Calcolo Scientifico

Tommaso Baiocchi

Anno Accademico 2025-26

Indice

1 Introduzione

Lo scopo di questo corso è sviluppare strategie numeriche efficaci per la soluzione di due classi di problemi, spesso incontrate nelle applicazioni:

Sistemi lineari della forma $Ax = b$, dove A è o quadrata e invertibile, o data come problema dei minimi quadrati $\min \|Ax - b\|_2$, con A rettangolare e non necessariamente di rango massimo.

Problemi agli autovalori della forma $Av = \lambda v$ per qualche $v \neq 0$. A volte, tutti gli autovalori e autovettori sono ricercati. In altri casi, solo alcuni di essi sono rilevanti. Esempi includono quelli con modulo più grande o più piccolo, o racchiusi in qualche regione $\Omega \subseteq \mathbb{C}$.

In entrambi i casi abbiamo bisogno di differenziare il nostro approccio per problemi che sono "piccoli" o "grandi". Nel primo caso, saranno applicabili i cosiddetti **metodi diretti**. Nel secondo, quando la matrice A è così grande che è impossibile memorizzarla a meno che non abbia qualche struttura particolare, avremo bisogno di impiegare tecniche di proiezione per ridurre la dimensionalità del problema. I metodi in quest'ultima categoria sono noti come **metodi iterativi**.

2 Discretizzazione alle differenze di problemi differenziali

Consideriamo il problema di Cauchy

$$\begin{cases} u''(x) = f(x), & x \in (a, b), \\ u(a) = \alpha, \\ u(b) = \beta, \end{cases} \tag{2.1}$$

per una funzione incognita $u : [a, b] \to \mathbb{R}$, e alcune condizioni al contorno $\alpha, \beta \in \mathbb{R}$. Approssimiamo la soluzione $u(x)$ di (2.1) mediante i suoi valori nei punti discreti $x_j = a + \frac{j(b-a)}{n+1}$, per $j = 1, \ldots, n$. Si noti che, i punti x_j sono equispaziati su una griglia sul segmento lineare $[a, b]$, e la distanza tra due punti adiacenti è $h := \frac{b-a}{n+1}$. In pratica, cerchiamo un vettore $\hat{\mathbf{u}} \in \mathbb{R}^n$, tale che

$$\hat{\mathbf{u}}_j \approx u(x_j), \quad j = 1, \ldots, n.$$

L'idea è di esprimere $\hat{\mathbf{u}}$ come la soluzione di un sistema lineare che rappresenta una controparte discreta di (2.1). Per esempio, valutando (2.1) in ogni punto x_j otteniamo le n equazioni $u''(x_j) = f(x_j)$, dove possiamo facilmente ottenere i termini noti valutando $f(x)$. Tuttavia, abbiamo ancora bisogno di chiarire come trattare le valutazioni della derivata seconda della soluzione e come relazionare queste con il vettore $\hat{\mathbf{u}}$.

2.1 Discretizzazione di operatori differenziali

Un'idea naturale per approssimare la derivata prima a partire da valutazioni della funzione è utilizzare i rapporti incrementali. Ad esempio, possiamo fare uso delle espressioni

$$D_+u(x_j) = \frac{u(x_{j+1}) - u(x_j)}{h}, \quad D_-u(x_j) = \frac{u(x_j) - u(x_{j-1})}{h},$$

che convergono a $u'(x_j)$ per $h \to 0$ (ovvero quando aumentiamo il numero n di punti della griglia all'interno di $[a, b]$), con un errore $\mathcal{O}(h)$.

Più in generale, possiamo ricavare formule di approssimazione di questo tipo combinando sviluppi di Taylor di u valutati nei vari punti che vogliamo coinvolgere.

Esempio 1 Calcoliamo una formula di approssimazione per $u'(x_j)$ che richieda solo la valutazione di u in x_{j-1} e x_{j+1}. Sviluppando u in x_j, e valutando lo sviluppo in x_{j+1} e x_{j-1}, otteniamo:

$$u(x_{j+1}) = u(x_j) + u'(x_j)h + \frac{u''(x_j)}{2}h^2 + \mathcal{O}(h^3),$$
$$u(x_{j-1}) = u(x_j) - u'(x_j)h + \frac{u''(x_j)}{2}h^2 + \mathcal{O}(h^3).$$

Sottraendo la 2' equazione dalla 1' e isolando $u'(x_j)$ avremo $u'(x_j) = \frac{u(x_{j+1}) - u(x_{j-1})}{2h} + \mathcal{O}(h^2)$. Questo ci porta alla formula

$$D_0u(x_j) = \frac{u(x_{j+1}) - u(x_{j-1})}{2h} \approx u'(x_j),$$

che è anche nota come *approssimazione alle differenze finite centrate*. L'errore associato tende a zero come $\mathcal{O}(h^2)$.

L'approccio utilizzato nell'esempio precedente può essere reso sistematico eseguendo i seguenti passi:

- **Selezionare** i $k > 1$ punti che vogliamo coinvolgere nella formula.
- **Calcolare** lo sviluppo di Taylor troncato in x_j di grado $k - 1$, e valutarlo in tutti i punti selezionati nel passo precedente.
- **Considerare** una combinazione lineare con coefficienti incogniti degli k sviluppi troncati.
- **Ricavare** i coefficienti della combinazione lineare (che equivale a ricavare la formula), imponendo che il fattore che moltiplica $u'(x_j)$ sia uguale a 1, e che tutti gli altri fattori, che moltiplicano le altre derivate di u in x_j, siano 0.

Esempio 2. Per trovare un'approssimazione di $u'(x_j)$ che si basi solo su $u(x_j), u(x_{j-1}), u(x_{j-2})$, consideriamo gli sviluppi di Taylor centrati in x_j:

$$\begin{aligned}
u(x_j) &= u(x_j) \\
u(x_{j-1}) &= u(x_j) - u'(x_j)h + \frac{u''(x_j)}{2}h^2 + \mathcal{O}(h^3) \\
u(x_{j-2}) &= u(x_j) - 2u'(x_j)h + \frac{u''(x_j)}{2}(2h)^2 + \mathcal{O}(h^3) \\
&= u(x_j) - 2u'(x_j)h + 2u''(x_j)h^2 + \mathcal{O}(h^3)
\end{aligned}$$

La combinazione lineare descritta sopra diventa:

$$\begin{aligned} c_1 u(x_j) &= c_1 u(x_j) \\ c_2 u(x_{j-1}) &= c_2 u(x_j) - c_2 u'(x_j)h + c_2 \frac{u''(x_j)}{2} h^2 + \mathcal{O}(h^3) \\ c_3 u(x_{j-2}) &= c_3 u(x_j) - 2c_3 u'(x_j)h + 2c_3 u''(x_j)h^2 + \mathcal{O}(h^3) \end{aligned}$$

Quindi otteniamo:

$$c_1 u(x_j)+c_2 u(x_{j-1})+c_3 u(x_{j-2}) = (c_1+c_2+c_3)u(x_j)+(-c_2 h-2c_3 h)u'(x_j)+\left(\frac{c_2}{2}h^2 + 2c_3 h^2\right) u''(x_j)+\mathcal{O}(h^3)$$

Imponiamo che questa combinazione approssimi $u'(x_j)$:

$$\begin{cases} c_1 + c_2 + c_3 = 0 & \text{(coefficiente di } u(x_j) = 0) \\ -c_2 h - 2c_3 h = 1 & \text{(coefficiente di } u'(x_j) = 1) \\ \frac{c_2}{2} h^2 + 2c_3 h^2 = 0 & \text{(coefficiente di } u''(x_j) = 0) \end{cases}$$

Dividendo la seconda equazione per h e la terza per h^2, otteniamo il sistema:

$$\begin{cases} c_1 + c_2 + c_3 = 0 \\ -c_2 - 2c_3 = \frac{1}{h} \\ \frac{c_2}{2} + 2c_3 = 0 \end{cases} \quad \Rightarrow \quad \begin{cases} c_1 + c_2 + c_3 = 0 \\ c_2 + 2c_3 = -\frac{1}{h} \\ c_2 + 4c_3 = 0 \end{cases}$$

che porta alla formula

$$u'(x_j) \approx D_2 u(x_j) = \frac{3u(x_j) - 4u(x_{j-1}) + u(x_{j-2})}{2h}.$$

Lo stesso approccio può essere applicato per approssimare la derivata seconda, modificando il sistema lineare imponendo che il coefficiente di $u''(x_j)$ sia uguale a 1 e gli altri a 0. Per esempio, se consideriamo un'approssimazione della forma $u''(x_j) \approx c_1 u(x_j) + c_2 u(x_{j+1}) + c_3 u(x_{j-1})$ e combiniamo lo sviluppo di Taylor troncato al grado 2 otteniamo il sistema lineare

$$\begin{cases} c_1 + c_2 + c_3 = 0 \\ c_2 - c_3 = 0 \\ c_2 + c_3 = \frac{2}{h^2} \end{cases}$$

che porta alla formula

$$u''(x_j) \approx D^2 u(x_j) = \frac{u(x_{j-1}) - 2u(x_j) + u(x_{j+1})}{h^2}, \tag{2.2}$$

con un errore associato che tende a 0 come $\mathcal{O}(h^2)$. L'approccio si adatta analogamente a derivate di ordine superiore.

2.2 Discretizzazione del problema di Poisson 1D

Abbiamo ora tutti gli ingredienti per associare un sistema lineare al problema di Cauchy (2.1), che è anche noto come *problema di Poisson*. Più specificamente, valutiamo $u''(x) = f(x)$ in ogni punto della griglia e sostituiamo $u''(x_j)$ con l'approssimazione alle differenze finite in (2.2); questo produce il sistema lineare di equazioni

$$\begin{cases} \dfrac{\widehat{\mathbf{u}}_{j-1} - 2\widehat{\mathbf{u}}_j + \widehat{\mathbf{u}}_{j+1}}{h^2} = f(x_j) \\ \widehat{\mathbf{u}}_0 = \alpha, \quad \widehat{\mathbf{u}}_{n+1} = \beta \end{cases} \quad , \qquad j = 1, \ldots, n.$$

Infine, riscriviamo quest'ultimo in forma matriciale come $T^{(h)}\widehat{\mathbf{u}} = \mathbf{f}^{(h)}$ dove

$$T^{(h)} := \frac{1}{h^2}\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix} \in \mathbb{R}^{n\times n}, \quad \mathbf{f}^{(h)} := \begin{bmatrix} f(x_1) - \dfrac{\alpha}{h^2} \\ f(x_2) \\ \vdots \\ f(x_{n-1}) \\ f(x_n) - \dfrac{\beta}{h^2} \end{bmatrix} \in \mathbb{R}^n. \tag{2.3}$$

Per ricapitolare, risolvere (2.3) fornisce un'approssimazione $\widehat{\mathbf{u}}$ del vettore $\mathbf{u} \in \mathbb{R}^n$ contenente le valutazioni $\mathbf{u}_j = u(x_j)$ della vera soluzione di (2.1) sulla griglia equispaziata; quanto è buona questa approssimazione? Idealmente, vorremmo avere $\|\mathbf{u} - \widehat{\mathbf{u}}\| = \mathcal{O}(h^2)$ per una certa norma matriciale. Per dare un'analisi dell'errore rigorosa, introduciamo alcune nozioni.

Definizione Sia $A^{(h)}\widehat{\mathbf{u}} = \mathbf{b}^{(h)}$ il sistema lineare risultante dalla discretizzazione di un'equazione differenziale lineare con un metodo alle differenze finite su una griglia equispaziata relativa al parametro h, e sia $\mathbf{u}$ il vettore contenente le valutazioni della vera soluzione sulla griglia. Chiamiamo *errore di troncamento locale* il vettore

$$\tau^{(h)} := A^{(h)}\mathbf{u} - \mathbf{b}^{(h)},$$

e *errore globale* il vettore

$$\mathbf{e}^{(h)} := \mathbf{u} - \widehat{\mathbf{u}}.$$

Osservazione Si noti che, sottraendo $A^{(h)}\widehat{\mathbf{u}} = \mathbf{b}^{(h)}$ da $A^{(h)}\mathbf{u} = \mathbf{b}^{(h)} + \tau^{(h)}$, otteniamo

$$A^{(h)}\mathbf{e}^{(h)} = \tau^{(h)} \quad \Rightarrow \quad \mathbf{e}^{(h)} = (A^{(h)})^{-1}\tau^{(h)},$$

il che significa che l'errore globale è la soluzione dell'equazione differenziale discretizzata dove l'errore di troncamento locale sostituisce il termine noto.

Una volta fissati il problema di Cauchy e la griglia, l'errore di troncamento locale dipende solo dalla formula alle differenze finite utilizzata per discretizzare l'operatore differenziale. Per esempio, nel caso di (2.1) con la formula di approssimazione (2.2) otteniamo

$$\tau_j = (T^{(h)}\mathbf{u} - f^{(h)})_j = \frac{1}{h^2}(\mathbf{u}_{j-1} - 2\mathbf{u}_j + \mathbf{u}_{j+1}) - \mathbf{f}_j$$

Infatti se sviluppiamo in serie di Taylor ogni termine centrato in x_j:

$$\begin{aligned}
\mathbf{u}_{j-1} &= u(x_{j-1}) = u(x_j - h) = u(x_j) - hu'(x_j) + \frac{h^2}{2}u''(x_j) - \frac{h^3}{6}u'''(x_j) + \frac{h^4}{24}u^{(4)}(x_j) + \mathcal{O}(h^5) \\
\mathbf{u}_j &= u(x_j) \\
\mathbf{u}_{j+1} &= u(x_{j+1}) = u(x_j + h) = u(x_j) + hu'(x_j) + \frac{h^2}{2}u''(x_j) + \frac{h^3}{6}u'''(x_j) + \frac{h^4}{24}u^{(4)}(x_j) + \mathcal{O}(h^5)
\end{aligned}$$

Calcoliamo la combinazione alle differenze finite:

$$\begin{aligned}
\mathbf{u}_{j-1} - 2\mathbf{u}_j + \mathbf{u}_{j+1} = &\left[u(x_j) - hu'(x_j) + \frac{h^2}{2}u''(x_j) - \frac{h^3}{6}u'''(x_j) + \frac{h^4}{24}u^{(4)}(x_j)\right] \\
&- 2u(x_j) \\
&+ \left[u(x_j) + hu'(x_j) + \frac{h^2}{2}u''(x_j) + \frac{h^3}{6}u'''(x_j) + \frac{h^4}{24}u^{(4)}(x_j)\right] + \mathcal{O}(h^5)
\end{aligned}$$

Semplificando i termini otteniamo quindi

$$\mathbf{u}_{j-1} - 2\mathbf{u}_j + \mathbf{u}_{j+1} = h^2u''(x_j) + \frac{h^4}{12}u^{(4)}(x_j) + \mathcal{O}(h^5)$$

Dividendo per h^2:

$$\frac{1}{h^2}(\mathbf{u}_{j-1} - 2\mathbf{u}_j + \mathbf{u}_{j+1}) = u''(x_j) + \frac{h^2}{12}u^{(4)}(x_j) + \mathcal{O}(h^3)$$

Ma dall'equazione differenziale originale sappiamo che $u''(x_j) = f(x_j) = \mathbf{f}_j$, quindi:

$$\begin{aligned}
\tau_j &= \frac{1}{h^2}(\mathbf{u}_{j-1} - 2\mathbf{u}_j + \mathbf{u}_{j+1}) - \mathbf{f}_j \\
&= \left[u''(x_j) + \frac{h^2}{12}u^{(4)}(x_j) + \mathcal{O}(h^3)\right] - u''(x_j) \\
&= \frac{h^2}{12}u^{(4)}(x_j) + \mathcal{O}(h^3)
\end{aligned}$$

Pertanto $\tau_j = \mathcal{O}(h^2)$.

Per ottenere un limite superiore sull'errore globale possiamo scrivere:

$$\|\mathbf{e}^{(h)}\| = \|(T^{(h)})^{-1}\tau^{(h)}\| \leq \|(T^{(h)})^{-1}\|\|\tau^{(h)}\|.$$

La disuguaglianza precedente dice che per garantire la convergenza alla vera soluzione quando $h \to 0$, è sufficiente assicurare che il prodotto $\|(T^{(h)})^{-1}\|\|\tau^{(h)}\|$ tenda a zero. Questo motiva le seguenti definizioni.

Definizione Sia $A^{(h)}\widehat{\mathbf{u}} = \mathbf{b}^{(h)}$ la discretizzazione di un'equazione differenziale lineare con un metodo alle differenze finite su una griglia equispaziata relativa al parametro h. Il metodo alle differenze finite si dice *consistente*, rispetto a una norma vettoriale $\|\cdot\|$, se

$$\lim_{h\to 0} \|\tau^{(h)}\| = 0$$

ed è *stabile* rispetto alla norma matriciale indotta se esistono $C, h_0 \in \mathbb{R}^+$ tali che

$$\|(A^{(h)})^{-1}\| \leq C < \infty, \qquad \forall h < h_0.$$

Infine, il metodo si dice *convergente* se

$$\lim_{h\to 0} \|\mathbf{e}^{(h)}\| = 0.$$

Osservazione È facile vedere che *consistente* + *stabile* $\Rightarrow$ *convergente*. Inoltre, quando il metodo è stabile, l'ordine di convergenza coincide con quello dell'errore di troncamento locale, infatti dall'equazione fondamentale $\mathbf{e}^{(h)} = (A^{(h)})^{-1}\tau^{(h)}$, prendendo le norme otteniamo:

$$\|\mathbf{e}^{(h)}\| = \|(A^{(h)})^{-1}\tau^{(h)}\| \leq \|(A^{(h)})^{-1}\| \cdot \|\tau^{(h)}\|$$

Se il metodo è *consistente*, allora $\|\tau^{(h)}\| \to 0$ per $h \to 0$.

Se il metodo è *stabile*, allora $\|(A^{(h)})^{-1}\| \leq C < \infty$ per h sufficientemente piccolo.

Combinando queste due proprietà:

$$\|\mathbf{e}^{(h)}\| \leq C \cdot \|\tau^{(h)}\| \to 0 \quad \text{per } h \to 0$$

Quindi il metodo è convergente.

Per quanto riguarda l'ordine di convergenza: se $\|\tau^{(h)}\| = \mathcal{O}(h^p)$ e il metodo è stabile, allora:

$$\|\mathbf{e}^{(h)}\| \leq C \cdot \mathcal{O}(h^p) = \mathcal{O}(h^p)$$

Questo significa che l'errore globale decade con lo stesso ordine p dell'errore di troncamento locale. Nel nostro caso specifico del problema di Poisson, poiché $\tau_j = \mathcal{O}(h^2)$ e il metodo è stabile, otteniamo $\|\mathbf{e}^{(h)}\| = \mathcal{O}(h^2)$.

2.3 Stabilità rispetto alla norma infinito per il problema di Poisson 1D

Dimostriamo che la norma infinito di $(T^{(h)})^{-1}$ è limitata superiormente da una costante indipendente da h (e da n). Per prima cosa, riscriviamo $T^{(h)} = -\frac{2}{h^2}B^{(h)}$, dove

$$B^{(h)} = I - C^{(h)}, \qquad C^{(h)} := \frac{1}{2}\begin{bmatrix} 0 & 1 & & & \\ 1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 0 & 1 \\ & & & 1 & 0 \end{bmatrix}.$$

Per i teoremi di Gershgorin abbiamo che $\rho(C^{(h)}) < 1$, il che implica

$$(T^{(h)})^{-1} = -\frac{h^2}{2}(B^{(h)})^{-1} = -\frac{h^2}{2}\sum_{j\geq 0}(C^{(h)})^j.$$

infatti da $\rho(C^{(h)}) < 1$ la serie geometrica di matrici $\sum_{j\geq 0}(C^{(h)})^j$ converge a $(I - C^{(h)})^{-1} = (B^{(h)})^{-1}$

Poiché $C^{(h)}$ è non negativa elemento per elemento, anche $(B^{(h)})^{-1}$ lo è; questo significa che la norma infinito di $(B^{(h)})^{-1}$ è ottenuta moltiplicando per il vettore e di tutti uno, ovvero

$$\|(T^{(h)})^{-1}\|_\infty = \frac{h^2}{2}\|(B^{(h)})^{-1}\|_\infty = \frac{h^2}{2}\|(B^{(h)})^{-1}e\|_\infty.$$

infatti per una matrice non negativa A, la norma infinito $\|A\|_\infty$ è il massimo della somma delle righe, che si ottiene proprio moltiplicando per il vettore di tutti uno.

Per stimare $(B^{(h)})^{-1}e$, introduciamo i vettori

$$p = \begin{bmatrix} 1 \\ 2 \\ 3 \\ \vdots \\ n \end{bmatrix}, \qquad s = \begin{bmatrix} 1 \\ 4 \\ 9 \\ \vdots \\ n^2 \end{bmatrix},$$

e, con un calcolo diretto, osserviamo che

$$B^{(h)}p = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \frac{n+1}{2} \end{bmatrix}$$

$$B^{(h)}s = \begin{bmatrix} -1 \\ -1 \\ \vdots \\ -1 \\ -1 + \frac{(n+1)^2}{2} \end{bmatrix} = -e + (n+1)B^{(h)}p.$$

Spiegazione dei calcoli

- Per $B^{(h)}p$: per le righe interne $j = 2, \ldots, n-1$ abbiamo:

$$(B^{(h)}p)_j = p_j - \frac{1}{2}(p_{j-1} + p_{j+1}) = j - \frac{1}{2}((j-1) + (j+1)) = 0$$

- Per la prima riga: $1 - \frac{1}{2}(0+2) = 0$
- Per l'ultima riga: $n - \frac{1}{2}(n-1+0) = \frac{n+1}{2}$

Pertanto $(B^{(h)})^{-1}e = -s + (n+1)p$, il che implica

$$\|(B^{(h)})^{-1}\|_\infty = \max_{j=1,\ldots,n} |(n+1)j - j^2| \leq \frac{(n+1)^2}{4},$$

e a sua volta

$$\|(T^{(h)})^{-1}\|_\infty \leq \frac{h^2}{2} \cdot \frac{(n+1)^2}{4} = \frac{(b-a)^2}{8},$$

che dimostra la stabilità del metodo, infatti dal calcolo della norma infinito

$$(B^{(h)})^{-1}e = -s + (n+1)p = \begin{bmatrix} -1 + (n+1)\cdot 1 \\ -4 + (n+1)\cdot 2 \\ -9 + (n+1)\cdot 3 \\ \vdots \\ -n^2 + (n+1)\cdot n \end{bmatrix} = \begin{bmatrix} (n+1) - 1^2 \\ 2(n+1) - 2^2 \\ 3(n+1) - 3^2 \\ \vdots \\ n(n+1) - n^2 \end{bmatrix}$$

Quindi per $j = 1, \ldots, n$:

$$[(B^{(h)})^{-1}e]_j = j(n+1) - j^2 = -j^2 + (n+1)j$$

Questa è una parabola concava verso il basso. Il massimo si trova nel vertice:

$$j_{max} = \frac{n+1}{2}, \quad \text{valore massimo} = \frac{(n+1)^2}{4}$$

Dunque per sostituzione finale

$$\|(T^{(h)})^{-1}\|_\infty = \frac{h^2}{2}\|(B^{(h)})^{-1}e\|_\infty \leq \frac{h^2}{2} \cdot \frac{(n+1)^2}{4} = \frac{(b-a)^2}{8}$$

La maggiorazione è indipendente da h e n, quindi il metodo è stabile.

2.4 Problema di Poisson 2D

È abbastanza naturale generalizzare la discretizzazione di (2.1) al problema di Cauchy bidimensionale:

$$\begin{cases} \dfrac{\partial^2 u(x,y)}{\partial x^2} + \dfrac{\partial^2 u(x,y)}{\partial y^2} = f(x,y) & (x,y) \in \Omega := [a,b] \times [a,b] \\ u(x,y) = u_0(x,y) & (x,y) \in \partial\Omega \end{cases}, \tag{2.4}$$

per una funzione incognita $u : \Omega \to \mathbb{R}$, e una data funzione $u_0(x,y) : \partial\Omega \to \mathbb{R}$.
Consideriamo la griglia quadrata uniforme di punti

$$\{(x_i, y_j) = (a + ih, a + jh) : \ i,j = 1, \ldots, n\} \subset \Omega,$$

dove, ancora, $h = \frac{b-a}{n+1}$. Quindi, cerchiamo un'approssimazione del vettore $\mathbf{u} \in \mathbb{R}^{n^2}$ contenente le valutazioni della vera soluzione di (2.4) sulla griglia quadrata, con un ordinamento lessicografico per gli indici (i,j):

$$\mathbf{u} := \begin{bmatrix} u(x_1,y_1) \\ u(x_1,y_2) \\ \vdots \\ u(x_1,y_n) \\ u(x_2,y_1) \\ \vdots \\ u(x_2,y_n) \\ \vdots \\ u(x_n,y_1) \\ \vdots \\ u(x_n,y_n) \end{bmatrix} \in \mathbb{R}^{n^2}.$$

Per ottenere approssimazioni delle derivate seconde che coinvolgano solo valutazioni di $u(x,y)$ sulla griglia possiamo impiegare (2.2) considerando una delle due variabili come fissa; questo significa:

$$\begin{aligned}\frac{\partial^2 u(x_i,y_j)}{\partial x^2} &\approx \frac{u(x_{i-1},y_j)-2u(x_i,y_j)+u(x_{i+1},y_j)}{h^2},\\ \frac{\partial^2 u(x_i,y_j)}{\partial y^2} &\approx \frac{u(x_i,y_{j-1})-2u(x_i,y_j)+u(x_i,y_{j+1})}{h^2}.\end{aligned} \tag{2.5}$$

Mediante (2.5), approssimiamo l'equazione $\frac{\partial^2 u(x_i,y_j)}{\partial x^2}+\frac{\partial^2 u(x_i,y_j)}{\partial y^2}=f(x_i,y_j)$ con

$$\frac{u(x_{i-1},y_j)-4u(x_i,y_j)+u(x_{i+1},y_j)+u(x_i,y_{j-1})+u(x_i,y_{j+1})}{h^2}=f(x_i,y_j),$$

per $i,j=1,\dots,n$ (quindi per ogni punto della griglia). Impilando queste equazioni in un unico sistema lineare otteniamo

$$T_{2d}^{(h)}\tilde{\mathbf{u}}=\mathbf{f}_{2d}^{(h)}, \tag{2.6}$$

dove

$$T_{2d}^{(h)}=\frac{1}{h^2}\begin{bmatrix} M & I & & & \\ I & M & I & & \\ & I & M & \ddots & \\ & & \ddots & \ddots & I \\ & & & I & M \end{bmatrix}\in\mathbb{R}^{n^2\times n^2},\quad M=\begin{bmatrix} -4 & 1 & & & \\ 1 & -4 & 1 & & \\ & 1 & -4 & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & -4 \end{bmatrix}\in\mathbb{R}^{n\times n}.$$

Analogamente al caso 1D, il termine noto tiene conto delle condizioni al contorno

$$\begin{aligned} u_{n+1,j} &= u(b,y_j)=u_0(b,y_j),\\ u_{i,n+1} &= u(x_i,b)=u_0(x_i,b),\\ u_{0,j} &= u(a,y_j)=u_0(a,y_j),\\ u_{i,0} &= u(x_i,a)=u_0(x_i,a),\end{aligned}$$

in modo che

$$\mathbf{f}_{2d}^{(h)} = \begin{bmatrix} f(x_1,y_1) - \dfrac{u_0(a,y_1)}{h^2} - \dfrac{u_0(x_1,a)}{h^2} \\ f(x_1,y_2) - \dfrac{u_0(a,y_2)}{h^2} \\ \vdots \\ f(x_1,y_n) - \dfrac{u_0(a,y_n)}{h^2} - \dfrac{u_0(x_1,b)}{h^2} \\ f(x_2,y_1) - \dfrac{u_0(x_2,a)}{h^2} \\ f(x_2,y_2) \\ \vdots \\ f(x_2,y_{n-1}) \\ f(x_2,y_n) - \dfrac{u_0(x_2,b)}{h^2} \\ \vdots \\ f(x_n,y_n) - \dfrac{u_0(b,y_n)}{h^2} - \dfrac{u_0(x_n,b)}{h^2} \end{bmatrix} \in \mathbb{R}^{n^2}.$$

Spiegazione della struttura del termine noto:

- **Punti interni** (es: $f(x_2, y_2)$): Nessuna correzione al contorno
- **Punti sul bordo sinistro** ($x = a$): Sottrazione di $\dfrac{u_0(a, y_j)}{h^2}$
- **Punti sul bordo destro** ($x = b$): Sottrazione di $\dfrac{u_0(b, y_j)}{h^2}$
- **Punti sul bordo inferiore** ($y = a$): Sottrazione di $\dfrac{u_0(x_i, a)}{h^2}$
- **Punti sul bordo superiore** ($y = b$): Sottrazione di $\dfrac{u_0(x_i, b)}{h^2}$
- **Punti d'angolo**: Doppia correzione

2.5 Stabilità rispetto alla norma 2 per il problema di Poisson 2D

Per fornire un altro esempio di risultati di convergenza per la discretizzazione di equazioni differenziali, per vettori che rappresentano valutazioni di funzioni sulla griglia quadrata $n \times n$, consideriamo la norma 2 scalata:

$$\|u\|_{l_2} := h\|u\|_2 = \frac{b-a}{n+1}\sqrt{\sum_{j=1}^{n^2} |u_j|^2}.$$

Spiegazione della norma scalata:

- La norma standard $\|u\|_2 = \sqrt{\sum_{j=1}^{n^2} |u_j|^2}$ non è appropriata per l'analisi di convergenza

- Quando $h \to 0$ (cioè $n \to \infty$), il numero di punti n^2 cresce, quindi $\|u\|_2$ diverge
- Il fattore h compensa la densità dei punti sulla griglia
- In 2D, l'area elementare è h^2, ma nella norma usiamo h perché:

$$h\|u\|_2 = h\sqrt{\sum_{j=1}^{n^2} |u_j|^2} = \sqrt{h^2 \sum_{j=1}^{n^2} |u_j|^2}$$

- Questo corrisponde a un'approssimazione della norma L^2 integrale

Si noti che $\|\cdot\|_{l_2}$ induce la consueta norma matriciale 2, e che

$$\lim_{h\to 0} \|u\|_{l_2} = \sqrt{\int_\Omega |u(x,y)|^2 \, dx \, dy}.$$

Questa norma ci permette di studiare il comportamento dell'errore quando il passo della griglia tende a zero, garantendo che le stime siano indipendenti dal numero di punti di discretizzazione.

Prima di fornire una stima di $\|(T_{2d}^{(h)})^{-1}\|_2$, dobbiamo introdurre la nozione di prodotto di Kronecker, che sarà utile per scoprire le proprietà spettrali di $T_{2d}^{(h)}$.

Definizione Siano $A \in \mathbb{C}^{m\times n}$,$B \in \mathbb{C}^{k\times p}$, chiamiamo *prodotto di Kronecker di A con B* la matrice

$$A \otimes B := \begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \vdots & \vdots & & \vdots \\ a_{m1}B & a_{m2}B & \dots & a_{mn}B \end{bmatrix} \in \mathbb{C}^{mk\times np}.$$

Esempio

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \otimes \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} 1 \cdot \begin{bmatrix} a & b \\ c & d \end{bmatrix} & 2 \cdot \begin{bmatrix} a & b \\ c & d \end{bmatrix} \\ 3 \cdot \begin{bmatrix} a & b \\ c & d \end{bmatrix} & 4 \cdot \begin{bmatrix} a & b \\ c & d \end{bmatrix} \end{bmatrix} = \begin{bmatrix} a & b & 2a & 2b \\ c & d & 2c & 2d \\ 3a & 3b & 4a & 4b \\ 3c & 3d & 4c & 4d \end{bmatrix}$$

Il prodotto di Kronecker gode delle seguenti proprietà:

- $(A \otimes B)^* = A^* \otimes B^*$,
- se A, B sono matrici quadrate invertibili

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1},$$

- se le matrici coinvolte hanno dimensioni compatibili, vale

$$(A \otimes B) \cdot (C \otimes D) \cdot (E \otimes F) = (ACE \otimes BDF).$$

Con un calcolo diretto, troviamo che la matrice che sorge dalla discretizzazione del problema di Poisson 2D è collegata con quella associata al problema 1D tramite la seguente relazione:

$$T_{2d}^{(h)} = I \otimes T^{(h)} + T^{(h)} \otimes I. \tag{2.7}$$

Spiegazione della relazione (2.7)

- $I \otimes T^{(h)}$: rappresenta la derivata seconda nella direzione x
- $T^{(h)} \otimes I$: rappresenta la derivata seconda nella direzione y
- Dimensione: se $T^{(h)} \in \mathbb{R}^{n\times n}$, allora $T_{2d}^{(h)} \in \mathbb{R}^{n^2\times n^2}$

L'equazione (2.7) collega anche gli autovalori di $T_{2d}^{(h)}$ con quelli di $T^{(h)}$, come spiegato nel prossimo risultato.

Lemma *Gli autovalori di* $T_{2d}^{(h)}$ *sono dati da*

$$\lambda_i(T^{(h)}) + \lambda_j(T^{(h)}), \quad i,j = 1,\dots,n,$$

dove $\lambda_i(T^{(h)})$ *denota l'i-esimo autovalore di* $T^{(h)}$.

Dimostrazione. Le due matrici $I \otimes T^{(h)}$ e $T^{(h)} \otimes I$ commutano:

$$(I \otimes T^{(h)})(T^{(h)} \otimes I) = T^{(h)} \otimes T^{(h)} = (T^{(h)} \otimes I)(I \otimes T^{(h)})$$

Dunque sono simultaneamente diagonalizzabili, quindi la loro somma ha come autovalori la somma degli autovalori.
Siano v_i autovettori di $T^{(h)}$ con autovalori $\lambda_i(T^{(h)})$, e w_j autovettori di $T^{(h)}$ con autovalori $\lambda_j(T^{(h)})$.

Consideriamo il prodotto tensore $v_i \otimes w_j$:

$$(A \otimes B)(v_i \otimes w_j) = (Av_i) \otimes (Bw_j) = (\lambda_i(A)v_i) \otimes (\lambda_j(B)w_j) = \lambda_i(A)\lambda_j(B)(v_i \otimes w_j)$$

Nel nostro caso specifico, per $T_{2d}^{(h)} = I \otimes T^{(h)} + T^{(h)} \otimes I$:

$$\begin{aligned} T_{2d}^{(h)}(v_i \otimes w_j) &= (I \otimes T^{(h)})(v_i \otimes w_j) + (T^{(h)} \otimes I)(v_i \otimes w_j) \\ &= (Iv_i) \otimes (T^{(h)}w_j) + (T^{(h)}v_i) \otimes (Iw_j) \\ &= v_i \otimes (\lambda_j(T^{(h)})w_j) + (\lambda_i(T^{(h)})v_i) \otimes w_j \\ &= \lambda_j(T^{(h)})(v_i \otimes w_j) + \lambda_i(T^{(h)})(v_i \otimes w_j) \\ &= (\lambda_i(T^{(h)}) + \lambda_j(T^{(h)}))(v_i \otimes w_j) \end{aligned}$$

Quindi $v_i \otimes w_j$ è autovettore di $T_{2d}^{(h)}$ con autovalore $\lambda_i(T^{(h)}) + \lambda_j(T^{(h)})$. □

Siamo pronti per studiare $\|(T_{2d}^{(h)})^{-1}\|_2$.
Per prima cosa osserviamo che $T_{2d}^{(h)}$ è una matrice simmetrica, e così è la sua inversa; quindi, vale

$$\|(T_{2d}^{(h)})^{-1}\|_2 = \rho((T_{2d}^{(h)})^{-1}) = \frac{1}{\min_{i,j} |\lambda_i(T^{(h)}) + \lambda_j(T^{(h)})|} = \frac{1}{2\min_i |\lambda_i(T^{(h)})|},$$

dove l'ultima uguaglianza segue dal fatto che $T^{(h)}$ è simmetrica e definita negativa. Infine, abbiamo

$$\frac{1}{\min_i |\lambda_i(T^{(h)})|} = \rho((T^{(h)})^{-1}) \leq \|(T^{(h)})^{-1}\|_\infty \leq \frac{(b-a)^2}{8},$$

e questo implica

$$\|(T_{2d}^{(h)})^{-1}\|_2 \leq \frac{(b-a)^2}{16}.$$

2.6 Integrazione di problemi dipendenti dal tempo

In questa sezione discutiamo un altro approccio per risolvere l'equazione differenziale

$$\begin{cases} \dfrac{\partial}{\partial t}u(t,x) - \dfrac{\partial^2}{\partial x^2}u(t,x) = 0, \quad t \in [0, t_{\max}],\ x \in [0,1] \\ u(0,x) \equiv u_0(x), \\ u(t,0) = u(t,1) = 0. \end{cases}$$

La discussione in questa sezione si applicherebbe a problemi più generali della forma $\frac{\partial}{\partial t}u + Lu = f$, dove L è un operatore differenziale definito positivo con appropriate condizioni al contorno, e f è una funzione nota.

Il metodo presentato in questa sezione è talvolta noto come "metodo delle linee": discretizziamo la PDE nello spazio, lasciando solo una variabile continua (il tempo). Quindi, l'equazione differenziale ordinaria (ODE) risultante ad alta dimensione viene integrata con un metodo numerico appropriato. In pratica, possiamo basarci sulla discretizzazione alle differenze finite descritta nella sezione precedente, e ottenere la seguente ODE:

$$\begin{cases} \mathbf{u}' = T^{(h)}\mathbf{u}, \\ \mathbf{u}(0) = \mathbf{u}_0. \end{cases}$$

Nell'equazione precedente, abbiamo le seguenti quantità:

$\mathbf{u}(t)$ Il vettore dipendente dal tempo contenente la valutazione della soluzione al tempo t in tutti i punti della griglia $x_1, \ldots, x_n$.

$T^{(h)}$ La matrice tridiagonale che discretizza l'azione della derivata seconda, con passo di discretizzazione $h = 1/(n+1)$.

Consideriamo due possibili modi per discretizzare la ODE precedente nel tempo: i metodi di Eulero esplicito e implicito. Fissiamo una discretizzazione temporale con passo Δt, tale che possiamo definire $t_0 = 0$ e $t_i = i \cdot \Delta t$; facciamo variare i da 0 a $N \approx t_{\max}/\Delta t$. I metodi producono una sequenza di approssimazioni $\mathbf{u}^{(i)} \approx \mathbf{u}(t_i)$ definite dalle seguenti identità:

$$\mathbf{u}^{(i+1)} = \mathbf{u}^{(i)} + \Delta t \left(T^{(h)} \mathbf{u}^{(i)} \right) \tag{2.10}$$

$$\mathbf{u}^{(i+1)} = \mathbf{u}^{(i)} + \Delta t \left(T^{(h)} \mathbf{u}^{(i+1)} \right) \tag{2.11}$$

L'equazione (2.10) fornisce il metodo di Eulero esplicito, mentre l'equazione (2.11) fornisce la variante implicita. La differenza chiave è che il primo ci permette di calcolare l'iterata successiva $\mathbf{u}^{(i+1)}$ mediante una formula esplicita, mentre il secondo richiede di risolvere un'equazione dove $\mathbf{u}^{(i+1)}$ è l'incognita. In pratica, per questa ODE lineare le iterazioni di Eulero esplicito e implicito possono essere riscritte come segue:

$$\mathbf{u}^{(i+1)} = (I + \Delta t \, T^{(h)}) \mathbf{u}^{(i)}, \qquad \mathbf{u}^{(i+1)} = (I - \Delta t \, T^{(h)})^{-1} \mathbf{u}^{(i)}.$$

Quale metodo dovremmo preferire? Per rispondere a questa domanda, ricordiamo che poiché stiamo discretizzando un operatore definito negativo, ci aspettiamo che anche $T^{(h)}$ sia definita negativa; infatti, dal teorema di Gershgorin, sappiamo che lo spettro di $T^{(h)}$ è racchiuso nell'intervallo $[-4/h^2, 0]$. Poiché l'ODE è lineare e $T^{(h)}$ può essere diagonalizzata come $T^{(h)} = QD^{(h)}Q^*$, possiamo scrivere esplicitamente la soluzione al tempo t_i come

$$\begin{aligned}
\mathbf{u}(t_i) = e^{t_i T^{(h)}} \mathbf{u}_0 &= \left(I + t_i T^{(h)} + \frac{1}{2} t_i^2 (T^{(h)})^2 + \dots \right) \mathbf{u}_0 \\
&= \left(I + t_i Q D^{(h)} Q^* + \frac{1}{2} t_i^2 (Q D^{(h)} Q^*)^2 + \dots \right) \mathbf{u}_0 \\
&= Q \left(I + t_i D^{(h)} + \frac{1}{2} t_i^2 (D^{(h)})^2 + \dots \right) Q^* \mathbf{u}_0 \\
&= Q \begin{bmatrix} e^{t_i \lambda_1^{(h)}} & & \\ & \ddots & \\ & & e^{t_i \lambda_n^{(h)}} \end{bmatrix} Q^* \mathbf{u}_0
\end{aligned}$$

Poiché tutti gli autovalori $\lambda_i^{(h)}$ sono reali e negativi, la soluzione tende a zero per $t_i \to \infty$. È naturale chiedere che la soluzione prodotta dallo schema di integrazione numerica abbia la stessa proprietà. Sfruttando ancora una volta la diagonalizzazione di $T^{(h)}$ possiamo scrivere la soluzione per Eulero esplicito:

$$\mathbf{u}^{(i)} = (I + \Delta t \, T^{(h)})^i \mathbf{u}_0 = Q \begin{bmatrix} (1 + \Delta t \lambda_1)^i & & \\ & \ddots & \\ & & (1 + \Delta t \lambda_n)^i \end{bmatrix} Q^* \mathbf{u}_0.$$

Assumendo che $\mathbf{u}_0$ possa essere arbitrario, l'espressione sopra è limitata per $i \to \infty$ se e solo se, per tutti gli autovalori di $T^{(h)}$, abbiamo $|1 + \Delta t \lambda_i| < 1$; poiché tutti gli autovalori sono reali e negativi, questo è equivalente alla condizione di stabilità $\Delta t < 2 \cdot (\max_i |\lambda_i|)^{-1}$. Poiché il più grande autovalore in modulo è vicino a $-4/h^2$, questa condizione è equivalente a imporre $\Delta t \lesssim h^2/2$, a meno di termini di ordine superiore in h. In conclusione, affinché la soluzione discreta rimanga limitata, abbiamo bisogno di soddisfare questa condizione (stretta) sul passo temporale, che è impraticabile nella maggior parte delle situazioni. Si noti che scegliere un passo temporale che non soddisfa questo vincolo farà andare la soluzione discreta all'infinito (in modulo) esponenzialmente veloce, e sarà quindi assolutamente inutile dal punto di vista del modello.

D'altra parte, effettuando la stessa analisi per lo schema di Eulero implicito, otteniamo

$$\mathbf{u}^{(i)} = (I - \Delta t\, T^{(h)})^{-i} \mathbf{u}_0 = Q \begin{bmatrix} (1 - \Delta t \lambda_1)^{-i} & & \\ & \ddots & \\ & & (1 - \Delta t \lambda_n)^{-i} \end{bmatrix} Q^* \mathbf{u}_0,$$

che è limitata se e solo se $|1 - \Delta t \lambda| > 1$ per tutti gli autovalori λ di $T^{(h)}$. Tuttavia, sappiamo che tutti i λ sono reali e strettamente negativi, e quindi questa condizione è banalmente vera: il metodo di Eulero implicito è stabile (cioè, restituisce soluzioni limitate) per tutte le scelte di Δt.

Questo esempio mostra che due delle principali sfide dell'algebra lineare numerica che esploreremo nei prossimi capitoli sono importanti per l'analisi delle PDE:

- Calcolare **autovalori**, per essere in grado di costruire metodi stabili e caratterizzare comportamenti a lungo termine delle soluzioni.
- Risolvere **sistemi lineari di grandi dimensioni**, per essere in grado di applicare metodi impliciti (per i quali Eulero implicito è il rappresentante più semplice).

Quindi, le PDE saranno spesso la fonte più naturale di esempi e casi di test (sebbene non saranno l'unica) per i metodi sviluppati nel resto del corso.

3 Problemi agli autovalori non simmetrici

Il problema agli autovalori (standard) può essere formulato come la ricerca di tutti gli scalari λ tali che $Av = \lambda v$, per qualche $v \neq 0$; molto spesso, siamo interessati anche agli autovettori destri o sinistri. Dalle prime lezioni di algebra lineare, sappiamo che il problema può essere riformulato come il calcolo delle radici del polinomio caratteristico:

$$p(\lambda) := \det(\lambda I - A).$$

Questa caratterizzazione può portare a un primo algoritmo tentativo per calcolare gli autovalori di una matrice A:

1. Determinare il polinomio $p(\lambda)$ calcolando il determinante (ciò è fattibile tramite una variante della fattorizzazione LU).
2. Usare qualche iterazione funzionale per calcolare tutte le radici.

3. Se sono necessari anche gli autovettori, calcolarli trovando una base per il nucleo di $A - \lambda I$.

Questo approccio, sebbene teoricamente valido, ha diversi svantaggi "numerici". Come può il nostro metodo essere inaccurato? La risposta a questa questione è sottile ma fondamentale per lo sviluppo di metodi numerici stabili. Quello che stiamo facendo è trasformare un problema in un altro (un problema agli autovalori in uno di ricerca delle radici di un polinomio), attraverso una mappa Γ:

$$\Gamma : \mathbb{C}^{n\times n} \to \mathbb{C}[\lambda], \quad A \mapsto \det(\lambda I - A)$$

Non possiamo garantire che piccole perturbazioni nei dati di ingresso di un problema corrispondano a piccole perturbazioni nei dati di ingresso dell'altro: piccole variazioni nei coefficienti di $p(\lambda)$ possono causare grandi cambiamenti nelle entrate della matrice originale A.

Poiché lavoriamo con l'aritmetica in floating point, introdurre errori di arrotondamento è inevitabile: abbiamo bisogno di assicurarci che qualsiasi algoritmo che sviluppiamo sia stabile sotto perturbazioni, e quindi costruire una teoria delle perturbazioni significativa per analizzarli.

3.1 Teoria delle perturbazioni per problemi agli autovalori

Studieremo ora l'effetto delle perturbazioni sugli spettri delle matrici. Questo argomento è strettamente correlato con il numero di condizionamento.

Definizione Sia A una matrice $n \times n$, e λ un autovalore in $\Lambda(A)$; allora, il *numero di condizionamento di* λ, denotato da $\kappa(A, \lambda)$, è definito da

$$\kappa(A, \lambda) := \lim_{h\to 0} \frac{\sup_{\|\delta A\|\le h} \min\{|\mu - \lambda| \mid \mu \in \Lambda(A + \delta A)\}}{h}.$$

In generale, il numero di condizionamento può essere finito o infinito. Si noti che la definizione di numero di condizionamento dipende dalla scelta della norma. Spesso questa sarà la norma spettrale, per la quale usiamo la notazione $\kappa_2(A, \lambda)$.

Teorema *Sia A una matrice complessa $n \times n$. Allora, esistono n funzioni continue $\lambda_i : \mathbb{C}^{n\times n} \to \mathbb{C}$ tali che*

$$\Lambda(A + \delta A) = \{\lambda_1(A + \delta A), \ldots, \lambda_n(A + \delta A)\}$$

Dimostrazione. Iniziamo notando che $p(\lambda) := \det(\lambda I - A)$ ha coefficienti che sono funzioni continue delle entrate di A. Quindi, è sufficiente dimostrare che le radici di $p(\lambda)$ sono funzioni continue dei suoi coefficienti.

Siano $\lambda_1, \ldots, \lambda_r$ gli autovalori di A, con le loro molteplicità m_i. Selezioniamo un $\epsilon > 0$ abbastanza piccolo affinché gli insiemi $B(\lambda_i, \epsilon)$ siano disgiunti; in particolare questo implica che $p(\lambda)$ non si annulla sul bordo di $\partial B(\lambda_i, \epsilon)$. Grazie al teorema dei residui abbiamo

$$m_i := \frac{1}{2\pi i} \int_{\partial B(\lambda_i,\epsilon)} \frac{p'(z)}{p(z)}\, dz, \tag{3.1}$$

e la funzione p'/p è una funzione continua e limitata di z e dei coefficienti di $p(z)$ sull'insieme compatto $\mathcal{S}_\epsilon := \cup_{i=1}^{r} \partial B(\lambda_i, \epsilon)$. Pertanto, possiamo selezionare δ tale che per ogni perturbazione δp con norma del vettore dei coefficienti limitata da δ, vale

$$\max_{z \in \mathcal{S}_\epsilon} \left| \frac{p'(z) + \delta p'(z)}{p(z) + \delta p(z)} - \frac{p'(z)}{p(z)} \right| \leq \frac{1}{2\epsilon}$$

Se calcoliamo la formula integrale (3.1) per il polinomio perturbato $p(z) + \delta p(z)$ abbiamo che il numero di radici contate con molteplicità all'interno di ogni $B(\lambda_i, \epsilon)$ non può cambiare di più di $\frac{1}{2}$. Essendo un intero, questo implica che il numero non cambia, e quindi le radici non possono sfuggire dalle palle $B(\lambda_i, \epsilon)$, il che conclude la dimostrazione. □

Caratterizziamo ora il numero di condizionamento per autovalori semplici, cioè di molteplicità geometrica 1.

Teorema Sia $A \in \mathbb{C}^{n \times n}$, e λ un autovalore semplice. Allora,

$$\kappa_2(A, \lambda) = \frac{\|v\|_2 \|w\|_2}{|w^* v|},$$

dove w e v sono rispettivamente gli autovettori sinistro e destro relativi a λ.

Dimostrazione. Poiché l'autovalore è semplice, possiamo fare uno sviluppo al primo ordine; assumendo che $Av = \lambda v$ possiamo scrivere

$$(A + \delta A)(v + \delta v) = (\lambda + \delta\lambda)(v + \delta v).$$

Riorganizzando gli addendi ignorando i termini del secondo ordine si ottiene

$$\delta A v + A\delta v - \lambda \delta v = \delta\lambda v + \mathcal{O}(\|\delta A\|_2^2).$$

sviluppiamo i prodotti:

$$Av + A\delta v + \delta A v + \delta A \delta v = \lambda v + \lambda \delta v + \delta\lambda v + \delta\lambda\delta v$$

Sappiamo che $Av = \lambda v$, quindi semplifichiamo:

$$A\delta v + \delta A v + \delta A \delta v = \lambda \delta v + \delta\lambda v + \delta\lambda\delta v$$

Trascuriamo i termini del secondo ordine ($\delta A \delta v$ e $\delta\lambda\delta v$):

$$A\delta v + \delta A v = \lambda \delta v + \delta\lambda v$$

Riorganizziamo:

$$A\delta v - \lambda \delta v + \delta A v = \delta\lambda v$$

$$(A - \lambda I)\delta v + \delta A v = \delta\lambda v$$

Ora moltiplichiamo a sinistra per w^* (l'autovettore sinistro):

$$w^*(A - \lambda I)\delta v + w^*\delta A v = w^*(\delta\lambda v)$$

Ma $w^*(A - \lambda I) = 0$ perché $w^*A = \lambda w^*$, quindi:

$$w^*\delta A v = \delta\lambda(w^*v)$$

Isoliamo $\delta\lambda$:

$$\delta\lambda = \frac{w^*\delta A v}{w^*v} + \mathcal{O}(\|\delta A\|^2)$$

Prendendo le norme:

$$|\delta\lambda| \leq \frac{\|w^*\|_2\|\delta A\|_2\|v\|_2}{|w^*v|} = \frac{\|w\|_2\|v\|_2}{|w^*v|}\|\delta A\|_2$$

Per mostrare che il bound è ottimale, consideriamo:

$$\delta A = h\frac{wv^*}{\|v\|_2\|w\|_2}$$

Allora:

$$w^*\delta A v = w^*\left(h\frac{wv^*}{\|v\|_2\|w\|_2}\right)v = h\frac{(w^*w)(v^*v)}{\|v\|_2\|w\|_2} = h\frac{\|w\|_2^2\|v\|_2^2}{\|v\|_2\|w\|_2} = h\|v\|_2\|w\|_2$$

che conclude la dimostrazione prendendo il limite per $h \to 0$. □

Vale la pena menzionare alcuni esempi di numeri di condizionamento di autovalori per classi speciali di matrici.

- Se $A = A^*$ allora gli autovettori sinistro e destro coincidono, e quindi $\kappa_2(A, \lambda) = 1$.

- Se A è un blocco di Jordan, gli autovettori sinistro e destro sono ortogonali; anche se il Teorema non copre questo caso, un'applicazione diretta della formula dà $\frac{1}{0}$, e infatti in questo caso il numero di condizionamento è uguale a ∞.

Esercizio Dimostrare che se una matrice è normale, cioè $AA^* = A^*A$, allora il numero di condizionamento dei suoi autovalori è uguale a 1 (come nel caso speciale delle matrici simmetriche menzionato sopra).

Soluzione. Sia A una matrice normale con $AA^* = A^*A$, e sia λ un autovalore semplice di A con autovettore destro v e autovettore sinistro w.

Per matrici normali, vale la proprietà fondamentale che gli autovettori sinistri e destri coincidono a meno di coniugio complesso. Più precisamente, se $Av = \lambda v$, allora $A^*v = \bar{\lambda}v$ (poiché A è normale).

Quindi l'autovettore sinistro w soddisfa $w^*A = \lambda w^*$, e possiamo prendere $w = v$.

Calcoliamo ora il numero di condizionamento:

$$\kappa_2(A,\lambda) = \frac{\|v\|_2\|w\|_2}{|w^*v|} = \frac{\|v\|_2\|v\|_2}{|v^*v|} = \frac{\|v\|_2^2}{\|v\|_2^2} = 1$$

□

Esercizio Dimostrare che per un blocco di Jordan, il numero di condizionamento dell'autovalore è uguale a $+\infty$. In particolare, le funzioni autovalore sono continue ma non C^1: cosa si può dire sulla loro regolarità?

Soluzione. Consideriamo un blocco di Jordan $J_n(\lambda_0)$ di dimensione n:

$$J_n(\lambda_0) = \begin{bmatrix} \lambda_0 & 1 & & \\ & \lambda_0 & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_0 \end{bmatrix}$$

L'autovettore destro v è:

$$v = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

L'autovettore sinistro w (autovettore di $J_n(\lambda_0)^*$) è:

$$w = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

Calcoliamo il prodotto scalare:

$$w^*v = 0 \cdot 1 + \cdots + 0 \cdot 0 + 1 \cdot 0 = 0$$

Applicando la formula del numero di condizionamento:

$$\kappa_2(J_n(\lambda_0), \lambda_0) = \frac{\|v\|_2\|w\|_2}{|w^*v|} = \frac{1 \cdot 1}{0} = \infty$$

Per quanto riguarda la regolarità: le funzioni autovalore sono sempre continue (per il Teorema), ma nel caso di autovalori multipli come nei blocchi di Jordan, la mappa $A \mapsto \lambda(A)$ non è differenziabile. In particolare, è solo Lipschitz ma non di classe C^1. Questo significa che esiste una costante C tale che:

$$|\delta\lambda| \leq C\|\delta A\|_2$$

ma la derivata non esiste in senso classico. □

Enunciamo ora un risultato che limita la distanza tra gli autovalori di A e $A + \delta A$.

Teorema (Bauer-Fike) *Sia $A \in \mathbb{C}^{n\times n}$ una matrice diagonalizzabile con matrice di autovettori V:*

$$V^{-1}AV = D = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}.$$

Allora, per ogni $\delta A \in \mathbb{C}^{n\times n}$ e autovalore μ di $A+\delta A$, esiste un autovalore λ_i che soddisfa $|\lambda_i - \mu| \leq \kappa(V)\|\delta A\|$, dove $\|\cdot\|$ è una qualsiasi norma matriciale subordinata indotta da una norma assoluta[1]*.*

Dimostrazione. Sia $\mu \in \Lambda(A+\delta A)$; se μ è autovalore di A, il teorema è banalmente vero, altrimenti consideriamo la matrice singolare

$$V^{-1}(A + \delta A - \mu I)V = (D - \mu I) + V^{-1}\delta AV.$$

infatti poiché μ è autovalore di $A + \delta A$, la matrice $A + \delta A - \mu I$ è singolare. Moltiplicando per V^{-1} a sinistra e V a destra otteniamo una matrice ancora singolare.

Grazie alla non singolarità di $D - \mu I$ (poiché μ non è autovalore di A), possiamo fattorizzare:

$$V^{-1}(A + \delta A - \mu I)V = (D - \mu I)\left[I + (D - \mu I)^{-1}V^{-1}\delta AV\right]$$

abbiamo scritto la matrice come prodotto di due matrici. Poiché il prodotto è singolare e $D - \mu I$ è invertibile, deve essere singolare il secondo fattore.

Quindi $I+(D-\mu I)^{-1}V^{-1}\delta AV$ è singolare, e pertanto -1 appartiene allo spettro di $(D-\mu I)^{-1}V^{-1}\delta AV$. Per il teorema di Hirsch (che lega il raggio spettrale alla norma), abbiamo:

$$1 \leq \rho\left((D - \mu I)^{-1}V^{-1}\delta AV\right) \leq \|(D - \mu I)^{-1}V^{-1}\delta AV\|$$

Il raggio spettrale è sempre minore o uguale alla norma, e se -1 è autovalore, allora il raggio spettrale è almeno 1.

Maggioriamo ulteriormente usando le proprietà delle norme subordinate:

$$\|(D - \mu I)^{-1}V^{-1}\delta AV\| \leq \|(D - \mu I)^{-1}\| \cdot \|V^{-1}\| \cdot \|\delta A\| \cdot \|V\|$$

Ricordando che $\kappa(V) = \|V\| \cdot \|V^{-1}\|$, otteniamo:

$$1 \leq \|(D - \mu I)^{-1}\| \cdot \kappa(V) \cdot \|\delta A\|$$

[1] Una norma assoluta è una norma per cui la proprietà componente per componente $|x_i| > |y_i|$ implica $\|x\| > \|y\|$. Per tali norme abbiamo $\|D\| = \max_i |d_{ii}|$ per ogni matrice diagonale D.

Poiché $\|\cdot\|$ è una norma subordinata assoluta, per una matrice diagonale $D - \mu I$ vale:

$$\|(D-\mu I)^{-1}\| = \max_{i=1,\dots,n} \frac{1}{|\lambda_i - \mu|} = \frac{1}{\min_{i=1,\dots,n} |\lambda_i - \mu|}$$

Sostituendo nell'equazione precedente:

$$1 \leq \frac{1}{\min_{i=1,\dots,n} |\lambda_i - \mu|} \cdot \kappa(V) \cdot \|\delta A\|$$

che conclude la dimostrazione. □

Applicando il teorema di Bauer-Fike a una matrice normale con la norma spettrale si ottiene il limite superiore

$$|\lambda_i - \mu| \leq \|\delta A\|_2,$$

poiché le matrici normali sono diagonalizzate da matrici unitarie o ortogonali con numero di condizionamento uguale a 1. Questo risultato è più forte del fatto che il numero di condizionamento per tali matrici è uguale a 1, poiché non è coinvolta alcuna approssimazione del primo ordine.

Definizione Sia $\lambda \in \mathbb{C}$, e $v \in \mathbb{C}^n$. L'*errore all'indietro di* λ, v *come autocoppia di* A è definito come

$$BE(A, \lambda, v) := \min\{\|\delta A\| \mid (A + \delta A)v = \lambda v\}.$$

Analogamente, l'*errore all'indietro di* λ *come autovalore di* A è definito come

$$BE(A, \lambda) := \min\{\|\delta A\| \mid \lambda \in \Lambda(A + \delta A)\}.$$

Chiaramente, abbiamo $BE(A, \lambda) \leq BE(A, \lambda, v)$, per ogni scelta di v. L'errore all'indietro della coppia eigen può essere facilmente calcolato a posteriori, in contrasto con l'errore in avanti.

Teorema Sia $A \in \mathbb{C}^{n\times n}$ una matrice quadrata, e λ, v una coppia eigen candidata. Allora, per la norma spettrale $\|\cdot\|_2$,

$$BE_2(A, \lambda, v) = \frac{\|Av - \lambda v\|_2}{\|v\|_2}.$$

Dimostrazione. La dimostrazione procede in due parti: prima mostriamo una disuguaglianza, poi dimostriamo che è raggiungibile. Sia δA una perturbazione qualsiasi di A tale che λ e v siano autovalore e autovettore di $A + \delta A$. Allora abbiamo:

$$(A + \delta A)v = \lambda v \implies Av - \lambda v = -\delta A v$$

Prendendo le norme e usando le proprietà delle norme subordinate:

$$\|Av - \lambda v\|_2 = \|\delta A v\|_2 \leq \|\delta A\|_2 \cdot \|v\|_2$$

Da cui ricaviamo:

$$\|\delta A\|_2 \geq \frac{\|Av - \lambda v\|_2}{\|v\|_2}$$

Poiché questo vale per ogni δA che soddisfa la condizione, abbiamo:

$$BE_2(A, \lambda, v) \geq \frac{\|Av - \lambda v\|_2}{\|v\|_2}$$

Definiamo il residuo: $r := Av - \lambda v$ e consideriamo la perturbazione $\delta A := -\frac{rv^*}{\|v\|_2^2}$.

Verifichiamo che questa perturbazione funziona:

$$\begin{aligned}
(A + \delta A)v &= Av + \delta Av \\
&= Av - \frac{rv^*}{\|v\|_2^2}v \\
&= Av - r\frac{v^*v}{\|v\|_2^2} \\
&= Av - r\frac{\|v\|_2^2}{\|v\|_2^2} \quad (\ v^*v = \|v\|_2^2) \\
&= Av - r \\
&= Av - (Av - \lambda v) \\
&= \lambda v
\end{aligned}$$

Quindi λ, v è effettivamente una coppia eigen di $A + \delta A$.

Calcoliamo ora la norma di δA:

$$\|\delta A\|_2 = \left\| -\frac{rv^*}{\|v\|_2^2} \right\|_2 = \frac{\|rv^*\|_2}{\|v\|_2^2}$$

Per una matrice di rango 1 della forma xy^*, la norma spettrale è $\|xy^*\|_2 = \|x\|_2\|y\|_2$. Sostituendo:

$$\|\delta A\|_2 = \frac{\|r\|_2\|v\|_2}{\|v\|_2^2} = \frac{\|r\|_2}{\|v\|_2} = \frac{\|Av - \lambda v\|_2}{\|v\|_2}$$

Abbiamo quindi costruito una perturbazione δA che raggiunge esattamente il valore $\frac{\|Av-\lambda v\|_2}{\|v\|_2}$, dimostrando che:

$$BE_2(A, \lambda, v) = \frac{\|Av - \lambda v\|_2}{\|v\|_2}$$

□

Una caratterizzazione simile può essere enunciata per l'errore all'indietro di un autovalore.

Teorema *Sia $A \in \mathbb{C}^{n\times n}$ una matrice quadrata, e λ un autovalore candidato. Allora, per la norma spettrale $\|\cdot\|_2$, abbiamo*

$$\mathrm{BE}_2(A,\lambda) = \|(A-\lambda I)^{-1}\|_2^{-1}, \qquad \forall\ \lambda \notin \Lambda(A)$$

Dimostrazione. La dimostrazione procede in due parti.

Sia δA una perturbazione tale che $(A+\delta A)v = \lambda v$ per qualche $v \neq 0$. Allora:

$$(A+\delta A)v = \lambda v \implies (A-\lambda I)v = -\delta A v$$

Poiché $\lambda \notin \Lambda(A)$, la matrice $A-\lambda I$ è invertibile, quindi:

$$v = -(A-\lambda I)^{-1}\delta A v$$

Prendendo le norme:

$$\|v\|_2 = \|(A-\lambda I)^{-1}\delta A v\|_2 \leq \|(A-\lambda I)^{-1}\|_2 \cdot \|\delta A\|_2 \cdot \|v\|_2$$

Dividendo entrambi i membri per $\|v\|_2$ (che è non nullo):

$$1 \leq \|(A-\lambda I)^{-1}\|_2 \cdot \|\delta A\|_2$$

Da cui:

$$\|\delta A\|_2 \geq \frac{1}{\|(A-\lambda I)^{-1}\|_2} = \|(A-\lambda I)^{-1}\|_2^{-1}$$

Poiché questo vale per ogni δA tale che $\lambda \in \Lambda(A+\delta A)$, abbiamo:

$$\mathrm{BE}_2(A,\lambda) \geq \|(A-\lambda I)^{-1}\|_2^{-1}$$

Consideriamo adesso v e w tali che:

$$(A-\lambda I)^{-1}v = w, \quad \|v\|_2 = \|(A-\lambda I)^{-1}\|_2^{-1}, \quad \|w\|_2 = 1$$

Tali vettori esistono perché $\|(A-\lambda I)^{-1}\|_2 = \max_{\|x\|_2=1}\|(A-\lambda I)^{-1}x\|_2$, quindi il massimo è raggiunto.

Da $(A-\lambda I)^{-1}v = w$ ricaviamo:

$$(A - \lambda I)w = v$$

Ora consideriamo l'errore all'indietro per la coppia (λ, w):

$$\mathrm{BE}_2(A, \lambda, w) = \frac{\|(A - \lambda I)w\|_2}{\|w\|_2} = \frac{\|v\|_2}{1} = \|(A - \lambda I)^{-1}\|_2^{-1}$$

Ma per definizione abbiamo:

$$\mathrm{BE}_2(A, \lambda) \leq \mathrm{BE}_2(A, \lambda, w)$$

poiché l'errore all'indietro per l'autovalore è il minimo su tutti i possibili autovettori. Quindi:

$$\mathrm{BE}_2(A, \lambda) \leq \|(A - \lambda I)^{-1}\|_2^{-1}$$

Combinando le due disuguaglianze, otteniamo l'uguaglianza

$$\mathrm{BE}_2(A, \lambda) = \|(A - \lambda I)^{-1}\|_2^{-1}$$

□

Abbiamo enfatizzato come trasformare un problema agli autovalori in uno di ricerca delle radici di un polinomio sia generalmente una cattiva idea. L'alternativa più naturale che perseguiremo presto è costruire una sequenza di matrici

$$A_0 := A \to A_1 := F(A_0) \to \ldots \to A_{k+1} = F(A_k) \to \ldots$$

tale che tutte le matrici siano simili, $\lim_k A_k$ sia calcolabile con sufficiente accuratezza, e gli autovalori possano essere letti dal limite. Per esempio, possiamo chiedere che il limite sia triangolare superiore o diagonale. Affinché tutto questo funzioni, dobbiamo assicurarci che la trasformazione $A_{k+1} = F(A_k)$ non peggiori il numero di condizionamento degli autovalori. Non tutte le similitudini sono adatte allo scopo, ma questo è vero quando usiamo matrici unitarie o ortogonali.

Esercizio Dimostrare che se Q è unitaria, allora i numeri di condizionamento per gli autovalori di A e QAQ^* coincidono, cioè

$$\mathrm{BE}_2(A, \lambda) = \mathrm{BE}_2(QAQ^*, \lambda) \quad \text{e} \quad \mathrm{BE}_2(A, \lambda, v) = \mathrm{BE}_2(QAQ^*, \lambda, Qv)$$

Soluzione. Dimostriamo separatamente le due uguaglianze.
Per il Teorema sappiamo che per $\lambda \notin \Lambda(A)$:

$$\mathrm{BE}_2(A, \lambda) = \|(A - \lambda I)^{-1}\|_2^{-1}$$

Calcoliamo ora $\mathrm{BE}_2(QAQ^*, \lambda)$:

$$\begin{aligned}
\mathrm{BE}_2(QAQ^*, \lambda) &= \|(QAQ^* - \lambda I)^{-1}\|_2^{-1} \\
&= \|(QAQ^* - \lambda QIQ^*)^{-1}\|_2^{-1} \quad (\text{poiché } QIQ^* = I) \\
&= \|Q(A - \lambda I)^{-1}Q^*\|_2^{-1} \\
&= \|(A - \lambda I)^{-1}\|_2^{-1} \quad (\text{per l'invarianza unitaria della norma 2})
\end{aligned}$$

Quindi $\mathrm{BE}_2(A,\lambda) = \mathrm{BE}_2(QAQ^*,\lambda)$.

$$\mathrm{BE}_2(A,\lambda,v) = \frac{\|Av-\lambda v\|_2}{\|v\|_2}$$

Calcoliamo ora $\mathrm{BE}_2(QAQ^*,\lambda,Qv)$:

$$\begin{aligned}\mathrm{BE}_2(QAQ^*,\lambda,Qv) &= \frac{\|(QAQ^*)(Qv)-\lambda(Qv)\|_2}{\|Qv\|_2} = \frac{\|QA(Q^*Q)v-\lambda Qv\|_2}{\|Qv\|_2} = \frac{\|Q(Av-\lambda v)\|_2}{\|Qv\|_2}\\ &= \frac{\|Av-\lambda v\|_2}{\|v\|_2} \quad \text{(per l'invarianza unitaria della norma 2)}\\ &= \mathrm{BE}_2(A,\lambda,v)\end{aligned}$$

□

Osservazione Questo risultato è molto importante perché giustifica l'uso di trasformazioni unitarie negli algoritmi per il calcolo degli autovalori. Le trasformazioni unitarie preservano il numero di condizionamento degli autovalori, a differenza di trasformazioni di similitudine più generali che potrebbero peggiorare la stabilità numerica.

3.2 Il metodo delle potenze

Introduciamo il primo metodo per il calcolo degli autovalori: il metodo delle potenze. Sia A una matrice qualsiasi. Consideriamo la successione di vettori definita, per qualsiasi scelta di v_0, come segue:

$$v^{(k+1)} = \frac{Av^{(k)}}{\|Av^{(k)}\|_2}, \quad k \geq 0, \quad \lambda_k = (v^{(k)})^* Av^{(k)} \quad v^{(0)} \text{ assegnato.} \tag{3.2}$$

A meno del fattore di normalizzazione, il vettore $v^{(k)}$ soddisfa $v^{(k)} = A^k v^{(0)}$.
Sotto opportune condizioni, i termini $(\lambda_k, v^{(k)})$ convergono a una coppia eigen dominante di A. Aggiungiamo l'ipotesi che A sia diagonalizzabile, con autovalori

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|.$$

Allora, possiamo riscrivere l'iterazione come segue

$$w^{(k)} = \gamma_k D^k w^{(0)}, \qquad D := \begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{bmatrix}, \qquad \gamma_k := \frac{1}{\|VD^k w^{(0)}\|}$$

Qui stiamo diagonalizzando $A = VDV^{-1}$ e definendo $w^{(k)} = V^{-1}v^{(k)}$. Il vettore $w^{(k)}$ rappresenta le coordinate di $v^{(k)}$ nella base degli autovettori. Questo produce la seguente espressione esplicita per $w^{(k)}$:

$$w^{(k)} = \gamma_k \lambda_1^k \begin{bmatrix} w_1^{(0)} \\ \left(\frac{\lambda_2}{\lambda_1}\right)^k w_2^{(0)} \\ \vdots \\ \left(\frac{\lambda_n}{\lambda_1}\right)^k w_n^{(0)} \end{bmatrix}$$

In particolare

- Il fattore λ_1^k cresce più rapidamente di tutti gli altri poiché $|\lambda_1| > |\lambda_i|$ per $i > 1$
- I termini $\left(\frac{\lambda_i}{\lambda_1}\right)^k$ tendono a 0 per $k \to \infty$ poiché $|\frac{\lambda_i}{\lambda_1}| < 1$
- Solo la prima componente $w_1^{(k)}$ sopravvive asintoticamente
- Il fattore γ_k normalizza il vettore mantenendo $\|v^{(k)}\|_2 = 1$

Poiché γ_k è scelto per normalizzare $v^{(k)}$, abbiamo che se $w_1^{(0)} \neq 0$ tutte le componenti in $w^{(k)}$ tendono a zero per $k \to \infty$, e $w^{(k)}$ converge a un multiplo di e_1 con velocità $(\frac{\lambda_2}{\lambda_1})^k$. Poiché $v^{(k)} = Vw^{(k)}$, concludiamo che $v^{(k)}$ converge a un autovettore relativo a λ_1, e conseguentemente $\lambda^{(k)} = (v^{(k)})^* A v^{(k)}$ converge a λ_1 con la stessa velocità lineare:

$$\lim_{k\to\infty} \lambda_k = \lim_{k\to\infty} \frac{(v^{(k)})^* A v^{(k)}}{(v^{(k)})^* v^{(k)}} = \frac{v_1^*(\overline{w}_1^{(0)} A v_1 w_1^{(0)})}{\overline{w}_1^{(0)} w_1^{(0)} v_1^* v_1} = \lambda_1$$

Si noti che la condizione $w_1^{(0)} \neq 0$ è generica, nel senso che se scegliamo $v^{(0)}$ a caso (rispetto a qualsiasi misura di probabilità assolutamente continua) allora abbiamo $w_1^{(0)} \neq 0$ con probabilità 1. In teoria, se scegliamo $v^{(0)}$ tale che $w_1^{(0)} = 0$, questa condizione dovrebbe continuare a valere durante le iterazioni. Tuttavia, lavorando in aritmetica floating point si introdurranno perturbazioni che ci riporteranno al caso generico.

3.3 Velocità di convergenza del metodo delle potenze

Analizziamo formalmente la convergenza dell'iterazione delle potenze rispetto all'autovettore dominante v_1. Si noti che, anche nel caso in cui λ_1 sia semplice, l'autovettore dominante è definito a meno di una costante; in particolare, ha poco senso misurare quantità come $\|v_1 - v^{(k)}\|$ poiché, per esempio, se $v^{(k)} \to -v_1$ non rileveremmo alcuna convergenza. Il punto chiave è quantificare quanto sono collineari i vettori v_1 e $v^{(k)}$, e questo richiede di introdurre funzioni trigonometriche di un angolo tra due vettori.

Definizione Dati $x, y \in \mathbb{C}^n$, $x \neq 0$, definiamo le proiezioni ortogonali su span(x) e sul suo complemento come

$$\Pi_x(y) = \frac{xx^*}{\|x\|_2^2} y, \qquad \Pi_x^\perp(y) = \left(I - \frac{xx^*}{\|x\|_2^2}\right) y.$$

Inoltre definiamo il sin, cos e tan dell'angolo tra x e y come segue:

$$\sin\theta(x,y) = \frac{\|\Pi_x^{\perp}(y)\|_2}{\|y\|_2} = \frac{\min_{z\in\mathrm{span}(x)}\|y-z\|_2}{\|y\|_2},$$
$$\cos\theta(x,y) = \frac{\|\Pi_x(y)\|_2}{\|y\|_2} = \frac{|x^*y|}{\|x\|_2\|y\|_2},$$
$$\tan\theta(x,y) = \frac{\sin\theta(x,y)}{\cos\theta(x,y)} = \frac{\|\Pi_x^{\perp}(y)\|_2}{\|\Pi_x(y)\|_2}.$$

Osservazione Vale $\sin\theta(x,y)^2 + \cos\theta(x,y)^2 = 1$, $\forall x, y \in \mathbb{C}^n \setminus \{0\}$.
Osservazione Le funzioni trigonometriche per vettori sono commutative rispetto ai due ingressi, invarianti per riscalamento, e non cambiano se applichiamo la stessa matrice unitaria a entrambi x e y. In particolare, quando analizziamo la convergenza del metodo delle potenze possiamo considerare l'iterazione semplificata $v^{(k)} = A^k v^{(0)}$ poiché il passo di normalizzazione non ha influenza sulla collinearità dell'iterata rispetto a v_1.

Prima di enunciare il risultato principale, assumiamo che A sia diagonalizzabile e che $V^{-1}AV = D := \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Quindi, consideriamo la sequenza ausiliaria

$$y^{(0)} = V^{-1}v^{(0)}, \qquad y^{(k)} = Dy^{(k-1)} = D^k y^{(0)} \qquad \Longrightarrow \qquad y^{(k)} = V^{-1}v^{(k)}$$

e analizziamo quanto è collineare $y^{(k)}$ rispetto a $e_1 = V^{-1}v_1$, che è l'autovettore dominante per D. Partizionando a blocchi $y^{(k)}$ e D come

$$y^{(k)} = \begin{bmatrix} y_1^{(k)} \\ y_2^{(k)} \end{bmatrix}, \qquad D = \begin{bmatrix} \lambda_1 & \\ & D_2 \end{bmatrix}, \qquad y_1^{(k)} \in \mathbb{C}, \quad y_2^{(k)} \in \mathbb{C}^{n-1}$$

e analizziamo quanto è collineare $y^{(k)}$ rispetto a $e_1 = V^{-1}v_1$, che è l'autovettore dominante per D. Partizionando a blocchi $y^{(k)}$ e D come

$$y^{(k)} = \begin{bmatrix} y_1^{(k)} \\ y_2^{(k)} \end{bmatrix}, \qquad D = \begin{bmatrix} \lambda_1 & \\ & D_2 \end{bmatrix}, \qquad y_1^{(k)} \in \mathbb{C}, \quad y_2^{(k)} \in \mathbb{C}^{n-1},$$

vediamo che

$$y^{(k)} = D^k y^{(0)} = \begin{bmatrix} \lambda_1^k y_1^{(0)} \\ D_2^k y_2^{(0)} \end{bmatrix} = \lambda_1^k \begin{bmatrix} y_1^{(0)} \\ \left(\frac{D_2}{\lambda_1}\right)^k y_2^{(0)} \end{bmatrix}.$$

Inoltre, abbiamo $\| \left(\frac{D_2}{\lambda_1}\right)^k \|_2 = \left|\frac{\lambda_2}{\lambda_1}\right|^k$ (poiché D_2 è diagonale) e

$$\Pi_{e_1}^{\perp}(y^{(k)}) = \begin{bmatrix} 0 \\ y_2^{(k)} \end{bmatrix}, \qquad \Pi_{e_1}(y^{(k)}) = \begin{bmatrix} y_1^{(k)} \\ 0 \end{bmatrix}.$$

Mettendo tutto insieme, abbiamo

$$\tan\theta(e_1, y^{(k)}) = \frac{\|y_2^{(k)}\|_2}{\left|y_1^{(k)}\right|} \leq \left|\frac{\lambda_2}{\lambda_1}\right|^k \frac{\|y_2^{(0)}\|_2}{\left|y_1^{(0)}\right|} = \left|\frac{\lambda_2}{\lambda_1}\right|^k \tan\theta(e_1, y^{(0)}). \tag{3.3}$$

Siamo pronti per enunciare il risultato principale sulla convergenza del metodo delle potenze.
Teorema *Sia $A \in \mathbb{C}^{n\times n}$ diagonalizzabile con matrice di autovettori V, e autovalore dominante λ_1 tale che $|\lambda_1| > |\lambda_2|$. Se $v^{(0)} \in \mathbb{C}^n$ è tale che $u_1^* v^{(0)} \neq 0$, per un autovettore sinistro dominante u_1, allora la k-esima iterata del metodo delle potenze, partendo da $v^{(0)}$, verifica*

$$\sin\theta(v_1, v^{(k)}) \leq \kappa(V) \left|\frac{\lambda_2}{\lambda_1}\right|^k \frac{\sin\theta(v_1, v^{(0)})}{\cos\theta(e_1, V^{-1}v^{(0)})}.$$

Dimostrazione. Notiamo che $u_1^* v^{(0)} \neq 0$ implica $\cos\theta(e_1, y^{(0)}) \neq 0$, quindi il membro destro di (3.3) è ben definito. Dalla disuguaglianza (3.3) abbiamo:

$$\tan\theta(e_1, y^{(k)}) \leq \left|\frac{\lambda_2}{\lambda_1}\right|^k \tan\theta(e_1, y^{(0)}).$$

Osserviamo che per qualsiasi angolo θ, vale $\sin\theta \leq \tan\theta$, quindi:

$$\sin\theta(e_1, y^{(k)}) \leq \tan\theta(e_1, y^{(k)}) \leq \left|\frac{\lambda_2}{\lambda_1}\right|^k \tan\theta(e_1, y^{(0)}). \tag{1}$$

Ora consideriamo $\sin\theta(v_1, v^{(k)})$. Poiché $v_1 = Ve_1$ e $v^{(k)} = Vy^{(k)}$, abbiamo:

$$\sin\theta(v_1, v^{(k)}) = \sin\theta(Ve_1, Vy^{(k)}) = \frac{\min_{z\in\mathrm{span}(y^{(k)})} \|Ve_1 - Vz\|_2}{\|Ve_1\|_2}.$$

Per qualsiasi $z \in \mathrm{span}(y^{(k)})$, possiamo maggiorare:

$$\|Ve_1 - Vz\|_2 = \|V(e_1 - z)\|_2 \leq \|V\|_2 \|e_1 - z\|_2.$$

Prendendo il minimo su $z \in \mathrm{span}(y^{(k)})$:

$$\min_{z\in\mathrm{span}(y^{(k)})} \|Ve_1 - Vz\|_2 \leq \|V\|_2 \min_{z\in\mathrm{span}(y^{(k)})} \|e_1 - z\|_2 = \|V\|_2 \|\Pi_{y^{(k)}}^{\perp}(e_1)\|_2.$$

Ma $\|\Pi_{y^{(k)}}^{\perp}(e_1)\|_2 = \sin\theta(e_1, y^{(k)})\|e_1\|_2 = \sin\theta(e_1, y^{(k)})$, quindi:

$$\sin\theta(v_1, v^{(k)}) \leq \frac{\|V\|_2}{\|Ve_1\|_2} \sin\theta(e_1, y^{(k)}). \tag{2}$$

Combinando (1) e (2):

$$\sin\theta(v_1, v^{(k)}) \leq \frac{\|V\|_2}{\|Ve_1\|_2} \left|\frac{\lambda_2}{\lambda_1}\right|^k \tan\theta(e_1, y^{(0)}).$$

Osserviamo che:

$$\tan\theta(e_1, y^{(0)}) = \frac{\sin\theta(e_1, y^{(0)})}{\cos\theta(e_1, y^{(0)})}.$$

Inoltre

$$\sin\theta(e_1, y^{(0)}) = \frac{\min_{z\in\text{span}(y^{(0)})} \|e_1 - z\|_2}{\|e_1\|_2} = \min_{z\in\text{span}(y^{(0)})} \|e_1 - z\|_2.$$

Ma

$$\min_{z\in\text{span}(y^{(0)})} \|e_1 - z\|_2 = \min_{z\in\text{span}(y^{(0)})} \|V^{-1}(Ve_1 - Vz)\|_2 \leq \|V^{-1}\|_2 \min_{z\in\text{span}(y^{(0)})} \|Ve_1 - Vz\|_2.$$

E

$$\min_{z\in\text{span}(y^{(0)})} \|Ve_1 - Vz\|_2 = \sin\theta(v_1, v^{(0)})\|Ve_1\|_2,$$

quindi

$$\sin\theta(e_1, y^{(0)}) \leq \|V^{-1}\|_2 \sin\theta(v_1, v^{(0)})\|Ve_1\|_2. \tag{3}$$

Sostituendo (3) nell'espressione precedente:

$$\sin\theta(v_1, v^{(k)}) \leq \frac{\|V\|_2}{\|Ve_1\|_2} \left|\frac{\lambda_2}{\lambda_1}\right|^k \frac{\|V^{-1}\|_2 \sin\theta(v_1, v^{(0)})\|Ve_1\|_2}{\cos\theta(e_1, y^{(0)})}.$$

Semplificando $\|Ve_1\|_2$:

$$\sin\theta(v_1, v^{(k)}) \leq \|V\|_2\|V^{-1}\|_2 \left|\frac{\lambda_2}{\lambda_1}\right|^k \frac{\sin\theta(v_1, v^{(0)})}{\cos\theta(e_1, y^{(0)})}.$$

Ricordando che $\kappa(V) = \|V\|_2\|V^{-1}\|_2$ e che $\cos\theta(e_1, y^{(0)}) = \cos\theta(e_1, V^{-1}v^{(0)})$, otteniamo il risultato desiderato. □

Osservazione Alcune osservazioni sulle ipotesi del teorema precedente:

- A diagonalizzabile può essere rilassata assumendo λ_1 semplice.

- Anche $|\lambda_1| > |\lambda_2|$ non può essere rimossa; si consideri per esempio il caso $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ e un vettore iniziale che non è allineato né con $v_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ né con $v_2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$.

3.4 Il caso hermitiano

Nel caso in cui A è hermitiana possiamo mostrare che l'approssimante dell'autovalore dominante calcolato dal metodo delle potenze converge con il doppio del tasso di decadimento rispetto al caso generale. è istruttivo guardare la funzione quoziente di Rayleigh $\rho_A(x) = \frac{x^*Ax}{x^*x}$ che è tale che $\rho_A(v_1) = \lambda_1$. Se guardiamo il gradiente (considerando ρ_A come una funzione su vettori reali) abbiamo

Consideriamo $\rho_A(x) = \frac{x^*Ax}{x^*x}$. Calcoliamo il gradiente rispetto a x (considerando $x \in \mathbb{R}^n$ per semplicità).

Sia $N(x) = x^*Ax$ e $D(x) = x^*x$. Allora:

$$\nabla N(x) = (A + A^*)x, \quad \nabla D(x) = 2x$$

Usando la regola del quoziente:

$$\nabla \rho_A(x) = \frac{D(x)\nabla N(x) - N(x)\nabla D(x)}{[D(x)]^2}$$

Sostituendo:

$$\nabla \rho_A(x) = \frac{(x^*x)(A + A^*)x - (x^*Ax)(2x)}{(x^*x)^2}$$

$$\nabla \rho_A(x) = \frac{1}{x^*x}\left[(A + A^*)x - 2\rho_A(x)x\right]$$

In particolare, quando A è hermitiana v_1 (e qualsiasi altro autovettore) è un punto stazionario per ρ_A mentre non lo è quando $A \neq A^*$. Infatti, se $A = A^*$ e $x = v_1$ (autovettore):

$$\nabla \rho_A(v_1) = \frac{1}{v_1^*v_1}\left[2Av_1 - 2\lambda_1 v_1\right] = \frac{1}{v_1^*v_1}\left[2\lambda_1 v_1 - 2\lambda_1 v_1\right] = 0$$

Quindi, guardando lo sviluppo di Taylor di $\rho_A(x)$ abbiamo

$$|\rho_A(x) - \lambda_1| = |\rho_A(x) - \rho_A(v_1)| = \begin{cases} \mathcal{O}(\|x - v_1\|_2^2) & \text{se } A \text{ è hermitiana} \\ \mathcal{O}(\|x - v_1\|_2) & \text{altrimenti} \end{cases}.$$

Più formalmente, dimostriamo il seguente risultato.

Teorema Sia $A \in \mathbb{C}^{n\times n}$ hermitiana con autovalori $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n| \geq 0, v^{(0)} \in \mathbb{C}^n$ tale che $v_1^* v^{(0)} \neq 0$. Allora, la k-esima iterata del metodo delle potenze, partendo da $v^{(0)}$, verifica

$$\tan\theta(v_1, v^{(k)}) \leq \left|\frac{\lambda_2}{\lambda_1}\right|^k \tan\theta(v_1, v^{(0)}),$$

$$|\lambda_1 - \rho_A(v^{(k)})| \leq \max_{j=1,\ldots,n} |\lambda_1 - \lambda_j| \cdot \left|\frac{\lambda_2}{\lambda_1}\right|^{2k} [\tan\theta(v_1, v^{(0)})]^2$$

Dimostrazione. La disuguaglianza riguardante la convergenza dell'autovettore segue da (3.3) applicando una matrice unitaria di autovettori V ai vettori coinvolti nelle funzioni trigonometriche in entrambi i membri.
Per mostrare la seconda disuguaglianza assumiamo che il passo di normalizzazione nel metodo delle potenze non venga eseguito e che il vettore iniziale $v^{(0)}$ sia riscalato per ottenere $\|v^{(k)}\|_2 = 1$; si noti che, tutte queste assunzioni non causano perdita di generalità poiché il quoziente di Rayleigh è invariante per riscalamento (non nullo) dell'argomento. Sia $v^{(0)} = \sum_{j=1}^n a_j v_j$, allora abbiamo

$$v^{(k)} = A^k v^{(0)} = \sum_{j=1}^n a_j \lambda_j^k v_j$$

Calcoliamo il quoziente di Rayleigh:

$$\rho_A(v^{(k)}) = (v^{(k)})^* A v^{(k)} = \frac{(v^{(k)})^* A v^{(k)}}{(v^{(k)})^* v^{(k)}}$$

Sostituendo le espressioni:

$$(v^{(k)})^* A v^{(k)} = \left(\sum_{j=1}^n a_j \lambda_j^k v_j\right)^* A \left(\sum_{i=1}^n a_i \lambda_i^k v_i\right) = \sum_{j,i=1}^n a_j^* a_i \lambda_j^k \lambda_i^k v_j^* A v_i$$

Poiché $Av_i = \lambda_i v_i$ e $v_j^* v_i = \delta_{ij}$ (autovettori ortonormali per matrici hermitiane):

$$(v^{(k)})^* A v^{(k)} = \sum_{j=1}^n |a_j|^2 \lambda_j^{2k+1}$$

Analogamente

$$(v^{(k)})^* v^{(k)} = \sum_{j=1}^n |a_j|^2 \lambda_j^{2k}$$

Quindi

$$\rho_A(v^{(k)}) = \frac{\sum_{j=1}^n |a_j|^2 \lambda_j^{2k+1}}{\sum_{j=1}^n |a_j|^2 \lambda_j^{2k}}$$

In modo che

$$|\lambda_1 - \rho_A(v^{(k)})| = \left| \frac{\sum_{j=2}^n a_j^2 \lambda_j^{2k}(\lambda_j - \lambda_1)}{\sum_{j=1}^n a_j^2 \lambda_j^{2k}} \right|$$

infatti se sottraiamo λ_1 da entrambi i membri

$$\lambda_1 - \rho_A(v^{(k)}) = \lambda_1 - \frac{\sum_{j=1}^n a_j^2 \lambda_j^{2k+1}}{\sum_{j=1}^n a_j^2 \lambda_j^{2k}} = \frac{\sum_{j=1}^n a_j^2 \lambda_j^{2k} \lambda_1 - \sum_{j=1}^n a_j^2 \lambda_j^{2k+1}}{\sum_{j=1}^n a_j^2 \lambda_j^{2k}} = \frac{\sum_{j=1}^n a_j^2 \lambda_j^{2k}(\lambda_1 - \lambda_j)}{\sum_{j=1}^n a_j^2 \lambda_j^{2k}}$$

Ma per $j = 1$, il termine è zero ($\lambda_1 - \lambda_1 = 0$), quindi:

$$|\lambda_1 - \rho_A(v^{(k)})| = \left| \frac{\sum_{j=2}^n a_j^2 \lambda_j^{2k}(\lambda_1 - \lambda_j)}{\sum_{j=1}^n a_j^2 \lambda_j^{2k}} \right|$$

Ora maggioriamo

$$|\lambda_1 - \rho_A(v^{(k)})| \leq \frac{\sum_{j=2}^n |a_j|^2 |\lambda_j|^{2k} |\lambda_1 - \lambda_j|}{|a_1|^2 |\lambda_1|^{2k} + \sum_{j=2}^n |a_j|^2 |\lambda_j|^{2k}}$$

Poiché $|\lambda_j| \leq |\lambda_2|$ per $j \geq 2$, abbiamo

$$|\lambda_1 - \rho_A(v^{(k)})| \leq \frac{\sum_{j=2}^n |a_j|^2 |\lambda_2|^{2k} |\lambda_1 - \lambda_j|}{|a_1|^2 |\lambda_1|^{2k}} \leq \frac{\max_{j=1,\ldots,n} |\lambda_1 - \lambda_j|}{|a_1|^2} \sum_{j=2}^n |a_j|^2 \left| \frac{\lambda_2}{\lambda_1} \right|^{2k}$$

Osserviamo che

$$\tan\theta(v_1, v^{(0)}) = \frac{\|\Pi_{v_1}^{\perp}(v^{(0)})\|_2}{\|\Pi_{v_1}(v^{(0)})\|_2} = \frac{\sqrt{\sum_{j=2}^n |a_j|^2}}{|a_1|}$$

Quindi

$$[\tan\theta(v_1, v^{(0)})]^2 = \frac{\sum_{j=2}^n |a_j|^2}{|a_1|^2}$$

Sostituendo

$$|\lambda_1 - \rho_A(v^{(k)})| \leq \max_{j=1,\ldots,n} |\lambda_1 - \lambda_j| \left| \frac{\lambda_2}{\lambda_1} \right|^{2k} [\tan\theta(v_1, v^{(0)})]^2$$

□

3.5 Iterazione per sottospazi

Come generalizzazione naturale del metodo delle potenze, possiamo considerare l'iterazione su sottospazi invece che su vettori. Matematicamente, desideriamo selezionare un sottospazio iniziale $\mathcal{U}_0 \subseteq \mathbb{C}^n$, e poi costruire una sequenza di sottospazi come segue:

$$\mathcal{U}_{k+1} := A\mathcal{U}_k = \{Ax \mid x \in \mathcal{U}_k\}.$$

Nel caso dell'iterazione vettoriale, abbiamo convergenza a un autovettore; questo può essere reinterpretato come convergenza a una base di un sottospazio unidimensionale, ponendo $\mathcal{U}_k := \text{span}(v_k)$.

Per sottospazi di dimensione superiore, la convergenza a un autovettore è sostituita dalla convergenza a un sottospazio invariante. Ricordiamo che, dato un operatore lineare A, un sottospazio invariante è uno che soddisfa $A\mathcal{U} \subseteq \mathcal{U}$. Se U è una matrice le cui colonne generano $\mathcal{U}$, la proprietà di essere un sottospazio invariante di dimensione p può essere riformulata come

$$AU = UR, \qquad R \in \mathbb{C}^{p\times p}. \tag{3.4}$$

Si noti che se $Rw = \lambda w$ allora Uw è un autovettore relativo a λ per A:

$$AUw = URw = \lambda Uw \implies \lambda \in \Lambda(A).$$

Quindi, trovare un sottospazio invariante descritto come in (3.4) è utile per calcolare autovalori selezionati.

Non tutte le basi sono numericamente adatte per rappresentare sottospazi. Data una base U, abbiamo che qualsiasi vettore in $\mathcal{U}$ può essere scritto come $v = Uw$, dove w è il vettore delle coordinate nella base scelta:

$$v = w_1 u^{(1)} + \ldots + w_k u^{(k)} = \begin{bmatrix} u^{(1)} & \ldots & u^{(k)} \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_k \end{bmatrix}.$$

Dobbiamo assicurarci che piccole perturbazioni nei dati di input per questa rappresentazione (ad esempio, il vettore w), corrispondano a piccole variazioni nell'output (il vettore v). Una scelta naturale per raggiungere questo obiettivo è prendere U ortogonale. Ciò garantisce

$$\|U(w + \delta w) - Uw\|_2 = \|U\delta w\|_2 = \|\delta w\|_2,$$

grazie all'invarianza unitaria della norma euclidea.

Data una qualsiasi matrice base V di dimensione $n \times k$, possiamo sempre renderla ortogonale (o unitaria) calcolando una fattorizzazione QR economy-size:

$$V = QR \implies \text{colspan}(V) = \text{colspan}(Q)$$

che vale perché $\det(R) \neq 0$, poiché V ha rango massimo. La matrice Q è calcolata attraverso una sequenza di k riflettori di Householder, ciascuno dei quali annulla gli elementi sottodiagonali nella i-esima colonna. In dettaglio, iniziamo determinando un riflettore $P_1 = I - \beta_1 u_1 u_1^*$ tale che $P_1(Ve_1) = r_{11}e_1$, che produce

$$P_1 V = \begin{bmatrix} r_{11} & \times & \cdots & \times \\ 0 & \times & \cdots & \times \\ \vdots & \vdots & \cdots & \vdots \\ 0 & \times & \cdots & \times \end{bmatrix}.$$

La matrice P_1 è una perturbazione di rango 1 della matrice identità, quindi il costo computazionale di $P_1 V$ è $\mathcal{O}(nk)$ flop (operazioni in floating point). Poi, le colonne rimanenti possono essere ridotte in forma triangolare superiore calcolando matrici simili $P_2, \ldots, P_k$, con un costo computazionale totale di $\mathcal{O}(nk^2)$ (più precisamente, quando $k = n$ solo $k-1$ riflettori sono necessari, mentre k sono necessari in tutti gli altri casi).

Ora abbiamo tutti gli strumenti per descrivere l'iterazione per sottospazi, partendo da una base generica $n \times k$ per $U^{(0)}$. Il corrispondente pseudocodice è descritto dal seguente algoritmo.

1: **procedure** IterazioneSottospazi($A, U^{(0)}$)
2: **for** $k = 0, 1, \ldots$ **do**
3: $W^{(k+1)} \leftarrow AU^{(k)}$
4: $U^{(k+1)} R^{(k+1)} \leftarrow W^{(k+1)}$ (fattorizzazione QR)
5: $Y^{(k+1)} \leftarrow (U^{(k+1)})^* A U^{(k+1)}$
6: **end for**
7: **end procedure**

Quest'ultimo introduce la quantità $Y^{(k+1)} = (U^{(k+1)})^* A U^{(k+1)}$, che assume il ruolo del termine $(v^{(k)})^* A v^{(k)}$ che avevamo nel metodo delle potenze. Si osservi che se $U^{(k)}$ è una base per un sottospazio invariante, questo implica

$$AU^{(k)} = U^{(k)} Y^{(k)} \implies \Lambda(Y^{(k)}) \subseteq \Lambda(A),$$

con $U^{(k)} w$ che sono gli autovettori, se $Y^{(k)} w = \lambda w$. Quindi, quando A è grande, possiamo usare gli autovalori della matrice (piccola) $Y^{(k)}$ come approssimazione dei suoi autovalori (più grandi). Anche le approssimazioni agli autovettori sono così ottenute.

Un teorema di convergenza per l'iterazione per sottospazi richiederebbe l'angolo tra sottospazi,uno strumento che non abbiamo ancora introdotto. Quindi, ci limiteremo a comprendere la convergenza degli autovalori di $Y^{(k)}$ verso quelli di A, che dipende da λ_{p+1}/λ_p.

Teorema Sia A una matrice $n \times n$ diagonalizzabile, con $V^{-1} A V = D$, e $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Sia $U^{(0)} \in \mathbb{C}^{n \times p}$ una matrice con colonne ortogonali. Se gli autovalori, ordinati per magnitudine, soddisfano

$$|\lambda_1| \geq \cdots \geq |\lambda_p| > |\lambda_{p+1}| \geq \cdots \geq |\lambda_n|,$$

e $V^{-1} U^{(0)}$ ha un minore invertibile nelle prime p righe, allora l'iterazione per sottospazi definita nell'Algoritmo produce una sequenza di matrici $Y^{(k)}$ il cui spettro converge a $\{\lambda_1, \ldots, \lambda_p\}$ con velocità $(\lambda_{p+1}/\lambda_p)^k$.

Dimostrazione. La definizione di iterazione per sottospazi implica che l'iterata $U^{(k)}$ è una base ortogonale di $A^k U^{(0)}$. Se quest'ultima è una matrice a rango pieno, questo determina completamente lo spazio colonna di $U^{(k)}$.

Possiamo scrivere $A^k U^{(0)}$ sfruttando la diagonalizzazione di A, usando l'ipotesi sull'invertibilità della sottomatrice $p \times p$ in alto di $V^{-1} U^{(0)}$:

$$U^k = A^k U^{(0)} = V D^k V^{-1} U^{(0)} =: V D^k \begin{bmatrix} X_0 \\ X_1 \end{bmatrix}, \quad \det X_0 \neq 0.$$

Si noti che questo implica che $A^k U^{(0)}$ ha rango pieno per ogni k. Partizionando D come $D_1 \oplus D_2$, con D_1 contenente gli autovalori $\lambda_1, \ldots, \lambda_p$, otteniamo che

$$\text{colspan}(U^{(k)}) = \text{colspan}\left(V \begin{bmatrix} D_1^k X_0 \\ D_2^k X_1 \end{bmatrix} \right) = \text{colspan}\left(V \begin{bmatrix} I_p \\ D_2^k X_1 X_0^{-1} D_1^{-k} \end{bmatrix} \right),$$

dove abbiamo usato che $\text{colspan}(AB) = \text{colspan}(A)$ per qualsiasi matrice invertibile B. Studiamo $\|D_2^k X_1 X_0^{-1} D_1^{-k}\|$:

- D_2 contiene gli autovalori $\lambda_{p+1}, \ldots, \lambda_n$ con $|\lambda_i| < |\lambda_p|$
- D_1 contiene gli autovalori $\lambda_1, \ldots, \lambda_p$ con $|\lambda_i| \geq |\lambda_p|$
- Quindi: $\|D_2^k\|_2 \sim |\lambda_{p+1}|^k$ e $\|D_1^{-k}\|_2 \sim |\lambda_p|^{-k}$
- Il prodotto ha norma: $\|D_2^k X_1 X_0^{-1} D_1^{-k}\|_2 \leq \|D_2^k\|_2 \|X_1 X_0^{-1}\|_2 \|D_1^{-k}\|_2 \sim \mathcal{O}\left(\left| \frac{\lambda_{p+1}}{\lambda_p} \right|^k \right)$

Poiché $\left| \frac{\lambda_{p+1}}{\lambda_p} \right| < 1$ per ipotesi, $\|D_2^k X_1 X_0^{-1} D_1^{-k}\|$ converge a zero con velocità $(\lambda_{p+1}/\lambda_p)^k$. Dunque si ha intuitivamente che

$$\text{colspan}(U^{(k)}) \to \text{colspan}\left(V \begin{bmatrix} I_p \\ 0 \end{bmatrix} \right),$$

che sono gli autovettori relativi a $\lambda_1, \ldots, \lambda_p$. Formalizzare questa affermazione richiederebbe angoli tra sottospazi; ora dimostriamo l'affermazione sugli autovalori di $Y^{(k)}$.

Sia v_j l'autovettore per λ_j in A. Allora, definendo $w_j^{(k)} := (U^{(k)})^* v_j$ abbiamo:

$$\begin{aligned} Y^{(k)} w_j^{(k)} &= (U^{(k)})^* A U^{(k)} (U^{(k)})^* v_j \\ &= (U^{(k)})^* A [U^{(k)} (U^{(k)})^*] v_j \end{aligned}$$

Ora aggiungiamo e sottraiamo $(U^{(k)})^* A v_j$

$$\begin{aligned} &= (U^{(k)})^* A v_j - (U^{(k)})^* A v_j + (U^{(k)})^* A [U^{(k)} (U^{(k)})^*] v_j \\ &= (U^{(k)})^* (\lambda_j v_j) - (U^{(k)})^* A [v_j - U^{(k)} (U^{(k)})^* v_j] \\ &= \lambda_j (U^{(k)})^* v_j - (U^{(k)})^* A (I - U^{(k)} (U^{(k)})^*) v_j \end{aligned}$$

Quindi otteniamo

$$Y^{(k)} w_j^{(k)} = \lambda_j w_j^{(k)} - (U^{(k)})^* A (I - U^{(k)} (U^{(k)})^*) v_j.$$

Osservazioni

- Il termine $U^{(k)} (U^{(k)})^*$ è il proiettore ortogonale sullo spazio colonna di $U^{(k)}$

- $I - U^{(k)}(U^{(k)})^*$ è il proiettore ortogonale sul complemento ortogonale
- Il termine $(I - U^{(k)}(U^{(k)})^*)v_j$ rappresenta la componente di v_j ortogonale a $U^{(k)}$
- Quando $U^{(k)}$ si avvicina allo spazio degli autovettori, questo termine tende a zero

Prendendo le norme spettrali, possiamo maggiorare il residuo per la coppia eigen $\lambda_j, w_j^{(k)}$ come segue:

$$\|Y^{(k)}w_j^{(k)} - \lambda_j w_j^{(k)}\|_2 \leq \|A\|_2\|(I - U^{(k)}(U^{(k)})^*)v_j\|_2 = \|A\|_2 \min_{z\in\text{colspan } U^{(k)}} \|v_j - z\|_2,$$

dove nell'ultimo passaggio abbiamo usato la caratterizzazione della proiezione ortogonale come minimizzazione della norma euclidea della differenza. Possiamo fare una scelta esplicita per z, ponendo

$$z = V\begin{bmatrix} I_p \\ D_2^k X_1 X_0^{-1} D_1^{-k}\end{bmatrix} e_j \implies z - v_j = V\begin{bmatrix} 0_p \\ D_2^k X_1 X_0^{-1} D_1^{-k}\end{bmatrix} e_j.$$

Prendendo le norme, si ottiene la maggiorazione

$$\|Y^{(k)}w_j^{(k)} - \lambda_j w_j^{(k)}\|_2 \leq \|A\|_2\|V\|_2\|X_1X_0^{-1}\|_2\|D_2^k\|_2\|D_1^{-k}\|_2 \sim \mathcal{O}\left(\left|\frac{\lambda_{p+1}}{\lambda_p}\right|^k\right).$$

Quindi, λ_j è un autovalore approssimato di $Y^{(k)}$ con errore all'indietro maggiorato come sopra, grazie ad un precedente Teorema. La conclusione segue per un argomento di continuità dello spettro, combinato con il fatto che $Y^{(k)}$ è diagonalizzabile per k sufficientemente grande, e quindi la dipendenza è almeno di classe C^1. □

3.6 Iterazione simultanea

Un vantaggio chiave dell'iterazione per sottospazi è che, mentre si esegue l'algoritmo con sottospazi di dimensione p, si stanno in realtà eseguendo simultaneamente tutte le iterazioni per $p' = 1, \ldots, p$.

Si noti che, se W è una matrice alta e stretta, la sua fattorizzazione QR in forma economica contiene incorporate tutte le fattorizzazioni QR in forma economica per W' che includono le prime p' colonne di W:

$$W = QR \implies W\begin{bmatrix} I_{p'} \\ 0\end{bmatrix} = QR\begin{bmatrix} I_{p'} \\ 0\end{bmatrix} = \left(Q\begin{bmatrix} I_{p'} \\ 0\end{bmatrix}\right)\left(\begin{bmatrix} I_{p'} & 0\end{bmatrix} R\begin{bmatrix} I_{p'} \\ 0\end{bmatrix}\right).$$

Quindi, se restringiamo le matrici $U^{(k)}$ e $Y^{(k)}$ generate dall'iterazione per sottospazi considerando solo le prime p' colonne di $U^{(k)}$ e il minore principale $p' \times p'$ di $Y^{(k)}$, otteniamo l'iterazione per sottospazi di dimensione p' iniziata dalle prime p' colonne di $U^{(0)}$.

Una conseguenza immediata di questa osservazione è il seguente risultato.

Teorema Sia A una matrice diagonalizzabile con autovalori ordinati come $|\lambda_1| > \ldots > |\lambda_n|$, e si considerino le matrici $U^{(k)}$ generate dall'iterazione per sottospazi iniziata da $U^{(0)} = I_n$. Allora, se i minori principali $p \times p$ della matrice inversa degli autovettori V^{-1} sono tutti invertibili, la sequenza $Y^{(k)}$ converge, a meno di scalatura per matrici unitarie diagonali, a una forma di Schur di A.

Dimostrazione. È sufficiente combinare tutte le osservazioni che abbiamo fatto finora. L'ipotesi su V^{-1} garantisce la convergenza di tutte le iterazioni simultanee per sottospazi per $p = 1, \ldots, n$. Di conseguenza, le matrici unitarie $U^{(k)}$ convergono a una base ortogonale generata dagli autovettori relativi a $\lambda_1, \ldots, \lambda_n$, il che a sua volta implica la convergenza di $Y^{(k)}$ a una forma di Schur.

La base degli autovettori è determinata in modo unico a meno di un fattore di scala delle colonne per un numero complesso di modulo 1, da cui segue la tesi. □

Esercizio Si mostri che le assunzioni del Teorema falliscono per matrici reali con autovalori complessi, ma nondimeno la dimostrazione può essere modificata per garantire la convergenza alla forma di Schur reale, con blocchi 2×2 sulla diagonale.

Soluzione. Per matrici reali con autovalori complessi, le assunzioni del Teorema falliscono perché:

- Gli autovalori complessi occorrono in coppie coniugate $\lambda, \bar{\lambda}$ con $|\lambda| = |\bar{\lambda}|$
- La condizione di ordinamento $|\lambda_1| > \ldots > |\lambda_n|$ non può essere soddisfatta per coppie complesse coniugate
- La matrice degli autovettori V contiene elementi complessi, quindi V^{-1} non è reale

Tuttavia, la dimostrazione può essere modificata come segue:

- Invece di convergere a singoli autovettori complessi, l'algoritmo converge ai sottospazi invarianti 2-dimensionali generati dalle parti reale e immaginaria delle coppie di autovettori complessi coniugati
- La matrice $Y^{(k)}$ converge a una *forma di Schur reale* con blocchi 1×1 per autovalori reali e blocchi 2×2 per coppie di autovalori complessi coniugati
- Ogni blocco 2×2 sulla diagonale rappresenta una coppia coniugata di autovalori complessi
- La velocità di convergenza per autovalori complessi è determinata dal rapporto $|\lambda_{p+1}|/|\lambda_p|$, dove le coppie complesse sono trattate come aventi lo stesso modulo

Questa modifica funziona perché l'iterazione per sottospazi preserva naturalmente l'aritmetica reale quando iniziata con matrici iniziali reali, e la fattorizzazione QR di matrici reali produce matrici ortogonali reali. □

3.7 L'iterazione QR

Riformuliamo ora l'iterazione simultanea per sottospazi in un modo che sarà molto più adatto al calcolo efficiente. Da un lato, l'iterazione simultanea per sottospazi fornisce un'approssimazione della forma di Schur, come originariamente desiderato. Dall'altro lato, lo fa a un costo elevato: la velocità di convergenza è lenta (governata dal rapporto minimo tra due autovalori consecutivi) e il costo per iterazione è cubico.

Ricordiamo che il nostro obiettivo è progettare un'iterazione matriciale che produca una sequenza di matrici che sono simili, attraverso matrici unitarie o ortogonali. Infatti, l'iterazione simultanea iniziata con $U^{(0)} = I_n$ costruisce tale sequenza:

$$Y^{(k+1)} = (U^{(k+1)})^* A U^{(k+1)} = (U^{(k+1)})^* U^{(k)} \underbrace{(U^{(k)})^* A U^{(k)}}_{Y^{(k)}} (U^{(k)})^* U^{(k+1)} = (Z^{(k)})^* Y^{(k)} Z^{(k)},$$

dove abbiamo posto $Z^{(k)} := (U^{(k)})^* U^{(k+1)}$. Inoltre, guardando alla linea 4 dell'Algoritmo di iterazione dei sottospazi, vediamo che

$$Z^{(k)} R^{(k+1)} = Y^{(k)},$$

significa che $Z^{(k)}$ è il fattore Q di una fattorizzazione QR della matrice $Y^{(k)}$. Infine, si osservi che per ottenere il coniugato di una matrice quadrata rispetto al suo fattore Q è sufficiente calcolare il prodotto RQ della sua fattorizzazione QR; nel nostro contesto questo si legge come

$$Y^{(k+1)} = (Z^{(k)})^* Y^{(k)} Z^{(k)} = R^{(k+1)} Z^{(k)}.$$

Pertanto, se troviamo un modo per costruire le matrici $Z^{(k)}$ direttamente, possiamo riformulare l'iterazione in un modo più conveniente. Per raggiungere questo obiettivo, dobbiamo prima ricordare alcuni fatti rilevanti riguardanti la fattorizzazione QR di una matrice A.

Teorema Sia $A \in \mathbb{C}^{m \times n}$ una matrice a rango pieno con $m \geq n$, e $Q_1 R_1 = Q_2 R_2 = A$ due fattorizzazioni QR in forma economica. Allora, esiste una matrice diagonale unitaria D tale che $Q_1 = Q_2 D$.

Dimostrazione. Poiché R_1 e R_2 devono essere matrici $n \times n$ invertibili, possiamo riorganizzare le due fattorizzazioni scrivendo:

$$D := Q_2^* Q_1 = R_2 R_1^{-1}$$

Dall'equazione sopra concludiamo che D è triangolare superiore. Inoltre, usando che lo spazio colonna di Q_1 è incluso in quello di Q_2 (e viceversa), abbiamo anche che D è una matrice quadrata unitaria (o ortogonale). Una matrice unitaria triangolare superiore deve essere diagonale con elementi diagonali di modulo 1. Per concludere, usiamo ancora che lo spazio colonna di Q_1 è incluso in quello di Q_2 per ottenere:

$$Q_1 = Q_2 Q_2^* Q_1 = Q_2 D.$$

□

Le osservazioni che abbiamo usato per definire $Z^{(k)}$ permettono di costruire l'iterazione QR, descritta nell'Algoritmo che segue:

1: **procedure** QR(A)
2: $Y^{(0)} \leftarrow A$
3: **for** $k = 0, 1, \ldots$ **do**
4: $Z^{(k)}, R^{(k)} \leftarrow \mathrm{QR}(Y^{(k)})$ (fattorizzazione QR)
5: $Y^{(k+1)} \leftarrow R^{(k)} Z^{(k)}$
6: **end for**
7: **end procedure**

Tale algoritmo è lontano dall'essere pratico per le seguenti ragioni:

- La convergenza dipende dal fatto che gli autovalori abbiano moduli diversi, e può essere molto lenta per autovalori raggruppati.
- Ogni iterazione ha un costo cubico (sia le fattorizzazioni QR che la moltiplicazione matrice-matrice contribuiscono a questo), e anche nello scenario ottimistico in cui sono sufficienti $O(n)$ iterazioni, questo produrrebbe comunque un algoritmo $O(n^4)$.
- Diverse ipotesi che abbiamo fatto spesso non sono soddisfatte. Ad esempio, tutte le matrici reali con autovalori complessi coniugati hanno autovalori con $|\lambda_p| = |\lambda_{p+1}|$.

La prossima sezione sarà dedicata a modificare l'algoritmo per renderlo pratico.

3.8 Shifting e deflation

Consideriamo il seguente problema modello: abbiamo una matrice A con autovalori che soddisfano le seguenti disuguaglianze:

$$|\lambda_1| \geq \ldots \geq |\lambda_{n-1}| > |\lambda_n|, \qquad \frac{|\lambda_n|}{|\lambda_{n-1}|} = \epsilon \ll 1.$$

Alla luce dell'analisi precedente, ci aspettiamo che dopo k iterazioni del metodo QR otteniamo una matrice $Y^{(k)}$ della forma

$$Y^{(k)} = \left[\begin{array}{c|c} \hat{Y}^{(k)} & w^{(k)} \\ \hline 0 & \lambda_n^{(k)} \end{array}\right] \qquad \begin{cases} |\lambda_n^{(k)} - \lambda_n| \sim \mathcal{O}(\epsilon^k) \\ \|w^{(k)}\| \sim \mathcal{O}(\epsilon^k) \end{cases}.$$

Se ϵ è sufficientemente piccolo, dopo poche iterazioni avremo che $\|w^{(k)}\|$ sarà dell'ordine della precisione di macchina, e quindi possiamo considerare la matrice leggermente perturbata

$$Y^{(k)} + \delta Y^{(k)} = \left[\begin{array}{c|c} \hat{Y}^{(k)} & 0 \\ \hline 0 & \lambda_n^{(k)} \end{array}\right],$$

che ha esattamente $\lambda_n^{(k)}$ come autovalore. Poiché questa matrice è unitariamente simile a A, questo corrisponde all'iterazione QR esatta con la matrice $A+\delta A$ con $\delta A = U^{(k)} \delta Y^{(k)} (U^{(k)})^*$, che ha norma

spettrale uguale a $\|w^{(k)}\|$. Quindi, possiamo decidere che $\lambda_n^{(k)}$ è un autovalore approssimato di A con un piccolo errore all'indietro, e continuare l'iterazione sulla matrice più piccola $(n-1)\times(n-1)$ $\hat{Y}^{(k)}$. Questa procedura è chiamata *deflation*.

In generale, non c'è motivo di assumere che λ_n sia molto più piccolo del resto dello spettro, e quindi di essere in questa situazione favorevole. Si scopre che possiamo sempre modificare leggermente il problema agli autovalori per farlo accadere.

Supponiamo di avere un certo shift $\sigma \in \mathbb{C}$ tale che $\sigma \approx \lambda_n$; allora, la matrice shiftata $A - \sigma I$ ha $\lambda_n - \sigma$ come autovalore di modulo più piccolo (se assumiamo che σ sia più vicino a λ_n che a qualsiasi altro autovalore). Applicando un passo dell'iterazione QR alla matrice shiftata si otterrà

$$\begin{aligned} Y_\sigma^{(0)} &= A - \sigma I \\ Z_\sigma^{(0)} R_\sigma^{(0)} &= Y_\sigma^{(0)} \\ Y_\sigma^{(1)} &= (Z_\sigma^{(0)})^* Y_\sigma^{(0)} Z_\sigma^{(0)} = (Z_\sigma^{(0)})^* A Z_\sigma^{(0)} - \sigma I, \end{aligned}$$

dove abbiamo denotato con $Y_\sigma^{(k)}$ l'iterazione ottenuta partendo da $A - \sigma I$. Questa osservazione può essere generalizzata a un numero arbitrario di passi attraverso il seguente risultato.

Lemma Sia $Y_\sigma^{(k)}$ la sequenza di matrici generata dall'iterazione QR iniziata con $A - \sigma I$. Allora, se denotiamo con $Z_\sigma^{(k)}$ la matrice ortogonale della fattorizzazione QR al passo k,

$$(Z_\sigma^{(0)} \dots Z_\sigma^{(k-1)})^* A (Z_\sigma^{(0)} \dots Z_\sigma^{(k-1)}) = Y_\sigma^{(k)} + \sigma I, \qquad \forall k \geq 0.$$

Dimostrazione. La definizione dell'iterazione QR fornisce

$$(Z_\sigma^{(0)} \dots Z_\sigma^{(k-1)})^* (A - \sigma I)(Z_\sigma^{(0)} \dots Z_\sigma^{(k-1)}) = Y_\sigma^{(k)}.$$

La tesi segue spostando σI al membro destro, e ricordando che le matrici $Z^{(i)}$ sono unitarie. □

Concludiamo che, se abbiamo a disposizione una buona approssimazione $\sigma \approx \lambda_n$, possiamo far convergere l'iterazione QR (shiftata) in pochi passi a una forma dove λ_n può essere "deflazionato".

3.9 Riduzione di Hessenberg

Un'osservazione chiave per ridurre il costo dell'iterazione è preprocessare la matrice per renderla "il più triangolare superiore possibile". Chiaramente, il passo dell'algoritmo deve lavorare con matrici unitarie ed essere una similitudine.

Definizione Una matrice H è in *forma di Hessenberg* se ha elementi nulli sotto la prima sottodiagonale, cioè se $H_{ij} = 0$ per tutti $i > j + 1$.

La riduzione della matrice alla forma di Hessenberg può essere calcolata con $O(n^3)$ flop usando riflettori di Householder.

Lemma Sia A una qualsiasi matrice complessa $n \times n$, con $n \geq 2$. Allora, esistono una matrice di Hessenberg superiore H e $n-2$ riflettori di Householder P_j per $j = 1, \dots, n-2$, tali che

$$P_{n-2} \dots P_1 A P_1^* \dots P_{n-2}^* = H.$$

Le matrici H e $P := P_{n-2} \dots P_1$ possono essere calcolate da A con $O(n^3)$ flop.

Dimostrazione. La dimostrazione presenta un algoritmo per calcolare H e P_j con la complessità asintotica richiesta. Una dimostrazione più formale può essere ottenuta usando l'induzione. Sia $\hat{P}_1$ un riflettore di Householder $(n-1) \times (n-1)$ tale che

$$\hat{P}_1 A_{2:n,1} = \begin{bmatrix} x \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

dove x è usato per denotare un elemento generico non nullo. Allora, se definiamo $P_1 := I_1 \oplus \hat{P}_1$ che denota la matrice blocco diagonale ottenuta ponendo lo scalare 1, cioè l'identità 1×1, sopra a sinistra e $\hat{P}_1$ in basso a destra, la matrice $P_1 A P_1^*$ ha il seguente schema di sparsità:

$$A^{(1)} := P_1 A P_1^* = \begin{bmatrix} x & x & x & \dots & x \\ x & x & x & \dots & x \\ 0 & x & x & \dots & x \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & x & x & \dots & x \end{bmatrix},$$

e può essere calcolata in $O(n^2)$ flop sfruttando la struttura del riflettore di Householder. Seguendo la stessa idea, P_2 può essere definita per avere

$$P_2 = I_2 \oplus \hat{P}_2, \quad \hat{P}_2 A^{(1)}_{3:n,2} = \begin{bmatrix} x \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Calcolare $P_2 A^{(1)} P_2^*$ metterà la seconda colonna in "forma di Hessenberg", e non deteriorerà la struttura della prima colonna, grazie alla presenza di I_2 in alto. Continuando il processo si ottiene la matrice di Hessenberg superiore richiesta $H = A^{(n-2)}$. □

Preprocessare la matrice A per essere in forma di Hessenberg superiore porta due vantaggi chiave all'iterazione QR:

- Per una matrice di Hessenberg superiore, la fattorizzazione QR $Z^{(k)} R^{(k)} = Y^{(k)}$ e l'iterata successiva $Y^{(k+1)} := R^{(k)} Z^{(k)}$ possono essere calcolate con $O(n^2)$ flop.

- La struttura di Hessenberg è preservata dalle iterazioni QR, e quindi il beneficio di cui sopra non è limitato al primo passo.

Introduciamo ora le *rotazioni di Givens*, che sono matrici unitarie con un obiettivo simile ai riflettori di Householder, ma che agiscono elemento per elemento, rendendo più facile preservare la sparsità.

Definizione Una rotazione di Givens che agisce sulle righe (k, l) è una matrice della forma G tale che, per alcuni $c, s \in \mathbb{C}$ con $|c|^2 + |s|^2 = 1$,

$$G = \begin{bmatrix} I_{k_1} & & & & \\ & c & s & & \\ & & I_{k_2} & & \\ & -\bar{s} & \bar{c} & & \\ & & & & I_{k_3} \end{bmatrix},$$

e tale che gli elementi $c, s, -\bar{s}, \bar{c}$ si trovino sulle righe e colonne k o l. Queste trasformazioni sono unitarie con $\det(G) = 1$.

La proprietà $|c|^2 + |s|^2$ permette di interpretare c e s come coseni e seni (complessi), e questa è la ragione per chiamare queste trasformazioni "rotazioni". Spesso, considereremo $l = k + 1$, che permette di cercare la forma semplificata

$$G = I_{k_1} \oplus \hat{G} \oplus I_{k_2}, \qquad \hat{G} := \begin{bmatrix} c & s \\ -\bar{s} & \bar{c} \end{bmatrix}.$$

Usiamo ora le rotazioni di Givens per calcolare una fattorizzazione QR di una matrice di Hessenberg superiore in tempo quadratico.

Lemma Sia H una matrice di Hessenberg superiore $n \times n$. Allora, esistono $n-1$ rotazioni di Givens $G_1, \ldots, G_{n-1}$ con G_i che agisce sulle righe i e $i+1$, tali che

$$H = G_1^* \ldots G_{n-1}^* R = QR,$$

con R triangolare superiore. Le matrici Q e R possono essere calcolate con $\mathcal{O}(n^2)$ flop.

Dimostrazione. Dimostriamo il risultato per induzione, mostrando che la costruzione richiede al più $8n^2$ operazioni in aritmetica floating point. Il risultato è banalmente vero per $n = 1$, poiché H è già triangolare superiore, e possiamo semplicemente porre $Q = 1$ come prodotto vuoto di 0 rotazioni.

Assumiamo che il risultato sia vero per $n-1$, e consideriamo una rotazione G_1 che opera sulle righe 1 e 2 tale che:

$$G_1 \begin{bmatrix} H_{11} \\ H_{21} \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} \times \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Abbiamo allora

$$G_1 H = \left[\begin{array}{c|ccc} \times & \times & \ldots & \times \\ \hline & \multicolumn{3}{c}{\hat{H}} \end{array}\right],$$

con $\hat{H}$ una matrice di Hessenberg superiore $(n-1) \times (n-1)$. Per induzione, abbiamo $\hat{H} = \hat{G}_1^* \ldots \hat{G}_{n-2}^* \hat{R}$, e ponendo $G_i := 1 \oplus \hat{G}_{i-1}$ per $i = 2, \ldots, n-1$, otteniamo

$$H = G_1^* G_2^* \ldots G_{n-1}^* \underbrace{\left[\begin{array}{c|ccc} \times & \times & \ldots & \times \\ \hline & \multicolumn{3}{c}{\hat{R}} \end{array}\right]}_{=:R} = G_1^* G_2^* \ldots G_{n-1}^* R = QR.$$

Un calcolo diretto mostra che moltiplicare una rotazione per una matrice richiede $4n$ operazioni in floating point, ottenere R e Q da $\hat{R}$ e $\hat{Q} := \hat{G}_1^* \dots \hat{G}_{n-2}^*$ richiede 2 prodotti per G_1. In aggiunta, abbiamo 4 operazioni in virgola mobile in più per trovare G_1 e calcolare $G_1 H e_1$, che produce il costo totale

$$8n + 4 + 8(n-1)^2 = 8n^2 - 8n + 12 \sim \mathcal{O}(n^2)$$

□

3.10 Calcolo di autovettori e sottospazi invarianti

L'iterazione QR discussa nelle sezioni precedenti permette di costruire una sequenza di matrici simili $Y^{(k)}$ che, sotto opportune ipotesi, convergono a una forma di Schur di $Y^{(0)} = A$. La forma di Schur finale T è sufficiente per determinare gli autovalori (dobbiamo solo leggere gli elementi diagonali) e nel caso di autovalori multipli anche i corrispondenti blocchi di Jordan.

Il recupero degli autovettori è più complesso e viene eseguito in due passi:

- Prima determiniamo gli autovettori w della matrice triangolare superiore T, corrispondenti agli autovalori $\lambda_i := T_{ii}$ per $i = 1, \dots, n$.
- Poi recuperiamo gli autovettori del problema originale usando la relazione $Q^* A Q = T$ e ponendo $v = Qw$.

Se $Tw = \lambda w$ allora $AQ = QT$ implica $Av = AQw = QTw = \lambda Qw = \lambda v$, e quindi il secondo passo caratterizza completamente gli autovettori di A a partire da quelli di T.

Per calcolare gli autovettori della matrice triangolare superiore, facciamo l'ipotesi che λ_i sia semplice, e ci basiamo sulla seguente osservazione:

$$T - \lambda_i I = \left[\begin{array}{c|c|c} T_1 & x & \\ \hline & 0 & \\ \hline & & T_2 \end{array}\right]$$

con T_1 non singolare e triangolare superiore. Dobbiamo determinare un vettore nel nucleo destro della matrice sopra, che può essere fatto imponendo:

$$(T - \lambda_i I) w = 0 \qquad w = \begin{bmatrix} y \\ 1 \\ 0 \end{bmatrix}$$

dove w è partizionato per corrispondere alla struttura a blocchi identificata in T. Allora, risolviamo l'equazione ponendo $T_1 y = -x$. Quindi, y (e di conseguenza w) è determinato risolvendo un sistema lineare triangolare superiore, che costa al più $\mathcal{O}(n^2)$ flop. Questo deve essere ripetuto per tutti gli autovalori, producendo un costo totale di $\mathcal{O}(n^3)$.

Una tecnica simile può essere usata per trovare basi ortogonali per sottospazi invarianti corrispondenti a un sottoinsieme $\{\lambda_1, \dots, \lambda_k\} \subseteq \Lambda(A)$ di tutti gli autovalori di A. Supponiamo di essere particolarmente fortunati, e che la forma di Schur calcolata dall'iterazione QR soddisfi

$$Q^*AQ = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}, \qquad T_{11} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_k \end{bmatrix},$$

con T_{22} contenente tutti gli altri autovalori. Allora, il sottospazio invariante in considerazione è generato dalle prime k colonne di Q, che formano una base ortogonale per esso. Il nostro problema è facilmente risolto.

Tuttavia, non c'è una ragione particolare per cui questo dovrebbe accadere: l'iterazione QR può calcolare gli autovalori in qualsiasi ordine, e abbiamo poco controllo sul processo. Se gli autovalori finiscono nella posizione "sbagliata", possiamo semplicemente riordinarli, per spingere quelli di interesse in cima alla matrice.

Il problema può essere ridotto al caso 2×2 che è risolto dal seguente Lemma.

Lemma Sia T una matrice triangolare superiore con due autovalori distinti $t_{11} = \lambda_1 \neq \lambda_2 = t_{22}$; sia G una rotazione di Givens tale che, per qualche $\alpha \in \mathbb{C}$,

$$G \begin{bmatrix} t_{12} \\ \lambda_2 - \lambda_1 \end{bmatrix} = \begin{bmatrix} \alpha \\ 0 \end{bmatrix},$$

Allora, la matrice GTG^* è triangolare superiore con autovalori elencati nell'ordine opposto.

Dimostrazione. Si noti che, per costruzione, abbiamo

$$G(T - \lambda_1 I)G^* = G \begin{bmatrix} 0 & t_{12} \\ 0 & \lambda_2 - \lambda_1 \end{bmatrix} G^* = \begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix} G^* = \begin{bmatrix} \times & \times \\ 0 & 0 \end{bmatrix},$$

dove come al solito abbiamo usato $\times$ per denotare lo schema di sparsità nella matrice. Applicando la stessa trasformazione a T si ottiene

$$GTG^* = \begin{bmatrix} \times & \times \\ 0 & 0 \end{bmatrix} + \lambda_1 I = \begin{bmatrix} \lambda_2 & \times \\ 0 & \lambda_1 \end{bmatrix},$$

dove l'elemento in posizione $(1, 1)$ è determinato essere esattamente λ_2 perché gli autovalori di T non cambiano per similitudine. □

Il Lemma può essere impiegato per scambiare due autovalori λ_i, λ_{i+1} di una matrice triangolare superiore $n \times n$ più grande considerando una rotazione su due righe consecutive. Usando il fatto che le trasposizioni generano tutte le permutazioni, concludiamo che qualsiasi permutazione degli autovalori è possibile, ed è facilmente ottenuta mediante ripetute applicazioni del Lemma.

3.11 Double shifting e la forma di Schur reale

Se la matrice A è reale, calcolare la forma di Schur con shift complessi può essere indesiderabile, a causa del costo aggiuntivo dell'aritmetica complessa. Chiaramente, non c'è speranza di trovare la forma di Schur con aritmetica reale se la matrice ha autovalori complessi.

Possiamo, tuttavia, limitare la nostra attenzione alla forma di Schur reale:

Definizione Una matrice T è in *forma di Schur reale* se è a blocchi triangolare superiore con blocchi diagonali T_{ii} tali che o T_{ii} è una matrice reale 1×1, o una matrice 2×2 della forma

$$T_{ii} = \begin{bmatrix} a & b \\ -b & a \end{bmatrix},$$

che ha $a \pm ib$ come autovalori.

Anche se una matrice in forma di Schur reale non è in senso stretto triangolare superiore, i suoi autovalori sono immediatamente leggibili dai blocchi diagonali senza alcun calcolo. Inoltre, gli argomenti usati per trovare gli autovettori e i sottospazi invarianti dalla forma di Schur possono essere facilmente adattati.

La struttura reale può essere mantenuta durante tutte le iterazioni con il seguente trucco; se uno shift σ è determinato (per esempio dalla strategia di shifting di Wilkinson), procediamo come segue:

- *Se* $\sigma \in \mathbb{R}$, procediamo con l'iterazione QR standard.
- *Se* $\sigma \in \mathbb{C} \setminus \mathbb{R}$, consideriamo il polinomio a coefficienti reali

  $$p(z) = (z - \sigma)(z - \overline{\sigma})$$

  e calcoliamo $p(Y^{(k)})e_1$.

Nel secondo caso, consideriamo le due rotazioni necessarie per trasformare $p(Y^{(k)})e_1$ in un multiplo di e_1, e applichiamo queste rotazioni a $Y^{(k)}$. La struttura di Hessenberg superiore può essere ripristinata usando una tecnica nota come *bulge-chasing*.

Un'iterazione di questa forma costa circa il doppio di un'iterazione con shift singolo. Tuttavia, la convergenza può essere collegata all'iterazione per sottospazi applicata a $p(Y^{(k)})$, e quindi possiamo aspettarci che gli autovalori vicini a σ e $\overline{\sigma}$ siano ben approssimati insieme.

4 Problemi agli autovalori simmetrici e SVD

I problemi agli autovalori simmetrici sono intrinsecamente più facili di quelli non simmetrici, e permettono di dimostrare risultati e caratterizzazioni molto più forti. In questa sezione, discutiamo l'iterazione QR tridiagonale e lo schema divide-et-impera.

Vedremo poi che c'è una stretta relazione tra il problema agli autovalori simmetrico e la decomposizione ai valori singolari (SVD), una fattorizzazione potente sia teorica che algoritmica.

4.1 Iterazione QR tridiagonale

Se applichiamo l'iterazione QR a una matrice simmetrica, alcune osservazioni possono essere fatte, che sono riassunte dal seguente Lemma.

Lemma Sia $A = A^*$ una matrice simmetrica o hermitiana, e $Y^{(k)}$ le iterate QR applicate dopo la riduzione di Hessenberg $Y^{(0)} = Q^*AQ$. Allora, tutte le matrici $Y^{(k)}$ sono tridiagonali.

Dimostrazione. Si noti che $Y^{(k)}$ sono unitariamente simili a A, quindi esiste Q_k ortogonale (o unitaria) tale che $Q_k^*AQ_k = Y^{(k)}$. Quindi, $Y^{(k)} = (Y^{(k)})^*$ sono tutte simmetriche (o hermitiane).

Tutte le $Y^{(k)}$ sono matrici di Hessenberg e quest'ultime hanno solo una sottodiagonale diversa da zero. La simmetria implica che tutte le $Y^{(k)}$ hanno solo una superdiagonale non nulla, e sono quindi tridiagonali. □

Ridurre una matrice simmetrica A alla forma tridiagonale non è più economico che ridurre una matrice generale alla forma di Hessenberg superiore, dobbiamo ancora applicare le rotazioni sulla matrice completa, per un costo totale di $\mathcal{O}(n^3)$ flop. Se A è tridiagonale, tuttavia, questa struttura è facilmente sfruttata nell'iterazione QR se sono desiderati solo gli autovalori. Infatti, calcolare $Y^{(k+1)}$ da $Y^{(k)}$ richiede i seguenti passi:

- *Trovare uno shift appropriato* σ (costo: $\mathcal{O}(1)$ flop).
- *Determinare una rotazione tale che* $Y^{(k)}e_1 - \sigma e_1$ *sia un multiplo di* e_1 (costo: $\mathcal{O}(1)$ flop).
- *Applicare le rotazioni alla matrice fino al fondo* (costo: applicare $\mathcal{O}(n)$ rotazioni).

L'ultimo punto è la parte costosa, e nel caso non strutturato ogni rotazione costa $\mathcal{O}(n)$ flop. Nel caso tridiagonale, la struttura tridiagonale-più-bulge è preservata durante tutto il processo di inseguimento, e quindi una rotazione può essere applicata a costo $\mathcal{O}(1)$. Riassumendo, possiamo eseguire l'iterazione QR tridiagonale con $\mathcal{O}(n)$ flop per iterazione, per un costo totale di $\mathcal{O}(n^2)$ flop.

Remark Calcolare l'autovettore nel caso tridiagonale è molto più costoso: le rotazioni devono essere applicate alle matrici Q che rappresentano il cambio di base, e questo richiede $O(n)$ flop per iterazione. Il costo totale del metodo è ancora $O(n^3)$ flop.

4.2 Teorema di Courant-Fischer

Come abbiamo visto analizzando il metodo delle potenze, nel caso hermitiano c'è una relazione tra autovalori, autovettori e il quoziente di Rayleigh. Qui forniamo un potente strumento teorico, noto come *teorema min-max di Courant-Fischer*, che caratterizza gli autovalori come valori ottimali del quoziente di Rayleigh su sottospazi.

Teorema (Courant-Fischer) Sia $A \in \mathbb{C}^{n\times n}$ una matrice hermitiana con autovalori $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Allora:

$$\max_{\substack{U\subset\mathbb{C}^n \\ \dim(U)=k}} \min_{\substack{x\in U \\ x\neq 0}} \frac{x^*Ax}{x^*x} = \lambda_k,$$

$$\min_{\substack{U\subset\mathbb{C}^n \\ \dim(U)=k}} \max_{\substack{x\in U \\ x\neq 0}} \frac{x^*Ax}{x^*x} = \lambda_{n-k+1},$$

per $k = 1, 2, \ldots, n$.

Dimostrazione. Dimostriamo solo la parte max-min poiché la min-max è completamente analoga. Siano $v_1, \ldots, v_n$ una base ortonormale di $\mathbb{C}^n$ composta da autovettori di A e sia $S := \operatorname{colspan}(v_k, \ldots, v_n)$. Allora, per ogni $U \subset \mathbb{C}^n$ di dimensione k abbiamo che $S \cap U \neq \{0\}$; più specificamente, esiste $x \in S \cap U$, tale che $x = \sum_{j=k}^n c_j v_j$ e $x \neq 0$. Questo implica

$$\frac{x^*Ax}{x^*x} = \frac{\sum_{j=k}^{n} |c_j|^2 \lambda_j}{\sum_{j=k}^{n} |c_j|^2} \leq \lambda_k.$$

Questo prova che, per ogni U, il minimo del quoziente di Rayleigh è minore o uguale a λ_k, il che implica che il massimo su tutti i possibili U del minimo quoziente di Rayleigh è anche limitato superiormente da λ_k. Per ottenere la tesi, è sufficiente mostrare che per almeno una scelta di U, il valore λ_k corrisponde al minimo del quoziente di Rayleigh. Questo accade quando si considera $U = \text{colspan}(v_1, \ldots, v_k)$. □

Corollario Sia $A \in \mathbb{C}^{n\times n}$ una matrice hermitiana con autovalori $\alpha_1 \geq \cdots \geq \alpha_n$, $Q \in \mathbb{C}^{n\times(n-1)}$ tale che $Q^*Q = I_{n-1}$, e $B = Q^*AQ \in \mathbb{C}^{(n-1)\times(n-1)}$ con autovalori $\beta_1 \geq \cdots \geq \beta_{n-1}$. Allora,

$$\alpha_1 \geq \beta_1 \geq \alpha_2 \geq \beta_2 \geq \cdots \geq \beta_{n-1} \geq \alpha_n,$$

e diciamo che gli autovalori di A sono interlacciati con quelli di B (*interlacing property*).

Dimostrazione. Alla luce del Teorema precedente abbiamo

$$\beta_k = \max_{\substack{U\subset\mathbb{C}^{n-1}\\ \dim(U)=k}} \min_{\substack{x\in U\\ x\neq 0}} \frac{x^*Bx}{x^*x} = \min_{\substack{x\in\tilde{U}\\ x\neq 0}} \frac{x^*Q^*AQx}{x^*Q^*Qx},$$

dove $\tilde{U}$ è un sottospazio k-dimensionale di $\mathbb{C}^{n-1}$ dove il massimo è raggiunto. Sia

$$\widehat{U} = Q\tilde{U} = \{y \in \mathbb{C}^n : y = Qx, \text{ per qualche } x \in \tilde{U}\},$$

allora $\dim(\widehat{U}) = k$ e

$$\beta_k = \min_{\substack{x\in\tilde{U}\\ x\neq 0}} \frac{x^*Q^*AQx}{x^*Q^*Qx} = \min_{\substack{y\in\widehat{U}\\ y\neq 0}} \frac{y^*Ay}{y^*y} \leq \max_{\substack{\widehat{U}\subset\mathbb{C}^n\\ \dim(\widehat{U})=k}} \min_{\substack{y\in\widehat{U}\\ y\neq 0}} \frac{y^*Ay}{y^*y} = \alpha_k.$$

La disuguaglianza $\beta_{k-1} \geq \alpha_k$ è ottenuta applicando lo stesso argomento alle matrici $-A$ e $-B$. □

Corollario Sia $A \in \mathbb{C}^{n\times n}$ una matrice hermitiana con autovalori $\alpha_1 \geq \cdots \geq \alpha_n$ e sia $B \in \mathbb{C}^{m\times m}$ una sottomatrice principale di A, per $m \leq n$, con autovalori $\beta_1 \geq \cdots \geq \beta_m$. Allora

$$\alpha_j \geq \beta_j \geq \alpha_{j+(n-m)},$$

per $j = 1, \ldots, m$.

Dimostrazione. Dimostriamo per induzione su $n - m$.

Caso base: $n - m = 1$ (cioè $m = n - 1$).
In questo caso, B è una sottomatrice principale $(n-1)\times(n-1)$ di A. Possiamo scrivere A come:

$$A = \begin{bmatrix} B & c \\ c^* & a \end{bmatrix},$$

dove $c \in \mathbb{C}^{n-1}$ e $a \in \mathbb{R}$. Per il Corollario precedente, gli autovalori di A e B sono interlacciati:

$$\alpha_1 \geq \beta_1 \geq \alpha_2 \geq \beta_2 \geq \cdots \geq \beta_{n-1} \geq \alpha_n.$$

Da questa catena di disuguaglianze, per $j = 1, \ldots, n-1$ abbiamo:

- $\alpha_j \geq \beta_j$ (dalla disuguaglianza sinistra)
- $\beta_j \geq \alpha_{j+1} = \alpha_{j+(n-(n-1))}$ (dalla disuguaglianza destra)

Quindi il caso base è verificato.

Passo induttivo: Supponiamo che il risultato sia vero per tutte le sottomatrici principali di dimensione $m+1$ di una matrice hermitiana di dimensione n, e dimostriamolo per sottomatrici di dimensione m.

Sia B una sottomatrice principale $m \times m$ di A. Possiamo considerare una sottomatrice principale C di dimensione $(m+1) \times (m+1)$ che contiene B come sottomatrice principale. Più precisamente, possiamo scrivere:

$$C = \begin{bmatrix} B & d \\ d^* & c \end{bmatrix},$$

dove $d \in \mathbb{C}^m$ e $c \in \mathbb{R}$.

Siano $\gamma_1 \geq \cdots \geq \gamma_{m+1}$ gli autovalori di C. Per l'ipotesi induttiva (applicata a C come sottomatrice principale di A), abbiamo:

$$\alpha_j \geq \gamma_j \geq \alpha_{j+(n-(m+1))} \quad \text{per } j = 1, \ldots, m+1.$$

Ora, applicando il caso base a C e alla sua sottomatrice principale B, otteniamo l'interlacciamento:

$$\gamma_1 \geq \beta_1 \geq \gamma_2 \geq \beta_2 \geq \cdots \geq \beta_m \geq \gamma_{m+1}.$$

Combinando le due catene di disuguaglianze:

- Per la disuguaglianza sinistra: $\alpha_j \geq \gamma_j \geq \beta_j$
- Per la disuguaglianza destra: $\beta_j \geq \gamma_{j+1} \geq \alpha_{(j+1)+(n-(m+1))} = \alpha_{j+(n-m)}$

Quindi abbiamo dimostrato che:

$$\alpha_j \geq \beta_j \geq \alpha_{j+(n-m)} \quad \text{per } j = 1, \ldots, m.$$

□

Corollario Siano A, B, C matrici hermitiane con autovalori ordinati $\alpha_j, \beta_j, \gamma_j$ e tali che $A = B + C$. Allora vale:

$$\beta_j + \gamma_{n-j+i} \leq \alpha_i \leq \beta_k + \gamma_{i-k+1},$$

per $1 \leq k \leq i \leq j \leq n$.

4.3 Decomposizione ai Valori Singolari

Introduciamo ora un'importante fattorizzazione per una matrice rettangolare generica A, che è chiamata *decomposizione ai valori singolari (SVD)*. L'idea dietro questa fattorizzazione è decomporre qualsiasi operatore lineare come il prodotto di tre matrici, qui riportate per il caso $m \geq n$:

$$A = U\Sigma V^*, \quad \Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ & 0 & \end{bmatrix},$$

Le matrici U, V sono unitarie, Σ è reale e diagonale, e $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$. Qui con "diagonale" intendiamo che Σ può essere rettangolare, ma ha elementi non nulli solo sulle entrate diagonali Σ_{ii}. Il caso riportato sopra è per $m \geq n$, ma la definizione analoga può essere data per $n \geq m$.

Geometricamente, possiamo interpretare questa fattorizzazione come la decomposizione dell'azione di A in un'isometria, seguita da un ridimensionamento (non negativo) degli assi, e poi ancora da un'isometria. La fattorizzazione può essere usata per fornire diverse soluzioni esplicite a problemi computazionali.
Dimostriamo ora che la decomposizione ai valori singolari esiste per qualsiasi matrice.

Teorema Sia $A \in \mathbb{C}^{m\times n}$ con $m \geq n$. Allora, esistono due matrici unitarie quadrate U, V di dimensione $m \times m$ e $n \times n$, rispettivamente, e una matrice $m \times n$ Σ con diagonale non negativa con elementi decrescenti e zero altrove, tali che $A = U\Sigma V^*$. Se A è reale, U e V possono essere scelte reali anch'esse.

Dimostrazione. Dimostriamo questo risultato per induzione su n; sia $n = 1$, e m arbitrario. Allora, A è un vettore colonna e possiamo porre

$$U = \begin{bmatrix} \frac{1}{\|A\|_2}A & B \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \|A\|_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad V = \begin{bmatrix} 1 \end{bmatrix}$$

dove $B \in \mathbb{C}^{m\times(m-1)}$ è un completamento di $\frac{1}{\|A\|_2}A$ a una base ortonormale di $\mathbb{C}^m$. Per verifica diretta, abbiamo $A = U\Sigma V^*$, e le tre matrici soddisfano tutti i requisiti per essere una SVD di A.

Assumiamo ora che il risultato sia valido per $n-1$ (e m arbitrario). Allora, per definizione di norma spettrale esiste un vettore v_1 di norma unitaria tale che

$$w = Av_1, \qquad \|w\|_2 = \|A\|_2.$$

Se $A = 0$ la SVD è ottenuta in modo banale, quindi possiamo assumere che $\|w\|_2 \neq 0$ e definire le matrici $\hat{U}, \hat{V}$ come segue:

$$\hat{U} := \begin{bmatrix} \frac{w}{\|w\|_2} & w_2 & \cdots & w_m \end{bmatrix}, \quad \hat{V} := \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix},$$

dove $w_2, \ldots, w_m$ e $v_2, \ldots, v_n$ sono scelti come qualsiasi completamento unitario della prima colonna. Affermiamo ora che la matrice $\hat{U}^* A \hat{V}$ ha la seguente forma:

$$\hat{U}^* A \hat{V} = \begin{bmatrix} \|A\|_2 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \hat{A} & \\ 0 & & & \end{bmatrix}.$$

Il fatto che l'elemento in posizione $(1,1)$ sia uguale a $\|A\|_2$ può essere verificato direttamente:

$$(\hat{U}^* A \hat{V})_{11} = (\hat{U} e_1)^T A (\hat{V} e_1) = \frac{1}{\|w\|_2} w^* A v_1 = \frac{w^* w}{\|w\|_2} = \|w\|_2 = \|A\|_2. \tag{4.3}$$

Se qualsiasi altro elemento nella prima colonna o riga fosse diverso da zero, allora la matrice A avrebbe una colonna o riga con norma euclidea strettamente maggiore di $\|A\|_2$, che è una contraddizione. Quindi, la struttura di sparsità in (4.3) è una conseguenza immediata di $(\hat{U}^* A \hat{V})_{11} = \|A\|_2$.

Possiamo ora usare l'ipotesi induttiva per ottenere una SVD di $\hat{A} = \tilde{U} \tilde{\Sigma} \tilde{V}^*$ e scrivere la seguente decomposizione per A:

$$A = \hat{U} \begin{bmatrix} 1 & 0 \\ 0 & \tilde{U} \end{bmatrix} \begin{bmatrix} \|A\|_2 & 0 \\ 0 & \tilde{\Sigma} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \tilde{V}^* \end{bmatrix} \hat{V}^*.$$

Chiamando le matrici unitarie date dal prodotto delle prime due matrici e delle ultime due rispettivamente U e V, e ponendo $\sigma_1 := \|A\|_2$ e $\sigma_i := \tilde{\sigma}_{i-1}$ per $i > 1$, questa è una SVD della matrice A. L'unico fatto rimanente da verificare è che i valori singolari sono in ordine decrescente, cioè $\tilde{\sigma}_1 \leq \|A\|_2$. A questo scopo, possiamo notare che per una matrice diagonale, la norma è il massimo dei moduli degli elementi diagonali, il che a sua volta implica $\max\{\|A\|_2, \tilde{\sigma}_1\} = \|A\|_2$ e quindi $\tilde{\sigma}_1 \leq \|A\|_2$. □

Una SVD di A non è necessariamente unica. Osserviamo che, data qualsiasi matrice diagonale unitaria D, possiamo scalare diagonalmente U e V per ottenere $A = U\Sigma V^* = UD\Sigma D^* V^*$. Poiché $UD \neq U$ (a meno che $D = I$), abbiamo un numero infinito di diverse decomposizioni ai valori singolari.

4.3.1 Proprietà della SVD

Presentiamo ora alcune proprietà essenziali della decomposizione ai valori singolari.

Lemma Sia $A = U\Sigma V^*$ una SVD di $A \in \mathbb{C}^{m \times n}$ con $m \geq n$. Allora,

(i) La matrice simmetrica definita positiva $A^* A$ è diagonalizzata da V: $V^* A^* A V = \Sigma^* \Sigma = D$, e ha σ_i^2 come autovalori per $i = 1, \ldots, n$.

(ii) La matrice simmetrica definita positiva AA^* è diagonalizzata da U: $U^*AA^*U = \Sigma\Sigma^* = D$, e ha $m-n$ autovalori zero, e gli altri uguali a σ_i^2.

(iii) La seguente matrice simmetrica M ha $\pm\sigma_i$ e $m-n$ zeri come autovalori:

$$M = \begin{bmatrix} 0 & A^* \\ A & 0 \end{bmatrix} \implies \Lambda(M) = \{\pm\sigma_i \mid \sigma_i \text{ valore singolare di } A\}.$$

Dimostrazione. La dimostrazione di (i) e (ii) è ottenuta mediante un calcolo diretto. Per quanto riguarda M, facciamo la seguente osservazione:

$$\begin{bmatrix} V^* & 0 \\ 0 & U^* \end{bmatrix} \begin{bmatrix} 0 & A^* \\ A & 0 \end{bmatrix} \begin{bmatrix} V & 0 \\ 0 & U \end{bmatrix} = \begin{bmatrix} 0 & V^*A^*U \\ U^*AV & 0 \end{bmatrix} = \begin{bmatrix} 0 & \Sigma^* \\ \Sigma & 0 \end{bmatrix} =: M_\Sigma.$$

Poiché M e M_Σ sono simili, hanno gli stessi autovalori, dunque mostriamo che gli autovalori di M_Σ sono $\pm\sigma_i$ e gli $m-n$ zeri. Consideriamo la permutazione π di $\{1,\dots,m+n\}$ tale che

$$\pi(i) = \begin{cases} \frac{i+1}{2} & i \equiv 1 \pmod 2 \text{ e } i \le 2n \\ \frac{i}{2} + n & i \equiv 0 \pmod 2 \text{ e } i \le 2n \\ i & 2n < i \le m+n \end{cases}.$$

Se Π è la matrice di permutazione associata a π, calcolando $\Pi^* M_\Sigma \Pi$ si ottiene una matrice a blocchi diagonali della seguente forma:

$$\Pi^* M_\Sigma \Pi = \begin{bmatrix} \Sigma_1 & & & \\ & \ddots & & \\ & & \Sigma_n & \\ & & & 0_{m-n} \end{bmatrix}, \quad \Sigma_i := \begin{bmatrix} 0 & \sigma_i \\ \sigma_i & 0 \end{bmatrix}.$$

Gli autovalori delle matrici 2×2 Σ_i sono esattamente $\pm\sigma_i$, quindi si ha la tesi. □

Osservazione Dalla definizione della SVD deriva immediatamente l'invarianza dei valori singolari sotto trasformazioni unitarie: $\sigma_i(A) = \sigma_i(QA) = \sigma_i(AZ)$ per qualsiasi scelta di Q, Z unitarie o ortogonali nel caso reale.

lemma Sia A una matrice $m \times n$, con $m \ge n$ e SVD $A = U\Sigma V^*$. Allora, valgono le seguenti identità:

$$\|A\|_2 = \sigma_1(A), \quad \|A\|_F = \sqrt{\sigma_1^2 + \ldots + \sigma_n^2}.$$

Dimostrazione. La tesi segue notando che, essendo la norma spettrale e di Frobenius invarianti sotto trasformazioni unitarie, abbiamo

$$\|A\|_{2/F} = \|U\Sigma V^*\|_{2/F} = \|\Sigma\|_{2/F},$$

e per la definizione delle norme spettrale e di Frobenius. □

Esercizio Si mostri che, se una norma matriciale $\|\cdot\|$ è invariante sotto trasformazioni unitarie, allora può essere scritta nella forma $\|A\| = f(\sigma_1(A), \ldots, \sigma_n(A))$ per qualche f.

Soluzione dell'Esercizio 4.4.6. Sia $\|\cdot\|$ una norma matriciale invariante sotto trasformazioni unitarie, cioè tale che $\|QA\| = \|A\|$ e $\|AZ\| = \|A\|$ per tutte le matrici unitarie Q e Z di dimensioni appropriate.

Data qualsiasi matrice $A \in \mathbb{C}^{m\times n}$, consideriamo la sua SVD: $A = U\Sigma V^*$, dove U e V sono unitarie e Σ è la matrice dei valori singolari. Per l'invarianza della norma sotto trasformazioni unitarie, abbiamo:

$$\|A\| = \|U\Sigma V^*\| = \|\Sigma\|.$$

Ora, consideriamo due permutazioni qualsiasi Π e Π' delle righe e colonne di Σ. Poiché le matrici di permutazione sono unitarie, abbiamo:

$$\|\Pi\Sigma\Pi'\| = \|\Sigma\|.$$

Questo implica che la norma dipende solo dai valori singolari $\sigma_1, \ldots, \sigma_n$, ma non dal loro ordine o dalla struttura esatta di Σ. In altre parole, esiste una funzione $f : \mathbb{R}^n \to \mathbb{R}$ tale che:

$$\|A\| = f(\sigma_1(A), \ldots, \sigma_n(A)),$$

e questa funzione deve essere simmetrica nelle sue variabili (invariante per permutazioni degli argomenti) a causa dell'invarianza sotto permutazioni spiegata sopra. □

4.3.2 Il teorema di Eckart-Young-Mirsky

La decomposizione ai valori singolari fornisce una risposta esplicita e costruttiva al problema dell'approssimazione di rango basso di trovare B di rango al più k che minimizzi $\|A - B\|$, rispetto alla norma spettrale o di Frobenius.

Questo ha applicazioni nella compressione dei dati (una matrice di rango basso è molto più economica da memorizzare di una piena), nell'analisi dei dati e molto altro.

Teorema [Eckart-Young-Mirsky] Sia $A \in \mathbb{C}^{m\times n}$, e $A = U\Sigma V^*$ la sua SVD. Sia A_k definita come segue:

$$A_k := U\Sigma_k V^*, \quad \Sigma_k := \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \\ & & \end{bmatrix}$$

dove Σ_k è uguale a Σ con $\sigma_{k+1}, \ldots, \sigma_{\min\{m,n\}}$ posti a zero. Allora, valgono le seguenti:

(i) La matrice A_k soddisfa $\sigma_{k+1} = \|A - A_k\|_2 \leq \|A - B\|_2$ per qualsiasi matrice B con rango minore o uguale a k.

(ii) La matrice A_k soddisfa

$$\sqrt{\sigma_{k+1}^2 + \cdots + \sigma_{\min\{m,n\}}^2} = \|A - A_k\|_F \leq \|A - B\|_F$$

per qualsiasi matrice B con rango minore o uguale a k.

Per dimostrare questo risultato, procediamo come segue:

- Dimostriamo l'affermazione (i) per $\|\cdot\|_2$;
- la usiamo per mostrare un Lemma ausiliario sui valori singolari di $A_1 + A_2$, la somma di due matrici arbitrarie;
- usiamo il Lemma per dimostrare il risultato per (ii).

Dimostrazione del Teorema per $\|\cdot\|_2$. Prima verifichiamo che $\|A - A_k\|_2 = \sigma_{k+1}$. Per semplicità, assumiamo che $m \geq n$, l'altro caso può essere ottenuto trasponendo A. Usando la SVD, otteniamo

$$A - A_k = U(\Sigma - \Sigma_k)V^* = U \begin{bmatrix} 0 & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ & & & \sigma_{k+1} & & \\ & & & & \ddots & \\ & & & & & \sigma_n \end{bmatrix} V^*.$$

Prendendo le norme si ottiene $\|A - A_k\|_2 = \|\Sigma - \Sigma_k\|_2 = \sigma_{k+1}$, dove abbiamo usato l'invarianza della norma spettrale sotto trasformazioni unitarie, e che la norma 2 di una matrice diagonale è il massimo dei moduli degli elementi diagonali.

Per concludere dobbiamo verificare che, per qualsiasi matrice B di rango al più k, $\|A - B\|_2 \geq \sigma_{k+1}$. Scegliamo un vettore v di norma unitaria dal sottospazio $\mathrm{Ker}(B) \cap \mathrm{colspan}\{v_1, \ldots, v_{k+1}\}$ dove $v_j := Ve_j$ sono le colonne di V. Poiché $\mathrm{Ker}(B)$ ha dimensione almeno $n - k$, l'intersezione dei sottospazi ha dimensione almeno 1. Possiamo scrivere tale vettore v in coordinate rispetto alle colonne di V:

$$v = \sum_{j=1}^{k+1} \alpha_j v_j = V\alpha, \qquad \alpha := \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_{k+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad \sum_{j=1}^{k+1} |\alpha_j|^2 = 1.$$

Questo produce un'espressione esplicita per $(A - B)v$, della forma

$$(A-B)v = Av = U\Sigma V^* v = U\Sigma\alpha = U \begin{bmatrix} \sigma_1\alpha_1 \\ \vdots \\ \sigma_{k+1}\alpha_{k+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

La norma euclidea di $(A-B)v$ può essere limitata inferiormente come segue:

$$\|(A-B)v\|_2^2 = \sum_{j=1}^{k+1} \sigma_j^2 |\alpha_j|^2 \geq \sigma_{k+1}^2 \sum_{j=1}^{k+1} |\alpha_j|^2 = \sigma_{k+1}^2.$$

Poiché $\|A-B\|_2 \geq \|(A-B)v\|_2$ per qualsiasi $\|v\|_2 = 1$, questo prova l'affermazione. □

lemma [Weyl] Siano A_1, A_2 due matrici di dimensioni compatibili, e $A = A_1 + A_2$. Allora, per qualsiasi $i, j \geq 0$,

$$\sigma_{i+j+1}(A) \leq \sigma_{i+1}(A_1) + \sigma_{j+1}(A_2),$$

dove poniamo $\sigma_k(A) = 0$ per qualsiasi k maggiore della dimensione più piccola di A.

Dimostrazione del Lemma di Weyl. Grazie al Teorema sappiamo che esistono due matrici $A_{i,1}$ e $A_{j,2}$ di rango rispettivamente al più i e j , che rappresentano l'SVD troncata all'ordine i per A_1 e all'ordine j per A_2 e tali che

$$\|A_1 - A_{i,1}\|_2 = \sigma_{i+1}(A_1), \qquad \|A_2 - A_{j,2}\|_2 = \sigma_{j+1}(A_2).$$

Se poniamo $B := A_{i,1} + A_{j,2}$ abbiamo che $\text{rank}(B) \leq i + j$, e quindi, ancora in virtù del Teorema,

$$\begin{aligned} \sigma_{i+j+1}(A) &\leq \|A - B\|_2 = \|A_1 - A_{i,1} + A_2 - A_{j,2}\|_2 \\ &\leq \|A_1 - A_{i,1}\|_2 + \|A_2 - A_{j,2}\|_2 = \sigma_{i+1}(A_1) + \sigma_{j+1}(A_2). \end{aligned}$$

□

Dimostrazione del Teorema per $\|\cdot\|_F$. Per dimostrare la seconda parte del teorema iniziamo verificando $\|A - A_k\|_F^2 = \sigma_{k+1}^2 + \ldots + \sigma_n^2$. Questo segue dallo stesso argomento usato per la norma spettrale, ricordando che il quadrato della norma di Frobenius è la somma dei quadrati degli elementi in una matrice.

Prendiamo ora B come qualsiasi matrice di rango al più k, e affermiamo che

$$\|A - B\|_F^2 \geq \|A - A_k\|_F^2 = \sigma_{k+1}^2 + \ldots + \sigma_n^2.$$

Notiamo che possiamo scrivere

$$\|A-B\|_F^2 = \sum_{l=1}^{n} \sigma_l(A-B)^2$$

Decomponendo $A = (A-B)+B$ e usando il Lemma di Weyl con $i = l-1$ e $j = k$ otteniamo

$$\sigma_{l+k}^2(A) \leq \sigma_l^2(A-B) + 2\sigma_l(A-B)\sigma_{k+1}(B) + \sigma_{k+1}^2(B) = \sigma_l^2(A-B).$$

Usando questa disuguaglianza nell'identità precedente si ottiene

$$\|A-B\|_F^2 = \sum_{l=1}^{n} \sigma_l^2(A-B) \geq \sum_{l=1}^{n-k} \sigma_{l+k}^2(A).$$

□

5 PageRank

Quando utilizziamo un motore di ricerca tipo Google per avere informazioni su un certo argomento ci viene fornita in risposta una lista numerosa di pagine, generalmente migliaia o centinaia di migliaia, che contengono le parole chiave che abbiamo richiesto. Queste pagine vengono ordinate in base alla loro importanza in modo che nei primi posti troviamo quelle che sono certamente più significative e in fondo alla lista si trovano quelle pagine che non hanno una grande rilevanza. In questo modo il motore di ricerca ci permette di evitare di passare in rassegna tutte le migliaia di pagine, impresa che sarebbe umanamente impossibile.

Ma come viene stabilito se una pagina è più importante di un'altra? Con quale criterio vengono ordinate le pagine senza dover entrare dentro il loro contenuto?

Nei motori di ricerca di molti anni fa l'importanza veniva calcolata in base al numero di volte con cui la parola cercata compariva nei documenti presenti nella pagina. Per cui in testa alla lista venivano messi i documenti che contenevano il numero più alto di occorrenze della parola cercata e in fondo alla lista i documenti che contenevano una volta sola la parola chiave. Questo criterio sembrava rispondere pienamente alle esigenze di allora. Questo metodo si rivelò però inefficiente e vulnerabile. Sono stati Sergey Brin e Larry Page, fondatori di Google, a rivoluzionare il modo di attribuire un rango alle pagine del Web indipendentemente dal loro contenuto. La loro idea si basa su un modello matematico particolare e utilizza la teoria di Perron-Frobenius delle matrici non negative. Questa teoria risale ai primi del 1900 quando il mondo di internet non veniva nemmeno immaginato dai più brillanti scrittori di fantascienza. Naturalmente sia Oskar Perron che Georg Frobenius, matematici tedeschi, quando hanno inventato il teorema che va sotto il loro nome non pensavano lontanamente alle applicazioni che esso avrebbe avuto in futuro. La consistenza del modello e l'esistenza e unicità della soluzione è infatti garantita dal teorema di Perron-Frobenius.

Il problema del calcolo della soluzione è un aspetto non trascurabile della questione. La soluzione infatti può essere vista come l'autovettore dominante di una matrice di N righe e di N colonne dove N è uguale al numero di pagine esistenti sul Web. Attualmente il valore di N è di circa 10 miliardi.

Se usassimo i metodi standard per risolvere questo problema, pur usando i più veloci computer disponibili attualmente, dovremmo aspettare milioni di anni prima di conoscere la soluzione. Il metodo di calcolo dell'importanza delle pagine web che viene attualmente usato si basa su un adattamento del metodo delle potenze che viene chiamato algoritmo di PageRank.

Assumiamo di avere N pagine in rete e numeriamole con gli interi da 1 a N. Per descrivere il World-Wide Web è utile usare un grafo orientato in cui i nodi rappresentano le pagine presenti sul Web e gli archi orientati descrivono le connessioni di tali pagine. Più precisamente un arco collega il nodo i col nodo j se la pagina i contiene un link alla pagina j.

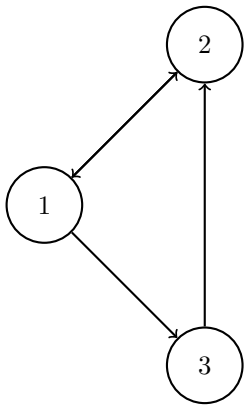

Figura 1: Grafo associato ad un Web costituito da 3 pagine

Ad esempio se il nostro WWW fosse fatto da 3 pagine in cui la pagina 1 punta alla 2 e alla 3, la pagina 2 punta alla 1 e la 3 punta alla 2, allora il grafo sarebbe quello dato in figura 1.

Un grafo orientato può essere univocamente descritto da una matrice di *adiacenza* $H = (h_{i,j})$ di dimensione $N \times N$ in cui $h_{i,j} = 1$ se c'è un arco orientato che collega il nodo i col nodo j (se la pagina i contiene un link alla pagina j) mentre $h_{i,j} = 0$ altrimenti.

La matrice di adiacenza associata al grafo di sopra è

$$H = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

Alcuni possibili criteri per definire l'importanza di una pagina:

1. una pagina è più importante se ha un numero maggiore di link ad altre pagine;
2. una pagina è importante se riceve un numero maggiore di link da altre pagine.

Si vede subito che il criterio 1 non è valido. Se così fosse, basterebbe riempire la propria pagina di un numero arbitrariamente grande di link ad altre pagine per renderla più importante.

Anche il secondo criterio, sebbene più sensato, non è immune da truffa. Non è infatti complicato costruire un numero arbitrario di pagine fittizie che contengono un link alla propria pagina per

poterla rendere più importante. Inoltre in un modello sensato non dovrebbe dare troppa importanza essere puntati da tante pagine di livello trascurabile mentre sarebbe più rilevante essere puntati da (poche) pagine di importanza elevata.

Un criterio più corretto che cattura queste ultime considerazioni è il seguente:
Una pagina i che punta altre pagine, ad esempio $j_1, j_2, \ldots, j_k$, distribuisce la sua importanza in parti uguali alle pagine $j_1, j_2, \ldots, j_k$, e quindi dà $1/k$ della sua importanza alle pagine che punta.

In questo modello, se denotiamo con $d_i = \sum_{j=1}^{N} h_{ij}$, supponendo $d_i \neq 0$ per $i = 1, \ldots, N$, e se indichiamo con w_j l'importanza della pagina j, vale allora

$$w_j = \sum_{i=1}^{N} w_i \frac{h_{ij}}{d_i}, \quad j = 1, \ldots, N.$$

Ad esempio nel caso dell'esempio (1) si ha

$$\begin{aligned} w_1 &= w_2 \\ w_2 &= \frac{1}{2} w_1 + w_3 \\ w_3 &= \frac{1}{2} w_1 \end{aligned}$$

Si osserva che questo non è altro che un problema di autovalori e autovettori formulato nel seguente modo. Posto $e = (1, 1, \ldots, 1)^T$, $d = (d_i) = He$, e $D = \mathrm{diag}(d)$ si ha

$$w^T M = w^T, \quad M = D^{-1} H$$

dove $w^T = (w_1, \ldots, w_N)$.

5.1 Problemi nella formulazione

Elenchiamo alcuni problemi che si incontrano in questa formulazione.

1. Cosa succede se $d_i = 0$ per qualche i? Questo succede nei casi in cui ci sono pagine che non puntano a nulla. Il problema non è insolito, infatti ci possono essere pagine che non hanno link a nulla. I nodi che hanno questa caratteristica sono chiamati *dangling nodes*.
2. Esiste sempre una soluzione?
3. La soluzione è unica (a meno di multipli scalari)?
4. La soluzione è positiva?
5. Come si può calcolare?

Si osserva che i dangling nodes sono individuati per il fatto che essi corrispondono alle righe di H con tutti gli elementi nulli. Per poter trattare il caso in cui esistano dei dangling nodes si introduce una leggera modifica al modello. Più precisamente si sostituisce la matrice iniziale di adiacenza H con una nuova matrice $\widehat{H}$ che coincide con H dappertutto eccetto che nelle righe tutte nulle

in cui gli elementi di $\widehat{H}$ vengono posti tutti uguali a 1. Dal punto di vista modellistico è come assumere che un documento che nel modello originale non cita nessun altro documento nel web, nel nuovo modello modificato va a citare tutti i documenti presenti. Quindi distribuisce $1/N$ della sua importanza uniformemente a tutti.

La matrice $\hat{H}$ viene quindi scritta come

$$\hat{H} = H + ue^T \tag{2}$$

dove u è il vettore con componente 1 in corrispondenza dei dangling nodes e con componente zero altrove.

In seguito denoteremo con M la matrice

$$M = \hat{D}^{-1}\hat{H}, \quad \hat{D} = \text{diag}(\hat{d}), \quad \hat{d} = \hat{H}e. \tag{3}$$

Possiamo dare subito risposta affermativa alla domanda 2 osservando che $Me = e$ e quindi 1 è autovalore, quindi w è un qualsiasi autovettore sinistro corrispondente all'autovalore 1.

Per rispondere alle altre domande dobbiamo riportare alcuni risultati classici della teoria di Perron-Frobenius delle matrici non negative.

5.2 Teorema di Perron-Frobenius

Riportiamo il teorema di Perron-Frobenius:

Teorema (Perron-Frobenius) Sia A una matrice $n \times n$ di elementi non negativi. Allora esiste un autovalore λ di A tale che $\lambda = \rho(A) \geq 0$. Esistono un autovettore destro x e sinistro y corrispondenti a λ con componenti non negative. Se inoltre A è irriducibile allora λ è semplice e gli autovettori x e y hanno componenti positive. Se infine A ha elementi positivi allora λ è l'unico autovalore di modulo massimo.

Si osserva che in base al teorema di Perron-Frobenius ogni soluzione ha sempre componenti non negative come è giusto che sia. Però la sola condizione di nonnegatività non garantisce l'unicità della soluzione (a meno di multipli scalari). Mentre con la condizione di irriducibilità la soluzione è unica.

È facile costruire reti di pagine interconnesse che hanno una matrice di adiacenza riducibile. Quindi il modello così come è stato introdotto non è ancora adeguato.

Si osserva ancora che nel caso di matrici irriducibili e non negative possono esistere altri autovalori che hanno lo stesso modulo del raggio spettrale. Questo crea dei seri problemi dal punto di vista algoritmico.

Esempio: La matrice

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

ha autovalori $1, i, -1, -i$.

Il teorema di Perron-Frobenius esclude però che esistano blocchi di Jordan, relativi al raggio spettrale, di dimensione maggiore di 1.

5.3 Modello di PageRank modificato

Per far fronte ai problemi discussi, il modello del PageRank descritto nella precedente sezione viene così modificato. La matrice M viene sostituita con la matrice

$$A = \gamma M + (1 - \gamma)ev^T, \quad 0 < \gamma < 1, \tag{4}$$

dove v è un arbitrario vettore a componenti non negative tale che $v^T e = 1$ e γ è un parametro, di solito si sceglie $\gamma = 0.85$. In questo modo la matrice A ha elementi positivi. La soluzione quindi esiste, è unica (a meno di multipli) e $\rho(A)$ è l'unico autovalore di modulo 1.

Dal punto di vista modellistico è come se l'importanza di una pagina fosse ripartita in due parti: una frazione γ viene distribuita in base ai link come nel modello originale, la frazione complementare $1 - \gamma$ viene distribuita a tutte le altre pagine secondo un criterio dato dal vettore v. Se ad esempio $v = (1/n)e$ allora la distribuzione è fatta in modo uniforme a tutte le pagine del Web.