# 1 Tracce di analisi numerica

aggiornamento: 14 aprile 2014

(!) indica un argomento fondamentale, (F) un argomento facoltativo, (∗) un argomento o dimostrazione impegnativi, (NR) una dimostrazione non richiesta; per approfondimenti di analisi numerica, si vedano ad es. V. Comincioli: Analisi Numerica - McGraw-Hill cartaceo e Apogeo e-book, G. Rodriguez: Algoritmi Numerici - Pitagora, e le dispense online di A. Sommariva in http://www.math.unipd.it/∼alvise/didattica con le referenze citate; per approfondimenti di analisi funzionale si veda ad es. A.N. Kolmogorov, S.V. Fomin: Elementi di teoria delle funzioni e di analisi funzionale - Mir; per le basi sulle differenze finite per equazioni differenziali si veda ad es. A. Quarteroni, F. Saleri: Introduzione al calcolo scientifico - Springer.

## 1.1 Teoria dell'approssimazione

### 1.1.1 Densità e migliore approssimazione

- (!) sia $(X, \|\cdot\|)$ uno spazio funzionale normato, ad esempio $(C[a,b], \|\cdot\|_\infty)$ oppure $(L^2(-\pi,\pi), \|\cdot\|_2)$, e $S_0 \subset S_1 \subset \ldots \subset S_n \subset \ldots$ una successione crescente di sottospazi di dimensione finita $N_n = dim(S_n)$, ad esempio i polinomi $\mathbb{P}_n = \langle 1, x, x^2, \ldots, x^n\rangle$ con $N_n = n+1$, oppure i polinomi trigonometrici reali $\mathbb{T}_n = \mathbb{T}_n^{\mathbb{R}} = \langle 1, \cos(x), \sin(x), \ldots, \cos(nx), \sin(nx)\rangle$ o complessi $\mathbb{T}_n^{\mathbb{C}} = \langle 1, \exp(ix), \exp(-ix), \ldots, \exp(inx), \exp(-inx)\rangle$ con $N_n = 2n+1$ (si verifichi che $\mathbb{T}_n^{\mathbb{C}} \supset \mathbb{T}_n^{\mathbb{R}}$).

  Si dimostri in generale che $E_n(f) = \inf_{p\in S_n} \|p - f\| \to 0$, $n \to \infty$ se e solo se $\bigcup_n S_n$ è denso in $X$.

- (!) *teoremi di densità di Weierstrass*: ogni funzione continua in $[a,b]$ è limite uniforme di una successione di polinomi (NR); ogni funzione continua e periodica in $[-\pi,\pi]$ è limite uniforme di una successione di polinomi trigonometrici (NR). Dedurne la densità di $\mathbb{P} = \bigcup_n \mathbb{P}_n$ in $(C[a,b], \|\cdot\|_2)$ e (F∗) di $\mathbb{T} = \bigcup_n \mathbb{T}_n$ in $(C[-\pi,\pi], \|\cdot\|_2)$.

- l'*inf* del primo esercizio è in realtà un minimo (in generale, dato uno spazio normato e un suo sottospazio di dimensione finita $S = \langle \phi_1, \ldots, \phi_N\rangle$, $dim(S) = N$, esiste in $S$ almeno un *elemento di distanza minima* da $f \in X$); sugg.: $\inf_{p\in S}\{\|p - f\|\} \le \|0 - f\| = \|f\|$ ed essendo in dimensione finita le palle chiuse ... .

- (!) il polinomio in $\mathbb{P}_n$ di migliore approssimazione uniforme per $f \in C[a,b]$, che indichiamo con $p_n^*$, è unico (NR) ma è difficile da calcolare (algoritmo di Remez (NR)); si può però dimostrare (NR) il seguente *teorema di Jackson*: se

$f \in C^r[a,b]$, $r \geq 0$, allora

$$E_n(f) = \|p_n^* - f\|_\infty \leq c_r \, osc(f^{(r)}; (b-a)/n) \, \frac{(b-a)^r}{n(n-1)\dots(n-r+1)} \ , \ n > r$$

con $c_r > 0$ indipendente da $a, b, f$ (dove l'oscillazione di passo $h > 0$ di una funzione $g \in C[a,b]$ è definita come $\max_{|x-y| \leq h} \{|g(x) - g(y)|\}$).

Se $f \in C^{r+\alpha}[a,b]$, $0 < \alpha \leq 1$, cioè $f^{(r)}$ è Hölderiana di esponente $\alpha$, allora $E_n(f) = \mathcal{O}\left(n^{-(r+\alpha)}\right)$ (dove $g$ si dice Hölderiana di esponente $\alpha$ se esiste una costante $L$ tale che $|g(x) - g(y)| \leq L|x-y|^\alpha$, e Lipschitziana se $\alpha = 1$).

Si può inoltre dimostrare (NR) che se $f$ è analitica (olomorfa) in un aperto del piano complesso $\Omega \supset [a,b]$, esiste $\theta \in (0,1)$ tale che $E_n(f) = \|p_n^* - f\|_\infty = \mathcal{O}(\theta^n)$, e che se $f$ è intera ($\Omega = \mathbb{C}$) la convergenza diventa superlineare, $\limsup_{n\to\infty}(E_n(f))^{1/n} = 0$.

- (!) se $X$ è uno spazio euclideo con la norma $\|f\|_X = (f,f)^{1/2}$ indotta da un prodotto scalare $(\cdot,\cdot)$, l'elemento (unico) di distanza minima è la *proiezione ortogonale* $\pi_S = \pi_S f$ di $f$ su $S$ (approssimazione ai minimi quadrati), che si calcola risolvendo il sistema $G\{c_j\} = \{(f, \phi_i)\}$ dove $G = \{g_{ij}\} = \{(\phi_j, \phi_i)\}$ è la matrice di Gram di una qualsiasi base $\{\phi_j\}$; sugg.: considerando per semplicità il caso reale, si tratta di minimizzare nelle variabili $\{c_j\}$ la funzione quadratica $\|f - \sum c_j \phi_j\|^2$, il cui gradiente è ... (si osservi che la matrice di Gram è simmetrica e definita positiva). In generale, $\|f - \phi\|$ è minima in $S$ se e solo se $\|f - \phi + \lambda g\|^2 \geq \|f - \phi\|^2$ per ogni $g \in S$ e $\lambda > 0$, da cui si arriva alla disuguaglianza $\lambda^2\|g\|^2 + 2\lambda Re(f - \phi, g) \geq 0$, ... ; si osservi poi che $(f - \phi, g) = 0$ per ogni $g \in S$ se e solo se $f - \phi$ è ortogonale a tutti gli elementi di base, ... .

- se è nota una base ortogonale $\{\phi_j\}$ di $S$, allora $c_j = (f, \phi_j)/(\phi_j, \phi_j)$: ad esempio, le basi canoniche (si veda il primo esercizio) di $\mathbb{T}_n$ sono ortogonali.

- (F$*$) si esplori il legame tra i risultati precedente e l'approssimazione discreta ai minimi quadrati in uno spazio funzionale di dimensione finita: in generale, dato un insieme di nodi $\{x_k,\ 1 \leq k \leq M\}$, $M > N$, e la matrice rettangolare di tipo Vandermonde $V = \{\phi_j(x_k)\}$, chi è la matrice $\overline{V}^t V$ del sistema delle equazioni normali? inoltre, la soluzione delle equazioni normali con la fattorizzazione $V = QR$ corrisponde al procedimento di ortonormalizzazione di $\{\phi_j\}$ rispetto al prodotto scalare discreto ... .

- se $\{\phi_j\}_{j\geq 1}$ è una successione ortogonale in $X$, la serie $\sum c_j \phi_j$ (serie di Fourier generalizzata) converge a $f$ se e solo se $\bigcup_k \langle \phi_1, \dots, \phi_k \rangle$ è denso in $X$ (si veda il primo esercizio). ($*$) Esempio: le serie trigonometriche, visto che $\mathbb{T}$ è denso in $(C[-\pi,\pi], \|\cdot\|_2)$ (si veda il secondo esercizio) e anche in $(L^2(-\pi,\pi), \|\cdot\|_2)$ (NR).

- (F) dall'uguaglianza $\|\pi_S - f\|^2 + \|\pi_S\|^2 = \|f\|^2$ (teorema di Pitagora) si ottiene che data una successione di sottospazi di dimensione finita $S_0 \subset S_1 \subset \dots \subset$

$S_n \subset \ldots$, si ha che $\lim_{n\to\infty} \|\pi_{S_n} - f\| = 0$ se e solo se $\lim_{n\to\infty} \|\pi_{S_n}\|^2 = \|f\|^2$ (identità di Parseval).

- lo sviluppo di $f \in C[a,b]$ in serie di polinomi ortogonali rispetto ad una misura $d\mu = w(x)dx$ su $(a,b)$, $w \in L^1_+(a,b)$, è convergente in $\|\cdot\|_{2,w}$ (si veda l'inizio della sezione 1.1.3: i polinomi si ottengono ortogonalizzando con Gram-Schmidt la base canonica, quindi i primi $n+1$ sono una base di $\mathbb{P}_n$, e per il t. di densità di Weierstrass ...); d'altra parte si dimostra facilmente la catena di disuguaglianze $\|f - \pi_{\mathbb{P}_n}\|_{2,w} \le \|f - p_n^*\|_{2,w} \le \sqrt{\|w\|_1}\,\|f - p_n^*\|_\infty$, che permette di stimare la velocità di convergenza della serie col t. di Jackson.

- (F) se i coefficienti dello sviluppo trigonometrico di $f$ (proiezione ortogonale su $\mathbb{T}_n$) vengono approssimati tramite una formula di quadratura su $2n+1$ punti equispaziati $x_k = -\pi + \pi k/n$, $0 \le k \le 2n$, ovvero $(f, \phi_j) = \int_{-\pi}^{\pi} f(x)\exp(-ijx)dx \approx \sum_{k=0}^{2n} w_k f(x_k)\exp(-ijx_k)$, $-n \le j \le n$, questi possono essere ottenuti tramite il famoso algoritmo FFT (Fast Fourier Transform), che calcola la *trasformata discreta di Fourier* di un generico vettore complesso $\{\alpha_0, \ldots, \alpha_{M-1}\}$, ovvero le $M$ somme $\sum_{k=0}^{M-1} \alpha_k \exp(-2\pi ikh/M)$, $0 \le h \le M-1$, con $\mathcal{O}(M \log M)$ invece che $\mathcal{O}(M^2)$ operazioni aritmetiche.

### 1.1.2 Interpolazione

- (!) dato un sottospazio di dimensione finita $S = \langle \phi_1, \ldots, \phi_N \rangle \subset C(K)$ (lo spazio delle funzioni continue su un compatto $K \subset \mathbb{R}^d$, $d \ge 1$), un insieme di punti $\{x_1, \ldots, x_N\} \subset K$ si dice *unisolvente* per il problema di interpolazione se $det(V) \neq 0$, dove $V = V(x_1, \ldots, x_N)$ è la matrice quadrata di tipo *Vandermonde* $V = \{v_{ij}\} = \{\phi_j(x_i)\}$; si controlli che data $f \in C(K)$ la funzione (unica) in $S$ che interpola $f$ su $\{x_1, \ldots, x_N\}$ si può scrivere in *forma di Lagrange*

$$Lf(x) = L_{\{x_i\}} f(x) = \sum_{i=1}^{N} f(x_i)\,\ell_i(x)$$

dove $\langle \ell_1, \ldots, \ell_N \rangle = S$ è la base di funzioni "cardinali"

$$\ell_i(x) = \ell_{x_i}(x) = \frac{det(V(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_N))}{det(V(x_1, \ldots, x_N))}\ , \quad \ell_i(x_k) = \delta_{ik}$$

(che non dipende dalla base di partenza, perché ...).

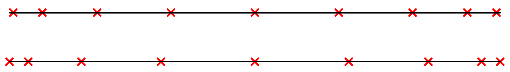
- nel caso polinomiale, $S = \mathbb{P}_n$, in una variabile $\ell_i(x) = \prod_{k\neq i}(x - x_k)/(x_i - x_k)$ e la matrice $V$ è non singolare se e solo se i punti di interpolazione sono distinti; in $d$ variabili, $d \ge 2$, un insieme di $N = dim(\mathbb{P}_n) = \binom{n+d}{d}$ punti distinti non è in generale unisolvente, si faccia un esempio nel piano dove $\mathbb{P}_n = \langle x^h y^k,\ 0 \le h + k \le n \rangle$ e $N = (n+1)(n+2)/2$ (sugg.: se i punti stanno su una retta, una circonferenza o in generale una curva algebrica ...).

- (!) dato un insieme unisolvente di punti di interpolazione, la norma dell'operatore lineare e continuo di interpolazione $L : (C(K), \|\cdot\|_\infty) \to S$, $f \mapsto Lf = \sum_{i=1}^N f(x_i)\,\ell_i$, è definita da $\|L\| = \sup_{f\neq 0} \|Lf\|_\infty/\|f\|_\infty = \sup_{\|f\|_\infty=1} \|Lf\|_\infty$, e si ha
$$\|L\| \leq \max_{x\in K} \sum_{i=1}^N |\ell_i(x)|$$
((∗) in realtà $\|L\| = \max_{x\in K} \sum_{i=1}^N |\ell_i(x)|$; si noti che $\|L\|$ dipende solo dai punti di interpolazione). Si studi il ruolo di $\|L\|$ per la *stabilità dell'interpolazione* (risposta dell'interpolante agli errori su $f$).

- (!) nel caso polinomiale ($S = \mathbb{P}_n$), $\Lambda_n = \|L\|$ si chiama *costante di Lebesgue* dei punti di interpolazione ed è invariante per trasformazioni affini $t = \sigma(x) = Ax + b$ con $A$ invertibile ((F∗) sugg.: si osservi che fissata $g \in C(\sigma(K))$, detta $f(x) = g(\sigma(x))$ per l'unicità dell'interpolante si ha $L_{\{\sigma(x_i)\}} g(t) = L_{\{x_i\}} f(\sigma^{-1}(t))$, ...). In una variabile, $K = [a, b]$, $N = n + 1$, è noto che (NR): i punti equispaziati $x_i = a + i(b - a)/n$, $0 \leq i \leq n$, hanno una costante di Lebesgue che cresce esponenzialmente in $n$, $\Lambda_n \sim 2^n/(en \log n)$; invece ad esempio i *punti di Chebyshev*
$$x_i = \frac{b-a}{2} \cos\left(\frac{(2i+1)\pi}{2n+2}\right) + \frac{b+a}{2}\,, \quad 0 \leq i \leq n$$
(che sono gli zeri del polinomio $T_{n+1}((2x-b-a)/(b-a))$ dove $T_n(t) = \cos(n \arccos(t))$ è il polinomio di Chebyshev di grado $n$ per $t \in [-1, 1]$), oppure i *punti di Chebyshev-Lobatto*
$$x_i = \frac{b-a}{2} \cos\left(\frac{i\pi}{n}\right) + \frac{b+a}{2}\,, \quad 0 \leq i \leq n$$
hanno $\Lambda_n = \mathcal{O}(\log n)$ (questa crescita è quasi ottimale perché si può dimostrare (NR) che la costante di Lebesgue per l'interpolazione su un intervallo ha crescita *almeno* logaritmica nel grado). In figura i 9 punti di Chebyshev (sopra) e Chebyshev-Lobatto (sotto) per grado 8 (si noti che i secondi comprendono gli estremi dell'intervallo mentre i primi sono interni).

- (!) vale la seguente *stima fondamentale per l'errore di interpolazione*
$$\|f - Lf\|_\infty \leq (1 + \|L\|) \min_{p\in S} \|p - f\|_\infty$$
(sugg.: la disuguaglianza in realtà è valida per qualsiasi *operatore di proiezione* (lineare e continuo), cioè tale che $Lp = p$ per ogni $p \in S$, ...). La stima mostra che controllando $\|L\|$ si controlla non solo la stabilità, ma anche la discrepanza fra l'interpolazione e la migliore approssimazione.

  Data una successione di sottospazi $S_0 \subset S_1 \subset \ldots \subset S_n \subset \ldots \subset (C(K), \|\cdot\|_\infty)$ e di operatori di proiezione $L_n : (C(K), \|\cdot\|_\infty) \to S_n$, la stima fondamentale mostra che se $\lim_{n\to\infty} \|L_n\|\, E_n(f) = 0$ allora $L_n f$ converge uniformemente ad $f$ per ogni fissata funzione continua su $K$. Nel caso polinomiale univariato, si



studi la velocità di convergenza dell'interpolazione sui punti di tipo Chebyshev (sugg.: in base al t. di Jackson ...).

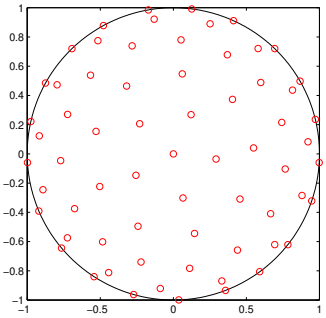
- (F) interpolazione polinomiale bivariata di tipo *prodotto tensoriale*: si consideri lo spazio dei polinomi bivariati prodotto tensoriale di grado $\leq n$, $\mathbb{P}_n^1 \otimes \mathbb{P}_n^1 = \langle x^h y^k,\ 0 \leq h,k \leq n \rangle$ (si osservi che $\mathbb{P}_n \subset \mathbb{P}_n^1 \otimes \mathbb{P}_n^1 \subset \mathbb{P}_{2n}$ e che $dim(\mathbb{P}_n^1 \otimes \mathbb{P}_n^1) = (n+1)^2$). In qualsiasi rettangolo $[a,b]\times[c,d]$, l'insieme prodotto $\{x_1,\ldots,x_{n+1}\}\times\{y_1,\ldots,y_{n+1}\}$, dove $\{x_i\} \subset [a,b]$ e $\{y_j\} \subset [c,d]$ sono insiemi di $n+1$ punti distinti, è unisolvente per l'interpolazione in $\mathbb{P}_n^1 \otimes \mathbb{P}_n^1$ e l'interpolante si scrive

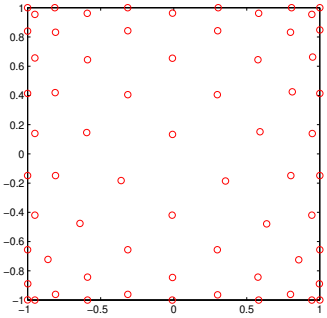
$$Lf(x,y) = \sum_{i,j=1}^{n+1} f(x_i,y_j)\ell_i(x)\ell_j(y)$$

dove $\ell_i$, $\ell_j$ sono i polinomi elementari di Lagrange in $x$ e $y$ (si verifichi direttamente che $Lf(x)$ è interpolante; perché è unico?). Si calcoli $\|L\|$; come converrà scegliere i punti di interpolazione?

- (F) abbiamo visto che per la qualità dell'interpolazione è importante che $\|L\|$ non cresca troppo rapidamente al crescere della dimensione del sottospazio, e questo dipende dalla scelta dei punti di interpolazione. Una buona scelta è data dai *punti di Fekete*, ovvero punti che massimizzano $|det(V)|$ su $K^N$ (perché esistono sempre tali punti?) I punti di Fekete per l'interpolazione polinomiale sono noti teoricamente solo in due casi univariati (intervallo reale e circonferenza complessa), in generale vanno calcolati risolvendo numericamente un difficile problema di ottimizzazione. Si dimostri che utilizzando i punti di Fekete vale la sovrastima $\|L\| \leq N$.

  In figura $N = 66$ punti di Fekete per l'interpolazione polinomiale bivariata di grado $n = 10$ su un cerchio e un quadrato, approssimati con un metodo di ottimizzazione numerica vincolata in $2N = 132$ variabili (si noti che i punti non sono distribuiti uniformemente ma si infittiscono al bordo del dominio; in questo esempio si ha $\|L\| \approx 7.62$ per il cerchio e $\|L\| \approx 5.32$ per il quadrato).





### 1.1.3 Polinomi ortogonali

- (!) polinomi ortogonali in $(a, b)$ (dove eventualmente $a = -\infty$ e/o $b = +\infty$) rispetto ad una misura $d\mu = w(x)dx$ con $w \in L^1_+(a, b)$ (cioè quasi ovunque non negativa e integrabile in $(a, b)$; nel caso di intervalli non limitati si deve anche chiedere $x^j w(x) \in L^1(a, b)$ per ogni $j \geq 0$): si ortogonalizza la base canonica $\{x^j\}_{j\geq 0}$ con il procedimento di Gram-Schmidt rispetto al prodotto scalare $(f, g) = \int_a^b f(x)g(x)w(x)dx$ (definito in generale in $L^2_w(a, b) = \{f : \int_a^b |f(x)|^2 w(x)dx < \infty\}$), ottenendo una base ortogonale $\{\phi_j\}_{j\geq 0}$.

  (!) *osservazione importante*: per come agisce il procedimento di ortogonalizzazione di Gram-Schmidt, si ha che $\langle \phi_0, \ldots, \phi_n \rangle = \langle 1, \ldots, x^n \rangle = \mathbb{P}_n$ e di conseguenza $\phi_n$ *è ortogonale a qualsiasi polinomio di grado* $< n$.

- *relazione di ricorrenza*: si può dimostrare (NR) che ogni famiglia di polinomi ortogonali soddisfa una relazione di ricorrenza a tre termini del tipo

$$\phi_{n+1}(x) = \alpha_n(x - \beta_n)\phi_n(x) + \gamma_n\phi_{n-1}(x) \ , \quad n \geq 0$$

  con opportuni coefficienti $\alpha_n$, $\beta_n$, $\gamma_n$, dove $\phi_{-1} \equiv 0$ e $\phi_0 \equiv 1$.

- (!) *teorema sugli zeri dei polinomi ortogonali*: si consideri una misura $d\mu = w(x)dx$ con funzione peso $w$ non negativa, continua e integrabile in $(a, b)$: gli zeri dei corrispondenti polinomi ortogonali sono tutti reali, semplici e contenuti nell'intervallo aperto $(a, b)$.
  (sugg. $(*)$: se uno zero $\xi$ di $\phi_n$ non stesse in $(a, b)$, allora se reale $\phi_n(x) = (x - \xi)q(x)$, se complesso $\phi_n(x) = (x - \xi)(x - \overline{\xi})q(x)$, in ogni caso il polinomio $\phi_n(x)q(x)$ sarebbe sempre non negativo o non positivo in $(a, b)$ ...). Si osservi che il teorema resta valido se la funzione peso è (quasi ovunque) non negativa e integrabile in $(a, b)$, e positiva su un sottoinsieme di misura positiva.

- alcune famiglie classiche di polinomi ortogonali:

  - Jacobi: $(a, b) = (-1, 1)$, $w(x) = (1+x)^\alpha(1-x)^\beta$, $\alpha, \beta > -1$; casi particolari notevoli i *polinomi di Chebyshev* $T_n(x) = \cos(n \arccos(x))$ per $w(x) = 1/\sqrt{1 - x^2}$ ($\alpha = \beta = -1/2$) e i *polinomi di Legendre* per $w(x) \equiv 1$ ($\alpha = \beta = 0$)

- Laguerre: $(a,b) = (0,+\infty)$, $w(x) = e^{-x}$
- Hermite: $(a,b) = (-\infty,+\infty)$, $w(x) = e^{-x^2}$

- i polinomi di Chebyshev $T_n(t) = \cos(n \arccos(t))$, $t \in [-1,1]$, soddisfano la relazione di ricorrenza $T_{n+1}(t) = 2tT_n(t) - T_{n-1}(t)$ (come si vede direttamente tramite note identità trigonometriche). Inoltre si può dimostrare (NR) che godono dell'importante *proprietà min-max* : $2^{1-n}T_n(t)$ è il polinomio monico (coeff. direttore = 1) di grado $n$ con la minima norma infinito ($= 2^{1-n}$) su $[-1,1]$. Si mostri utilizzando tale proprietà che interpolando sugli zeri di $T_{n+1}((2x-b-a)/(b-a))$ si minimizza la norma infinito del fattore polinomiale nella formula dell'errore $f(x) - L_n f(x) = (x-x_0)\dots(x-x_n)f^{(n+1)}(\xi)/(n+1)!$ per $f \in C^{n+1}[a,b]$, e si ha la stima

$$\|f - L_n f\|_\infty \leq 2\,\frac{\|f^{(n+1)}\|_\infty}{(n+1)!}\,\left(\frac{b-a}{4}\right)^{n+1}$$

### 1.1.4 Quadratura

- una formula di quadratura su $n+1$ punti distinti $x_0,\dots,x_n \in [a,b]$ è una somma pesata di valori della funzione integranda

$$\varphi_n(f) = \sum_{j=0}^{n} w_j f(x_j) \approx \varphi(f) = \int_a^b f(x)w(x)dx\ ,\quad w \in L^1(a,b)$$

  in cui in punti $\{x_j\} = \{x_j^{(n)}\}$ sono detti *nodi di quadratura* e i coefficienti $\{w_j\} = \{w_j^{(n)}\}$ *pesi di quadratura*; si dimostri che $\varphi$, $\varphi_n : (C[a,b], \|\cdot\|_\infty) \to \mathbb{R}$ sono funzionali lineari e continui, e vale

$$\|\varphi\| = \sup_{f\neq 0} \frac{|\varphi(f)|}{\|f\|_\infty} = \|w\|_1 = \int_a^b |w(x)|dx\ ,\quad \|\varphi_n\| = \sum_{j=0}^{n} |w_j|$$

  (sugg.: le disuguaglianze $\leq$ sono immediate; per dimostrare l'uguaglianza, nel secondo caso si prenda $f$ continua tale che $f(x_j) = sgn(x_j)$, ...).

- una formula di quadratura si dice *algebrica* (o anche interpolatoria) se è esatta (cioè il risultato della formula coincide con l'integrale) sui polinomi di $\mathbb{P}_n$: si dimostri che una formula è algebrica se e solo se $w_j = \int_a^b \ell_j(x)w(x)dx$. Il massimo $m_n \geq n$ tale che una formula algebrica è esatta su $\mathbb{P}_{m_n}$ si dice *grado di esattezza* della formula (si può dimostrare che $m_n \leq 2n+1$ (NR)). (F) Si verifichi anche che per una formula algebrica $\|\varphi_n\| \leq \|w\|_1 \Lambda_n$ (dove $\Lambda_n$ è la costante di Lebesgue dei punti di interpolazione/quadratura).

- una formula di quadratura si dice *composta* (di grado $s$) se è ottenuta integrando un'interpolante polinomiale a tratti di grado $s$ sui nodi (dove $x_0 = a$ e $x_n = b$)

per $n$ multiplo di $s$: si dimostri che le formule composte hanno effettivamente la forma di somme pesate descritte sopra (e sono esatte su $\mathbb{P}_s$). Chi sono i pesi della formula composta dei trapezi ($s = 1$) e delle parabole ($s = 2$) nel caso di passo costante $h = (b-a)/n$ e $w(x) \equiv 1$?

- (!) *formule gaussiane*: le formule algebriche corrispondenti agli zeri del polinomio ortogonale $\phi_{n+1}$ di grado $n+1$ rispetto a $d\mu = w(x)dx$ con $w$ positiva, continua e integrabile, si chiamano formule gaussiane; tali formule sono *esatte su* $\mathbb{P}_{2n+1}$ ed hanno *pesi positivi*.
  (sugg. $(*)$: per l'esattezza si osservi che dato $p \in \mathbb{P}_{2n+1}$ di grado maggiore di $n$, si ha $p = \phi_{n+1}q + r$, dove $q$ ed $r$ hanno grado minore di $n+1$, e utilizzando l'ortogonalità e il teorema sugli zeri dei polinomi ortogonali ...; per la positività basta osservare che la formula è esatta su $\ell_j^2(x)$, ...)

- una caratteristica delle formule gaussiane è quella di isolare certe singolarità di una funzione integranda nella funzione peso e quindi nei pesi di quadratura, integrando in tal modo una funzione molto più regolare (si pensi ad esempio alle formule di Gauss-Jacobi nel caso di singolarità agli estremi di tipo potenza frazionaria).

- una formula di quadratura si dice *convergente* su una fissata funzione $f$ se $\lim_{n\to\infty} \varphi_n(f) = \varphi(f)$ e convergente su un dato sottospazio se converge su ogni fissata funzione del sottospazio (nel linguaggio dell'analisi funzionale si tratta della convergenza puntuale della successione di funzionali). Si dimostri che le formule algebriche sono convergenti se la corrispondente interpolazione polinomiale converge uniformemente, e che le formule composte di grado $s$ sono convergenti su $C^{s+1}[a,b]$ purché $\max \Delta x_j \to 0$, $n \to \infty$.

- (F) la formula algebrica per $w(x) \equiv 1$ costruita sui punti di Chebyshev-Lobatto, detta di Clenshaw-Curtis, converge su $C^s[a,b]$, $s > 0$ (ed ha pesi positivi (NR)); allo stesso modo si comporta la formula algebrica per $w(x) \equiv 1$ costruita sui punti di Chebyshev classici, cioè gli zeri di $T_{n+1}((2x-b-a)/(b-a))$, detta formula di Fejér (da non confondersi con la formula gaussiana sugli stessi nodi, detta di Gauss-Chebyshev, che corrisponde a $w(x) = (1-x^2)^{-1/2}$).

- (!) *teorema di Polya-Steklov* (risultato fondamentale sulla quadratura): una formula di quadratura converge su $C[a,b]$ se e solo se

  $(i)$ converge su $\mathbb{P}$

  $(ii)$ $\exists K > 0$ tale che $\sum_{j=0}^{n} |w_j| \le K \ \forall n$

  (sugg. $(*)$: la necessità di $(ii)$ viene dal t. di Banach-Steinhaus (NR): una successioni di funzionali lineari e continui su uno spazio normato completo è limitata in norma se e solo se è limitata puntualmente; per la sufficienza di $(i)$ e $(ii)$, si cominci a far vedere che per il t. di densità di Weierstrass, $\forall \varepsilon > 0$ $\exists p_\varepsilon \in \mathbb{P}$ tale che $|\varphi(f) - \varphi_n(f)| \le (\|\varphi_n)\| + \|\varphi\|)\varepsilon + |\varphi(p_\varepsilon) - \varphi_n(p_\varepsilon)|$, ...).

- (!) *corollari*: una formula algebrica è convergente su $C[a,b]$ se e solo se la somma dei moduli dei pesi è limitata; una formula a *pesi positivi* è convergente su $C[a,b]$ se e solo se è convergente su $\mathbb{P}$ (quest'ultimo garantisce ad esempio che le formule gaussiane, dei trapezi, delle parabole, di Clenshaw-Curtis sono convergenti su $C[a,b]$); le formule di Newton-Cotes, che sono le formule algebriche su nodi equispaziati, non sono convergenti su $C[a,b]$, si può infatti far vedere (NR) che la somma dei moduli dei corrispondenti pesi non è limitata. Pensare ad un esempio di funzione su cui le formule di Newton-Cotes sono convergenti (sugg.: esistono casi di convergenza uniforme dell'interpolazione polinomiale su nodi equispaziati?)

- si controlli che l'ipotesi $(ii)$ del t. di Polya-Steklov garantisce la *stabilità* della formula di quadratura, studiando l'effetto di errori sui valori della funzione integranda. Di conseguenza le formule convergenti su $C[a,b]$ sono anche stabili, mentre le formule di Newton-Cotes sono instabili (per rendere rigorosa quest'ultima affermazione, si consideri la formula perturbata da errori nel campionamento di $f$, $\tilde{f}(x_j) = f(x_j) + \varepsilon_j$ con $\varepsilon_j = \varepsilon\, sgn(w_j)$, ...).

- (F$*$) generalizzazione: una formula di quadratura a pesi positivi convergente sui polinomi converge sulle funzioni continue a tratti in $[a,b]$, e nel caso di $w(x) \equiv 1$ sulle funzioni Riemann-integrabili in $[a,b]$.
  (sugg.: siccome $f$ ha un numero finito di punti di discontinuità, $\forall \varepsilon > 0$ si possono trovare due funzioni continue $f_1 \le f \le f_2$ tali che $\varphi(f_2 - f_1) \le \varepsilon$, allora detto $e_n = \varphi_n - \varphi$ il funzionale errore si ha $e_n(f_1) - \varepsilon \le \varphi_n(f_1) - \varphi(f_2) \le e_n(f) \le$ ... ; per le funzioni Riemann-integrabili si ricorra alle funzioni a gradino).

- (!) *teorema di Stieltjes* (sulla velocità di convergenza delle formule algebriche): per una formula algebrica con grado di esattezza $m_n$ che soddisfi l'ipotesi $(ii)$ del t. di Polya-Steklov vale la seguente stima dell'errore di quadratura

$$|\varphi(f) - \varphi_n(f)| \le (K + \|w\|_1) E_{m_n}(f)$$

  (sugg.: si utilizzi il polinomio $p^*_{m_n}$ di migliore approssimazione uniforme per $f$ in $\mathbb{P}_{m_n}$). Questa stima mostra che l'errore delle formule gaussiane è $\mathcal{O}(E_{2n+1})$.

- si può ottenere una formula di quadratura su qualsiasi intervallo $[a,b]$ con funzione peso $w(x)$, utilizzando una formula di quadratura su $[-1,1]$ con funzione peso $w(\sigma(t))$, $x = \sigma(t) = (b-a)t/2 + (b+a)/2$; detti $\{t_j\}$ e $\{w_j\}$ i nodi e pesi per $w(\sigma(t))$, avremo nodi $x_j = \sigma(t_j)$ e pesi $(b-a)w_j/2$ (sugg.: tramite il cambio di variabile $x = \sigma(t)$ l'integrale diventa ...). Cosa si può dire del grado di esattezza di questa formula?

- i pesi di una formula algebrica (dati i nodi), $\omega = (w_0, \ldots, w_n)^t$, possono essere calcolati risolvendo il sistema lineare

$$V^t \omega = m \ , \quad m = (m_0, \ldots, m_n)^t \ ,$$

dove gli $m_j = \int_a^b \phi_j(x)\, w(x)\, dx$ sono i "momenti" di una base polinomiale $\{\phi_j\}$ e $V^t = \{\phi_i(x_j)\}$ è la trasposta della matrice di Vandermonde in quella base (sugg.: la formula è esatta su ogni $p \in \mathbb{P}_n$ se e solo se è esatta su tutti gli elementi di una base, ...). Si controlli che se la base è ortonormale rispetto a $d\mu = w(x)\, dx$ allora $m_j = 0$, $j > 0$, $m_0 = \sqrt{\|w\|_1}$. (F) Si calcolino i momenti della base di Chebyshev su $[-1, 1]$ per $d\mu = dx$ (da cui si possono ricavare i pesi delle formule di Clenshaw-Curtis e di Fejér visto che la matrice di Vandermonde nella base di Chebyshev risulta molto meglio condizionata della matrice di Vandermonde nella base monomiale canonica).

- (F) cubatura (formule prodotto): qual'è la struttura delle formule di integrazione numerica di una funzione di due variabili ottenute integrando un'interpolante polinomiale tensoriale (si veda l'ultimo esercizio sull'interpolazione) su un rettangolo $[a, b] \times [c, d]$? come converrà scegliere i nodi $\{x_i\} \times \{y_j\}$?

## 1.2 Algebra lineare numerica

- (!) *teorema fondamentale di invertibilità*: data una norma matriciale in $\mathbb{C}^{m\times m}$ tale che $\|AB\| \le \|A\|\,\|B\|$ (come sono tutte le norme indotte da norme vettoriali, ovvero $\|A\| = \sup_{x\neq 0} \|Ax\|/\|x\| = \sup_{\|x\|=1} \|Ax\|$), se $\|A\| < 1$ allora $I - A$ è invertibile e si ha

$$(I-A)^{-1} = \sum_{j=0}^{\infty} A^j \ , \quad \|(I-A)^{-1}\| \le \frac{1}{1-\|A\|}$$

  (sugg.: $\|\sum_{j=m}^{n} A^j\| \le \sum_{j=m}^{n} \|A\|^j$ e la serie geometrica $\sum \|A\|^j$ è convergente e quindi di Cauchy, ... ; nel caso $\|A\|$ sia indotta da una norma vettoriale, la dimostrazione di invertibilità e la stima si possono fare in modo più diretto osservando che $\|(I-A)x\| \ge \|x\| - \|Ax\| > 0$, ..., e detta $S = (I-A)^{-1}$ si ha $1 = \|S(I-A)\| \ge \|S\| - \|AS\| \ge$ ...).

- si verifichi che $\|A\| = \max_{i,j} |a_{ij}|$ è una norma matriciale ma non soddisfa la disuguaglianza $\|AB\| \le \|A\|\,\|B\|$ per ogni coppia di matrici.

- localizzazione rozza degli autovalori: data una norma matriciale come sopra, gli autovalori di $A \in \mathbb{C}^{m\times m}$ stanno in $\mathcal{C}[0, \|A\|]$ (il cerchio complesso chiuso di centro 0 e raggio $\|A\|$).
  (sugg.: se $\lambda \in \mathbb{C}$, $|\lambda| > \|A\|$, scrivendo $(\lambda I - A) = \lambda(I - A/\lambda)$, ...).

- (!) localizzazione fine degli autovalori (*teorema dei cerchi Gershgorin*): gli autovalori di $A \in \mathbb{C}^{m\times m}$ stanno in $\bigcup_{i=1}^{m} \mathcal{C}[a_{ii}, \sum_{j\neq i} |a_{ij}|]$.

  (sugg.: se $\lambda \in \mathbb{C}$ e $\lambda \notin \bigcup_{i=1}^{m} \mathcal{C}[a_{ii}, \sum_{j\neq i} |a_{ij}|]$, scrivendo $A = D + E$, dove $D$ è la matrice diagonale che coincide con la parte diagonale di $A$, si ha $\lambda I - A = (\lambda I - D) - E = (\lambda I - D)(I - (\lambda I - D)^{-1}E)$, dove $\|(\lambda I - D)^{-1}E)\|_\infty < 1$, ...).

- si deduca dal teorema di Gershgorin che una matrice quadrata *diagonalmente dominante in senso stretto*, ovvero tale che $|a_{ii}| > \sum_{j\neq i} |a_{ij}|\ \forall i$, è invertibile.

- applicazione (stime di condizionamento): dato il sistema quadrato $Ax = b$ con $det(A) \neq 0$ e il sistema perturbato $(A+\delta A)(x+\delta x) = b+\delta b$, se $\|k(A)\|\,\|\delta A\| < \|A\|$ (dove $k(A) = \|A\|\,\|A^{-1}\|$ è l'*indice di condizionamento* della matrice in una norma matriciale indotta da una norma vettoriale), vale la stima

$$\frac{\|\delta x\|}{\|x\|} \le \frac{k(A)}{1 - k(A)\|\delta A\|/\|A\|} \left( \frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right)$$

  (sugg.: partendo da $(A+\delta A)\delta x = \delta b - \delta A x$ e osservando che $(A+\delta A) = A(I + A^{-1}\delta A)$ con $\|A^{-1}\delta A\| \le \|A^{-1}\|\,\|\delta A\| < 1$, ...).

- applicazione (cond. suff. per la convergenza delle approssimazioni successive): un sistema quadrato della forma $x = Bx + c$ con $\|B\| < 1$ (in una norma matriciale indotta da una norma vettoriale) ha soluzione unica che si può ottenere

come limite della successione di approssimazioni successive $x_{n+1} = Bx_n + c$ $n \geq 0$, a partire da un qualsiasi vettore iniziale $x_0$.
(sugg.: il sistema ha soluzione unica se e solo se $I - B$ è invertibile, ...).

- (!) *cond. nec. e suff. per la convergenza delle approssimazioni successive*: il metodo delle approssimazioni successive $x_{n+1} = Bx_n + c$, $n > 0$, converge alla soluzione di $x = Bx + c$ per qualsiasi scelta dei vettori $x_0$ e $c$ se e solo se $\rho(B) < 1$ (dove $\rho(B)$ è il *raggio spettrale* della matrice quadrata $B$, ovvero il max dei moduli degli autovalori).
(sugg.: supponendo per semplicità che $B$ sia diagonalizzabile, $B = Q^{-1}\Lambda Q$ dove $\Lambda$ è la matrice diagonale degli autovalori di $A$ ((NR) il caso generale che si può trattare con la forma canonica di Jordan), si avrà $\sum B^j = Q^{-1}(\sum \Lambda^j)Q$, ...).

- (!) dato uno splitting di una matrice quadrata, $A = P - N$, con $det(P) \neq 0$, il sistema $Ax = b$ si può scrivere nella forma $x = Bx + c$ dove $B = P^{-1}N$ e $c = P^{-1}b$. Esempi di corrispondenti metodi delle approssimazioni successive nell'ipotesi $a_{ii} \neq 0 \, \forall i$ sono (posto $A = D - (E + F)$, dove $D$ è la parte diagonale di $A$ ed $-E, -F$ le parti triangolare inferiore e superiore di $A - D$)

  - il metodo di Jacobi: $P = D$, $N = E + F$
  - il metodo di Gauss-Seidel: $P = D - E$, $N = F$

  Si scrivano per componenti tali metodi, e si dimostri che il metodo di Jacobi è convergente per matrici diagonalmente dominanti in senso stretto. Si può dimostrare (NR) che anche il metodo di Gauss-Seidel converge in tale ipotesi nonché per matrici simmetriche definite positive.

- il metodo delle approssimazioni successive si può riscrivere come

$$x_{n+1} = (I - P^{-1}A)x_n + P^{-1}b = x_n + P^{-1}r(x_n)$$

  (dove $r(x_n) = b - Ax_n$ è il vettore *residuo* al passo $n$-esimo). Il ruolo della matrice $P^{-1}$ può essere visto come quello di *precondizionatore*: l'azione di $P^{-1}$ è efficace quando $P^{-1} \approx A^{-1}$, nel senso che gli autovalori di $P^{-1}A$ si accumulano intorno ad 1 (e nel contempo dato un vettore $v$, il calcolo di $z = P^{-1}v$, ovvero la soluzione del sistema $Pz = v$, ha basso costo computazionale). Vari metodi introducono un parametro di rilassamento $\alpha$,

$$x_{n+1} = x_n + \alpha P^{-1}r(x_n)$$

  che aumenti l'efficacia del precondizionatore (cercando di diminuire o addirittura minimizzare il raggio spettrale di $B(\alpha) = I - \alpha P^{-1}A$).

- (!) metodi di discesa: risolvere un sistema lineare $Ax = b$ con $A$ *simmetrica* e *definita positiva* è equivalente a risolvere il problema di minimo globale

$$\min_{x \in \mathbb{R}^m} F(x) \, , \quad F(x) = \frac{1}{2} x^t Ax - x^t b$$

(sugg.: $\nabla x^t Ax = Ax + A^t x$, ...). I metodi di discesa corrispondono a costruire un'iterazione del tipo

$$x_{n+1} = x_n + \alpha_n d_n \ , \quad n \geq 0$$

per diverse scelte della direzione di discesa $d_n$, dove il parametro $\alpha_n$ viene scelto in modo di minimizzare $F(x_{n+1})$. Si mostri che

$$\alpha_n = \frac{d_n^t r(x_n)}{d_n^t A d_n}$$

Il *metodo del gradiente* corrisponde ad usare la direzione di discesa localmente più ripida, $d_n = -\nabla F(x_n) = r(x_n)$. Si può dimostrare (NR) che per il metodo del gradiente vale la stima dell'errore $\|x - x_n\|_2 = \mathcal{O}(\theta^n)$, dove $\theta = (k_2(A) - 1)/(k_2(A) + 1) < 1$; per matrici mal condizionate $\theta \approx 1$, diventa quindi importante utilizzare un buon precondizionatore, che è equivalente (NR) nel caso $P^{-1}$ sia simmetrica e definita positiva ad applicare il metodo al sistema $P^{-1}Ax = P^{-1}b$ con $k_2(P^{-1}A) \ll k_2(A)$ (pur non essendo in generale la matrice $P^{-1}A$ simmetrica e definita positiva).

- test di arresto dello step: l'errore del metodo delle approssimazioni successive con $\rho(B) < 1$, supposta $B$ diagonalizzabile (ovvero $B = Q^{-1}\Lambda Q$ con $\Lambda$ matrice diagonale degli autovalori di $B$) si può stimare come

$$\|x - x_n\| \leq \frac{k(Q)}{1 - \rho(B)} \|x_{n+1} - x_n\|$$

purché la norma matriciale indotta dalla norma vettoriale soddisfi $\|D\| = \max\{|d_{ii}|\}$ per qualsiasi matrice diagonale. Si faccia un esempio in cui ha senso arrestare le iterazioni quando lo step $\|x_{n+1} - x_n\| \leq \varepsilon$, dove $\varepsilon$ è una tolleranza prefissata (sugg.: se $B$ è simmetrica, ...).

- (!) test di arresto del residuo: dato un qualsiasi metodo iterativo *convergente* per la soluzione di un sistema lineare non singolare $Ax = b$ con $b \neq 0$ (approssimazioni successive, gradiente, ...), si mostri che vale la seguente stima dell'errore relativo

$$\frac{\|x - x_n\|}{\|x\|} \leq k(A) \frac{\|r(x_n)\|}{\|b\|}$$

- *teorema di Bauer-Fike* (sul condizionamento del problema degli autovalori): data una matrice complessa diagonalizzabile $A \in \mathbb{C}^{m \times m}$, $A = Q^{-1}\Lambda Q$ con $\Lambda$ matrice diagonale degli autovalori di $A$, e una perturbazione matriciale $E$, detto $\mu$ un autovalore fissato di $A + E$, si ha la stima (NR)

$$\min_{\lambda \in \sigma(A)} |\lambda - \mu| \leq k_2(Q) \|E\|_2$$

dove $\sigma(A)$ è lo spettro di $A$ e $k_2(Q) = \|Q\|_2 \|Q^{-1}\|_2$ (da cui si vede che il problema degli autovalori per una matrice hermitiana è ottimamente condizionato).

(sugg.: se $\mu$ autovalore di $A+E$ non è autovalore di $A$, $A-\mu I$ è non singolare e $A+E-\mu I$ è invece singolare: allora esiste un vettore $z \neq 0$ tale che $Q(A+E-\mu I)Q^{-1}z = (\Lambda - \mu I + QEQ^{-1})z = 0$ e raccogliendo $\Lambda - \mu I$ si ottiene $z = -(\Lambda-\mu I)^{-1}QEQ^{-1}z$, da cui $\|z\|_2 \leq \ldots$).

- (!) *metodo delle potenze* per il calcolo degli autovalori estremali: data una matrice complessa diagonalizzabile $A \in \mathbb{C}^{m\times m}$, con un unico autovalore di modulo massimo (di qualsiasi molteplicità), e la successione di vettori $x_{n+1} = Ax_n$, $n \geq 0$ dove $x_0$ abbia componente non nulla nell'autospazio corrispondente, i *quozienti di Rayleigh* $R(x_n) = (Ax_n, x_n)/(x_n, x_n)$ (dove $(x, y)$ è il prodotto scalare euclideo di $x, y \in \mathbb{C}^m$) convergono a tale autovalore e $x_n/\|x_n\|_2$ tende ad un autovettore (normalizzato) corrispondente; come si può modificare il metodo per calcolare l'autovalore di modulo minimo quando $A$ è non singolare? (sugg.: si scriva $x_0$ nella base di autovettori, ...; per l'autovalore di modulo minimo, si osservi che gli autovalori di $A^{-1}$ sono ...).

  Il metodo modificato $z_{n+1} = Ay_n$, $y_{n+1} = z_{n+1}/\|z_{n+1}\|_2$, $n \geq 0$ a partire da $y_0 = x_0$ evita overflow e underflow in aritmetica di macchina quando l'autovalore di modulo massimo è molto grande o molto piccolo (si mostri che $y_n = x_n/\|x_n\|_2$). Perchè facendo una scelta random di $x_0$ ci si aspetta comunque convergenza in aritmetica di macchina?

  (F) cosa succede se l'autovalore di modulo massimo non è unico? come si può modificare il metodo per calcolare l'autovalore più vicino ad un valore prefissato?

- *metodo QR* per il calcolo dell'intero spettro: se gli autovalori di $A$ sono tutti distinti in modulo, si può dimostrare (NR) che la successione di matrici $\{A_n\}$ definita da

$$A_n = Q_n R_n \ , \quad A_{n+1} = R_n Q_n \ , \quad n \geq 0 \ ; \quad A_0 = A$$

  dove $Q_n$ è ortogonale (unitaria nel caso complesso) ed $R_n$ triangolare superiore, converge ad una matrice triangolare $T$; perchè $T$ ha gli stessi autovalori di $A$? (e quindi la diagonale di $A_n$ converge agli autovalori di $A$). Si osservi che se $A$ è simmetrica tali sono le $A_n$ da cui si vede che $T$ è una matrice diagonale. Il metodo può essere adattato al caso in cui ci siano autovalori con lo stesso modulo (NR).

- dato un polinomio $p(\lambda) = a_0 + a_1\lambda + \ldots + a_m\lambda^m$, $a_m \neq 0$, si vede facilmente per induzione che la matrice $(m+1)\times(m+1)$

$$C(p) = \begin{pmatrix} 0 & 0 & \ldots & 0 & -a_0/a_m \\ 1 & 0 & \ldots & 0 & -a_1/a_m \\ 0 & 1 & \ldots & 0 & -a_2/a_m \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & -a_{m-1}/a_m \end{pmatrix}$$

  detta "*matrice companion*" di $p$, ha polinomio caratteristico $\det(\lambda I - C(p)) = p(\lambda)/a_m$. È quindi possibile calcolare tutti gli zeri di un polinomio applicando il metodo QR (modificato per moduli non distinti) alla matrice companion.

## 1.3 Algebra non lineare numerica

- sia $\phi : (K \subseteq \mathbb{R}^m) \to K$, dove $K$ è un sottoinsieme chiuso (anche non limitato), una mappa tale che $\|\phi(x) - \phi(y)\| \leq \theta\|x - y\|$, $0 < \theta < 1$, per una qualche norma in $\mathbb{R}^m$ (*contrazione*): allora per il *metodo delle approssimazioni successive* (iterazione di punto fisso)

$$x_{n+1} = \phi(x_n) \ , \quad x_0 \in K$$

vale la disuguaglianza fondamentale

$$\|x_m - x_n\| \leq (1 + \theta + \ldots + \theta^{m-n-1}) \, \|x_{n+1} - x_n\| \ , \quad \forall m > n$$

(sugg.: $x_m - x_n = x_m - x_{m-1} + x_{m-1} - x_{m-2} + \ldots + x_{n+2} - x_{n+1} + x_{n+1} - x_n = \phi(x_{m-1}) - \phi(x_{m-2}) + \ldots + \phi(x_{n+1}) - \phi(x_n) + x_{n+1} - x_n$, ...).

- (!) *teorema delle contrazioni*: dato il sistema di punto fisso $x = \phi(x)$ con $\phi$ contrazione di $K$ in $K$, il metodo delle approssimazioni successive converge, per qualsiasi scelta di $x_0 \in C$, all'unico $\xi \in K$ tale che $\xi = \phi(\xi)$.
(sugg.: dalla disuguaglianza fondamentale e da $\|x_{n+1} - x_n\| \leq \theta^n \|x_1 - x_0\|$, si ricava che la successione $\{x_n\}$ è di Cauchy, ...; si osservi che l'enunciato è valido in qualsiasi spazio *normato completo*).

- una condizione sufficiente affinché $\phi$ sia una contrazione in $\|\cdot\|_\infty$ è che sia di classe $C^1(K)$, con $K$ chiusura di un aperto convesso e $\sup_{x \in K} \|J\phi(x)\|_\infty < 1$, dove $J\phi$ è la matrice jacobiana di $\phi$ (si utilizzi per componenti il teorema del valor medio in più variabili).

- un sistema lineare quadrato della forma $x = Bx + c$ con $\|B\| < 1$ (in una norma matriciale indotta da una norma vettoriale) è un caso particolare di sistema di punto fisso con una contrazione $\phi(x) = Bx + c$ definita su $K = \mathbb{R}^m$.

- convergenza locale: se $\phi$ è di classe $C^1$ in un intorno di $\xi$ punto fisso e $\|J\phi(\xi)\|_\infty < 1$, allora il metodo delle approssimazioni successive converge localmente a $\xi$ (sugg.: prendendo come $K$ una opportuna palla chiusa centrata in $\xi$ dove $\|J\phi(x)\|_\infty < 1$, ...).

- un altro risultato di convergenza locale: se $\phi \in C^1(B_\infty[x_0, r])$ (la palla chiusa di centro $x_0$ e raggio $r$ in $\|\cdot\|_\infty$) e $\theta = \max_{x \in B_\infty[x_0,r]} \|J\phi(x)\|_\infty < 1$, allora il metodo delle approssimazioni successive converge quando $\|x_1 - x_0\|_\infty \leq r(1 - \theta)$.
(sugg.: prendendo $K = B_\infty[x_0, r]$, si verifichi che $\phi(K) \subseteq K$, ...).

- (!) valgono le seguenti *stime dell'errore*:

  - *a priori*

$$\|\xi - x_n\| \leq \frac{\theta^n}{1 - \theta} \|x_1 - x_0\|$$

$$\|\xi - x_n\| \leq \theta^n \|\xi - x_0\|$$

– *a posteriori*

$$\|\xi - x_n\| \leq \frac{1}{1-\theta} \|x_{n+1} - x_n\|$$

- velocità di convergenza del metodo delle approssimazioni successive nelle ipotesi del teorema delle contrazioni: la convergenza è comunque *almeno lineare* visto che $\|\xi - x_{n+1}\| \leq \theta \|\xi - x_n\|$; se $\phi$ è $C^2$ in un intorno del punto fisso $\xi$ e $J\phi(\xi) = 0$ la convergenza diventa localmente *almeno quadratica*, ovvero $\|\xi - x_{n+1}\| \leq c\|\xi - x_n\|^2$ con una opportuna costante $c$ per $n$ abbastanza grande.
(sugg.: detta $B_2[\xi, r]$ una palla euclidea centrata in $\xi$ tale che $\phi \in C^2(B_2[\xi, r])$, utilizzando al formula di Taylor centrata in $\xi$ arrestata al secondo ordine si ha $x_{n+1} - \xi = \phi(x_n) - \phi(\xi) = J\phi(\xi)(x_n - \xi) + \varepsilon_n$ con $(\varepsilon_n)_i = \frac{1}{2}(x_n - \xi)^t H\phi_i(z_{n,i})(x_n - \xi)$, dove $H\phi_i$ è la matrice Hessiana della componente $\phi_i$ e $z_{n,i}$ sta nel segmento di estremi $x_n$ e $\xi$, da cui $|(\varepsilon_n)_i| \leq \frac{1}{2} \max_{1 \leq i \leq m} \max_{x \in B_2[\xi,r]} \|H\phi_i(x)\|_2 \|x_n - \xi\|_2^2$ e quindi $\|\varepsilon_n\|_2 \leq ...$).

- (!) stabilità del metodo delle approssimazioni successive: dato il seguente modello di metodo perturbato

$$\tilde{x}_{n+1} = \phi(\tilde{x}_n) + \varepsilon_{n+1} \ , \ \ n \geq 0$$

dove $\phi$ verifica le ipotesi del teorema delle contrazioni, vale la seguente stima per ogni $N > 0$

$$\max_{1 \leq n \leq N} \|\tilde{x}_n - x_n\| \leq \frac{1}{1-\theta} \max_{1 \leq n \leq N} \|\varepsilon_n\|$$

- si studi l'applicabilità del metodo delle approssimazione successive al sistema $x_1 = \arctan(x_1 + x_2)\sin(x_2)/10$, $x_2 = \cos(x_1/4) + \sin(x_2/4)$ e al sistema $2x_1^2 + x_2^2 = 5$, $x_1 + 2x_2 = 3$ (nel secondo caso si consideri la soluzione nel semipiano destro isolandola in un rettangolo opportuno tramite un'interpretazione grafica del sistema).

- (!) dato il sistema non lineare $f(x) = 0$, dove $f : (\Omega \subseteq \mathbb{R}^m) \to \mathbb{R}^m$ è un campo vettoriale differenziabile definito su un aperto $\Omega$ contenente $\xi$ tale che $f(\xi) = 0$, il *metodo di Newton* corrisponde alla *linearizzazione* iterativa

$$f(x_n) + J_n(x - x_n) = 0 \ , \ \ n \geq 0$$

a partire da un opportuno vettore iniziale $x_0$, dove $J_n = Jf(x_n)$ è la matrice Jacobiana (purché $x_n \in \Omega$ e $J_n$ sia invertibile ad ogni iterazione), ovvero

$$x_{n+1} = x_n - J_n^{-1} f(x_n) \ , \ \ n \geq 0$$

- (!) *velocità di convergenza* del metodo di Newton: se $f \in C^2(K)$ dove $K$ è la chiusura di un aperto convesso e limitato contenente $\xi$, in cui la Jacobiana di $f$ è invertibile, e supposto che le iterazioni $x_n$ siano tutte in $K$, posto $e_n = \|\xi - x_n\|_2$ vale la seguente stima (convergenza *almeno quadratica*)

$$e_{n+1} \leq c e_n^2 \ , \ n \geq 0 \ , \ \ c = \frac{\sqrt{m}}{2} \max_{x \in K} \|(Jf(x))^{-1}\|_2 \max_{1 \leq i \leq m} \max_{x \in K} \|Hf_i(x)\|_2$$

dove $Hf_i$ è la matrice Hessiana della componente $f_i$.
(sugg.: dalla formula di Taylor centrata in $x_n$ arrestata al secondo ordine, $0 = f(\xi) = f(x_n) + J_n(\xi - x_n) + \varepsilon_n$, e dalla definizione del metodo, si arriva a $\xi - x_{n+1} = -J_n^{-1}\varepsilon_n$, dove $(\varepsilon_n)_i = \frac{1}{2}(\xi - x_n)^t Hf_i(z_{n,i})(\xi - x_n)$, con $z_{n,i}$ che sta nel segmento di estremi $x_n$ e $\xi$, ...).

- (!) *convergenza locale* del metodo di Newton: se $f \in C^2(K)$ e $Jf(x)$ è invertibile in $K = B_2[\xi, r]$ (dove $\xi$ è soluzione del sistema, $f(\xi) = 0$), detta $c$ la costante dell'esercizio precedente, scelto $x_0$ tale che $e_0 < \min\{1/c, r\}$, il metodo di Newton è convergente e vale la seguente stima dell'errore

$$ce_n \leq (ce_0)^{2^n}, \quad n \geq 0$$

(sugg.: per induzione $e_{n+1} \leq (ce_n)e_n < e_n$ e quindi $x_{n+1} \in B_2[\xi, r]$, ...).

- nelle ipotesi di convergenza locale la stima *a posteriori* dell'errore con lo step $\|x_{n+1} - x_n\|$ è una buona stima (almeno per $n$ abbastanza grande)

$$e_n = \|\xi - x_n\| \approx \|x_{n+1} - x_n\|$$

(sugg.: si osservi che $f$ è localmente invertibile e che $Jf^{-1}(f(x)) = (Jf(x))^{-1}$, quindi $\xi - x_n = f^{-1}(f(\xi)) - f^{-1}(f(x_n)) \approx Jf^{-1}(f(x_n))(f(\xi) - f(x_n)) = ...$).

- il metodo di Newton corrisponde ad un'iterazione di punto fisso con $\phi(x) = x - (Jf(x))^{-1}f(x)$, da cui si deduce che se $f$ è $C^2$ in un intorno di $\xi$ la convergenza è localmente almeno quadratica perché $J\phi(\xi) = 0$ (sugg.: posto $(Jf(x))^{-1} = \{b_{ij}(x)\}$, si ha $\frac{\partial \phi_i}{\partial x_k}(x) = \frac{\partial}{\partial x_k}(x_i - \sum_j b_{ij}(x)f_j(x)) = \delta_{ik} - \sum_j \frac{\partial b_{ij}}{\partial x_k}(x)f_j(x) - \sum_j b_{ij}(x)\frac{\partial f_j}{\partial x_k}(x) = -\sum_j \frac{\partial b_{ij}}{\partial x_k}(x)f_j(x)$, ...).

## 1.4 Differenze finite per ODEs e PDEs

- (!) dato un problema ai valori iniziali $y' = f(t,y)$, $t \in [t_0, t_f]$; $y(t_0) = a$, con $f$ di classe $C^1$ tale che $|\partial f/\partial y| \leq L$, si dimostri che la "legge di propagazione" dell'errore $e_n = |y_n - y(t_n)|$ dei metodi di *Eulero esplicito*

$$y_{n+1} = y_n + hf(t_n, y_n) \ , \ \ 0 \leq n \leq N-1$$

ed *Eulero implicito*

$$y_{n+1} = y_n + hf(t_{n+1}, y_{n+1}) \ , \ \ 0 \leq n \leq N-1$$

su una discretizzazione a passo costante $h = (t_f - t_0)/N$, $t_n = t_0 + nh$, è del tipo

$$e_{n+1} \leq \alpha e_n + \delta_{n+1} \ , \ \ \delta_{n+1} = \delta_{n+1}(h) = \mathcal{O}(h^2)$$

dove $\alpha = \alpha(h) = (1 + hL)$ nel caso puramente Lipschitziano, oppure $\alpha = 1$ nel caso dissipativo ($-L \leq \partial f/\partial y \leq 0$) senza vincoli sul passo per Eulero implicito e con il vincolo $h \leq 2/L$ per Eulero esplicito.
(sugg. $(*)$: per Eulero esplicito si usi la sequenza ausiliaria $u_{n+1} = y(t_n) + hf(t_n, y(t_n))$ e la scrittura $y_{n+1} - y(t_{n+1}) = y_{n+1} - u_{n+1} + u_{n+1} - y(t_{n+1})$, $\delta_{n+1}(h) = |u_{n+1} - y(t_{n+1})|$, ricorrendo poi al teorema del valor medio per il primo addendo e alla formula di Taylor per il secondo ...; analogamente per Eulero implicito con la sequenza ausiliaria $u_{n+1} = y(t_n) + hf(t_{n+1}, y(t_{n+1}))$, ...).

- si deduca dall'esercizio precedente che l'errore globale dei metodi di Eulero è

$$\max_{0 \leq n \leq N} e_n \leq \alpha^N e_0 + \max_{1 \leq n \leq N} \{\delta_n\} \sum_{n=0}^{N-1} \alpha^n \leq c_1 e_0 + c_2 h \ ,$$

stimando le costanti $c_1$ e $c_2$ nei casi Lipschitziano e dissipativo

- (!) il *metodo di Crank-Nicolson* (o trapezoidale)

$$y_{n+1} = y_n + (h/2)[f(t_n, y_n) + f(t_{n+1}, y_{n+1})]$$

ha ordine di approssimazione locale $\delta_{n+1}(h) = \mathcal{O}(h^3)$ per $f \in C^2$ (sugg.: il metodo si ricava applicando alla rappresentazione integrale $y(t_{n+1}) - y(t_n) = \int_{t_n}^{t_{n+1}} y'(t)\, dt$ la formula del trapezio).

- (!) dato il problema test

$$y' = \lambda y + b(t) \ , \ \ t > 0 \ ; \ \ y(0) = y_0$$

dove $g \in C[0, +\infty)$ e $\lambda \in \mathbb{C}$, $Re\lambda \leq 0$ (problema "stiff"), qual'è l'effetto sulla soluzione di un errore $\varepsilon_0 = |y_0 - \tilde{y}_0|$ sul dato iniziale? si verifichi poi che la propagazione di un errore $\varepsilon_0$ sul dato iniziale per una discretizzazione a passo costante $h > 0$ della semiretta è del tipo

$$\varepsilon_n = (\phi(h\lambda))^n \varepsilon_0 \ , \ \ n > 0$$

per i metodi di Eulero (esplicito ed implicito) e per il metodo di Crank-Nicolson. Quale è la regione di stabilità di ciascun metodo, ovvero $\{z \in \mathbb{C} : \phi(z) \leq 1\}$? Quale dei tre metodi è stabile senza vincoli sul passo?

- (!) si estenda l'analisi dell'esercizio precedente al caso del sistema test

$$y' = Ay + b(t) \ , \quad t > 0 \ ; \quad y(0) = y_0$$

dove $y(t), y_0, b(t) \in \mathbb{R}^m$ e $A \in R^{m \times m}$ è una matrice costante diagonalizzabile con autovalori di parte reale non positiva (sistema "stiff") (sugg.: lavorando nella base di autovettori di $A$ ...).

- (!) dato il problema ai valori al contorno

$$u''(x) - cu(x) = f(x) \ , \quad x \in (a, b) \ ; \quad u(a) = u(b) = 0$$

dove $c$ è una costante positiva, si assuma che la soluzione $u$ sia di classe $C^4[a, b]$ e si consideri una discretizzazione dell'intervallo a passo costante $h = (b-a)/(n+1)$, con nodi $x_i = a+ih$, $0 \leq i \leq n+1$. Tramite l'approssimazione della derivata seconda ottenuta con le *differenze centrate*

$$u''(x) \approx \delta_h^2 u(x) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2}$$

si vede che il vettore $\{u(x_i)\}_{1 \leq i \leq n}$ soddisfa un sistema lineare del tipo

$$A\{u(x_i)\} = \{g(x_i)\} + \{\varepsilon_i\}$$

dove $A \in \mathbb{R}^{n \times n}$ è una matrice *tridiagonale* con $-2h^{-2} - c$ sulla diagonale principale e $h^{-2}$ sulle due diagonali adiacenti, e $|\varepsilon_i| \leq Mh^2$ (sugg.: si utilizzi la formula di Taylor di centro $x_i$ in $x_i - h = x_{i-1}$ e $x_i + h = x_{i+1}$). La matrice $A$ risulta *simmetrica* e *definita negativa*, con $\sigma(A) \subset [-(c + 4h^{-2}), -c]$ (si usi il teorema di Gershgorin).

Detto $\{u_i\}$ il vettore soluzione del sistema lineare

$$A\{u_i\} = \{g(x_i)\}$$

si provi che vale la stima

$$\frac{\|\{u_i\} - \{u(x_i)\}\|_2}{\sqrt{n}} = \mathcal{O}(h^2)$$

(sugg.: si ha che $\|A^{-1}\|_2 \leq 1/c$, ...); si osservi che il modo di misurare l'errore è giustificato da $\|\{u(x_i)\}\|_2/\sqrt{n} \approx \|u\|_{L^2(a,b)}/\sqrt{b-a}$. Il sistema può essere agevolmente risolto con il metodo di eliminazione di Gauss, che in questo caso ha complessità $\mathcal{O}(n)$ (perché?) Come va modificato il sistema per condizioni al contorno del tipo $u(a) = \alpha$, $u(b) = \beta$?

- nel caso di $c = 0$ (*equazione di Poisson* unidimensionale) si può ancora far vedere (NR) che $A$ è definita negativa, anzi per ogni $h$ vale $\sigma(A) \subset [-4h^{-2}, -\delta]$ con un opportuno $\delta > 0$ indipendente da $h$.

- (!) dato il problema ai valori al contorno

$$\Delta u(P) - cu(P) = f(P) \ , \quad P = (x, y) \in \Omega = (a, b) \times (c, d) \ ; \quad u|_{\partial\Omega} \equiv 0$$

dove $\Delta = \partial/\partial x^2 + \partial/\partial y^2$ è l'operatore laplaciano e $c$ è una costante positiva, si assuma che la soluzione $u$ sia di classe $C^4(\overline{\Omega})$ e si consideri una discretizzazione del rettangolo con passo $h = (b-a)/n$ nella direzione $x$ e $k = (d-c)/m$ nella direzione $y$, con nodi $P_{ij} = (x_i, y_j)$, $x_i = a + ih$, $0 \le i \le n+1$, $y_j = c + jk$, $0 \le j \le m+1$. Discretizzando l'operatore di Laplace tramite lo schema alle differenze "a croce"

$$\delta^2_{h,k} u(P) = \delta^2_{x,h} u(P) + \delta^2_{y,k} u(P) \approx \Delta u(P)$$

si vede che il vettore $\{u(P_{ij})\}_{1 \le i \le n, 1 \le j \le m}$ soddisfa un sistema lineare del tipo

$$A\{u(P_{ij})\} = \{g(P_{ij})\} + \{\varepsilon_{ij}\}$$

dove $A \in \mathbb{R}^{nm \times nm}$ e $|\varepsilon_{ij}| \le M_1 h^2 + M_2 k^2$.

La struttura della matrice dipende dalla numerazione dei nodi: utilizzando l'ordinamento *lessicografico* delle coppie $(i, j)$, si vede che la matrice risulta *simmetrica*, *tridiagonale a blocchi* con una diagonale di blocchi $n \times n$ che sono matrici tridiagonali con $-c - 2(h^{-2} + k^{-2})$ sulla diagonale principale e $h^{-2}$ sulle due diagonali adiacenti, e due diagonali adiacenti di blocchi $n \times n$ che sono matrici diagonali con $k^{-2}$ sulla diagonale principale. Inoltre $\sigma(A) \subset [-(c + 4(h^{-2} + k^{-2})), -c]$, quindi $A$ è *definita negativa* (questo è vero (NR) anche per $c = 0$, *equazione di Poisson* bidimensionale). Detto $\{u_{ij}\}$ il vettore soluzione del sistema lineare

$$A\{u_{ij}\} = \{g(P_{ij})\}$$

si provi che vale la stima

$$\frac{\|\{u_{ij}\} - \{u(P_{ij})\}\|_2}{\sqrt{nm}} = \mathcal{O}(h^2) + \mathcal{O}(k^2)$$

(si osservi che $\|\{u(P_{ij})\}\|_2/\sqrt{nm} \approx \|u\|_{L^2(\Omega)}/\sqrt{mis(\Omega)}$). In questo caso il metodo di eliminazione di Gauss non è conveniente (perché?), essendo $A$ fortemente *sparsa* (su ogni riga ci sono al massimo 5 elementi non nulli) e tendenzialmente di grande dimensione sono più adatti metodi iterativi, opportunamente precondizionati visto che $A$ è *mal condizionata*, $k_2(A) = \mathcal{O}(h^{-2} + k^{-2})$ (sugg.: si usi il teorema di Gershgorin per stimare il condizionamento di $A$).

- (!) *metodo delle linee* per l'equazione del calore: dato il problema evolutivo alle derivate parziali con condizioni iniziali e al contorno

$$\frac{\partial u}{\partial t}(P, t) = \sigma \Delta u(P, t) + g(P, t) \ , \quad (P, t) \in \Omega \times (0, +\infty)$$

$$u(P,0) = u_0(P) \ , \quad u(P,t)|_{P\in\partial\Omega} \equiv 0$$

nel caso unidimensionale con $P = x \in \Omega = (a,b)$ o bidimensionale con $P = (x,y) \in \Omega = (a,b) \times (c,d)$, la discretizzazione nelle variabili spaziali tramite $\Delta u \approx \delta^2 u$, posto rispettivamente $y(t) = \{u_i(t)\} \approx \{u(x_i,t)\}$ o $y(t) = \{u_{ij}(t)\} \approx \{u(P_{ij},t)\}$, porta ad un sistema di equazioni differenziali ordinarie nel tempo

$$y' = Ay + b(t) \ , \quad t > 0 \ ; \quad y(0) = y_0$$

dove $A$ è la matrice di discretizzazione del laplaciano vista sopra (tridiagonale o tridiagonale a blocchi). Si mostri che il metodo di Eulero esplicito è stabile con un passo temporale dell'ordine del quadrato dei passi spaziali, mentre i metodi di Eulero implicito e di Crank-Nicolson sono incondizionatamente stabili.
(sugg.: essendo la matrice $A$ simmetrica e definita negativa si tratta di un sistema stiff, ...).