspetto innocuo, è una funzione singolare: la sua derivata seconda non è limitata in 0. Al centro gli ordini di convergenza per la griglia uniforme ($x_j = \frac{j}{n+1}$) e a destra con una griglia graduata ($x_j = (\frac{j}{n+1})^\zeta$, $\zeta = 2$), in entrambi i casi per $n = 2^1, 2^2, \dots, 2^{16}$.*

**Esercizio ✎ 6.31** (Errore di quadratura)**.** Assumiamo che $q$ sia una funzione liscia e $\min_{x\in[a,b]} q(x) > 0$.

(i) Dimostrare che gli errori di quadratura commessi dall'approssimazione in (70) con $q_j = q(\frac{x_{j-1}+x_j}{2})$ hanno i seguenti ordini per $h \searrow 0$:

$$E_{j,j} := \left| \int_a^b q(x)\varphi_j^2(x)\,\mathrm{d}x - \frac{q(\frac{x_{j-1}+x_j}{2})h_j + q(\frac{x_j+x_{j+1}}{2})h_{j+1}}{3} \right| = \mathcal{O}(h^2),$$

$$E_{j-1,j} := \left| \int_a^b q(x)\varphi_{j-1}(x)\varphi_j(x)\,\mathrm{d}x - \frac{q(\frac{x_{j-1}+x_j}{2})h_j}{6} \right| = \mathcal{O}(h^3).$$

Suggerimento: usare l'espansione di Taylor di $q$ e ricordare come è stata ottenuta questa quadratura.

(ii) (Più difficile.) Mostrare che se la griglia è uniforme, cioè $h_j = h$ per ogni $j$, allora $E_{j,j} = \mathcal{O}(h^3)$.

(iii) Dedurre che l'errore relativo soddisfa $E_{j,j}/|A_{j,j}| = \mathcal{O}(h)$ e $E_{j-1,j}/|A_{j-1,j}| = \mathcal{O}(h^2)$ (e $E_{j,j}/|A_{j,j}| = \mathcal{O}(h^2)$ su una griglia uniforme).

(iv) Mostrare che se la matrice $\underline{\underline{\mathbf{A}}}$ è calcolata con la formula composita dei trapezi come nell'Esercizio 6.27 allora l'errore commesso è $\mathcal{O}(h)$, sia per i termini sulla diagonale che per gli altri.

Suggerimento: si può sfruttare il primo punto dell'esercizio.

(Notiamo che la formula dei trapezi in questo caso non raggiunge l'ordine classico $\mathcal{O}(h^3)$ come in [QSSG14, §8.1.2] perché l'integrando $q\varphi_j\varphi_k$ ha derivate proporzionali a $h^{-2}$.)

(v) Scrivere l'approssimazione di $\underline{\underline{\mathbf{A}}}$ ottenuta usando la formula dei rettangoli $\int_a^b g\,\mathrm{d}x \sim \sum_{j=1}^{n+1} h_j g(\frac{x_{j-1}+x_j}{2})$. Mostrare che l'errore commesso è $\mathcal{O}(h)$ per tutti i termini non-nulli della matrice.

Suggerimento: la matrice ottenuta differisce da quella in (70) solo nei valori 3 e 6 nei denominatori.

Attenzione a non confondersi: con la formula dei rettangoli approssimiamo il prodotto $q\varphi_j\varphi_k$ in un elemento con il suo valore nel punto medio. Con (70) invece approssimiamo $q$ con il suo valore nello stesso punto medio, ma integriamo esattamente il prodotto $\varphi_j\varphi_k$.

(vi) Ora consideriamo l'errore di quadratura nel calcolo del termine noto $\vec{\mathbf{F}}$. Mostrare che sia formula (71) che la formula dei trapezi nell'Esercizio 6.27 hanno errore assoluto $\mathcal{O}(h^2)$ e errore relativo $\mathcal{O}(h)$. Verificare che per il termine noto la formula dei rettangoli coincide con (71).

Mostrare che anche per il termine noto l'errore è $\mathcal{O}(h^3)$ se la griglia è uniforme.

(vii) Verificare quanto dimostrato con Matlab con una $q$ liscia a piacere.

**Esercizio ✎ 6.32** (Esattezza nodale)**.** Sia $u$ la soluzione del problema di Dirichlet (53) e $u_h$ l'approssimazione calcolata con il metodo degli elementi finiti lineari. Mostrare che se $q = 0$ e se il termine noto $\vec{\mathbf{F}}$ del sistema lineare è calcolato esattamente (cioè senza errori di quadratura), allora $u_h$ è "nodalmente esatta", cioè $u_h(x_j) = u(x_j)$ per ogni nodo $x_j$ della griglia computazionale.

Suggerimento: usare l'ortogonalità di Galerkin e scegliere una funzione test in modo furbo.

**Esercizio ✎ 6.33.** Siano $a < x_1 < x_2 < \cdots < x_n < b$ and $\psi_j(x) := \begin{cases} \frac{x-a}{x_j-a} & a \le x \le x_j, \\ \frac{b-x}{b-x_j} & x_j < x < b \end{cases}$ per $j = 1, \ldots, n$. Descrivere il metodo di Galerkin applicato al problema di Dirichlet (53) con base $\psi_1, \ldots, \psi_n$ per lo spazio discreto $V_h$ e confrontarlo con la versione avente per base le funzioni a tenda nodali $\varphi_1, \ldots, \varphi_n$, con la stessa scelta di nodi. In cosa differiscono e in cosa sono uguali? Quali proprietà della matrice del metodo sono diverse? Quale delle due soluzioni sarà più accurata? Quale richiede un costo computazionale minore?

### 6.6.2 Elementi finiti quadratici ($p = 2$)

Lo spazio discreto $S_0^2(\mathcal{T}_h)$ è lo spazio dei polinomi a tratti sulla griglia $\mathcal{T}_h$, continui, di grado al massimo 2, e che valgono 0 ai due estremi del dominio. La dimensione è $2n+1$. Un elemento di $S_0^2(\mathcal{T}_h)$ è determinato se conosciamo il suo valore in tre punti per ogni elemento. Quindi scegliamo come "gradi di libertà" i valori negli $n$ nodi $x_j$ e i valori negli $n+1$ punti medi $\frac{1}{2}(x_{j-1} + x_j)$ degli elementi: cioè scegliamo come basi gli elementi di $S_0^2(\mathcal{T}_h)$ che valgono 1 in uno di questi $2n+1$ punti e 0 nei rimanenti $2n$ punti. Le funzioni di base associate (cioè "che valgono 1 in") a nodi e punti medi sono rappresentate in Figura 38. Quelle associate ai punti medi sono supportate in un unico elemento e sono dette "funzioni a bolla" (*bubble functions*).

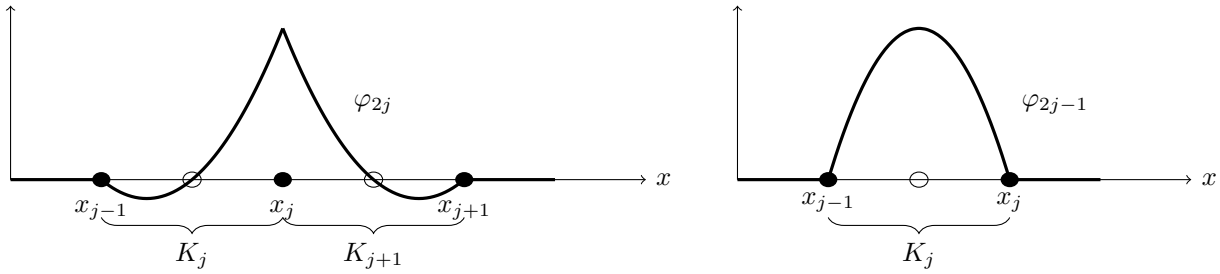



*Figura 38: Sinistra: la funzione di base associata al nodo $x_j$; ha supporto nei due elementi $K_j$ e $K_{j+1}$ ed è uguale a zero nei punti medi. Destra: la funzione di base associata al punto medio dell'elemento $K_j$; ha supporto solo in $K_j$ e vale zero in tutti i nodi della mesh. I pallini neri denotano i nodi che separano gli elementi, quelli bianchi i punti medi degli elementi.*

Le funzioni di base sono numerate chiamando $\varphi_{2j-1}$, $1 \le j \le n+1$ la funzione a bolla associata a $K_j$ e $\varphi_{2j}$ la funzione nodale associata a $x_j$, $1 \le j \le n$.

**Esercizio ✎ 6.34.** Verificare che il supporto della funzione $\varphi_{2j}$ interseca quello di $\varphi_k$ esattamente per $k = 2j-2, 2j-1, 2j+1$ e $2j+2$. Il supporto di $\varphi_{2j-1}$ interseca invece solo quello di $\varphi_{2j-2}$ e $\varphi_{2j}$.

Questo esercizio mostra che, con questo ordinamento degli elementi della base, la matrice $\underline{\underline{\mathbf{A}}}$ è **pentadiagonale**, cioè $A_{j,k} = 0$ se $|j-k| > 2$. I sistemi lineari pentadiagonali si possono risolvere con $\mathcal{O}(n)$ operazioni, estendendo la tecnica analizzata nel caso tridiagonale.

**Esercizio ✎ 6.35.** Scrivere l'espressione esplicita delle $\varphi_j$ e degli elementi di $\underline{\underline{\mathbf{A}}}$ e $\vec{\mathbf{F}}$.

**Esercizio ✎ 6.36.** Disegnare lo *sparsity pattern* (cioè la posizione degli elementi diversi da zero) della matrice $\underline{\underline{\mathbf{A}}}$ nel caso in cui gli elementi della base sono ordinati nel modo seguente: $\widetilde{\varphi}_1, \ldots, \widetilde{\varphi}_{n+1}$ sono le funzioni a bolla associate agli elementi $K_1, \ldots, K_{n+1}$, mentre $\widetilde{\varphi}_{n+2}, \ldots, \widetilde{\varphi}_{2n+1}$ sono le basi nodali associate ai nodi $x_1, \ldots, x_n$, rispettivamente. Come cambia la sparsità della matrice con questa scelta? E la struttura a bande?

In modo simile si possono costruire gli elementi di base di $S_0^p(\mathcal{T}_h)$ per $p > 2$. Fissato $p$ si sceglie una base composta da $n$ funzione nodali (una per ogni nodo) e da $(p-1)(n+1)$ funzioni a bolla. Al crescere di $p$, le funzioni a bolla devono essere scelte accuratamente per evitare numeri di condizionamento eccessivi. Una buona scelta è quella di prendere le lineari a tratti della sezione precedente come funzioni di base nodali, e i polinomi di Legendre integrati (opportunamente traslati e scalati) come funzioni a bolla. Chi vuole approfondire veda l'Esercizio 6.37.

Qual è il vantaggio di usare $V_h = S_0^2(\mathcal{T}_h)$ (o più in generale uno spazio di polinomi a tratti di grado maggiore) invece di $S_0^1(\mathcal{T}_h)$? La motivazione principale è che le proprietà di approssimazione sono migliori, quindi gli ordini di convergenza sono più alti. Affinché gli ordini di convergenza siano ottenuti è necessario usare una formula di quadratura sufficientemente accurata per l'approssimazione degli elementi di $\underline{\underline{\mathbf{A}}}$ e $\vec{\mathbf{F}}$.

**Esercizio ✎ 6.37** (Elementi finiti di grado $p$ arbitrario)**.** Fissiamo la solita griglia $\mathcal{T}_h$ e un grado polinomiale $p \in \mathbb{N}$. Per implementare il metodo degli elementi finiti con spazio discreto $V_h = S_0^p(\mathcal{T}_h)$, cioè polinomi a tratti di grado al più $p$, è necessario scegliere una "buona" base. Questo significa che gli elementi della base devono essere (1) facili da calcolare e manipolare, (2) dare una matrice più sparsa possibile (possibilmente a bande) e (3) con un numero di condizionamento accettabile.

Siano $\varphi_j$, $j = 1, \dots, n$ le funzioni a tenda definite in §6.6.1. Siano

$$M_{k,j}(x) := \begin{cases} M_k\big(-1 + \frac{2}{h_j}(x - x_{j-1})\big) & x \in K_j, \\ 0 & x \in [a,b] \setminus K_j, \end{cases} \qquad k = 1, \dots, p-1, \quad j = 1, \dots, n+1,$$

dove $M_k \in \mathbb{P}^{k+1}(-1,1)$ sono i polinomi di Legendre integrati definiti in §5.2, che soddisfano $M_k(\pm 1) = 0$.

(i) Mostrare che $\{\varphi_j\}_{j=1,\dots,n} \cup \{M_{k,j}\}_{k=1,\dots,p-1,\ j=1,\dots,n+1}$ è una base di $S_0^p(\mathcal{T}_h)$ (prima di tutto mostrare che è un sottoinsieme di $S_0^p(\mathcal{T}_h)$).

Notiamo che se $p = 2$ la base ottenuta è diversa da quella di Figura 38: qui gli elementi nodali sono lineari a tratti, là erano definiti dall'annullamento nei punti medi.

(ii) Mostrare che se gli elementi della base sono ordinati come

$$M_{1,1},\ M_{2,1},\ \dots,\ M_{p-1,1},\ \varphi_1,\ M_{1,2},\ \dots,\ M_{p-1,2},\ \varphi_2,\quad \dots\quad \varphi_n,\ M_{1,n+1},\ \dots,\ M_{p-1,n+1},$$

allora la matrice del metodo di Galerkin ha larghezza di banda uguale a $p$ (cioè $A_{j,k} = 0$ se $|j-k| > p$) e lo *sparsity pattern* della matrice è come nell'immagine a sinistra di Figura 39 (qui per $n = 5$ e $p = 4$).

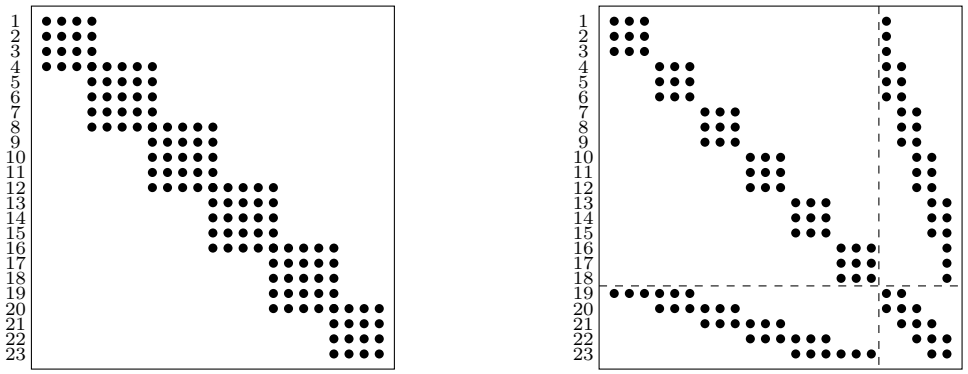


*Figura 39: Gli sparsity pattern delle matrici del metodo degli elementi finiti con grado $p = 4$ e $n = 5$, come descritto nell'Esercizio 6.37.*

(iii) Adesso consideriamo la stessa base di $S_0^p(\mathcal{T}_h)$ vista sopra ma ordinata nel modo seguente:

$$M_{1,1},\ M_{2,1},\ \dots,\ M_{p-1,1},\ M_{1,2},\ \dots,\ M_{p-1,2},\quad \dots\quad M_{1,n+1},\ \dots,\ M_{p-1,n+1},\ \varphi_1,\ \varphi_2, \dots,\ \varphi_n.$$

Mostrare che la matrice del metodo di Galerkin (che ha lo stesso numero di elementi non-zero di quella del caso (ii)) ha lo *sparsity pattern* a destra in Figura 39 (ancora per $n = 5$ e $p = 4$). In particolare non è a bande e la decomposizione LU produce matrici dense (fenomeno del *fill-in*).

(iv) Consideriamo il sistema lineare a blocchi

$$\underline{\underline{\mathbf{M}}}\vec{\mathbf{x}} = \vec{\mathbf{y}} \quad \text{con} \quad \underline{\underline{\mathbf{M}}} = \begin{pmatrix} \underline{\underline{\mathbf{M}}}_{\nwarrow} & \underline{\underline{\mathbf{M}}}_{\nearrow} \\ \underline{\underline{\mathbf{M}}}_{\swarrow} & \underline{\underline{\mathbf{M}}}_{\searrow} \end{pmatrix}, \quad \vec{\mathbf{x}} = \begin{pmatrix} \vec{\mathbf{x}}_{\uparrow} \\ \vec{\mathbf{x}}_{\downarrow} \end{pmatrix} \in \mathbb{R}^{d_1+d_2}, \quad \vec{\mathbf{y}} = \begin{pmatrix} \vec{\mathbf{y}}_{\uparrow} \\ \vec{\mathbf{y}}_{\downarrow} \end{pmatrix} \in \mathbb{R}^{d_1+d_2}.$$

Assumiamo che sia $\underline{\underline{\mathbf{M}}}$ che il primo blocco $\underline{\underline{\mathbf{M}}}_{\nwarrow} \in \mathbb{R}^{d_1 \times d_1}$ siano invertibili. Definiamo il **complemento di Schur** come la matrice $\underline{\mathbf{S}} := \underline{\underline{\mathbf{M}}}_{\searrow} - \underline{\underline{\mathbf{M}}}_{\swarrow}\underline{\underline{\mathbf{M}}}_{\nwarrow}^{-1}\underline{\underline{\mathbf{M}}}_{\nearrow} \in \mathbb{R}^{d_2 \times d_2}$. Mostrare che il vettore $\vec{\mathbf{x}}$ può essere calcolato un blocco alla volta risolvendo i sistemi lineari

$$\underline{\underline{\mathbf{M}}}_{\nwarrow}\,\vec{\mathbf{z}} = \vec{\mathbf{y}}_{\uparrow}, \qquad \underline{\mathbf{S}}\,\vec{\mathbf{x}}_{\downarrow} = \vec{\mathbf{y}}_{\downarrow} - \underline{\underline{\mathbf{M}}}_{\swarrow}\vec{\mathbf{z}} \qquad \text{e} \qquad \underline{\underline{\mathbf{M}}}_{\nwarrow}\,\vec{\mathbf{x}}_{\uparrow} = \vec{\mathbf{y}}_{\uparrow} - \underline{\underline{\mathbf{M}}}_{\nearrow}\vec{\mathbf{x}}_{\downarrow}. \tag{72}$$

Il significato di questa procedura è il seguente. Immaginiamo di avere un sistema lineare di dimensione $d_1 + d_2$, con $d_1 \gg d_2$ e di sapere calcolare facilmente la decomposizione LU del primo blocco $\underline{\underline{\mathbf{M}}}_{\nwarrow}$ (o più in generale di sapere risolvere velocemente sistemi lineari con matrice $\underline{\underline{\mathbf{M}}}_{\nwarrow}$). Allora possiamo calcolare

la soluzione $\vec{\mathbf{x}}$ risolvendo i tre sistemi in (72), due per $\underline{\underline{\mathbf{M}}}_{\nwarrow}$ e uno per il complemento di Schur $\underline{\underline{\mathbf{S}}}$ che ha dimensione $d_2$. Notiamo che il complemento di Schur richiede il calcolo di $\underline{\underline{\mathbf{M}}}_{\nwarrow}^{-1}\underline{\underline{\mathbf{M}}}_{\nearrow}$, che è possibile perché abbiamo assunto che conosciamo la decomposizione LU di $\underline{\underline{\mathbf{M}}}_{\nwarrow}$. Abbiamo risolto un sistema di dimensione $d_1 + d_2$ usando più volte la decomposizione LU del primo blocco e risolvendo un solo sistema di dimensione $d_2$.

(v) Applichiamo la tecnica del complemento di Schur al metodo di Galerkin con $V_h = S_0^p(\mathcal{T}_h)$ e la base del punto (iii) con $d_1 = (p-1)(n+1)$ e $d_2 = n$. Descrivere i sistemi lineari che devono essere risolti e spiegare perché sono poco costosi.

(vi) Ricordando da §5.2 che $M_k' = P_k$ e che polinomi di Legendre $P_k$ sono ortogonali in $L^2(-1,1)$, scrivere gli elementi della matrice degli elementi finiti in $S_0^p(\mathcal{T}_h)$ per il caso $q = 0$. Quali elementi valgono 0?

(vii) (💻 Difficile!) Implementare questo metodo per i problemi di Dirichlet visti precedentemente.

Per valutare la soluzione numerica ottenuta bisogna usare una procedura simile a quella usata per il metodo di collocazione. Per non "sprecare" l'accuratezza data dai polinomi di grado alto, è necessario usare una formula di quadratura di grado alto per calcolare il termine noto ed eventuali termini di reazione, ad esempio la quadratura di Gauss–Legendre (il cui codice è disponibile in §4.9.3).

### 6.6.3 Analisi del metodo degli elementi finiti: approssimazione e convergenza

Il Lemma di Céa 6.18, applicato al problema al bordo come in (69), garantisce che il metodo degli elementi finiti gode della quasi-ottimalità: l'errore commesso dipende solo da quanto bene lo spazio discreto $V_h$ approssima la soluzione $u$. Ci limitiamo al caso $p = 1$, quindi fissiamo $V_h = S_0^1(\mathcal{T}_h)$. Data $w \in H_0^1(a,b)$, un elemento di $V_h$ che approssima $w$ è l'interpolante $\Pi_h^1 w$ definito come l'unico elemento di $V_h$ con $(\Pi_h^1 w)(x_j) = w(x_j)$ per ogni $j = 1, \dots, n$. Ricordiamo che la quasi-ottimalità (69) vale in norma $H^1(a,b)$, quindi vogliamo controllare l'errore $\left\|w - \Pi_h^1 w\right\|_{H^1(a,b)}$. Per ottenere degli ordini di convergenza dobbiamo assumere $w \in H^2(a,b)$.

**Proposizione 6.38.** Sia $w \in H_0^1(a,b) \cap H^2(a,b)$. Sia $\mathcal{T}_h$ una griglia con meshwidth $h = \max_{j=1,\dots,n+1} h_j$, $V_h = S_0^1(\mathcal{T}_h)$ e $\Pi_h^1 : H_0^1(a,b) \to V_h$ l'operatore di interpolazione. Allora la norma $L^2$ e la (semi)norma $H^1$ dell'errore di interpolazione convergono quadraticamente e linearmente, rispettivamente, in $h$:

$$\left\|w - \Pi_h^1 w\right\|_{L^2(a,b)} \le h^2 \left\|w''\right\|_{L^2(a,b)}, \qquad \left|w - \Pi_h^1 w\right|_{H^1(a,b)} \le h \left\|w''\right\|_{L^2(a,b)}. \tag{73}$$

*Dimostrazione.* Sia $e = w - \Pi_h^1 w$ l'errore di interpolazione. Consideriamo un elemento $K_j = [x_{j-1}, x_j]$. Abbiamo $e(x_{j-1}) = e(x_j) = 0$, per la definizione di $\Pi_h^1$. La regolarità $e \in H^2(x_{j-1}, x_j)$ garantisce che $e$ è di classe $C^1$, quindi per il Teorema di Rolle esiste $\xi_j \in (x_{j-1}, x_j)$ per cui $e'(\xi_j) = 0$. Per il teorema fondamentale del calcolo, per ogni $x \in (x_{j-1}, x_j)$ vale

$$e'(x) = \int_{\xi_j}^{x} e''(s)\,\mathrm{d}s,$$

da cui, usando la disuguaglianza di Cauchy–Schwarz (55),

$$|e'(x)| \le \int_{x_{j-1}}^{x_j} |e''(s)|\,\mathrm{d}s \le \Big(\int_{x_{j-1}}^{x_j} 1^2\,\mathrm{d}s\Big)^{1/2}\Big(\int_{x_{j-1}}^{x_j} |e''(s)|^2\,\mathrm{d}s\Big)^{1/2} \le h_j^{1/2}\Big(\int_{x_{j-1}}^{x_j} |e''(s)|^2\,\mathrm{d}s\Big)^{1/2}$$

e integrando

$$\int_{x_{j-1}}^{x_j} |e'(x)|^2\,\mathrm{d}x \le \int_{x_{j-1}}^{x_j} h_j\Big(\int_{x_{j-1}}^{x_j} |e''(s)|^2\,\mathrm{d}s\Big)\,\mathrm{d}x \le h_j^2\Big(\int_{x_{j-1}}^{x_j} |e''(s)|^2\,\mathrm{d}s\Big).$$

Poiché $e = w - \Pi_h^1 w$ e $(\Pi_h^1 w)|_{K_j}$ è un polinomio lineare, abbiamo $e'' = w''$ Sommando sugli elementi e ricordando che $h = \max_{j=1,\dots,n+1} h_j$, troviamo la stima in seminorma $H^1$:

$$\left|w - \Pi_h^1 w\right|_{H^1(a,b)}^2 = \sum_{j=1}^{n+1} \int_{x_{j-1}}^{x_j} |e'(x)|^2\,\mathrm{d}x \le \sum_{j=1}^{n+1} h_j^2\Big(\int_{x_{j-1}}^{x_j} |w''(s)|^2\,\mathrm{d}s\Big) \le h^2 \left\|w''\right\|_{L^2(a,b)}^2.$$

(Attenzione: qui non avremmo potuto scrivere $\|e''\|_{L^2(a,b)}$ perché $e'$ è discontinua sui nodi e non possiamo scrivere $e''$ come una funzione a quadrato sommabile su $(a,b)$!)

Per la stima sulla norma $L^2$, usando $e(x_{j-1}) = 0$,

$$e(x) = \int_{x_{j-1}}^{x} e'(s)\,\mathrm{d}s \qquad \Rightarrow \qquad |e(x)| \le h_j^{1/2}\,\|e'\|_{L^2(x_{j-1},x_j)} \le h_j^{3/2}\,\|e''\|_{L^2(x_{j-1},x_j)} \quad \forall x \in (x_{j-1}, x_j),$$

e concludiamo come sopra. (Scrivere i dettagli per esercizio.) □

**Nota 6.39** (Ottimalità delle stime)**.** Le stime (73) sono ottimali nell'esponente di $h$ ma non nel coefficiente 1 che moltiplica il termine a destra. La stima in norma $L^2$ può essere migliorata di un fattore $1/\pi^2$ e quella in seminorma $H^1$ di un fattore $1/\pi$. Per dimostrare queste stime migliori si espande l'errore di interpolazione $e$ nell'elemento $K_j$ come serie di seni $e(x) = \sum_{\ell=1}^{\infty} a_\ell \sin\frac{\pi\ell(x-x_{j-1})}{h_j}$ e si calcolano esplicitamente le norme di $e$ in funzione dei coefficienti $a_\ell$. Per i dettagli si veda il Teorema 1.3 di [SF08].

Combinando le stime di approssimazione (73) e quelle di quasi-ottimalità (69) otteniamo una stima dell'errore del metodo agli elementi finiti: la norma $H^1(a,b)$ dell'errore converge a zero linearmente in $h$.

**Teorema 6.40.** Sia dato il problema al bordo (53), la griglia $\mathcal{T}_h$ con $h \le 1$ e lo spazio discreto $V_h = S_0^1(\mathcal{T}_h)$. Assumiamo che la soluzione $u \in H^2(a,b)$. Allora la soluzione $u_h \in V_h$ del metodo degli elementi finiti lineari converge a $u$ in norma $H^1(a,b)$ linearmente in $h$ e vale la stima

$$\|u - u_h\|_{H^1(a,b)} \le C_{qo}\sqrt{2}\,h\,\|u''\|_{L^2(a,b)}\,. \tag{74}$$

*Dimostrazione.* La stima (74) segue dalla quasi-ottimalità (69), la definizione della norma $H^1(a,b)$ e dalle stime di approssimazione (73):

$$\begin{aligned}\|u - u_h\|^2_{H^1(a,b)} &\le C_{qo}^2\,\big\|u - \Pi_h^1 u\big\|^2_{H^1(a,b)} \\ &= C_{qo}^2\big(\,\big\|u - \Pi_h^1 u\big\|^2_{L^2(a,b)} + \big|u - \Pi_h^1 u\big|^2_{H^1(a,b)}\,\big) \;\le C_{qo}^2(h^4 + h^2)\,\|u''\|^2_{L^2(a,b)}\,.\end{aligned}$$

□

Se confrontiamo la stima d'errore (74) per il metodo degli elementi finiti con la stima (27) per il metodo delle differenze finite notiamo che la prima vale per $u \in H^2(a,b)$, mentre la seconda richiede una soluzione molto più regolare, $u \in C^4(a,b)$.

In questa sezione abbiamo studiato l'errore commesso dal metodo assumendo che tutti gli integrali $A_{j,k}$ e $F_j$ siano calcolati esattamente. L'uso di una formula di quadratura introduce un errore; se la formula di quadratura è scelta in modo appropriato questo errore non modifica gli ordini di convergenza (l'influenza dell'errore di quadratura sul valore di $u_h$ è analizzata ad esempio in [SF08, Theorem 4.1]).

**Esercizio ✎ 6.41.** Considerare il problema con condizioni al bordo non omogenee $u(a) = \alpha$, $u(b) = \beta$ discretizzato come nella Nota 6.21: detta $\tilde{u}$ la funzione lineare che soddisfa le condizioni al bordo e $u_0 = u - \tilde{u}$, $u_{0,h}$ è l'approssimazione di $u_0$ ottenuta con il metodo di Galerkin e $u_h = \tilde{u} + u_{0,h}$. Mostrare che anche in questo caso vale la stima d'errore (74) per $u - u_h$.

**Nota 6.42.** Se il dato $f$ appartiene a $L^2(a,b)$, allora l'equazione differenziale in (53) si scrive $u'' = -f + qu$ e vale la stima di stabilità $\|u\|_{L^2(a,b)} \le \|u\|_{H^1(a,b)} \le (1 + C_P^2)\,\|f\|_{L^2(a,b)}$ dimostrata in (62). Combinando con il teorema otteniamo una stima dell'errore in dipendenza dai dati del problema:

$$\|u - u_h\|_{H^1(a,b)} \le Ch\,\|f\|_{L^2(a,b)}\,, \qquad \text{dove } C = \sqrt{2}C_{qo}\big(1 + (1 + C_P^2)\,\|q\|_{L^\infty(a,b)}\big).$$

**Nota 6.43** (Stima ottimale in norma $L^2$)**.** La Proposizione 6.38 mostra che $V_h$ contiene una funzione discreta che approssima $u$ con ordine lineare in norma $H^1$ e quadratico in norma $L^2$. Il Teorema 6.40 mostra la convergenza lineare del metodo degli elementi finiti sia in norma $H^1$ (ordine ottimale) che in norma $L^2$ (ordine subottimale). Dagli esperimenti numerici vediamo invece che la norma $L^2$ dell'errore converge con ordine quadratico, è possibile dimostrarlo? Apparentemente no, il metodo di Galerkin fornisce la quasi-ottimalità solo per la norma $H^1$. Ad esempio la soluzione del metodo degli elementi finiti per $-u'' = f$ è la miglior approssimazione di $u$ in seminorma $H^1$ ma non in norma $L^2$. La convergenza quadratica in norma $L^2(a,b)$ può essere dimostrata usando la cosiddetta tecnica di dualità, o "trucco di Nitsche", vedere [SF08, Teorema 1.5]:

$$\|u - u_h\|_{L^2(a,b)} \le Ch^2\,\|u''\|_{L^2(a,b)}\,.$$

**Nota 6.44** (Convergenza per $u \notin H^2(a,b)$)**.** Abbiamo visto in §6.3.1 un esempio di problema variazionale con soluzione $u \notin H^2(a,b)$, il caso con una delta di Dirac come termine di sorgente $f$. Questi problemi possono essere discretizzati con elementi finiti. La stima di quasi-ottimalità non cambia, ma la stima d'errore (74) non è applicabile perché $u'' \notin L^2(a,b)$. In questo caso si può dimostrare che $\|u - u_h\|_{H^1(a,b)} \to 0$ ma in generale non si può ottenere un ordine di convergenza, vedere [SF08, Teorema 1.4]. Nel caso del problema $-u'' = \delta$ visto in §6.3.1, si osserva numericamente che la norma $L^2$ e quella $H^1$ dell'errore convergono come $h^{3/2}$ e $h^{1/2}$. Questo può essere dimostrato con un analisi abbastanza complicata della regolarità della soluzione $u$, mostrando che questa sta esattamente "a metà strada" tra $H^1(a,b)$ e $H^2(a,b)$.

**Nota 6.45** (Stime a priori vs a posteriori)**.** La disuguaglianza (74) è detta stima "a priori", che significa che l'errore è controllato da una norma della soluzione esatta $u$ (cioè $\|u''\|_{L^2(a,b)}$). Esistono stime dette "a posteriori", in cui l'errore commesso è maggiorato da una funzione della soluzione discreta $u_h$. Queste sono utili perché permettono di stimare numericamente l'errore commesso e di migliorare l'approssimazione con delle iterazioni successive, ad esempio raffinando automaticamente la griglia nella parte del dominio dove l'errore è più grande. Si veda ad esempio [SM03, §14.5].

Finora abbiamo analizzato il metodo degli elementi finiti lineari, cioè con $V_h = S_h^1(\mathcal{T}_h)$. Come cambiano le stime d'errore per gradi polinomiali più alti $p > 1$? La stima di quasi-ottimalità non cambia, mentre quelle di approssimazione forniscono potenze di $h$ più alte, a condizione che $u$ sia sufficientemente regolare. Vale il seguente risultato: se la soluzione del problema al bordo soddisfa $u \in H^{s+1}(a,b)$ per $s \geq 1$ e il metodo di Galerkin viene applicato con $V_h = S_0^p(\mathcal{T}_h)$, allora

$$\|u - u_h\|_{H^1(a,b)} \leq C_1 h^{\min\{p,s\}} \|u\|_{H^{s+1}(a,b)},$$
$$\|u - u_h\|_{L^2(a,b)} \leq C_0 h^{1+\min\{p,s\}} \|u\|_{H^{s+1}(a,b)},$$

dove $C_0, C_1$ sono delle costanti positive indipendenti da $u$ e da $h$; [QSSG14, Proprietà 11.1]. Queste stime suggeriscono di usare un grado polinomiale alto solo quando la soluzione è sufficientemente regolare. Quando per un dato problema si considera una sequenza di discretizzazioni sempre più raffinate, cioè una sequenza di spazi discreti $V_h$ sempre più grandi, si parla di "convergenza in $h$" se gli spazi sono ottenuti considerando griglie sempre più fini con lo stesso grado polinomiale e "convergenza in $p$" (o metodo degli "elementi spettrali") se gli spazi sono definiti sulla stessa griglia ma hanno gradi polinomiali crescenti.

# 7 Equazione del calore

Finora ci siamo occupati di problemi al bordo per equazioni differenziali ordinarie. In quest'ultima sezione studiamo brevemente una delle più semplici equazioni alle derivati parziali.

Partendo dall'equazione di continuità e dalla legge di Fourier (o di Fick), nella Sezione 2.1 abbiamo derivato l'equazione del calore: data una sorgente di calore $f$, l'equazione alle derivate parziali (lineare, parabolica)

$$\frac{\partial u}{\partial t} - \Delta u = f$$

descrive l'evoluzione della temperatura $u$, funzione del tempo $t \in \mathbb{R}$ e della posizione $\mathbf{x} \in \mathbb{R}^n$. In questa sezione facciamo due ulteriori semplificazioni: ci limitiamo (1) a un dominio con una sola dimensione spaziale $n = 1$, e (2) all'equazione omogenea (cioè con $f = 0$). Ad esempio possiamo pensare all'evoluzione di una data temperatura iniziale in una barra metallica sottile e isolata dall'ambiente circostante nella sua lunghezza.

Studieremo alcune proprietà delle soluzioni di questa equazione usando la soluzione fondamentale e il metodo di Fourier, che sono due tecniche analitiche. Vedremo poi come approssimarle numericamente con il $\theta$-metodo, che combina differenze finite e metodi per equazioni differenziali ordinarie. Più dettagli si possono trovare su [TW05, §3, §4, §6.2, §10], [QSSG14, §12.1–3], [LeVeque07, §9] o [Iserles09, §16].

## 7.1 Problema ai valori iniziali su $\mathbb{R} \times \mathbb{R}^+$

Come sono fatte le soluzioni dell'equazione del calore? Consideriamo come primo esempio le soluzioni su una barra di lunghezza infinita, cioè sul dominio $x \in \mathbb{R}$ e $t > 0$. Se assegniamo la condizione iniziale $u(x,0) = u^0(x)$ con $u^0 \in C^0(\mathbb{R}) \cap L^\infty(\mathbb{R})$, la soluzione $u$ di

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0 \quad \text{in } \mathbb{R} \times \mathbb{R}^+, \qquad u(x,0) = u^0(x) \tag{75}$$

si può scrivere come una convoluzione nella variabile spaziale:

$$u(x,t) = \big(\Phi(\cdot,t) * u^0(\cdot)\big)(x) = \int_{\mathbb{R}} \Phi(x-y,t)u^0(y)\,\mathrm{d}y = \int_{\mathbb{R}} \frac{1}{\sqrt{4\pi t}} \mathrm{e}^{-\frac{(x-y)^2}{4t}} u^0(y)\,\mathrm{d}y \tag{76}$$

dove abbiamo usato la "soluzione fondamentale" (o *heat kernel*) $\Phi(x,t) := \frac{1}{\sqrt{4\pi t}}\mathrm{e}^{-\frac{x^2}{4t}}$, che è una funzione Gaussiana in $x$ con varianza $\sigma^2 = 2t$ che cresce in $t$, definita solo per $t > 0$.

**Esercizio ✎ 7.1.** Verificare che $u$ in (76) è soluzione dell'equazione del calore omogenea per $x \in \mathbb{R}$ e $t > 0$.

Per dimostrare che $u$ soddisfa la condizione iniziale, nel senso che $\lim_{(x,t)\to(x^0,0)} u(x,t) = u^0(x^0)$, si veda §2.3.1.b in [L.C. Evans, *Partial Differential Equations*, AMS, 2010]. Il teorema di derivazione sotto il segno di integrale implica che $u$ è di classe $C^\infty$ nel suo dominio $(x,t) \in \mathbb{R} \times (0,\infty)$. Questo è vero anche se il dato iniziale è meno regolare: l'equazione del calore ha un effetto "regolarizzante".

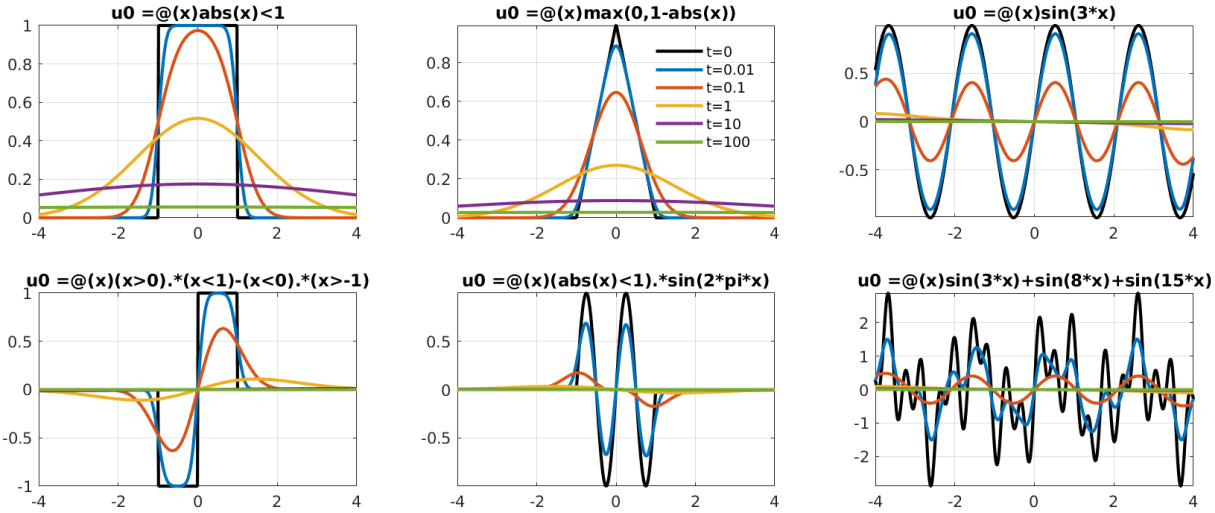



*Figura 40: La soluzione $u$ del problema (75) per alcuni valori fissati di $t$ e per $x \in (-4,4)$. Le sei figure corrispondono a diverse condizioni iniziali, scritte (come comando Matlab) nel titolo. I diversi colori corrispondono ai diversi istanti di tempo, come descritto nella legenda.*

Nella Figura 40 vediamo le soluzioni dell'equazione del calore per diversi valori iniziali $u^0$ (in nero). Vediamo che le soluzione decadono e si appiattiscono verso una costante. Questo ha senso fisicamente: immaginiamo una barra metallica omogenea isolata di lunghezza infinita. Se la temperatura iniziale è di 0°C ovunque tranne che in un segmento riscaldato a 100°C, dopo un certo intervallo di tempo la temperatura sarà tra 0°C e 100°C su un intervallo più ampio e diminuirà progressivamente (figura in alto a sinistra). Se invece un segmento della barra è a 100°C e un segmento adiacente ad esso e della stessa lunghezza è a −100°C, ci aspettiamo che la temperatura decada più velocemente a zero (figura in basso a sinistra). La temperatura converge a zero più velocemente se il dato iniziale è composto da tanti brevi segmenti a 100°C alternati ad altrettanti a −100°C piuttosto che da pochi segmenti più lunghi alle stesse temperature: le oscillazioni ad alta frequenza presenti in $u^0$ vengono smorzate più rapidamente di quelle a bassa frequenza.

Vediamo anche che un dato iniziale positivo a supporto compatto genera una temperatura strettamente positiva per ogni $x \in \mathbb{R}$ e ogni $t > 0$: questo significa che l'informazione si propaga a velocità infinita, questo aspetto dell'equazione del calore non è fisicamente plausibile (tuttavia il valore di $u(x,t)$ decresce in $x$ come una funzione Gaussiana, quindi per $t$ piccolo è "numericamente zero" a poca distanza dal supporto di $u^0$).

**Esercizio 💻 7.2.** Ricostruire con Matlab la Figura 40 usando l'espressione di $u$ in (76) e sperimentare l'evoluzione della soluzione dell'equazione del calore con diversi valori iniziali. Per approssimare la convoluzione con la soluzione discreta è necessario usare una formula di quadratura.

## 7.2 Il metodo di Fourier per l'equazione del calore

Descriviamo brevemente il metodo di **separazione delle variabili** sviluppato da J.B.J. Fourier per la risoluzione dell'equazione del calore su un dominio limitato. Per maggiori dettagli e un trattamento rigoroso si veda ad esempio [TW05, §3, §10].

Consideriamo l'equazione del calore omogenea sul dominio spaziotemporale rettangolare $(0,1)\times(0,T)$, cioè $0<x<1$, $0<t<T$, dove $T>0$ è un valore fissato. Come per ogni equazione differenziale, fissiamo delle condizioni al contorno sulla frontiera del dominio $(0,1)\times(0,T)$. Per semplicità consideriamo le condizioni di Dirichlet omogenee, cioè imponiamo che il valore della temperatura a $x=0$ e $x=1$ sia zero ad ogni istante dell'intervallo $(0,T)$. Infine imponiamo il valore iniziale $u^0$ al tempo $t=0$:

$$\begin{cases} \dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2} = 0 & \text{in } (0,1)\times(0,T) \\ u(x,0) = u^0(x) & x\in(0,1), \\ u(0,t) = u(1,t) = 0 & t\in(0,T). \end{cases} \tag{77}$$

Sul lato $t=T$ del dominio spazio-temporale non abbiamo imposto condizioni: la temperatura al tempo finale è parte dell'incognita che vogliamo determinare. Poiché abbiamo scelto l'equazione omogenea (con $f=0$) e le condizioni al bordo omogenee, l'unico termine forzante è il valore iniziale $u^0$.

Cerchiamo una soluzione scritta in forma "separabile", cioè come prodotto di due funzioni di una sola variabile: $u(x,t) = \mathsf{X}(x)\mathsf{T}(t)$. L'equazione del calore diventa $\mathsf{X}(x)\mathsf{T}'(t) = \mathsf{X}''(x)\mathsf{T}(t)$. Dividendo per $\mathsf{X}(x)\mathsf{T}(t)$ troviamo $\mathsf{T}'(t)/\mathsf{T}(t) = \mathsf{X}''(x)/\mathsf{X}(x) = \lambda$, dove il valore $\lambda$ deve essere indipendente da $x$ e da $t$. L'uguaglianza $\mathsf{X}''(x) = \lambda\mathsf{X}(x)$ indica che il fattore $\mathsf{X}(x)$ di $u(x,t) = \mathsf{X}(x)\mathsf{T}(t)$ deve essere autofunzione dell'operatore differenziale in spazio, cioè di $-\frac{\partial^2}{\partial x^2}$. Abbiamo già incontrato in §4.7 il problema agli autovalori $\mathsf{X}''(x) = \lambda\mathsf{X}(x)$ con le condizioni al bordo $\mathsf{X}(0) = \mathsf{X}(1) = 0$ e sappiamo che deve essere $\mathsf{X}(x) = \sin(\pi\ell x)$ e $\lambda = -(\pi\ell)^2$ per qualche $\ell\in\mathbb{N}$. L'espressione di $\mathsf{T}(t)$ segue immediatamente da $\mathsf{T}'(t) = \lambda\mathsf{T}(t) = -(\pi\ell)^2\mathsf{T}(t)$, cioè $\mathsf{T}(t) = \mathrm{e}^{-\pi^2\ell^2 t}$.

Abbiamo mostrato che se $u^0(x) = \sin(\pi\ell x)$ per $\ell\in\mathbb{N}$ allora $u(x,t) = \sin(\pi\ell x)\mathrm{e}^{-\pi^2\ell^2 t}$ è soluzione del problema ai valori iniziali e al bordo (77). Per linearità questo ci offre una formula per $u$ per ogni valore iniziale $u^0$ che si può scrivere come combinazione lineare di seni. La teoria delle serie di Fourier ([TW05, §9]) ci garantisce che questo è possibile per ogni $u^0\in L^2(0,1)$. Dato un valore iniziale $u^0\in L^2(0,1)$, il valore della soluzione $u(x,t)$ per $t>0$ si calcola usando i coefficienti di Fourier $\hat{u}_\ell$ di $u^0$ come

$$\hat{u}_\ell := 2\int_0^1 \sin(\pi\ell x)u^0(x)\,\mathrm{d}x, \qquad u^0(x) = \sum_{\ell=1}^{\infty}\hat{u}_\ell\sin(\pi\ell x), \qquad \boxed{u(x,t) = \sum_{\ell=1}^{\infty}\hat{u}_\ell\sin(\pi\ell x)\mathrm{e}^{-\pi^2\ell^2 t},} \tag{78}$$

dove la convergenza è intesa nel senso della norma $L^2$. Poiché non richiediamo convergenza puntuale, anche i valori iniziali $u^0$ diversi da zero agli estremi ($u^0(0)\neq 0$ oppure $u^0(1)\neq 0$) possono essere espansi in serie di seni; in questo caso la convergenza sarà molto lenta.

L'espressione di $u$ in (78) mostra che le componenti di $u$ corrispondenti a diverse frequenze del dato iniziale decadono in $t$ esponenzialmente con diverse velocità: **le componenti che oscillano più rapidamente decadono più velocemente**.

Notiamo un'analogia con la teoria dei sistemi di equazioni differenziali ordinarie lineari di primo grado $\frac{\partial}{\partial t}\vec{\mathbf{Y}} + \underline{\underline{\mathbf{A}}}\vec{\mathbf{Y}} = \vec{\mathbf{0}}$: l'evoluzione in $t$ della soluzione del problema dipende dalla scomposizione del dato iniziale in autofunzioni di $-\frac{\partial^2}{\partial x^2}$ o in autovettori di $\underline{\underline{\mathbf{A}}}$, ciascuna componente dipende da $t$ come un esponenziale $\mathrm{e}^{-\lambda t}$, dove $\lambda$ è il corrispondente autovalore.

**Esercizio ✎ 7.3.** Usare Proposizione 4.52 per dimostrare che ogni funzione $u^0\in L^2(0,1)$ può essere scritta come una serie di seni con i coefficienti $\hat{u}_\ell$ come in (78).

**Esercizio 💻 7.4.** Plottare la soluzione $u$ di (77) per diversi valori iniziali $u^0$, come in Figura 41.

Per riprodurre questi grafici è necessario scrivere la serie di Fourier di $u^0(x) = 1$ e di $u^0(x) = x$.

Per rappresentare funzioni di due variabili si possono usare i comandi `pcolor` (usato in Figura 41), `surf`, `mesh` o `contour`, combinati con `meshgrid`, oppure rappresentare diverse sezioni $t = t_1$, $t = t_2$,..., come in Figura 40.

**Esercizio ✎ 7.5.** Usare il metodo di Fourier per scrivere la soluzione del problema nel dominio illimitato $\mathbb{R}\times\mathbb{R}^+$ considerato nell'esempio precedente con dato iniziale $(2\pi/k)$-periodico $u^0(x) = \sin(kx)$ (oppure $u^0(x) = \cos(kx)$, o $u^0(x) = \mathrm{e}^{\mathrm{i}kx}$) per $k>0$.

Il dato iniziale generico $u^0\in L^2(\mathbb{R})$ non si può scrivere come combinazione lineare *discreta* (cioè una somma) di termini di periodo $2\pi/k$, poiché $k$ può prendere qualsiasi valore reale. È necessario prendere una combinazione lineare *continua* (cioè un integrale): $u^0(x) = \int_{\mathbb{R}}\hat{u}(k)\mathrm{e}^{\mathrm{i}kx}\,\mathrm{d}k$; la funzione di variabile reale $\hat{u}$ è detta "trasformata di Fourier" di $u$.

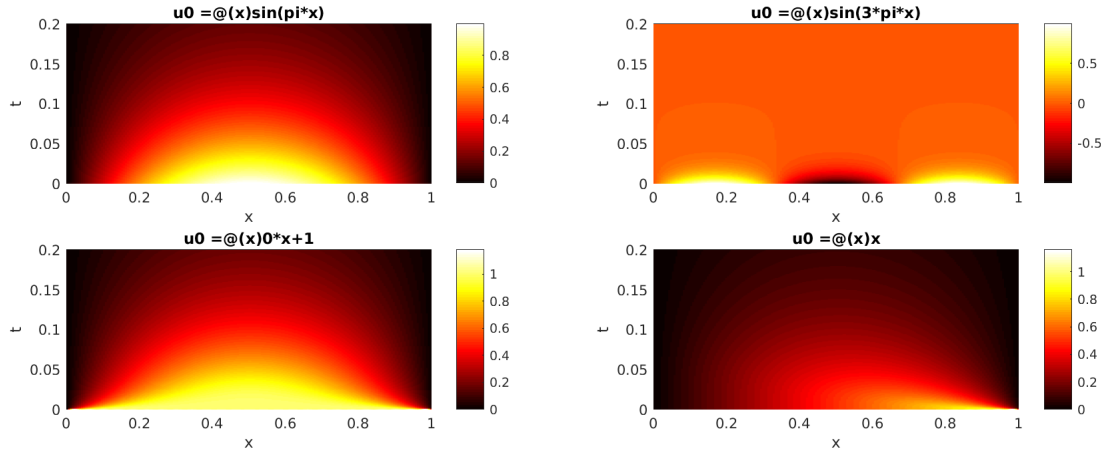



*Figura 41: La soluzione $u$ del problema (77) per $T = 0.2$ e diverse scelte delle condizioni iniziali $u^0$. Nei primi due esempi la soluzione è calcolata esattamente, nei secondi due la serie di Fourier (78) è troncata.*

**Nota 7.6.** Il metodo di Fourier si può estendere a molte situazioni più generali:

- altri domini spaziali, cioè intervalli diversi da $(0, 1)$;
- altre condizioni al bordo, ad esempio quelle di Neumann $\frac{\partial u}{\partial x}(0,t) = \frac{\partial u}{\partial x}(1,t) = 0$ per $0 < t < T$;
- il caso non omogeneo $\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = f$;
- il caso a coefficienti non-costanti $\frac{\partial u}{\partial t} - \frac{\partial}{\partial x}(K(x)\frac{\partial u}{\partial x}) = 0$, in cui i seni devono essere sostituiti dalle autofunzioni dell'operatore differenziale $\frac{\partial}{\partial x}(K(x)\frac{\partial}{\partial x}\cdot)$;
- il caso in cui sono presenti termini di reazione o trasporto $\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} + p\frac{\partial u}{\partial x} + qu = 0$;
- il caso a $n$-dimensionale $\frac{\partial u}{\partial t} - \Delta u = 0$ su un dominio $\Omega \times (0, T)$, $\Omega \subset \mathbb{R}^n$. In questo caso sono necessarie le autofunzioni del Laplaciano, note esplicitamente solo per $\Omega$ di forma molto semplice (rettangoli, dischi,...).

Al crescere della complessità del problema il metodo di Fourier diventa meno conveniente poiché le autofunzioni dell'operatore spaziale diventano più difficili da calcolare e possono richiedere un'approssimazione numerica.

Definiamo ora il funzionale dell'**energia** $E(t) := \frac{1}{2}\int_0^1 u^2(x,t)\,\mathrm{d}x$. (La parola "energia" non è usata in senso fisico: l'energia termica dell'oggetto è proporzionale all'integrale della temperatura $u$, non del suo quadrato.) Se $u$ è una soluzione di (77) sufficientemente regolare da poter scambiare derivazione e integrazione, abbiamo

$$\begin{aligned}\frac{d}{dt}E(t) &= \frac{1}{2}\frac{d}{dt}\int_0^1 u^2(x,t)\,\mathrm{d}x = \frac{1}{2}\int_0^1 \frac{\partial}{\partial t}u^2(x,t)\,\mathrm{d}x = \int_0^1 u(x,t)\frac{\partial u}{\partial t}(x,t)\,\mathrm{d}x = \int_0^1 u(x,t)\frac{\partial^2 u}{\partial x^2}(x,t)\,\mathrm{d}x \\ &= -\int_0^1 \Big(\frac{\partial u}{\partial x}(x,t)\Big)^2\,\mathrm{d}x + u(1,t)\frac{\partial u}{\partial x}(1,t) - u(0,t)\frac{\partial u}{\partial x}(0,t) = -\int_0^1 \Big(\frac{\partial u}{\partial x}(x,t)\Big)^2\,\mathrm{d}x \le 0, \qquad (79)\end{aligned}$$

cioè $E$ è una funzione non crescente di $t$. In particolare, l'unica soluzione di (77) con $u^0(x) = 0$ (e quindi $E(0) = 0$) è la soluzione costantemente nulla. Questo implica un risultato di **unicità**: il problema (77) ammette al più una soluzione. Per ogni dato iniziale $u^0$ che sappiamo espandere in serie di Fourier, la soluzione scritta in (78) è l'unica possibile. La disuguaglianza $\frac{d}{dt}E(t) \le 0$ dà anche una stima di dipendenza continua dai dati (**stabilità**): $\|u(\cdot,t)\|_{L^2(0,1)} \le \|u^0\|_{L^2(0,1)}$ per ogni $0 < t \le T$.

**Esercizio ✎ 7.7.** Usare il Lemma di Grönwall ($f'(t) \le g(t)f(t) \Rightarrow f(t) \le f(0)\mathrm{e}^{\int_0^t g(s)\,\mathrm{d}s}$) e (79) per dimostrare che l'energia decresce esponenzialmente in tempo: $E(t) \le E(0)\exp(-\frac{2}{C_P^2}t)$, dove $C_P$ è la costante di Poincaré dell'intervallo spaziale.

**Esercizio ✎ 7.8.** Mostrare che se $u$ è una soluzione di (77) e $t \in (0, T]$, allora vale $\int_0^1 (\frac{\partial u}{\partial x}(x,t))^2\,\mathrm{d}x \le \int_0^1 (\frac{\partial u}{\partial x}(x,0))^2\,\mathrm{d}x$, cioè anche la seminorma $|\cdot|_{H^1(0,1)}$ nella variabile spaziale è non crescente in $t$.

**Nota 7.9** (L'equazione del calore all'indietro)**.** La formula (78) ci permette di calcolare il valore di una soluzione dell'equazione del calore a tempo $t = T > 0$ dato il valore al tempo iniziale $t = 0$. Non è possibile fare il contrario, cioè risolvere l'equazione del calore all'indietro. Infatti la convergenza della serie $\sum_{\ell=1}^{\infty} \hat{u}_\ell \sin(\pi\ell x)\mathrm{e}^{-\pi^2\ell^2 T}$

non implica quella della serie $\sum_{\ell=1}^{\infty} \hat{u}_\ell \sin(\pi \ell x)$. Anche quando questa è convergente, ad esempio quando abbiamo un numero finito di termini, la moltiplicazione dei termini trigonometrici per gli esponenziali $\mathrm{e}^{\pi^2 \ell^2 T} \gg 1$ amplifica qualsiasi imprecisione in modo incontrollabile. Ad esempio per $T = 1$ tutti i coefficienti dopo il primo vengono moltiplicati per un valore oltre 30 volte maggiore dell'inverso della precisione macchina ($\mathrm{e}^{4\pi^2} \approx 1.4 \cdot 10^{17} > 30\epsilon_M^{-1}$). Questo è visibile anche nella disuguaglianza dell'energia (79): l'energia a un tempo futuro è controllata da quella a un tempo passato ma non viceversa. Si dice che il problema all'indietro è malposto (*ill-posed*): data $u$ al tempo finale $T$, non sempre esiste una soluzione dell'equazione del calore per $0 < t < T$, e quando esiste non non dipende con continuità dal dato. Qualsiasi metodo numerico usato per questo problema sarà instabile. Questo riflette l'irreversibilità dei fenomeni fisici di diffusione di sostanze e di diffusione termica.

**Nota 7.10** (Equazioni paraboliche in finanza)**.** L'equazione del calore, e più in generale le equazioni paraboliche, sono state sviluppate per modellare fenomeni fisici di diffusione. Un altro uso, oggi estremamente comune, è nella finanza per stabilire i prezzi di alcuni prodotti finanziari detti "derivati" (ad esempio le opzioni). Come la diffusione del calore o di una sostanza chimica dipende dal moto casuale di un grande numero di particelle, così il prezzo di un derivato dipende dalle fluttuazioni casuali dei prezzi dei cosiddetti "sottostanti", come azioni, materie prime, valute... La celebre equazione di Black–Scholes–Merton non è altro che un'equazione parabolica come (10), in cui l'incognita $u$ rappresenta il prezzo del derivato considerato. Nel caso delle opzioni più semplici il prezzo può essere calcolato esplicitamente usando la soluzione fondamentale come in (76) (formula di Black–Scholes). I derivati con una struttura più complicata richiedono un metodo numerico: i tre più comuni sono quelli delle differenze finite (come in §7.3), il metodo Monte Carlo e quello degli alberi binomiali.

## 7.3 Il metodo delle differenze finite in spazio e il $\theta$-metodo in tempo

Il metodo di Fourier è molto utile ma ha alcuni svantaggi, ad esempio: (i) non permette di trattare problemi non lineari, (ii) non è facilmente applicabile al caso di problemi a coefficienti variabili $\frac{\partial u}{\partial t} - \frac{\partial}{\partial x}(A(x)\frac{\partial u}{\partial x}) = 0$, (iii) gli integrali in (78) possono essere difficili da valutare, (iv) la serie di Fourier può richiedere molti termini per raggiungere un errore accettabile. Descriviamo quindi come estendere il metodo delle differenze finite al problema (77).

Per $n \in \mathbb{N}$ introduciamo i soliti nodi equispaziati $x_j = j/(n+1)$, $j = 0, \dots, n+1$. Approssimando la derivata in spazio con la differenza finita centrata otteniamo

$$\begin{aligned} &\frac{\partial U_j}{\partial t}(t) - \frac{U_{j+1}(t) - 2U_j(t) + U_{j-1}(t)}{h^2} = 0, \qquad j = 1, \dots, n \\ &U_0(t) = U_{n+1}(t) = 0, \\ &U_j(0) = u^0(x_j), \end{aligned}$$

dove, per ogni $j = 1, \dots, n$, la funzione $U_j(t)$ rappresenta l'approssimazione di $U(x_j, t)$. Questa è una **semidiscretizzazione**: solo la derivata in spazio è stata approssimata da una differenza finita, mentre la derivata in $t$ è ancora presente. In forma vettoriale:

$$\frac{\partial \vec{\mathbf{U}}}{\partial t} = -\underline{\underline{\mathbf{A}}}\vec{\mathbf{U}}, \qquad \vec{\mathbf{U}}(0) = \vec{\mathbf{U}}^0, \qquad \text{dove } \left(\vec{\mathbf{U}}^0\right)_j := u^0(x_j). \tag{80}$$

Qui $\underline{\underline{\mathbf{A}}}$ è la matrice tridiagonale $n \times n$ delle differenze finite in $x$ introdotta in (24) (con $q_j = 0$). Questo è un sistema $n$-dimensionale di equazioni differenziali ordinarie lineari del primo ordine. Questo sistema può essere approssimato usando qualsiasi metodo numerico per equazioni differenziali ordinarie. La tecnica di (1) semidiscretizzazione di un'equazione alle derivate parziali in spazio e tempo con un sistema di ODEs in tempo e (2) soluzione di questo con un metodo numerico per ODEs è detta **metodo delle linee** (*method of lines*). A questo punto sembrerebbe che l'analisi numerica di una PDE di questo tipo si riduca a quella dei metodi per ODEs: in realtà la discretizzazione in spazio e quella in tempo si influenzano a vicenda, quindi la teoria per PDEs si rivela più complicata di quella per ODEs.

Un metodo usato comunemente è il **theta-metodo** ($\theta$-metodo). Fissiamo un parametro $0 \le \theta \le 1$. Introduciamo dei tempi discreti $t^k = k\Delta t$ per un passo costante $\Delta t > 0$ e $k = 1, \dots, m$ e chiamiamo $\vec{\mathbf{U}}^k \in \mathbb{R}^n$ l'approssimazione numerica di $\vec{\mathbf{U}}(t^k)$ (cioè $U_j^k$ approssima $U(x_j, t^k)$). Approssimiamo (80) con[18]

$$\frac{\vec{\mathbf{U}}^{k+1} - \vec{\mathbf{U}}^k}{\Delta t} = -\underline{\underline{\mathbf{A}}}\big(\theta\vec{\mathbf{U}}^{k+1} + (1-\theta)\vec{\mathbf{U}}^k\big), \qquad k = 0, 1, \dots,$$

[18] Notare che [Iserles09, eq. (1.13)] scambia il ruolo di $\theta$ e $1-\theta$; qui stiamo seguendo la convenzione di [QSSG14, eq. (12.9)].

dove $\vec{\mathbf{U}}^0$ è definito come in (80). Possiamo riscrivere questa espressione in componenti come

$$\frac{U_j^{k+1}-U_j^k}{\Delta t}=\frac{1}{h^2}\Big(\theta(U_{j-1}^{k+1}-2U_j^{k+1}+U_{j+1}^{k+1})+(1-\theta)(U_{j-1}^k-2U_j^k+U_{j+1}^k)\Big),$$

oppure raccogliendo i termini con $\vec{\mathbf{U}}^{k+1}$ e quelli con $\vec{\mathbf{U}}^k$:

$$\boxed{\big(\underline{\underline{\mathbf{I}}}+\theta\Delta t\underline{\underline{\mathbf{A}}}\big)\vec{\mathbf{U}}^{k+1}=\big(\underline{\underline{\mathbf{I}}}-(1-\theta)\Delta t\underline{\underline{\mathbf{A}}}\big)\vec{\mathbf{U}}^k.} \tag{81}$$

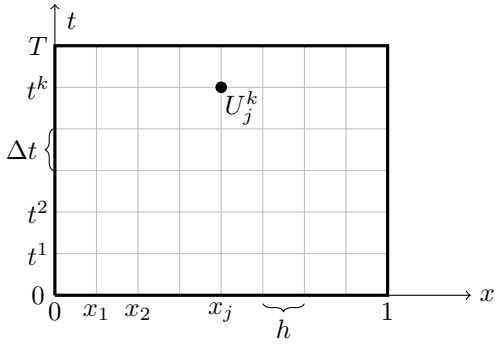



*Figura 42: Il θ-metodo corrisponde alla discretizzazione del dominio $(0,1)\times(0,T)$ con una griglia di nodi equispaziati a cui vengono associate le incognite $U_j^k$. Queste vengono calcolate attraverso un avanzamento in tempo (time-stepping): prima vengono calcolati i valori al primo livello temporale $U_1^1,\dots,U_j^1,\dots,U_n^1$, poi quelli al secondo livello $U_1^2,\dots,U_j^2,\dots,U_n^2$ e così via.*

Misuriamo l'**errore di troncamento** del $\theta$-metodo: se $u$ è sufficientemente liscia, denotando il valore della soluzione esatta nei nodi con $u_j^k=u(x_j,t^k)$,

$$\begin{aligned}
&\frac{u_j^{k+1}-u_j^k}{\Delta t}-\frac{1}{h^2}\Big(\theta(u_{j-1}^{k+1}-2u_j^{k+1}+u_{j+1}^{k+1})+(1-\theta)(u_{j-1}^k-2u_j^k+u_{j+1}^k)\Big)\\
&\quad=\frac{\partial u}{\partial t}(x_j,t^k)+\mathcal{O}(\Delta t)-\Big(\theta\frac{\partial^2 u}{\partial x^2}(x_j,t^{k+1})+\theta\mathcal{O}(h^2)+(1-\theta)\frac{\partial^2 u}{\partial x^2}(x_j,t^k)+(1-\theta)\mathcal{O}(h^2)\Big)\\
&\quad=\frac{\partial u}{\partial t}(x_j,t^k)-\frac{\partial^2 u}{\partial x^2}(x_j,t^k)+\theta\Big(\frac{\partial^2 u}{\partial x^2}(x_j,t^k)-\frac{\partial^2 u}{\partial x^2}(x_j,t^{k+1})\Big)+\mathcal{O}(\Delta t)+\mathcal{O}(h^2)\quad=\mathcal{O}(\Delta t+h^2).
\end{aligned}$$

Il $\theta$-metodo gode di un'accuratezza del secondo ordine in spazio e (almeno) del primo ordine in tempo. In particolare, se $\Delta t\lesssim h^2$, l'errore di troncamento è $\mathcal{O}(h^2)$. Il metodo è detto **consistente** perché l'errore di troncamento converge a zero per $h,\Delta t\to 0$. Vedremo che affinché il metodo converga è necessaria un'opportuna forma di stabilità.

### 7.3.1 Tre casi importanti

I tre esempi di $\theta$-metodo più interessanti corrispondono ai valori $\theta=0,1,1/2$.

Nel caso $\theta=0$ otteniamo il **metodo di Eulero esplicito**:

$$\vec{\mathbf{U}}^{k+1}=\big(\underline{\underline{\mathbf{I}}}-\Delta t\underline{\underline{\mathbf{A}}}\big)\vec{\mathbf{U}}^k,\qquad\qquad U_j^{k+1}=U_j^k+\frac{\Delta t}{h^2}(U_{j-1}^k-2U_j^k+U_{j+1}^k).$$

Questo è un metodo *esplicito*: per ogni tempo $t^{k+1}$ si possono calcolare i valori di $\vec{\mathbf{U}}^{k+1}$ come combinazioni lineari dei valori al tempo $t^k$ e non è necessario assemblare nessuna matrice. Nel metodo di Eulero esplicito il valore di $U_j^{k+1}$ dipende solo dal valore al tempo precedente nei tre nodi $x_{j-1}$, $x_j$ e $x_{j+1}$.

Se invece $0<\theta\le 1$ il metodo è *implicito*: $U_j^{k+1}$ dipende anche dai valori $U_{j-1}^{k+1}$ e $U_{j+1}^{k+1}$ allo stesso istante di tempo. Per calcolare $\vec{\mathbf{U}}^{k+1}$ **è necessario risolvere un sistema lineare $n\times n$ per ogni livello temporale**. Questo è chiaro dall'equazione (81): dato $\vec{\mathbf{U}}^k$ questo è un sistema lineare nell'incognita $\vec{\mathbf{U}}^{k+1}$, la cui matrice si riduce all'identità solo quando $\theta=0$. Poiché il sistema lineare è tridiagonale, la sua risoluzione ha costo $\mathcal{O}(n)$, solo leggermente più costoso del metodo esplicito. Sia $\underline{\underline{\mathbf{A}}}$ che la matrice $\underline{\underline{\mathbf{I}}}+\theta\Delta t\underline{\underline{\mathbf{A}}}$ del sistema lineare sono simmetriche e definite positive, quindi il sistema è non-singolare.

Nell'altro caso estremo $\theta = 1$ otteniamo il **metodo di Eulero implicito**:

$$(\underline{\underline{\mathbf{I}}} + \Delta t \underline{\underline{\mathbf{A}}})\vec{\mathbf{U}}^{k+1} = \vec{\mathbf{U}}^{k}, \qquad \frac{U_j^{k+1} - U_j^k}{\Delta t} = \frac{1}{h^2}(U_{j-1}^{k+1} - 2U_j^{k+1} + U_{j+1}^{k+1})$$

Possiamo quindi pensare al $\theta$-metodo come a una media pesata tra il metodo di Eulero implicito e quello esplicito.

**Esercizio ✎ 7.11.** Mostrare che il metodo di Eulero esplicito (risp., implicito) corrisponde alla discretizzazione dell'equazione del calore con differenze centrate in spazio e in avanti (risp., all'indietro) in tempo.

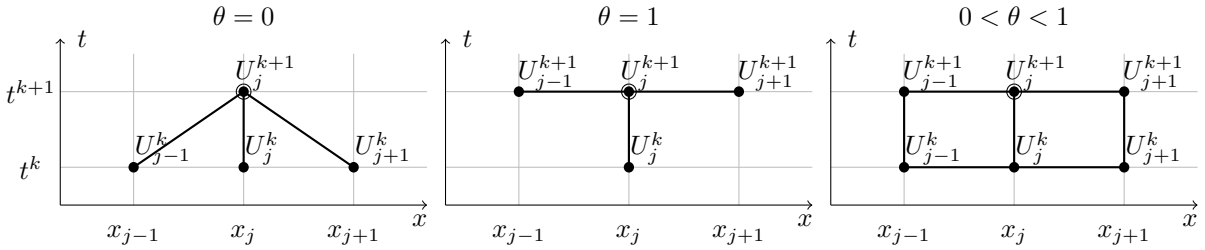



*Figura 43: Lo "stencil" del $\theta$-metodo con diversi valori di $\theta$.*
*A sinistra lo stencil del metodo di Eulero esplicito ($\theta = 0$): il valore di $U_j^{k+1}$ dipende solo dal valore al tempo precedente nei tre nodi $x_{j-1}$, $x_j$ e $x_{j+1}$.*
*Al centro lo stencil del metodo di Eulero implicito ($\theta = 1$): l'avanzamento in tempo mette in relazione $U_j^{k+1}$ con il valore nei nodi adiacenti allo stesso istante $t^{k+1}$ e con il valore nel solo nodo $x_j$ al tempo passato $t^k$.*
*A destra lo stencil del metodo con $0 < \theta < 1$: $U_j^{k+1}$ dipende dal valore nei tre nodi $x_{j-1}$, $x_j$ e $x_{j+1}$ sia al tempo precedente $t^k$ che al tempo presente $t^{k+1}$.*

Il terzo caso importante è il **metodo di Crank–Nicolson** o **metodo del trapezio**, con $\theta = 1/2$:

$$\Big(\underline{\underline{\mathbf{I}}} + \frac{\Delta t}{2}\underline{\underline{\mathbf{A}}}\Big)\vec{\mathbf{U}}^{k+1} = \Big(\underline{\underline{\mathbf{I}}} - \frac{\Delta t}{2}\underline{\underline{\mathbf{A}}}\Big)\vec{\mathbf{U}}^{k}, \quad \text{oppure}$$

$$-rU_{j-1}^{k+1} + (1+2r)U_j^{k+1} - rU_{j+1}^{k+1} = rU_{j-1}^k + (1-2r)U_j^k + rU_{j+1}^k, \qquad r := \frac{\Delta t}{2h^2}.$$

Si può dimostrare che il metodo di Crank–Nicolson ha errore di **troncamento quadratico in tempo**: $\mathcal{O}(h^2 + (\Delta t)^2)$ (mentre per $\theta \neq 1/2$ l'errore di troncamento è solamente $\mathcal{O}(h^2 + \Delta t)$).

**Esercizio 💻 7.12.** Implementare il $\theta$-metodo per:

- il dato iniziale $u^0(x) = 2\min\{x, 1-x\}$,
- il tempo $T = 0.1$,
- $n = 19$ nodi spaziali, cioè $h = 0.05$,
- $m = 10$, $m = 75$ e $m = 80$ intervalli temporali di lunghezza $\Delta t = T/m$,
- $\theta = 0$, $\theta = 1/2$ e $\theta = 1$.

Plottare l'approssimazione di $x \mapsto u(x, T)$ ottenuta. Cosa si osserva?

Usare (78) per mostrare che la soluzione esatta del problema ai valori iniziali è

$$u(x,t) = \frac{8}{\pi^2}\sum_{\ell=1}^{\infty}\frac{\sin(\pi\ell/2)}{\ell^2}\mathrm{e}^{-(\pi\ell)^2 t}\sin(\pi\ell x)$$

e confrontarla con i risultati ottenuti con il $\theta$-metodo.

I risultati dell'Esercizio 7.12 sono visibili nella Figura 44. Osserviamo che succede qualcosa di curioso. Per $\Delta t = 0.1/75 = 0.00133$ il metodo dà risultati accettabili per $\theta = 1$ e $\theta = 1/2$ ma non per $\theta = 0$. In questo caso (metodo esplicito) nella soluzione discreta sono presenti forti oscillazioni. Diminuendo di poco il passo temporale ($\Delta t = 0.1/80 = 0.00125$) il risultato del metodo di Eulero esplicito migliora improvvisamente. Se proviamo altri esempi ci accorgiamo che il parametro chiave è il **numero di Courant**

$$\mu := \frac{\Delta t}{h^2}.$$

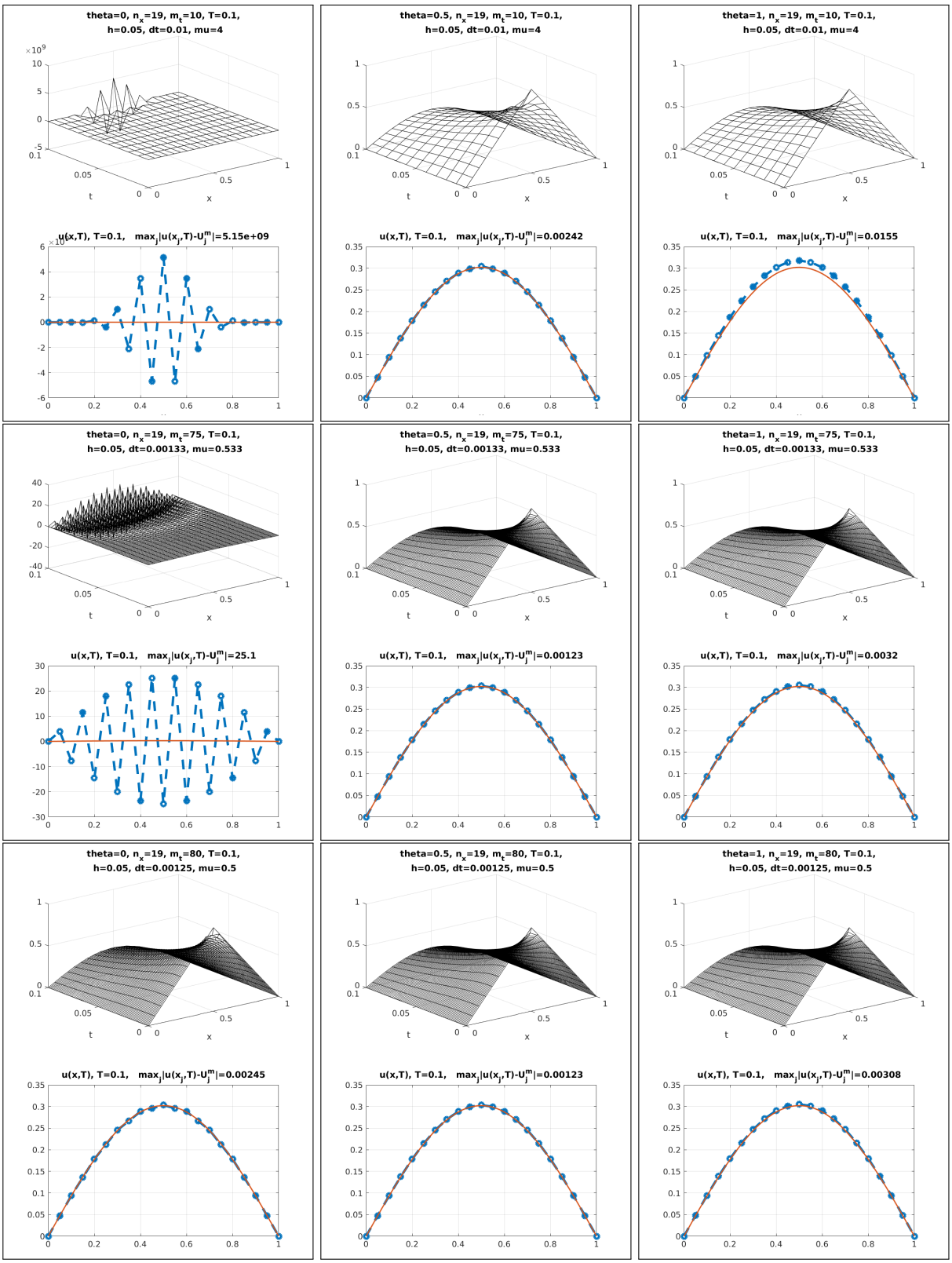



*Figura 44: L'approssimazione ottenuta dal θ-metodo della soluzione del problema ai valori iniziali (77) con $T = 0.1$ e $u^0(x) = 2\min\{x, 1-x\}$ descritto nell'Esercizio 7.12, al variare di $\theta$ ed $m$. Ogni pannello rappresenta in alto la soluzione numerica $U_j^k$ sul dominio spazio-temporale $(0,1) \times (0,T)$ e in basso il valore al tempo $t = T$ della soluzione esatta (linea continua in rosso) e della soluzione discreta (linea blu tratteggiata e markers circolari). Il titolo di ogni pannello mostra i parametri usati ("dt" rappresenta $\Delta t$, $n_x = 1/h$, $m_t = T/\Delta t$) e l'errore massimo commesso nei nodi al tempo finale.*
*Si nota che i metodi di Crank–Nicolson (seconda colonna) e quello di Eulero implicito (terza colonna) sono accurati anche per soli 10 passi temporali (prima riga). Al contrario il metodo di Eulero esplicito (prima colonna) è stabile per almeno 80 passi temporali (terza riga), mentre per un numero minore di passi sono presenti fortissime oscillazioni spurie (notare la scala dei grafici).*

Facendo diminuire $\Delta t$ e $h$ legati da $\Delta t = \mu h^2$ con $\mu$ costante, il metodo di Eulero esplicito converge solo per $\mu \leq 1/2$, mentre quello di Eulero implicito e quello di Crank–Nicolson sembrano convergere per ogni valore di $\mu$. Poiché la consistenza (indipendente da $\mu$) del $\theta$-metodo è garantita dall'errore di troncamento, questo è certamente un problema di mancanza di stabilità: studieremo la stabilità del $\theta$-metodo nella prossima sezione.

**Nota 7.13.** Se pensiamo a (81) come a un'approssimazione dell'ODE $\frac{\partial}{\partial t}\vec{\mathbf{U}} = -\underline{\underline{\mathbf{A}}}\vec{\mathbf{U}}$ possiamo aspettarci che sia stabile se $\lambda\Delta t$ appartiene alla **regione di assoluta stabilità** del $\theta$-metodo per ODEs, dove $\lambda$ varia tra gli autovalori di $-\underline{\underline{\mathbf{A}}}$ (vedere ad esempio [QSSG14, §10.3.3]). Tuttavia $\underline{\underline{\mathbf{A}}}$ dipende dalla discretizzazione spaziale: diminuendo $h$ la sua dimensione aumenta e gli autovalori cambiano.

Quello che abbiamo appena incontrato è un problema di **stiffness**: nel metodo esplicito la condizione sulla lunghezza del passo temporale $\Delta t$ che garantisce la stabilità è più restrittiva di quella necessaria per avere un troncamento piccolo. Questo è dovuto alle diverse scale temporali presenti nel problema. A livello numerico la stiffness può essere misurata come rapporto tra autovalore massimo e minimo delle matrici di avanzamento. A livello del problema continuo $\frac{\partial u}{\partial t} - \Delta u = 0$ abbiamo visto dal metodo di Fourier che le diverse frequenze decadono con scale temporali arbitrariamente diverse, quindi si può pensare che il problema continuo abbia "stiffness infinita". Per una discussione in proposito si veda [LeVeque07, §9.4].

**Esercizio 💻 7.14** (Equazione di Fisher)**.** L'evoluzione di una popolazione (di animali, esseri umani, batteri...) in un ambiente con risorse limitate può essere modellata, nei casi più semplici, dall'equazione di Fisher:

$$\frac{\partial u}{\partial t} - d\frac{\partial^2 u}{\partial x^2} - \alpha u(A - u) = 0.$$

Qui $d$ è un coefficiente di diffusione, $\alpha$ il tasso di crescita, $A$ la "capacità portante" dell'ambiente, e $u(x,t)$ la densità di popolazione. Questo è l'esempio più semplice di equazione di diffusione–reazione di evoluzione non-lineare. Per uno studio accurato e accessibile di questa equazione si veda [TW05, §11].

Fissiamo $d = A = \alpha = 1$ e consideriamo il problema ai valori iniziali su $(0,1) \times (0,T)$ con le condizioni al bordo come in (77).

- Scrivere ed implementare una discretizzazione del problema usando differenze finite in spazio e il metodo di Eulero esplicito in tempo.
- Implementare un metodo di avanzamento in tempo in cui il termine di diffusione $\frac{\partial^2 u}{\partial x^2}$ è discretizzato con il $\theta$-metodo, e il termine non lineare $u(1-u)$ è trattato in modo esplicito in tempo.

Se il dato iniziale è scelto come $u^0(x) = \cos^2(\pi x)$ è possibile confrontare i risultati ottenuti con quelli mostrati in [TW05, Fig. 11.2–11.3].

Nonostante l'equazione differenziale sia non lineare, possiamo approssimarla senza usare metodi di tipo Newton, come fatto nel caso ellittico in §4.8: i problemi di evoluzione sono spesso più semplici da approssimare di quelli stazionari. Questo è perché, tra i dati del problema, abbiamo a disposizione una prima approssimazione della soluzione: il dato iniziale.

### 7.3.2 Stabilità del θ-metodo

Esistono diversi metodi per studiare la stabilità del $\theta$-metodo. Una tecnica classica è quella di Von Neumann, che si basa su un espansione di Fourier (serie o trasformata a seconda che si considerino problemi su domini spaziali limitati §7.2 o illimitati §7.1) del dato iniziale e sull'evoluzione in $t$ delle frequenze discrete e continue; vedere ad esempio [TW05, §4.3] e [LeVeque07, §9.6]. Una seconda tecnica è quella di sfruttare le stime dell'energia come (79) a livello discreto; si veda [TW05, §4.5]. Qui, seguendo [LeVeque07, §9.5], facciamo un terzo tipo di analisi, basata sullo studio degli autovalori della matrice delle differenze finite spaziali $\underline{\underline{\mathbf{A}}}$ che già conosciamo da §4.7.

Il $\theta$-metodo in forma matriciale (81) si può riscrivere come

$$\vec{\mathbf{U}}^{k+1} = \underline{\underline{\mathbf{M}}}\vec{\mathbf{U}}^k, \tag{82}$$

dove $$\underline{\underline{\mathbf{M}}} := (\underline{\underline{\mathbf{I}}} + \theta\Delta t\underline{\underline{\mathbf{A}}})^{-1}(\underline{\underline{\mathbf{I}}} - (1-\theta)\Delta t\underline{\underline{\mathbf{A}}}).$$

Chiaramente avremo $\vec{\mathbf{U}}^k = \underline{\underline{\mathbf{M}}}^k\vec{\mathbf{U}}^0$.

Ora vogliamo formulare e dimostrare il teorema di equivalenza di Lax, che lega convergenza e stabilità, in una forma adatta al $\theta$-metodo per l'equazione del calore. A questo scopo introduciamo una definizione di stabilità adatta.

Assumiamo che i passi temporali e spaziali $\Delta t$ e $h$ vadano a zero simultaneamente e siano legati dalla relazione $\Delta t = \mu h^2$ per $\mu > 0$ costante.

**Definizione 7.15.** Un metodo nella forma (82) si dice **stabile nel senso di Lax–Richtmyer** se, per ogni tempo finale $T$ esiste una costante $C_T$ tale che per ogni $\Delta t > 0$ e per ogni $k \le T/\Delta t$, $k \in \mathbb{N}$, vale

$$\left\|\underline{\underline{\mathbf{M}}}^k\right\|_2 \le C_T.$$

La condizione $\left\|\underline{\underline{\mathbf{M}}}^k\right\|_2 \le C_T$ deve valere per $\Delta t$ piccolo a piacere. Poiché abbiamo assunto $\mu$ costante, al diminuire di $\Delta t$ anche $h$ diminuisce ($h = \sqrt{\Delta t/\mu}$) e la dimensione di $\underline{\underline{\mathbf{M}}}$ ($n = 1/h - 1$) conseguentemente aumenta. La condizione deve valere per tutte queste matrici di diverse dimensioni.

**Proposizione 7.16** (Teorema di equivalenza di Lax)**.** Un metodo nella forma (82) *consistente*, cioè con troncamento che converge a zero, è *convergente*, cioè $\sup_{j,k} |U_j^k - u(x_j, t^k)| \xrightarrow{h\to 0} 0$, se e solo se è *stabile* nel senso di Lax–Richtmyer.

Diamo solo uno sketch dell'implicazione consistenza+stabilità⇒convergenza. Sia $\vec{\mathbf{U}}^0, \dots, \vec{\mathbf{U}}^m \in \mathbb{R}^n$ la sequenza delle approssimazioni discrete ottenute con il $\theta$-metodo (81). Chiamiamo $\vec{\mathbf{u}}^0, \dots, \vec{\mathbf{u}}^m \in \mathbb{R}^n$ i valori della soluzione esatta nei nodi: $(\vec{\mathbf{u}}^k)_j = u_j^k = u(x_j, t^k)$ per $k = 0, \dots, m$ e $j = 1, \dots, n$. Denotiamo l'errore $\vec{\mathbf{e}}^k$ e il troncamento $\vec{\mathbf{T}}^k$ come

$$\vec{\mathbf{e}}^k := \vec{\mathbf{U}}^k - \vec{\mathbf{u}}^k, \qquad \vec{\mathbf{T}}^k := \frac{\vec{\mathbf{u}}^{k+1} - \underline{\underline{\mathbf{M}}}\vec{\mathbf{u}}^k}{\Delta t}.$$

L'errore evolve come

$$\vec{\mathbf{e}}^{k+1} = \vec{\mathbf{U}}^{k+1} - \vec{\mathbf{u}}^{k+1} = \underline{\underline{\mathbf{M}}}\vec{\mathbf{U}}^k - (\underline{\underline{\mathbf{M}}}\vec{\mathbf{u}}^k + \Delta t\vec{\mathbf{T}}^k) = \underline{\underline{\mathbf{M}}}\vec{\mathbf{e}}^k - \Delta t\vec{\mathbf{T}}^k,$$

quindi dopo $m$ passi

$$\vec{\mathbf{e}}^m = \underline{\underline{\mathbf{M}}}^m\vec{\mathbf{e}}^0 - \Delta t\sum_{k=1}^{m}\underline{\underline{\mathbf{M}}}^{m-k}\vec{\mathbf{T}}^{k-1}.$$

Se il metodo è inizializzato con il valore esatto delle condizioni iniziali $\vec{\mathbf{U}}^0 = \vec{\mathbf{u}}^0$, abbiamo $\vec{\mathbf{e}}^0 = \vec{\mathbf{0}}$. Da qui,

$$\|\vec{\mathbf{e}}^m\|_2 \le \Delta t\sum_{k=1}^{m}\left\|\underline{\underline{\mathbf{M}}}^{m-k}\right\|_2\left\|\vec{\mathbf{T}}^{k-1}\right\|_2 \le (m\Delta t)C_T\max_{k=1,\dots,m}\left\|\vec{\mathbf{T}}^{k-1}\right\|_2 \le TC_T\max_{k=1,\dots,m}\left\|\vec{\mathbf{T}}^{k-1}\right\|_2.$$

Questo converge a zero per $h \to 0$ se l'errore di troncamento $\vec{\mathbf{T}}^k$ converge a zero, cioè se il metodo è consistente.

Studiamo ora in quali casi (cioè per quali valori di $\theta$ e $\mu$) il $\theta$-metodo per l'equazione del calore è stabile nel senso di Lax–Richtmyer, quindi convergente. Nel caso del $\theta$-metodo per l'equazione del calore, $\underline{\underline{\mathbf{A}}}$, $\underline{\underline{\mathbf{M}}}$ e $\underline{\underline{\mathbf{M}}}^m$ sono matrici simmetriche; inoltre $\underline{\underline{\mathbf{A}}}$ è definita positiva. La norma $\|\cdot\|_2$ di una matrice simmetrica coincide con il suo raggio spettrale $\rho(\cdot)$, il massimo autovalore in valore assoluto. Per le matrici simmetriche, quindi diagonalizzabili, come $\underline{\underline{\mathbf{M}}}$ abbiamo

$$\left\|\underline{\underline{\mathbf{M}}}^m\right\|_2 = \rho(\underline{\underline{\mathbf{M}}}^m) = \rho(\underline{\underline{\mathbf{M}}})^m,$$

quindi affinché il metodo sia stabile è sufficiente che $\rho(\underline{\underline{\mathbf{M}}}) \le 1$ (non necessario, vedere [LeVeque07, eq. (9.22)]).

In §4.7 abbiamo dimostrato che gli autovalori di $\underline{\underline{\mathbf{A}}}$ sono

$$\lambda_\ell^h = \frac{2}{h^2}\Big(1 - \cos\frac{\pi\ell}{n+1}\Big) = \frac{2}{h^2}(1 - \cos\pi\ell h) \quad \in (0, 4/h^2) \qquad \ell = 1, \dots, n.$$

In particolare, usando $1 - \cos\epsilon = \frac{\epsilon^2}{2} + \mathcal{O}(\epsilon^4)$, si vede che l'autovalore minimo e quello massimo sono

$$\lambda_1^h = \frac{2}{h^2}(1 - \cos\pi h) = \pi^2 + \mathcal{O}(h^2), \qquad \lambda_n^h = \frac{2}{h^2}(1 - \cos\pi nh) = \frac{2}{h^2}(1 + \cos\pi h) = \frac{4}{h^2} - \pi^2 + \mathcal{O}(h^2).$$

**Esercizio ✎ 7.17.** Mostrare che gli autovalori di $\underline{\underline{\mathbf{M}}}$ sono

$$\delta_\ell := \frac{1 - (1-\theta)\Delta t\lambda_\ell^h}{1 + \theta\Delta t\lambda_\ell^h}, \qquad \ell = 1, \dots, n.$$

Suggerimento: usare la decomposizione spettrale $\underline{\underline{\mathbf{A}}} = \underline{\underline{\mathbf{O}}}^\top\underline{\underline{\mathbf{D}}}\underline{\underline{\mathbf{O}}}$, con $\underline{\underline{\mathbf{O}}}$ ortogonale e $\underline{\underline{\mathbf{D}}}$ diagonale.

Vogliamo verificare quando $|\delta_\ell| \leq 1$ per ogni $\ell$, che ci garantisce che il metodo è stabile. Poiché $\lambda_\ell^h > 0$, $\Delta t > 0$ e $0 \leq \theta \leq 1$ abbiamo sicuramente $\delta_\ell < 1$ quindi $\rho(\underline{\underline{\mathbf{M}}}) \leq 1$ se e solo se $\delta_\ell \geq -1$ per ogni $\ell$. Verifichiamo questa condizione nei diversi casi

- Nel caso del metodo di Eulero implicito $\theta = 1$, $\underline{\underline{\mathbf{M}}} = (\underline{\underline{\mathbf{I}}} + \Delta t \underline{\underline{\mathbf{A}}})^{-1}$, $\delta_\ell = (1 + \Delta t \lambda_\ell^h)^{-1}$ appartiene all'intervallo $(0, 1)$, quindi $\rho(\underline{\underline{\mathbf{M}}}) < 1$. **Il metodo di Eulero implicito è stabile per ogni** $\mu$.
- Nel caso del metodo di Eulero esplicito $\theta = 0$, $\underline{\underline{\mathbf{M}}} = \underline{\underline{\mathbf{I}}} - \Delta t \underline{\underline{\mathbf{A}}}$ e $\delta_\ell = 1 - \Delta t \lambda_\ell^h$. L'autovalore minimo è $\delta_n = 1 - \Delta t \lambda_n^h = 1 - 4\mu + \mathcal{O}(\Delta t)$ che sta in $[-1, 1)$ se $\mu \leq 1/2$. **Il metodo esplicito è stabile solo se** $\mu \leq 1/2$.
- Nel caso $\theta = 1/2$, $\delta_\ell = (2 - \Delta t \lambda_\ell^h)/(2 + \Delta t \lambda_\ell^h) > -1$, quindi **il metodo di Crank–Nicolson è stabile per ogni** $\mu$.
- Nel caso più generale si verifica facilmente che $\rho(\underline{\underline{\mathbf{M}}}) \leq 1$ se $\lambda_\ell \Delta t (1 - 2\theta) \leq 2$. Quindi il $\theta$-metodo con $1/2 \leq \theta \leq 1$ è stabile per ogni $\mu$, mentre con $0 \leq \theta < 1/2$ è stabile per $\mu \leq \frac{1}{2(1-2\theta)}$.

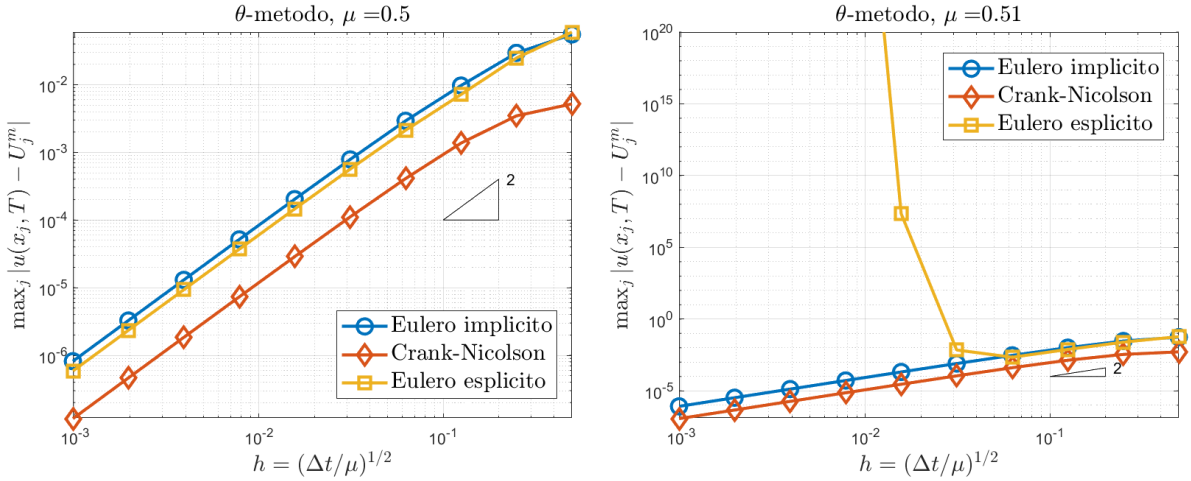



*Figura 45: L'errore (misurato nei nodi al tempo finale $T = 0.1$) del $\theta$-metodo per l'esempio dell'Esercizio 7.12 al variare di $h = 1/(n+1)$, $n = 2^1, \ldots, 2^{10}$. Qui $\Delta t$ è scelto in modo tale che il numero di Courant $\mu = \Delta t / h^2$ sia costante.*
*Per $\mu = 0.5$ i metodi di Eulero implicito ed esplicito e quello di Crank–Nicolson convergono quadraticamente in $h$ (sinistra). Per $\mu = 0.51$ il metodo di Eulero esplicito è instabile e l'errore diverge molto rapidamente (destra).*

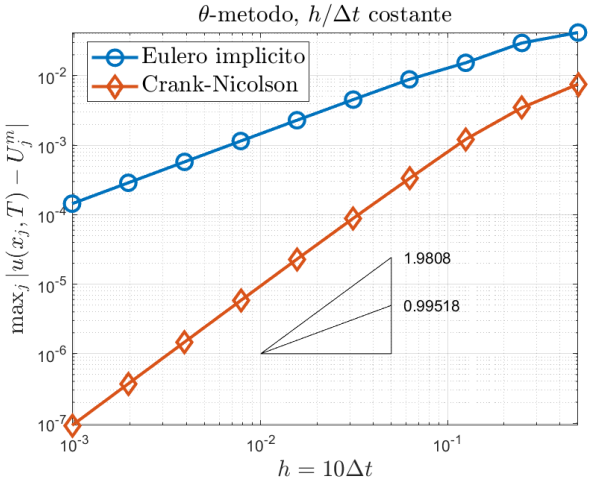



*Figura 46: L'errore dei metodi di Eulero implicito e di Crank–Nicolson per l'esempio dell'Esercizio 7.12 al variare di $h = 1/(n+1)$, $n = 2^1, \ldots, 2^{10}$, fissando $h = 10\Delta t$. I due metodi convergono con velocità lineare e quadratica in $h$, rispettivamente, a causa del diverso errore di troncamento.*

Il metodo esplicito richiede passi temporali $\Delta t \leq \frac{1}{2}h^2$, troppo piccoli per essere utilizzati in casi concreti. Il metodo di Eulero implicito non ha vincoli di stabilità così restrittivi ma richiede comunque $\Delta t = \mathcal{O}(h^2)$ per garantire l'accuratezza quadratica in $h$, per via dell'errore di troncamento. Il metodo di Crank–Nicolson è il più conveniente: non ha vincoli di stabilità e $\Delta t = \mathcal{O}(h)$ è sufficiente per avere accuratezza quadratica in $h$.

Questi fatti sono visibili nei grafici di convergenza dell'errore nelle Figure 45–46. La prima mostra che per $\Delta t = 0.5h^2$ i tre metodi si comportano in modo simile, mentre per $\Delta t = 0.51h^2$ il metodo esplicito diverge per la mancanza di stabilità. La seconda figura mostra che per $\Delta t = 0.1h$ il metodo di Crank–Nicolson mantiene la convergenza quadratica, mentre quello di Eulero implicito si riduce a convergenza lineare: infatti gli errori di troncamento sono rispettivamente $\mathcal{O}(h^2+(\Delta t)^2)$ e $\mathcal{O}(h^2+\Delta t)$. La conseguenza è che il metodo di Crank–Nicolson raggiunge un errore di circa $10^{-7}$ per $n = 1\,024$ nodi spaziali e $m = 1\,025$ passi temporali (Figura 46), mentre per ottenere un errore analogo con il metodo di Eulero implicito (con $h^2 = 0.5\Delta t$) richiede $n = 1\,024$ nodi spaziali e ben $m = 210\,125$ passi temporali (Figura 45).

**Nota 7.18.** Interpretiamo il problema (77) come il modello della diffusione di una sostanza con concentrazione iniziale $u^0(x) \geq 0$ di massa unitaria $\int_0^1 u^0(x)\,\mathrm{d}x = 1$. Vogliamo approssimare l'evoluzione di $u$ con un semplice modello "particellare" discreto. Fissiamo la solita griglia di nodi equispaziati in spazio $x_j = hj$, $j = 0, \dots, n+1$, e in tempo $t^k = k\Delta t$, $k = 0, \dots, m$. Assumiamo che ad ogni istante $t^k$ una "particella" della sostanza considerata si trovi in uno dei nodi $x_j$. Sia $P_j^0 = u^0(x_j)/\sum_{j'=1}^n u^0(x_{j'})$ la probabilità che la particella si trovi nel nodo $j$esimo al tempo iniziale. Fissiamo $0 < p < 1/2$ e assumiamo che se la particella si trova nel nodo $x_j$ al tempo $t^k$ allora ha probabilità $p$ di spostarsi a sinistra in $x_{j-1}$ all'istante successivo $t^{k+1}$, probabilità $p$ di andare nel nodo più a destra $x_{j+1}$, e probabilità $1-2p$ di rimanere in $x_j$. Quindi, denotando $P_j^k$ la probabilità di essere in $x_j$ al tempo $t^k$, vale

$$P_j^{k+1} = pP_{j-1}^k + (1-2p)P_j^k + pP_{j+1}^k, \qquad P_j^0 = \frac{u^0(x_j)}{\sum_{j'=1}^n u^0(x_{j'})}.$$

Più i nodi saranno vicini in spazio ($h \to 0$) più la probabilità $p$ di spostarsi sarà alta, più i nodi in tempo saranno vicini ($\Delta t \to 0$) più $p$ sarà bassa: scegliamo ad esempio $p = K\Delta t/h^2$ per qualche $K > 0$ (questa scelta delle potenze di $h$ e $\Delta t$ può essere giustificata). Sostituendo nell'equazione sopra otteniamo $\frac{1}{\Delta t}(P_j^{k+1} - P_j^k) = \frac{K}{h^2}(P_{j-1}^k - 2P_j^k + P_{j+1}^k)$. Questa non è altro che la discretizzazione già vista dell'equazione del calore $\frac{\partial P}{\partial t} = K\frac{\partial^2 P}{\partial x^2}$ con differenze finite in spazio e il metodo di Eulero esplicito in tempo. La distribuzione della sostanza al tempo finale sarà descritta da $P_j^m$ al variare di $j$. Secondo questo approccio l'equazione del calore appare come limite di un modello discreto, al contrario della derivazione (vista sopra) del $\theta$-metodo dal modello continuo. Il legame tra equazione del calore e moto casuale di particelle è molto profondo e compare addirittura nel famosissimo articolo di Einstein del 1905 sul moto Browniano.

**Nota 7.19** (Elementi finiti per l'equazione del calore)**.** In questa sezione abbiamo discretizzato i problemi ai valori iniziali per l'equazione del calore con differenze finite in spazio e il $\theta$-metodo in tempo. Un'alternativa consiste nel semidiscretizzare in spazio usando uno spazio di elementi finiti $V_h = \operatorname{span}\{\varphi_1, \dots, \varphi_n\} \subset H_0^1(0,1)$. In questo modo si ottiene un sistema di equazioni differenziali ordinarie nella variabile vettoriale $\vec{\mathbf{U}}(t) : [0,T] \to \mathbb{R}^n$, dove $u_h(x,t) = \sum_{j=1}^n U_j(t)\varphi_j(x)$ approssima la soluzione $u(x,t)$. Questo sistema di equazioni differenziali può essere discretizzato a sua volta con diversi metodi di avanzamento in tempo, ad esempio, anche in questo caso, il $\theta$-metodo. In questo caso, nell'espressione del $\theta$-metodo (81) la matrice $\underline{\underline{\mathbf{A}}}$ viene sostituita dalla "matrice di stiffness" degli elementi finiti con $A_{j,k} = \int_0^1 \varphi_k'\varphi_j'\,\mathrm{d}x$ (ad esempio (70) con $q = 0$), e la matrice identità $\underline{\underline{\mathbf{I}}}$ dalla "matrice di massa" $\underline{\underline{\mathbf{M}}}$ con $M_{j,k} = \int_0^1 \varphi_k\varphi_j\,\mathrm{d}x$. Questa tecnica è descritta in [QSSG14, §12.3].

Un altro modo, meno comune, di usare gli elementi finiti per l'equazione del calore consiste nell'utilizzare il metodo di Galerkin sia nella variabile spaziale che in quella temporale, scrivendo un'opportuna forma variazionale del problema (77). Per questa tecnica si veda [QSSG14, §12.4].

# Indice analitico


Autofunzione, 39

Boundary layer, 35
Buona posizione
  del metodo delle differenze finite (Dirichlet), 21
  del metodo delle differenze finite (Neumann), 26
  del metodo di Galerkin, 73
  del problema al bordo, 13
  del problema in forma debole, 70
  di un problema variazionale astratto, 68

Coercività, 68
Comando Matlab
  `case`, 30
  `condest`, 28
  `cond`, 28
  `eigs`, 39
  `eig`, 39
  `fft`, 61
  `full`, 28
  `ifft`, 61
  `norm`, 22
  `ode45`, 3
  `spalloc`, 28
  `sparse`, 28
  `spdiags`, 28
  `switch`, 30
  `tic`, `toc`, 62
  `polyfit`, 18
  `struct`, 30
Complemento di Schur, 82
Condizioni al bordo
  di Dirichlet, 10
  di Neumann, 15
  di Robin, 16
  miste, 27
  periodiche, 16
Consistenza, 22
Continuità
  di un funzionale lineare, 68
  di una forma bilineare, 68
Curse of dimensionality, 53

Delta di Dirac, 72
Densità, 7
Derivata debole, 66
DFT, 61
Differenze finite
  all'indietro, 16
  centrate, 16
  in avanti, 16
  per la derivata seconda, 17
Diffusione artificiale, 37
Dimensioni fisiche, 11
Disuguaglianza
  di Cauchy–Schwarz in $L^2(a,b)$, 66
  di Poincaré, 69

Ellitticità, 68
Equazione
  del calore, 9, 85
  del pendolo, 6
  di continuità
    in forma differenziale, 8
    in forma integrale, 8
  di diffusione–trasporto–reazione
    non-stazionaria, 9
    stazionaria, 9
  di Laplace, 9
  di Poisson, 9

FFT, 63
Flusso, 8
Formulazione debole di un p. al bordo, 67
Funzionale dell'energia (per il calore), 88
Funzione
  a bolla, 81
  a tenda, 77
  assolutamente continua, 66
  di Green, 13
    discreta, 25
  nodale, 77
  test, 67

IFT, 61

Legge
  dei grandi numeri, 50
  di Fick, 8
  di Fourier, 9
Lemma di Céa, 74

Matrice
  a predominanza diagonale, 26
  a predominanza diagonale stretta, 26
  circolante, 64
  di Fourier $\underline{\underline{\mathbf{W}}}$, 60
  monotona, 21
  sparsa, 28
  tridiagonale, 28, 30
Metodo
  $\theta$-metodo, 89
  delle differenze finite, 19
  aggiunto, 32
  degli elementi finiti, 76
  dell'energia, 11
  della quadratura Gaussiana, 52
  di collocazione, 55
  di collocazione spettrale, 55
  di Crank–Nicolson, 91
  di Eulero esplicito, 90
  di Eulero implicito, 90

di Fourier, 86
di Galerkin, 73
di Newton, 4, 46
di separazione delle variabili, 86
di shooting, 2
Monte Carlo, 49
pseudospettrale, 55
spettrale, 76
upwind, 37

Nodi di Chebyshev, 56
Norma
$H^k(a,b)$, 66
$L^\infty(a,b)$, 14
matriciale compatibile, 22
matriciale indotta, 22
Numero di Courant, 91
Numero di Péclet
globale, 35
locale, 36

Operatore autoaggiunto, 41
Ortogonalità di Galerkin, 74
Oscillazioni spurie, 36

Perturbazione
regolare, 35
singolare, 36
Polinomi di Legendre
$P_k$, 56
integrati $M_k$, 55
Principio del massimo
continuo, 11
discreto, 21
Principio di Ritz, 67
discreto, 74
in astratto, 68
Problema
ai limiti/al bordo/al contorno, 1
ai valori iniziali, 1
di Sturm–Liouville, 41
variazionale astratto, 68

Quantificazione dell'incertezza, 48
Quantità di interesse, 49
Quasi-ottimalità, 74

Radice $n$-sima dell'unità $\omega_n$, 60
Regola di quadratura
dei rettangoli, 80
dei trapezi, 59, 78
di Cavalieri–Simpson, 78
di Gauss (Gauss–Legendre), 52
Monte Carlo, 51

Semidiscretizzazione, 89
Soluzione fondamentale del calore, 86
Spazio
di polinomi a tratti $S^p(\mathcal{T}_h)$, 76
di polinomi a tratti $S^p_0(\mathcal{T}_h)$, 76
di probabilità, 48
di Sobolev, 66
Stabilità nel senso di Lax–Richtmyer, 94
Stimatore Monte Carlo, 49
Strato limite, 35

Teorema
centrale del limite, 51
dei cerchi di Gershgorin, 26
di equivalenza di Lax, 94
di Lax–Milgram, 68
Trasformata di Fourier
discreta (DFT), 61
veloce (FFT), 63
Troncamento, 22
nel caso non lineare, 47

Variabile aleatoria, 48
Viscosità numerica, 37

Starring:
Anderson, Philip Warren, 1923–2020, 43
Black, Fischer, 1938–1995, 89
Céa, Jean, 1932, 74
Cauchy, Augustin-Louis, 1789–1857, 13, 66
Cavalieri, Bonaventura Francesco, 1598–1647, 78
Chebyshev, Pafnutij L'vovič, 1821–1894, 56
Cooley, James William, 1926–2016, 63
Courant, Richard, 1888–1972, 91
Crank, John, 1916–2006, 91
Dirac, Paul Adrien Maurice, 1902–1984, 72
Dirichlet, Peter Gustav Lejeune, 1805–1859, 10
Einstein, Albert, 1879–1955, 96
Euler, Leonhard, 1707–1783, 90
Fick, Adolf Eugen, 1820–1901, 8
Fisher, Ronald Aylmer, 1890–1962, 93
Fourier, Jean Baptiste Joseph, 1768–1830, 9, 61, 86
Galerkin, Boris Grigoryevich, 1871–1945, 73
Gershgorin, Semyon Aronovich, 1901–1933, 26
Grönwall, Thomas Hakon, 1877-1932, 88
Green, George, 1793–1841, 13
Hermite, Charles, 1822–1901, 52
Hilbert, David, 1862–1943, 66
Jensen, Johan, 1859–1925, 70
Kutta, Martin Wilhelm, 1867–1944, 3
Laguerre, Edmond Nicolas, 1834–1886, 52
Laplace, Pierre-Simon, 1749–1827, 9
Lax, Peter David, 1926, 68, 94
Legendre, Adrien-Marie, 1752–1833, 55
Liouville, Joseph, 1809–1882, 41
Merton, Robert Cox, 1944, 89
Milgram, Arthur Norton, 1912–1961, 68
Neumann, Carl Gottfried, 1832–1925, 15
Newton, Isaac, 1642–1726, 4
Nicolson, Phyllis, 1917-1968, 91
Péclet, Jean Claude Eugène, 1793–1857, 35

Poincaré, Jules Henri, 1854–1912, 69
Poisson, Siméon-Denis, 1781–1840, 9
Richtmyer, Robert Davis, 1910–2003, 94
Ritz, Walther Heinrich Wilhelm, 1878–1909, 67
Robin, Victor Gustave, 1855–1897, 16
Rolle, Michel, 1652–1719, 83
Runge, Carl David Tolmé, 1856–1927, 3
Scholes, Myron Samuel, 1941, 89
Schur, Issai, 1875–1941, 82
Schwarz, Karl Hermann Amandus, 1843–1921, 66
Simpson, Thomas, 1710–1761, 78
Sobolev, Sergej L'vovič, 1908–1989, 66
Sturm, Jacques Charles François, 1803–1855, 41
Taylor, Brook, 1685–1731, 16
Tukey, John Wilder, 1915–2000, 63

# Indice

**1 Problemi ai limiti e metodo di shooting** **1**
1.1 Problemi ai valori iniziali e problemi ai limiti . . . 1
1.2 Metodi di shooting . . . 2
1.2.1 Il metodo di shooting combinato con il metodo di Newton . . . 4
1.2.2 Il problema del pendolo . . . 6

**2 Equazioni di diffusione, trasporto e reazione** **7**
2.1 Modelli di diffusione, trasporto e reazione . . . 7
2.1.1 Leggi di conservazione (equazioni di continuità) . . . 8
2.1.2 Leggi costitutive . . . 8
2.1.3 Equazioni di secondo grado . . . 9
2.1.4 Equazioni stazionarie . . . 9
2.2 Problemi al bordo lineari in una dimensione . . . 10
2.2.1 Il metodo dell'energia . . . 11
2.2.2 Principio del massimo . . . 11
2.2.3 Esistenza e unicità per il problema non omogeneo . . . 13
2.2.4 La funzione di Green . . . 13
2.2.5 Altre condizioni al bordo . . . 15

**3 Differenziazione numerica: le differenze finite** **16**
3.1 Differenze finite ed errore di troncamento . . . 16
3.2 Errore di arrotondamento . . . 17

**4 Il metodo delle differenze finite in una dimensione** **19**
4.1 Il metodo delle differenze finite per il problema di Dirichlet . . . 19
4.2 Invertibilità della matrice delle differenze finite . . . 20
4.3 Analisi dell'errore: troncamento, consistenza, stabilità e convergenza . . . 22
4.4 Discretizzazione del problema di Neumann . . . 25
4.5 Implementazione . . . 28
4.5.1 Risoluzione di sistemi tridiagonali . . . 30
4.5.2 Il caso periodico . . . 31
4.5.3 Altri usi delle differenze finite . . . 32
4.6 Problemi di diffusione–trasporto e metodo upwind . . . 34
4.6.1 Un problema modello di diffusione–trasporto . . . 34
4.6.2 Il metodo delle differenze finite per problemi con termine di trasporto . . . 35
4.6.3 Il metodo upwind . . . 37
4.7 Problemi agli autovalori . . . 39
4.7.1 Problemi di Sturm–Liouville . . . 41
4.8 Differenze finite per problemi non lineari . . . 46
4.9 Quantificazione dell'incertezza (uncertainty quantification) . . . 48
4.9.1 Metodo Monte Carlo . . . 49
4.9.2 Metodo della quadratura Gaussiana . . . 51
4.9.3 Confronto tra metodi Monte Carlo e di quadratura Gaussiana: un esempio concreto 53

**5 Il metodo di collocazione spettrale** **54**
5.1 Il metodo di collocazione in generale . . . 54
5.2 Il metodo di collocazione spettrale polinomiale . . . 55
5.3 Il metodo di collocazione spettrale trigonometrico . . . 57
5.3.1 La trasformata di Fourier discreta . . . 59
5.3.2 La FFT: la trasformata di Fourier veloce . . . 62

**6 Problemi variazionali e metodo di Galerkin** **65**
6.1 Formulazione debole di un problema al contorno . . . 65
6.2 Problemi variazionali astratti . . . 68
6.3 Formulazione variazionale astratta di problemi al bordo . . . 69
6.3.1 Esempi di problemi deboli che non ammettono soluzioni classiche . . . 70
Primo esempio . . . 71
Secondo esempio . . . 71
6.4 Il metodo di Galerkin . . . 72
6.4.1 Proprietà del metodo di Galerkin . . . 73
6.4.2 Il metodo di Galerkin per problemi al bordo . . . 74
6.5 Il metodo spettrale . . . 76
6.6 Il metodo degli elementi finiti . . . 76
6.6.1 Elementi finiti lineari ($p = 1$) . . . 77
6.6.2 Elementi finiti quadratici ($p = 2$) . . . 81
6.6.3 Analisi del metodo degli elementi finiti: approssimazione e convergenza . . . 83

**7 Equazione del calore** **85**
7.1 Problema ai valori iniziali su $\mathbb{R} \times \mathbb{R}^+$ . . . 85
7.2 Il metodo di Fourier per l'equazione del calore . . . 86
7.3 Il metodo delle differenze finite in spazio e il $\theta$-metodo in tempo . . . 89
7.3.1 Tre casi importanti . . . 90
7.3.2 Stabilità del $\theta$-metodo . . . 93

**Indice analitico** **97**

**Indice** **99**

**Bibliografia** **100**

# Bibliografia

[Iserles09] A. Iserles, *A First Course in the Numerical Analysis of Differential Equations*, Cambridge University Press, 2a ed., 2009.

[LeVeque07] R.J. LeVeque, *Finite Difference Methods for Ordinary and Partial Differential Equations. Steady-state and Time-dependent Problems*, SIAM 2007.

[QSSG14] A. Quarteroni, R. Sacco, F. Saleri, P. Gervasio, *Matematica Numerica*, Springer, 2014.

[SF08] G. Strang, G. Fix, *An Analysis of the Finite Element Method*, Wellesey–Cambridge press, 2008 (prima edizione del 1973).

[Süli06] E. Süli, *An Introduction to the Numerical Analysis of Partial Differential Equations*, 2005, dispense disponibili su `http://people.maths.ox.ac.uk/suli/nspde.ps`.

[SM03] E. Süli, D. Mayers, *An Introduction to Numerical Analysis*, Cambridge University Press, 2003.

[TBD18] L.N. Trefethen, Á. Birkisson, T.A. Driscoll, *Exploring ODEs*, SIAM, 2018.
Pdf e files Matlab su `http://people.maths.ox.ac.uk/trefethen/ExplODE/`

[TW05] A. Tveito, R. Winther, *Introduction to PDEs. A Computational Approach*, Springer 2005.