INDICE

1 Introduzione

2 Algebre di matrici

Un insieme $\mathcal{L} \subset \mathbb{C}^{n\times n}$ è un'algebra di matrici se $\mathcal{L}$ è un sottospazio vettoriale di $\mathbb{C}^{n\times n}$ ($\alpha, \beta \in \mathbb{C}$, $A, B \in \mathcal{L} \Rightarrow \alpha A + \beta B \in \mathcal{L}$) e il prodotto di matrici di $\mathcal{L}$ è ancora una matrice di $\mathcal{L}$ ($A, B \in \mathcal{L} \Rightarrow AB \in \mathcal{L}$). Vediamo degli esempi di algebre di matrici.

Esempio 1: Algebre di gruppo, matrici circolanti

Sia $\mathcal{G} = \{1, 2, \ldots, n\}$ un gruppo con elemento identico 1. Si può far corrispondere a $\mathcal{G}$ l'insieme di matrici

$$\mathcal{L} = \{A \in \mathbb{C}^{n\times n} : \, a_{i,j} = a_{ki,kj}, \, i, j, k \in \mathcal{G}\}.$$

Una prima cosa da osservare è che $\mathcal{L}$ ammette la seguente altra rappresentazione

$$\mathcal{L} = \{A \in \mathbb{C}^{n\times n} : \, a_{i,j} = a_{1,i^{-1}j}, \, i, j \in \mathcal{G}\},$$

dalla quale si deduce che una matrice in $\mathcal{L}$ è in particolare univocamente definita dalla sua prima riga, che può essere arbitraria. È evidente che $\mathcal{L}$ è un sottospazio vettoriale di $\mathbb{C}^{n\times n}$ di dimensione n. Verifichiamo che è chiuso rispetto alla moltiplicazione di matrici. Siano $A, B \in \mathcal{L}$ e $i, j, k \in \mathcal{G}$. Allora

$$[AB]_{i,j} = \sum_{s\in\mathcal{G}} [A]_{i,s}[B]_{s,j} = \sum_{s\in\mathcal{G}} [A]_{ki,ks}[B]_{ks,kj} = \sum_{r\in\mathcal{G}} [A]_{ki,r}[B]_{r,kj} = [AB]_{ki,kj},$$

il che da' la tesi.

Esercizio. L'insieme $\mathcal{L}$ è chiuso per inversione ? Cioè, $A \in \mathcal{L}$ non singolare implica $A^{-1} \in \mathcal{L}$?

Vediamo un esempio di algebra di gruppo. Sia $\mathcal{G}$ il gruppo ciclico di ordine n, cioè $\mathcal{G} = \{1, 2, \ldots, n\}$ con $i \leftrightarrow g^{i-1}$, $i \in \mathcal{G}$, dove g è un elemento generatore di $\mathcal{G}$ (si noti che $g^n = 1$). Studiamo la struttura dell'algebra di gruppo $\mathcal{L}$ corrispondente. Per definizione, per l'elemento generico di $A \in \mathcal{L}$ si deve avere

$$a_{i,j} = a_{1,i^{-1}j} = a_{1,(g^{i-1})^{-1}(g^{j-1})} = a_{1,g^{n+1-i}g^{j-1}} = \begin{cases} a_{1,g^{j-i}} = a_{1,j-i+1} & j \geq i \\ a_{1,g^{n+j-i}} = a_{1,n+j-i+1} & j < i \end{cases}.$$

È evidente che se $j - i$ è costante allora rimane invariato l'elemento (i, j) di A; in altre parole A deve essere una matrice di Toeplitz. Più precisamente, dalle uguaglianze ottenute segue che A è una matrice di Toeplitz definita univocamente dalla sua prima riga ed ha la seguente struttura:

$$A = \begin{bmatrix} & & & a_{1,k} & & \\ & & & & \ddots & \\ & & & & & a_{1,k} \\ a_{1,k} & & & & & \\ & \ddots & & & & \\ & & a_{1,k} & & & \end{bmatrix},$$

cioè ha, per ogni k, nelle posizioni $(1,k)$, $(2,k+1)$, $\ldots$, $(n+1-k,n)$, $(n+2-k,1)$, $\ldots$, $(n,k-1)$ sempre lo stesso elemento $a_{1,k}$. Una tale matrice A è detta *circolante* [1]. Ad esempio, nel caso $n=4$:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{14} & a_{11} & a_{12} & a_{13} \\ a_{13} & a_{14} & a_{11} & a_{12} \\ a_{12} & a_{13} & a_{14} & a_{11} \end{bmatrix}.$$

Dunque, l'algebra di gruppo $\mathcal{L}$ corrispondente al gruppo ciclico di ordine n coincide con l'insieme delle matrici circolanti $n \times n$. Tale $\mathcal{L}$ viene chiamata $\mathcal{C}$.

Esercizio. Dimostrare che l'algebra di gruppo corrispondente al gruppo ciclico è commutativa, ovvero che due generiche matrici circolanti dello stesso ordine commutano tra loro.

Esercizio [2]. Scrivere la matrice generica dell'algebra di gruppo $\mathcal{L}$ corrispondente al gruppo diedrale $\mathcal{G}$ di ordine 6: $\mathcal{G} = \{1,2,3,4,5,6\} = \{e, r, r^2, f, fr, fr^2\}$ dove r, f sono tali che $r^3 = f^2 = e$, $rfrf = e$. Osservare che $\mathcal{L}$ non è commutativa.

Esempio 2: Il commutatore di una matrice

Sia X una matrice $n \times n$ a elementi in $\mathbb{C}$. Sia

$$\mathcal{L} = \{A \in \mathbb{C}^{n\times n} : AX = XA\}.$$

È semplice mostrare che $\mathcal{L}$ è un'algebra di matrici chiusa per inversione. In generale le matrici di $\mathcal{L}$ non commutano tra loro (si prenda $X = I$).

Esercizio [3]. Siano χ e J le seguenti matrici $n \times n$

$$\chi = \begin{bmatrix} 0 & 1 & & & 1 \\ 1 & 0 & 1 & & \\ & 1 & 0 & \cdot & \\ & & \cdot & \cdot & 1 \\ 1 & & & 1 & 0 \end{bmatrix}, \quad J = \begin{bmatrix} & & & & 1 \\ & & & \cdot & \\ & & 1 & & \\ 1 & & & & \end{bmatrix} \tag{1}$$

(gli elementi non scritti si intendono zeri). Sia $\mathcal{C} + J\mathcal{C}$ l'insieme $\{A + JB : A, B \in \mathbb{C}^{n\times n} \cap \mathcal{C}\}$.

i) Osservare che $\{A \in \mathbb{C}^{n\times n} : AJ = JA\}$ coincide con l'insieme delle matrici *centrosimmetriche*.

ii) Dimostrare che $\mathcal{C} + J\mathcal{C} = \{A \in \mathbb{C}^{n\times n} : A\chi = \chi A\}$ ed osservare che $\mathcal{C} + J\mathcal{C}$ non è commutativo.

Esempio 3: Lo spazio dei polinomi in una matrice

Sia X una matrice $n \times n$ a elementi in $\mathbb{C}$. Dato un polinomio $p(t) = a_0 + a_1 t + \ldots + a_k t^k$, con il simbolo $p(X)$ intendiamo la matrice $a_0 I + a_1 X + \ldots + a_k X^k$. Sia

$$\mathcal{L} = \{p(X)\} = \{p(X) : p = \text{polinomi di grado } k,\ k \in \mathbb{N}\}.$$

È semplice mostrare che $\mathcal{L}$ è un'algebra di matrici commutativa la cui dimensione è data dal grado del polinomio minimo di X ed è quindi minore o uguale ad n. Si dimostra inoltre che $\mathcal{L}$ è chiusa per inversione, cioè se $A \in \mathcal{L}$ è non singolare, allora $A^{-1} \in \mathcal{L}$ (suggerimento: utilizzare il teorema di Cailey-Hamilton applicato ad A).

Un'algebra di questo tipo è l'algebra di gruppo $\mathcal{C}$ delle matrici circolanti. Dimostriamolo. Sia Π la matrice circolante $n \times n$ la cui prima riga è il vettore $[0\,1\,0\,\cdots\,0]$. È semplice osservare che la matrice $\sum_{k=1}^{n} a_k \Pi^{k-1}$ è circolante per ogni scelta degli a_k ($\{p(\Pi)\} \subset \mathcal{C}$), e che la generica matrice circolante, quella la cui prima riga è il generico vettore $[a_{11}\, a_{12}\, \cdots\, a_{1n}]$, si può scrivere nella forma $\sum_{k=1}^{n} a_{1,k} \Pi^{k-1}$ ($\mathcal{C} \subset \{p(\Pi)\}$). In altre parole, vale l'uguaglianza $\mathcal{C} = \{p(\Pi)\}$.

Esercizio. Scrivere la generica matrice degli spazi dei polinomi nelle matrici J e χ in (1).

Esercizio. Studiare l'insieme $\mathcal{C}_{-1}$ dei polinomi nella matrice Π_{-1} ottenuta modificando il valore dell'elemento $(n, 1)$ di Π da 1 a -1. $\mathcal{C}_{-1}$ è noto come lo spazio delle matrici (-1)-circolanti [1].

C'è una relazione tra il commutatore di X e lo spazio dei polinomi in X, studiata nei dettagli ad esempio in [4]. In particolare vale il seguente

Teorema 2.1 Sia X una matrice $n \times n$ a elementi in $\mathbb{C}$. Allora $\{p(X)\} \subset \{A : AX = XA\}$ e $\dim\{p(X)\} \leq n \leq \dim\{A : AX = XA\}$. Inoltre, gli spazi $\{p(X)\}$ e $\{A : AX = XA\}$ coincidono se e solo se $\dim\{p(X)\} = n$ se e solo se $\dim\{A : AX = XA\} = n$, e in tal caso X si dice *non derogatoria*.

Ci sono diverse condizioni equivalenti per la non derogatorietà di una matrice X. Ad esempio, una matrice è non derogatoria se e solo se ad ogni suo autovalore corrisponde un solo blocco di Jordan (ovvero, i polinomi minimo e caratteristico di X coincidono) [4], oppure se e solo se $\{p(X)\}$ ha la struttura di uno spazio di classe $\mathbb{V}$ [5]. Più semplicemente, ogni qual volta si ha

$$p \text{ polinomio}, \;\; p(X) = 0 \quad \Rightarrow \quad \deg p \geq n,$$

la matrice X è non derogatoria. Ad esempio, la matrice J in (1) soddisfa l'identità $J^2 - I = O$, quindi è derogatoria ($\forall\, n > 2$).

Esempio 4: Le matrici triangolari di Toeplitz
Sia Z la seguente matrice *lower-shift* $n \times n$

$$Z = \begin{bmatrix} 0 & & & & \\ 1 & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & 0 \end{bmatrix}. \tag{2}$$

Si noti che la moltiplicazione di Z per un vettore $\mathbf{v} = [v_0\, v_1\, \cdots\, v_{n-1}]^T \in \mathbb{C}^n$ sposta in giù le sue componenti, $Z\mathbf{v} = [0\, v_0\, v_1\, \cdots\, v_{n-2}]^T$. Sia $\mathcal{L}$ lo spazio delle matrici che commutano con Z. Studiamo nei dettagli tale spazio, che ovviamente è anche un'algebra. Sia $A \in \mathbb{C}^{n \times n}$ generica. Allora

$$AZ = \begin{bmatrix} a_{12} & \cdot & a_{1n} & 0 \\ \vdots & & \vdots & \vdots \\ a_{n2} & \cdot & a_{nn} & 0 \end{bmatrix}, \quad ZA = \begin{bmatrix} 0 & \cdots & 0 \\ a_{11} & \cdots & a_{1n} \\ \cdot & & \cdot \\ a_{n-11} & \cdots & a_{n-1n} \end{bmatrix}.$$

Imponendo l'uguaglianza di AZ con ZA si ottengono le condizioni $a_{12} = a_{13} = \ldots = a_{1n} = a_{2n} = \ldots = a_{n-1n} = 0$ e $a_{i,j+1} = a_{i-1,j}$, $i = 2, \ldots, n$, $j = 1, \ldots, n-1$, dalle quali si deduce la struttura di

$A \in \mathcal{L}$: A deve essere una matrice *triangolare inferiore di Toeplitz* del tipo

$$A = \begin{bmatrix} a_{11} & & & & \\ a_{21} & a_{11} & & & \\ a_{31} & a_{21} & a_{11} & & \\ \cdot & \cdot & \cdot & \cdot & \\ a_{n1} & \cdot & a_{31} & a_{21} & a_{11} \end{bmatrix}.$$

Ne segue in particolare che $\dim\{A : AZ = ZA\} = n$ e, quindi, per il Teorema 2.1, si ha anche l'identità $\{A : AZ = ZA\} = \{p(Z)\}$. Effettivamente, se si esaminano le potenze di Z ci si accorge che la matrice triangolare di Toeplitz A di cui sopra coincide con il polinomio $\sum_{k=1}^{n} a_{k1} Z^{k-1}$.

Osserviamo che l'inversa di una matrice triangolare inferiore di Toeplitz è ancora triangolare inferiore di Toeplitz (per quanto detto negli Esempi 2 e 3), ed è quindi anch'essa definita dalla sua prima colonna.

Esempio 5: Algebre di matrici simultaneamente diagonalizzabili

Sia $M \in \mathbb{C}^{n\times n}$ una matrice non singolare. Sia $\mathcal{L}$ lo spazio delle matrici simultaneamente diagonalizzate da M, cioè

$$\mathcal{L} = \operatorname{sd} M := \{MDM^{-1} : D = \text{matrici diagonali }, \ D_{ii} \in \mathbb{C}, \ \forall\, i\}.$$

È evidente che $\mathcal{L}$ è una algebra di matrici commutativa. Questo risultato segue anche dall'osservazione che $\mathcal{L}$ può essere rappresentata come l'insieme dei polinomi in una matrice X; è sufficiente scegliere $X = M\tilde{D}M^{-1}$ con $\tilde{D}$ matrice diagonale con elementi diagonali distinti. Non è vero il contrario, cioè non è vero in generale che uno spazio del tipo $\{p(X)\}$ sia esprimibile nella forma $\{MDM^{-1}\}$ per qualche matrice M non singolare. Ad esempio, non può essere vero per $\{p(Z)\}$, l'insieme delle matrici triangolari inferiori di Toeplitz, perché la matrice Z in (2) non è diagonalizzabile.

Esercizio [5], [7]. Sia $\mathbf{v} \in \mathbb{C}^n$ tale che $(M^T\mathbf{v})_i \neq 0, \forall\, i$. Dimostrare che $A \in \mathcal{L}$ è univocamente determinata dal vettore $\mathbf{v}^T A$, ovvero provare che $A \in \mathcal{L}$ se e solo se $A = Md(M^T A^T \mathbf{v})d(M^T\mathbf{v})^{-1}M^{-1}$, essendo $d(\mathbf{z})$ la matrice diagonale con elementi diagonali le componenti del vettore $\mathbf{z}$. In particolare, se $\mathbf{v} = \mathbf{e}_h$, allora ogni matrice di $\mathcal{L}$ è univocamente determinata dalla sua h-esima riga. Formulare analoghe affermazioni nel caso $(M^{-1}\mathbf{v})_i \neq 0, \ \forall\, i$ (in tal caso A è univocamente determinata dal vettore $A\mathbf{v}$).

Nel seguito descriveremo nei dettagli alcuni esempi di algebre $\{MDM^{-1}\}$ di matrici simultaneamente diagonalizzate da una matrice M. In tali esempi la matrice M è unitaria, $M^H = M^{-1}$, e definisce una trasformata discreta veloce, cioè ogni prodotto matrice-vettore $M\mathbf{z}$, $M^H\mathbf{z}$, $\mathbf{z} \in \mathbb{C}^n$, è calcolabile effettuando non più di $O(n \log n)$ operazioni aritmetiche. Inoltre mostreremo che per A in tali algebre oppure per A triangolare di Toeplitz le operazioni prodotto matrice-vettore $A\mathbf{f}$ e risoluzione sistema lineare $A\mathbf{x} = \mathbf{f}$ possono essere eseguite al basso costo computazionale di $O(n \log n)$ operazioni aritmetiche, utilizzando procedure alternative a quelle standard. Questa caratteristica può essere utilizzata per rendere più efficiente la risoluzione numerica di diversi problemi matematici, anche non di algebra lineare. Vedremo già in questo capitolo una prima interessante applicazione delle algebre di bassa complessità computazionale, nel calcolo dei numeri di Bernoulli [8].

2.1 Algebre di matrici simultaneamente diagonalizzate da trasformate discrete veloci unitarie

Ad ogni matrice unitaria M corrisponde l'algebra di matrici $\mathcal{L} = \{MDM^{-1}\}$, chiusa per inversione e per trasposizione coniugata. Se M è reale, le matrici di $\mathcal{L}$ sono simmetriche ed esiste una base per $\mathcal{L}$ costituita da matrici reali. Vedremo presto che quest'ultima affermazione può essere vera anche se M non è reale. Le algebre descritte nel seguito sono quelle corrispondenti a trasformate unitarie non reali di tipo Fourier, e reali di tipo Hartley e trigonometriche. Altre algebre, corrispondenti ad altre trasformate unitarie, sono di interesse. Menzioniamo in particolare quelle corrispondenti alle trasformazioni di Householder, che tra l'altro sono di complessità computazionale minima $O(n)$, di cui una applicazione in ottimizzazione numerica è descritta in [9].

Le algebre $\mathcal{C}$, $\mathcal{C}_{-1}$, $\mathcal{C}_\phi$, e la trasformata discreta di Fourier (DFT)

Sia Π la matrice $n \times n$ circolante con prima riga $[\,0\ 1\ 0 \cdots 0\,]$. Sia $\mathbf{e}$ è il vettore di $\mathbb{C}^n$ le cui componenti sono tutte uguali a 1. Allora $\Pi\mathbf{e} = \mathbf{e} = 1\mathbf{e}$. Inoltre, per $\omega \in \mathbb{C}$ si ha

$$\Pi \begin{bmatrix} 1 \\ \omega \\ \omega^2 \\ . \\ \omega^{n-1} \end{bmatrix} = \begin{bmatrix} \omega \\ \omega^2 \\ . \\ \omega^{n-1} \\ 1 \end{bmatrix} = \omega \begin{bmatrix} 1 \\ \omega \\ . \\ \omega^{n-2} \\ \omega^{n-1} \end{bmatrix},$$

dove l'ultima uguaglianza vale se $\omega^n = 1$. Più in generale, se $\omega^n = 1$, valgono le seguenti identità vettoriali

$$\Pi \begin{bmatrix} 1 \\ \omega^j \\ . \\ \omega^{(n-1)j} \end{bmatrix} = \begin{bmatrix} \omega^j \\ . \\ \omega^{(n-1)j} \\ 1 \end{bmatrix} = \omega^j \begin{bmatrix} 1 \\ \omega^j \\ . \\ \omega^{(n-1)j} \end{bmatrix},\ j = 0, 1, \ldots, n-1,$$

che, messe insieme, diventano l'identità matriciale $\Pi W = W D_{1\omega^{n-1}}$ coinvolgente le matrici $D_{1\omega^{n-1}}$ e W definite qui di seguito:

$$D_{1\omega^{n-1}} = \begin{bmatrix} 1 & & & \\ & \omega & & \\ & & . & \\ & & & \omega^{n-1} \end{bmatrix},\ W = \begin{bmatrix} 1 & 1 & \cdot & 1 & \cdot & 1 \\ 1 & \omega & & \omega^j & & \omega^{n-1} \\ . & . & & . & & . \\ 1 & \omega^{n-1} & . & \omega^{(n-1)j} & . & \omega^{(n-1)(n-1)} \end{bmatrix}.$$

Scegliendo ω anche tale che $\omega^j \neq 1$, $0 < j < n$, la matrice diagonale $D_{1\omega^{n-1}}$ viene ad avere sulla diagonale tutti gli autovalori di Π, che risultano dunque distinti, e le colonne della matrice W vengono ad essere corrispondenti autovettori, unitariamente ortogonali (Π è una matrice *normale* e autovettori corrispondenti ad autovalori distinti di una matrice normale sono unitariamente ortogonali [4]).

Un risultato più completo è riportato nella seguente

Proposizione 2.2 Sia $\omega \in \mathbb{C}$ tale che $\omega^n = 1$, $\omega^j \neq 1$ per $0 < j < n$, e $W \in \mathbb{C}^{n\times n}$ la matrice $W = (\omega^{(i-1)(j-1)})_{i,j=1}^n$. Allora $W^H W = nI$.

Dimostrazione. Poiché $|\omega| = 1$, $\overline{\omega} = \omega^{-1}$, si ha

$$[W^H W]_{ij} = [\overline{W} W]_{ij} = \sum_{k=1}^{n} [\overline{W}]_{ik} [W]_{kj} = \sum_{k=1}^{n} \overline{\omega}^{(i-1)(k-1)} \omega^{(k-1)(j-1)} = \sum_{k=1}^{n} \omega^{(k-1)(j-i)} = \sum_{k=1}^{n} (\omega^{j-i})^{k-1}.$$

Quindi $[W^H W]_{ij} = n$ se $i = j$, e $[W^H W]_{ij} = \frac{1-(\omega^{j-i})^n}{1-\omega^{j-i}} = 0$ se $i \neq j$ (si noti che l'ipotesi $\omega^j \neq 1$ per $0 < j < n$ è essenziale per rendere $1 - \omega^{j-i} \neq 0$ se $j \neq i$). $\square$

Da ora in poi si suppone ω tale che $\omega^n = 1$, $\omega^j \neq 1$ per $0 < j < n$. Per il risultato della Proposizione, possiamo dunque dire che la seguente matrice (simmetrica) *di Fourier*

$$F = \frac{1}{\sqrt{n}} W, \quad W = (\omega^{(i-1)(j-1)})_{i,j=1}^{n}, \ \omega^n = 1, \ \omega^j \neq 1, \ 0 < j < n, \tag{3}$$

è unitaria, i.e. $F^H F = I$.

Esercizio. Provare che $F^2 = J\Pi$ (J è la matrice di permutazione $J\mathbf{e}_k = \mathbf{e}_{n+1-k}$, $k = 1, \ldots, n$, già considerata in (1)). Ne segue che F^H si ottiene da F permutando le sue colonne (righe), infatti $F^H = J\Pi F = FJ\Pi$. Dimostrare che se λ è autovalore di F allora $\lambda \in \{1, -1, \mathbf{i}, -\mathbf{i}\}$, essendo $\mathbf{i}$ l'unità immaginaria.

L'identità matriciale soddisfatta da Π e da W può essere ovviamente riscritta in termini di F, cioè si ha che $\Pi F = F D_{1\omega^{n-1}}$. Quindi otteniamo l'uguaglianza

$$\Pi = F D_{1\omega^{n-1}} F^H$$

da cui segue che la matrice di Fourier diagonalizza la matrice Π ($F^H \Pi F$ è diagonale), o, più precisamente, che *le colonne della matrice di Fourier formano un sistema di n autovettori unitariamente ortonormali per la matrice Π con corrispondenti autovalori $1, \omega, \ldots, \omega^{n-1}$, essendo ω una radice n-esima principale dell'unità*. Ma se F diagonalizza Π, allora diagonalizza tutti i polinomi in Π, ovvero tutte le matrici circolanti $n \times n$. Più precisamente, chiamata $\mathcal{C}(\mathbf{a})$ la matrice circolante con prima riga $\mathbf{a}^T$, $\mathcal{C}(\mathbf{a}) = \sum_{k=1}^{n} a_k \Pi^{k-1}$, si ha che

$$\mathcal{C}(\mathbf{a}) = \sum_{k=1}^{n} a_k (F D_{1\omega^{n-1}} F^H)^{k-1} = F\big(\sum_{k=1}^{n} a_k D_{1\omega^{n-1}}^{k-1}\big) F^H = F \operatorname{diag}\big(\sum_{k=1}^{n} a_k \omega^{(j-1)(k-1)}, \ j = 1, \ldots, n\big) F^H.$$

Quindi, $\mathcal{C}(\mathbf{a}) = Fd(W\mathbf{a})F^H = \sqrt{n} Fd(F\mathbf{a})F^H = Fd(F^T\mathbf{a})d(F^T\mathbf{e}_1)^{-1}F^H$, dove $\mathbf{a} = \mathcal{C}(\mathbf{a})^T\mathbf{e}_1$.

Esercizio. Sia Π_{-1} la matrice $n \times n$ (-1)-circolante con prima riga $[0\,1\,0\,\cdots\,0]$. Procedendo analogamente al caso circolante cercare una matrice M unitaria tale che $\mathcal{C}_{-1} = \{MDM^{-1} : D = \text{diagonali}\}$. (Nota: M si ottiene da F moltiplicando le sue righe, dalla prima alla n-esima, rispettivamente per 1, ρ, …, ρ^{n-1}, dove ρ è una radice principale n-esima di -1). Più in generale, posto $\Pi_\phi = Z^T + \phi \mathbf{e}_n \mathbf{e}_1^T$, $\phi \in \mathbb{C}$, e considerato l'insieme $\mathcal{C}_\phi$ dei polinomi in Π_ϕ, dimostrare che $\mathcal{C}_\phi = \{MDM^{-1} : D = \text{diagonali}\}$ con $M = D_\phi F$ per una opportuna matrice diagonale D_ϕ, e che M è unitaria se $|\phi| = 1$. Osservare quindi che $\mathcal{C}_\phi(\mathbf{a}) = \sum_{k=1}^{n} a_k \Pi_\phi^{k-1}$, la matrice ϕ-circolante la cui prima riga è $\mathbf{a}^T$, ammette la seguente rappresentazione: $\mathcal{C}_\phi(\mathbf{a}) = D_\phi Fd(FD_\phi\mathbf{a})d(FD_\phi\mathbf{e}_1)^{-1}(D_\phi F)^{-1}$.

Esercizio. Sia T una matrice di Toeplitz $n \times n$, i.e. $T = (t_{i-j})_{i,j=1}^{n}$, per certi $t_k \in \mathbb{C}$. Mostrare che $T = A + B$ dove A è una matrice circolante e B è una matrice (-1)-circolante.

Proposizione 2.3 Dato $\mathbf{z} \in \mathbb{C}^n$, la complessità del prodotto matrice-vettore $F\mathbf{z}$ è al più $O(n \log n)$. Tale operazione è chiamata trasformata discreta di Fourier (DFT) di $\mathbf{z}$. Come conseguenza, sia il vettore prodotto matrice-vettore $\mathcal{C}(\mathbf{a})\mathbf{z}$ che la soluzione del sistema lineare $\mathcal{C}(\mathbf{a})\mathbf{x} = \mathbf{f}$, $\mathbf{f} \in \mathbb{C}^n$, sono calcolabili effettuando due DFT (dopo il pre-calcolo della DFT $F\mathbf{a}$), e, quindi, con al più $O(n \log n)$ operazioni aritmetiche. Analoghe affermazioni valgono più in generale per $\mathcal{C}_\phi(\mathbf{a})$.

Osservazione. Come conseguenza della Proposizione 2.3 e dell'ultimo esercizio, si può dire che anche il prodotto matrice di Toeplitz $n \times n$ per vettore è calcolabile con al più $O(n \log n)$ operazioni aritmetiche. Questo risultato ci permetterà di introdurre un metodo di costo $O(n \log n)$ per il calcolo della prima colonna dell'inversa di una generica matrice triangolare inferiore di Toeplitz, ovvero per la risoluzione dei sistemi triangolari di Toeplitz. Non è invece noto un algoritmo che risolve sistemi di Toeplitz generici $T\mathbf{x} = \mathbf{f}$ con al più $O(n \log n)$ operazioni aritmetiche, a meno che nel conto delle operazioni si omettano quelle fatte su T e non su $\mathbf{f}$, e questo è vero anche se si suppone T simmetrica (si vedano gli argomenti precondizionamento di matrici di Toeplitz e formule di dislocamento per l'inversa di matrici di Toeplitz, entrambi trattati più avanti). Più precisamente, il problema sta nel fatto che il calcolo del vettore (o dei vettori) che definiscono l'inversa di T richiede in generale più di $O(n \log n)$ operazioni aritmetiche, se T non è triangolare. Si veda ad esempio il caso, pur favorevole, in cui T è definita positiva, trattato in [10], in cui un solo vettore (la prima colonna di T^{-1}, come nel caso triangolare!) definisce T^{-1}.

Dimostrazione. Sia n divisibile per 2. Poiché l'elemento (i,k) di W è $\omega^{(i-1)(k-1)}$ e l'elemento k di $\mathbf{z} \in \mathbb{C}^n$ è z_k, si ha

$$\begin{aligned}(W\mathbf{z})_i &= \textstyle\sum_{k=1}^{n} \omega^{(i-1)(k-1)} z_k = \sum_{j=1}^{n/2} \omega^{(i-1)(2j-2)} z_{2j-1} + \sum_{j=1}^{n/2} \omega^{(i-1)(2j-1)} z_{2j} \\ &= \textstyle\sum_{j=1}^{n/2} (\omega^2)^{(i-1)(j-1)} z_{2j-1} + \sum_{j=1}^{n/2} \omega^{(i-1)(2(j-1)+1)} z_{2j} \\ &= \textstyle\sum_{j=1}^{n/2} (\omega^2)^{(i-1)(j-1)} z_{2j-1} + \omega^{i-1} \sum_{j=1}^{n/2} (\omega^2)^{(i-1)(j-1)} z_{2j}\,.\end{aligned}$$

Si noti che ω è di fatto una funzione di n, cioè la giusta notazione per ω dovrebbe essere ω_n. Allora $\omega^2 = \omega_n^2$ è tale che $(\omega_n^2)^{n/2} = 1$ e $(\omega_n^2)^i \neq 1$, $0 < i < n/2$; in altre parole $\omega_n^2 = \omega_{n/2}$ (ovvero ω_n^2 è radice $n/2$-esima principale di 1). Quindi, abbiamo le identità

$$(W_n\mathbf{z})_i = \sum_{j=1}^{n/2} \omega_{n/2}^{(i-1)(j-1)} z_{2j-1} + \omega_n^{i-1} \sum_{j=1}^{n/2} \omega_{n/2}^{(i-1)(j-1)} z_{2j},\ i = 1, 2, \ldots, n. \tag{4}$$

Ne segue che, per $i = 1, \ldots, \frac{n}{2}$,

$$(W_n\mathbf{z})_i = (W_{n/2} \begin{bmatrix} z_1 \\ z_3 \\ \cdot \\ z_{n-1} \end{bmatrix})_i + \omega_n^{i-1} (W_{n/2} \begin{bmatrix} z_2 \\ z_4 \\ \cdot \\ z_n \end{bmatrix})_i.$$

Inoltre, ponendo $i = \frac{n}{2} + k$, $k = 1, \ldots, \frac{n}{2}$, in (4), otteniamo

$$\begin{aligned}(W_n\mathbf{z})_{\frac{n}{2}+k} &= \sum_{j=1}^{n/2} \omega_{n/2}^{\frac{n}{2}(j-1)} \omega_{n/2}^{(k-1)(j-1)} z_{2j-1} + \omega_n^{\frac{n}{2}} \omega_n^{k-1} \sum_{j=1}^{n/2} \omega_{n/2}^{\frac{n}{2}(j-1)} \omega_{n/2}^{(k-1)(j-1)} z_{2j} \\ &= \sum_{j=1}^{n/2} \omega_{n/2}^{(k-1)(j-1)} z_{2j-1} - \omega_n^{k-1} \sum_{j=1}^{n/2} \omega_{n/2}^{(k-1)(j-1)} z_{2j} \\ &= (W_{n/2} \begin{bmatrix} z_1 \\ z_3 \\ \cdot \\ z_{n-1} \end{bmatrix})_k - \omega_n^{k-1} (W_{n/2} \begin{bmatrix} z_2 \\ z_4 \\ \cdot \\ z_n \end{bmatrix})_k, \ k = 1, \ldots, \tfrac{n}{2}.\end{aligned}$$

Quindi, si ha il seguente risultato

$$W_n\mathbf{z} = \begin{bmatrix} I & D_{1\omega_n^{\frac{n}{2}-1}} \\ I & -D_{1\omega_n^{\frac{n}{2}-1}} \end{bmatrix} \begin{bmatrix} W_{n/2} & O \\ O & W_{n/2} \end{bmatrix} Q_n\mathbf{z}, \quad D_{1\omega_n^{\frac{n}{2}-1}} = \begin{bmatrix} 1 & & & \\ & \omega_n & & \\ & & \cdot & \\ & & & \omega_n^{\frac{n}{2}-1} \end{bmatrix}, \tag{5}$$

dove Q_n è la matrice di permutazione $Q_n\mathbf{z} = [z_1 \ z_3 \ \cdot z_{n-1} \ z_2 \ z_4 \ \cdot z_n]^T$. Per la formula (5), che rappresenta W_n in termini di due matrici $W_{n/2}$, se c_n denota la complessità del prodotto matrice-vettore $F_n\mathbf{z}$, $F_n = \frac{1}{\sqrt{n}} W_n$, allora

$$c_n \leq 2c_{n/2} + rn, \ r \text{ costante},$$

e questo implica $c_n = O(n \log_2 n)$, se n è una potenza di 2. Nel caso n sia divisibile non per 2 ma per $b > 2$, con un procedimento simile a quello visto sopra si ottiene una rappresentazione di W_n in termini di b matrici $W_{n/b}$ (trovarla!). Se n è una potenza di b da tale rappresentazione si deduce un algoritmo per il calcolo di $W_n\mathbf{z}$ di costo $O(n \log_b n)$. □

Le algebre di tipo Hartley e le corrispondenti trasformate discrete

Ovviamente, ogni volta che una matrice $n \times n$ M, unitaria non singolare, definita per tutti gli n, soddisfa una identità del tipo

$$M_n = \Big[\text{ sparse matrix } \Big] \begin{bmatrix} M_{n/b} & & \\ & \cdot & \\ & & M_{n/b} \end{bmatrix} \Big[\text{ permutation matrix } \Big], \tag{6}$$

per ogni b divisore di n, si può dire che i prodotti matrice-vettore $M_n\mathbf{z}$ e $M_n^{-1}\mathbf{z}$ possono essere calcolati con al più $O(n \log n)$ operazioni aritmetiche, e, di conseguenza, l'algebra corrispondente ad M, $\mathcal{L} = \{MDM^{-1}\}$, acquista interesse. Come abbiamo già visto, l'identità di cui sopra è verificata per M =trasformata di Fourier F, e, quindi, più in generale, per $M = D_\phi F$. Ma essa è verificata anche per altre matrici M ed, in particolare, se M è una qualsiasi delle otto trasformate discrete unitarie reali *di tipo Hartley* (vedi [11]).

Sia $\operatorname{cas}\varphi = \cos\varphi + \sin\varphi$. Poniamo

$$H_n = \frac{1}{\sqrt{n}} (\operatorname{cas} \frac{2ij\pi}{n})_{i,j=0}^{n-1}, \quad K_n^T = \frac{1}{\sqrt{n}} (\operatorname{cas} \frac{(2i+1)j\pi}{n})_{i,j=0}^{n-1}, \quad G_n = \frac{1}{\sqrt{n}} (\operatorname{cas} \frac{(2i+1)(2j+1)\pi}{2n})_{i,j=0}^{n-1}.$$

Le matrici H_n, K_n^T, G_n sono matrici unitarie reali. Si noti che G_n è anche persimmetrica. Procedendo in maniera analoga al caso della matrice di Fourier, si può dimostrare che le matrici $V_{2n} = \sqrt{2n}H_{2n}$, $\sqrt{2n}K_{2n}^T$, $\sqrt{2n}G_{2n}$, $\sqrt{2n}K_{2n}$ possono tutte essere espresse in termini di due matrici V_n tramite la seguente identità

$$V_{2n} = S \begin{bmatrix} V_n & O \\ O & V_n \end{bmatrix} Q_{2n}, \quad S \text{ sparsa.}$$

È sufficiente infatti porre

$$S = \begin{bmatrix} I & X \\ I & -X \end{bmatrix}, \text{ con } \begin{cases} X = R_K & \text{se } V_n = \sqrt{n}H_n \\ X = R_\gamma & \text{se } V_n = \sqrt{n}K_n^T \end{cases},$$

$$S = \begin{bmatrix} X & Y \\ -YN & XN \end{bmatrix}, \text{ con } \begin{cases} X = R_+,\ Y = R_-,\ N = J & \text{se } V_n = \sqrt{n}G_n \\ X = \tilde{R}_+,\ Y = \tilde{R}_-,\ N = J\Pi_{-1} & \text{se } V_n = \sqrt{n}K_n \end{cases}.$$

dove $R_K = d(\mathbf{c}) + d(\mathbf{s})J\Pi$, $c_h = \cos\frac{h\pi}{n}$, $s_h = \sin\frac{h\pi}{n}$, $R_\gamma = d(\mathbf{c}) + d(\mathbf{s})J$, $c_h = \cos\frac{(2h+1)\pi}{2n}$, $s_h = \sin\frac{(2h+1)\pi}{2n}$, $R_\pm = d(\mathbf{c}) \pm d(\mathbf{s})J$, $c_h = \cos\frac{(2h+1)\pi}{4n}$, $s_h = \sin\frac{(2h+1)\pi}{4n}$, $\tilde{R}_\pm = d(\mathbf{c}) \pm d(\mathbf{s})J\Pi_{-1}$, $c_h = \cos\frac{h\pi}{2n}$, $s_h = \sin\frac{h\pi}{2n}$, e h ogni volta varia tra 0 e $n-1$.

Si noti che, più in generale, è possibile esprimere V_{mn} in termini di m matrici V_n [11].

Da queste considerazioni segue che le algebre $\mathcal{L} = \operatorname{sd}\frac{1}{\sqrt{n}}V_n$ sono costituite da matrici *di bassa complessità*, ovvero, se $A \in \mathcal{L}$ e $\mathbf{z} \in \mathbb{C}^n$ allora i vettori $A\mathbf{z}$, $\mathbf{z} \in \mathbb{C}^n$, e $\mathbf{x}$ tale che $A\mathbf{x} = \mathbf{f}$, $\mathbf{f} \in \mathbb{C}^n$, sono calcolabili con al più $O(n \log n)$ operazioni aritmetiche. Ha per questo interesse studiare la struttura delle matrici di tali algebre. Per farlo si possono utilizzare convenientemente gli spazi $\mathcal{C}_{\pm1}^S = \{A \in \mathcal{C}_{\pm1} : A = A^T\}$ e $\mathcal{C}_{\pm1}^{SK} = \{A \in \mathcal{C}_{\pm1} : A = -A^T\}$ [11], e così si ottengono le prime quattro algebre di tipo Hartley:

$$\operatorname{sd} H_n = \mathcal{C}^S + J\Pi\mathcal{C}^{SK}, \quad \operatorname{sd} K_n^T = \mathcal{C}^S + J\mathcal{C}^{SK}, \quad \operatorname{sd} G_n = \mathcal{C}_{-1}^S + J\mathcal{C}_{-1}^{SK}, \quad \operatorname{sd} K_n = \mathcal{C}_{-1}^S + J\Pi_{-1}\mathcal{C}_{-1}^{SK}.$$

Le rimanenti quattro algebre Hartley-type sono associate a trasformate parenti stretti di quelle già considerate [11]:

$$\operatorname{sd} H_n I_\eta^T = \mathcal{C}^S + J\Pi\mathcal{C}^S, \quad \operatorname{sd} K_n^T I_\eta = \mathcal{C}^S + J\mathcal{C}^S, \quad \operatorname{sd} G_n I_\mu = \mathcal{C}_{-1}^S + J\mathcal{C}_{-1}^S, \quad \operatorname{sd} K_n I_\mu^T = \mathcal{C}_{-1}^S + J\Pi_{-1}\mathcal{C}_{-1}^S,$$

$$I_\eta = \frac{1}{\sqrt{2}} \begin{bmatrix} \sqrt{2} & & & \\ & I & & J \\ & & \sqrt{2} & \\ & -J & & I \end{bmatrix}, \quad I_\mu = \frac{1}{\sqrt{2}} \begin{bmatrix} I & & -J & \\ & \sqrt{2} & & \\ J & & I & \end{bmatrix}.$$

Studiamo l'algebra $\gamma := \operatorname{sd} G_n$ un po' meglio. Possiamo innanzitutto osservare che per $n = 2 + 4s$ ciascuna riga di G_n ha almeno un elemento nullo, mentre per tutti gli altri valori di n si ha che $[G]_{1k} \neq 0$, $\forall\, k$. Ciò è come dire che le matrici di γ sono – come accade per le ϕ-circolanti – univocamente determinate da una loro riga (e in particolare dalla loro prima riga), ovvero vale la rappresentazione $\gamma = \{Gd(G^T\mathbf{z})d(G^T\mathbf{e}_1)^{-1}G^{-1} : \mathbf{z} \in \mathbb{C}^n\}$ se e solo se $n \neq 2 + 4s$. Per $n = 2 + 4s$, invece, una combinazione di più righe di $A \in \gamma$ è necessaria per definire univocamente A. Ad

esempio si osserva che la somma delle righe prima e ultima di $A \in \gamma$ definiscono A per ogni n, cioè, $\forall\, n$, vale la rappresentazione

$$\gamma = \{Gd(G^T\mathbf{z})d(G^T(\mathbf{e}_1+\mathbf{e}_n))^{-1}G^{-1} : \mathbf{z} \in \mathbb{C}^n\}.$$

La struttura delle matrici di γ è rivelata più chiaramente dall'identità $\gamma = \mathcal{C}_{-1}^S + J\mathcal{C}_{-1}^{SK}$, dimostrata in [11]. Con $\mathcal{C}_{-1}^S$ si intende l'algebra delle matrici (-1)-circolanti simmetriche $n \times n$, e con $\mathcal{C}_{-1}^{SK}$ si intende lo spazio delle matrici (-1)-circolanti antisimmetriche $n \times n$ (una matrice A è antisimmetrica se $A^T = -A$).

Esercizio. Dimostrare le uguaglianze $K_n = R_K H_n$ e $G_n = R_\gamma K_n^T$.

Esercizio. Si provi che lo spazio $\mathcal{C}_{-1}^S + J\mathcal{C}_{-1}^{SK}$ è una algebra di matrici commutativa. Si provi l'uguaglianza $\gamma = \mathcal{C}_{-1}^S + J\mathcal{C}_{-1}^{SK}$.

Esercizio. Sia $\mathcal{C}^S$ l'algebra delle matrici $n \times n$ circolanti simmetriche. Si provi che lo spazio $\eta = \mathcal{C}^S + J\mathcal{C}^S$ ammette la rappresentazione $\eta = \{A : A\chi = \chi A,\, AJ = JA\}$. L'algebra di matrici commutativa η è un altro esempio di algebra di Hartley [11]. Le matrici di η sono univocamente definite dalla loro prima riga, cioè esiste M unitaria reale tale che $\eta(\mathbf{a}) = Md(M^T\mathbf{a})d(M^T\mathbf{e}_1)^{-1}M^{-1}$, dove $\eta(\mathbf{a}) \in \eta$ e $\mathbf{e}_1^T\eta(\mathbf{a}) = \mathbf{a}^T$ $(M = K_n^T I_\eta)$.

Esercizio. È possibile scrivere una matrice di Toeplitz simmetrica come somma di due matrici Hartley-type? E una matrice triangolare di Toeplitz?

L'algebra delle matrici τ e la trasformata discreta seno (DST)

L'algebra di matrici τ, studiata nei dettagli in questa sezione, è solo una delle 16 note algebre *trigonometriche* o di *Jacobi* [12], [35], [37]. Allo stesso modo, la trasformata seno, che diagonalizza τ, è solo una delle 16 note trasformate discrete trigonometriche, di tipo seno e di tipo coseno. Tutte tali 16 trasformate hanno complessità $O(n \log_2 n)$ (anche se, in genere, non soddisfano esattamente una uguaglianza del tipo (6)).

Definiamo l'algebra τ introducendo una sua base. Si consideri la seguente matrice tridiagonale $n \times n$

$$Z + Z^T = \begin{bmatrix} 0 & 1 & & & & \\ 1 & 0 & 1 & & & \\ & 1 & 0 & \cdot & & \\ & & \cdot & \cdot & 1 & \\ & & & 1 & 0 & \end{bmatrix}. \tag{7}$$

Posto $J_1 = I$ e $J_2 = Z + Z^T$, si nota che $\mathbf{e}_1^T J_1 = \mathbf{e}_1^T$, $\mathbf{e}_1^T J_2 = \mathbf{e}_2^T$. Inoltre, poiché

$$(Z+Z^T)^2 = \begin{bmatrix} 1 & 0 & 1 & & & \\ 0 & 2 & 0 & 1 & & \\ 1 & 0 & 2 & \cdot & \cdot & \\ & 1 & \cdot & \cdot & \cdot & 1 \\ & & \cdot & \cdot & 2 & 0 \\ & & & 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & & & \\ 0 & 1 & 0 & 1 & & \\ 1 & 0 & 1 & \cdot & \cdot & \\ & 1 & \cdot & \cdot & \cdot & 1 \\ & & \cdot & \cdot & 1 & 0 \\ & & & 1 & 0 & 0 \end{bmatrix} + I,$$

abbiamo $\mathbf{e}_1^T((Z+Z^T)^2 - I) = [\,0\ 0\ 1\ 0 \cdots 0\,] = \mathbf{e}_3^T$. Si ponga allora $J_3 = (Z+Z^T)^2 - I = J_2(Z +$

$Z^T) - J_1$; si ha $\mathbf{e}_1^T J_3 = \mathbf{e}_3^T$. Più in generale, si ponga $J_{i+1} = J_i(Z + Z^T) - J_{i-1}$, $i = 2, 3, \ldots, n-1$. La matrice J_{i+1} è un polinomio in $Z + Z^T$ di grado i con la proprietà $\mathbf{e}_1^T J_{i+1} = \mathbf{e}_{i+1}^T$.

Per dimostrare tale proprietà delle J_k, supponiamo di sapere che $\mathbf{e}_1^T J_j = \mathbf{e}_j^T$, $j = 1, \ldots, i$ (ed effettivamente lo sappiamo per $i = 1, 2$); allora da ciò segue subito che

$$\mathbf{e}_1^T J_{i+1} = \mathbf{e}_1^T (J_i(Z + Z^T) - J_{i-1}) = (\mathbf{e}_i^T (Z + Z^T)) - \mathbf{e}_{i-1}^T = (\mathbf{e}_{i-1}^T + \mathbf{e}_{i+1}^T) - \mathbf{e}_{i-1}^T = \mathbf{e}_{i+1}^T.$$

Poiché le matrici $J_1, J_2, \ldots, J_n$ sono linearmente indipendenti, possiamo dire che esse generano l'insieme $\{p(Z + Z^T)\}$ di tutti i polinomi nella matrice $Z + Z^T$. Tale insieme è l'algebra τ.

Inoltre, poiché $\mathbf{e}_1^T J_k = \mathbf{e}_k^T$, ogni matrice di τ è univocamente definita dalla sua prima riga, e dati $a_k \in \mathbb{C}$ la matrice di τ la cui prima riga è $\mathbf{a}^T = [a_1 \ \cdots \ a_n]$ è $\tau(\mathbf{a}) = \sum_{k=1}^n a_k J_k$.

Troviamo una rappresentazione della matrice $\tau(\mathbf{a})$ in termini dei suoi autovalori e autovettori. Innanzitutto si osserva che valgono le seguenti uguaglianze vettoriali:

$$\begin{bmatrix} 0 & 1 & & & \\ 1 & 0 & 1 & & \\ & 1 & \cdot & \cdot & \\ & & \cdot & 0 & 1 \\ & & & 1 & 0 \end{bmatrix} \begin{bmatrix} \sin \frac{j\pi}{n+1} \\ \sin \frac{2j\pi}{n+1} \\ \cdot \\ \sin \frac{nj\pi}{n+1} \end{bmatrix} = 2\cos\frac{j\pi}{n+1} \begin{bmatrix} \sin \frac{j\pi}{n+1} \\ \sin \frac{2j\pi}{n+1} \\ \cdot \\ \sin \frac{nj\pi}{n+1} \end{bmatrix}, \ j = 1, \ldots, n.$$

Tali n uguaglianze possono essere riscritte come una unica uguaglianza matriciale $(Z + Z^T)S = SD$ dove S è la matrice

$$S_{ij} = \sqrt{\frac{2}{n+1}} \sin\frac{ij\pi}{n+1}, \quad i, j = 1, \ldots, n, \tag{8}$$

e D è la matrice diagonale con elementi diagonali distinti $D_{jj} = 2\cos\frac{j\pi}{n+1}$, $j = 1, \ldots, n$. Si noti che la matrice S, chiamata matrice seno, è reale, simmetrica e unitaria (provare a dimostrarlo!).

Esercizio. Sia $F_{2(n+1)}$ la matrice di Fourier di ordine $2(n+1)$. Dimostrare che $F_{2(n+1)}$ e la matrice seno S $n \times n$ verificano la seguente identità:

$$\mathbf{i}\big(I - F_{2(n+1)}^2\big)F_{2(n+1)} = \begin{bmatrix} 0 & \mathbf{0}^T & 0 & \mathbf{0}^T \\ \mathbf{0} & S & \mathbf{0} & -SJ \\ 0 & \mathbf{0}^T & 0 & \mathbf{0}^T \\ \mathbf{0} & -JS & \mathbf{0} & JSJ \end{bmatrix}$$

(si noti che $F_{2(n+1)}^2$ è una matrice di permutazione). Come conseguenza, una trasformata seno può essere calcolata effettuando una trasformata discreta di Fourier, ovvero con al più $O(n \log n)$ operazioni aritmetiche.

Quindi, le colonne della matrice seno S formano un sistema di autovettori unitariamente ortonormali per la matrice $Z + Z^T$. In altre parole, la matrice unitaria S diagonalizza $Z + Z^T$ e, ovviamente, diagonalizza ogni polinomio in $Z + Z^T$, i.e. ogni matrice di τ. Riassumendo, si ha che

$$Z + Z^T = SDS, \quad J_k = p_{k-1}(Z + Z^T), \quad \tau = \{S\big(\sum_{k=1}^n a_k p_{k-1}(D)\big)S : \ a_k \in \mathbb{C}\} = \operatorname{sd} S,$$

ed è evidente che la matrice di τ con prima riga $\mathbf{a}^T$ ammette la seguente rappresentazione

$$\tau(\mathbf{a}) = \sum_{k=1}^{n} a_k J_k = Sd(S^T\mathbf{a})d(S^T\mathbf{e}_1)^{-1}S^{-1}.$$

Da questa formula per $\tau(\mathbf{a})$ segue che prodotti matrice-vettore coinvolgenti matrici τ e la risoluzione di sistemi lineari con matrici dei coefficienti in τ hanno complessità al più $O(n \log n)$.

Concludiamo con una osservazione utile per capire rapidamente la struttura delle matrici di τ. Per il Teorema 2.1 ogni matrice A dello spazio τ deve commutare con la matrice $Z + Z^T$, o, più precisamente,

$$\tau = \{A \in \mathbb{C}^{n\times n} : \ A(Z + Z^T) = (Z + Z^T)A\}.$$

Ma ciò è equivalente a richiedere che gli elementi di A soddisfano le seguenti n^2 condizioni di *somma in croce*

$$a_{i,j-1} + a_{i,j+1} = a_{i-1,j} + a_{i+1,j}, \ i, j = 1, \ldots, n,$$

dove si suppone $a_{0,j} = a_{n+1,j} = a_{i,0} = a_{i,n+1} = 0$, $i, j = 1, \ldots, n$. Possiamo usare tali condizioni e il fatto che le matrici di τ sono sia simmetriche che persimmetriche per scrivere $\tau(\mathbf{a})$, per un generico vettore $\mathbf{a} = [a_1\, a_2\, \cdots\, a_n]^T$. Per esempio, per $n = 4$ ed $n = 5$,

$$\tau(\mathbf{a}) = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 \\ a_2 & a_1 + a_3 & a_2 + a_4 & a_3 \\ a_3 & a_2 + a_4 & a_1 + a_3 & a_2 \\ a_4 & a_3 & a_2 & a_1 \end{bmatrix}, \quad \tau(\mathbf{a}) = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 & a_5 \\ a_2 & a_1 + a_3 & a_2 + a_4 & a_3 + a_5 & a_4 \\ a_3 & a_2 + a_4 & a_1 + a_3 + a_5 & a_2 + a_4 & a_3 \\ a_4 & a_3 + a_5 & a_2 + a_4 & a_1 + a_3 & a_2 \\ a_5 & a_4 & a_3 & a_2 & a_1 \end{bmatrix},$$

$$J_1 = I, \ J_2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \ J_3 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \ J_4 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix},$$

e così via.

Esercizio. Provare che per n pari la matrice $J_2 = Z + Z^T$ è invertibile, e calcolare l'inversa.

Risoluzione. Sappiamo che se J_2 è invertibile, allora $J_2^{-1} \in \tau$ (dal fatto che J_2 commuta con $Z + Z^T$ segue che anche J_2^{-1} commuta con $Z + Z^T$). Quindi $J_2^{-1} = \tau(\mathbf{z})$ per qualche $\mathbf{z} \in \mathbb{C}^n$, e la tesi equivale a far vedere che esiste $\mathbf{z}$ per cui $\tau(\mathbf{z})J_2 = I$. Ma l'identità matriciale $\tau(\mathbf{z})J_2 = I$ è equivalente all'identità vettoriale $\mathbf{z}^T J_2 = \mathbf{e}_1^T$. Così, per esempio, per $n = 4$ abbiamo la condizione

$$[z_1\ z_2\ z_3\ z_4] \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} = [1\ 0\ 0\ 0],$$

che implica $z_1 = 0$, $z_2 = 1$, $z_3 = 0$, $z_4 = -1$, e, quindi,

$$J_2^{-1} = \begin{bmatrix} 0 & 1 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 1 & 0 \end{bmatrix} = J_2 - J_4.$$

Studiando i casi $n = 6, 8, \ldots$ si deduce la formula generale per J_2^{-1} (trovarla!).

Esercizio. Sia T una matrice di Toeplitz simmetrica $n \times n$, i.e. $T = (t_{|i-j|})_{i,j=1}^n$, per certi $t_k \in \mathbb{C}$. Mostrare che $T = A + B$ dove A è una matrice τ di ordine n e

$$B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & R & 0 \\ 0 & 0 & 0 \end{bmatrix}, \ R \in \tau \cap \mathbb{C}^{(n-2)\times(n-2)}.$$

Esercizio. Mostrare che la matrice $\mathbf{e}\mathbf{e}^T = (1)_{i,j=1}^n$ non appartiene a τ. Trovare una algebra n-dimensionale del tipo $\{p(X)\}$, $X = Z + Z^T + \varepsilon \mathbf{e}_1 \mathbf{e}_1^T + \varphi \mathbf{e}_n \mathbf{e}_n^T$, per cui $\mathbf{e}\mathbf{e}^T \in \{p(X)\}$.

Esercizio. Scrivere l'inversa della matrice di τ la cui prima riga è $[\,4\ 1\ 0\ 0 \cdots 0\,]$.

Classi di algebre di matrici: algebre di Hessenberg e di tipo Toeplitz più Hankel

Data X $n \times n$ di Hessenberg inferiore si consideri l'algebra $H_X = \{p(X)\}$ dei polinomi in X. Si può facilmente dimostrare che se $X_{i,i+1} \neq 0 \ \forall\, i$, allora H_X ha dimensione n (X è non derogatoria), è generata da matrici J_k tali che $\mathbf{e}_1^T J_k = \mathbf{e}_k^T$, $k = 1, \ldots, n$, e, quindi, per ogni $\mathbf{a} \in \mathbb{C}^n$ è ben definita $H_X(\mathbf{a})$, la matrice di H_X la cui prima riga è $\mathbf{a}^T$. Le algebre H_X sono chiamate *di Hessenberg* in [13], [14] ed ivi utilizzate per unificare diversi risultati particolari su formule *di dislocamento* per la rappresentazione efficiente di matrici *Toeplitz plus Hankel-like* (ottenuti negli anni 70-90) e per ottenerne di nuovi. È evidente che le algebre trigonometriche, come τ, e le algebre ϕ-circolanti, come $\mathcal{C}$, $\mathcal{C}_{-1}$, $\mathcal{C}_0$, sono particolari algebre di Hessenberg.

Anche le otto algebre di tipo Hartley non sono altro che esempi particolari di una classe di algebre significative, studiata in [15]. Lo studio in [15] nasce dall'osservazione che ogni algebra contenente la matrice $X = \Pi_\beta + \Pi_\beta^T + \varepsilon \mathbf{e}_1 \mathbf{e}_1^T + \varphi \mathbf{e}_n \mathbf{e}_n^T$ avrebbe potuto essere coinvolta in una rappresentazione efficiente dell'inversa di una matrice Toeplitz più Hankel $T + H$ e questo perché il rango di $(T+H)^{-1}X - X(T+H)^{-1}$ (il *rango di dislocamento* di $(T+H)^{-1}$, rispetto all'operatore commutatore in questo caso) è indipendente da n. Quindi, in [15] si ottengono esplicitamente, al variare di $\beta, \varepsilon, \varphi$, tutte le algebre $\mathcal{L}$ contenenti X (con base J_k tale che $\mathbf{e}_1^T J_k = \mathbf{e}_k^T$), e le si utilizzano in una formula di dislocamento generale e in formule più particolari, molto efficienti nella rapresentazione di matrici centrosimmetriche di tipo Toeplitz più Hankel.

... Vedi ... ??.

2.2 Complessità dei calcoli con le matrici triangolari di Toeplitz

Moltiplicare una matrice $n \times n$ triangolare inferiore di Toeplitz per un vettore non richiede più di $O(n \log n)$ operazioni aritmetiche. La stessa affermazione vale per la risoluzione di un sistema lineare di n equazioni la cui matrice dei coefficienti è triangolare inferiore di Toeplitz, e questo perché tale operazione può essere ricondotta al calcolo di $O(\log n)$ prodotti matrice-vettore dove la matrice è triangolare di Toeplitz e di dimensione ogni volta la metà. Tutto ciò sarà provato in questa sezione.

Moltiplicare una matrice triangolare inferiore di Toeplitz (t.i.T.) per un vettore

Ci sono almeno due modi per calcolare il prodotto di una matrice di Toeplitz $n \times n$ $T = (t_{i-j})_{i,j=1}^n$ per un vettore in al più $O(n \log n)$ operazioni aritmetiche, ed entrambi prevedono l'uso dell'algoritmo FFT illustrato nella Proposizione 2.3. Uno consiste nell'utilizzare la rappresentazione di T come somma di una matrice circolante e una matrice (-1)-circolante, vista in un esercizio. L'altro,

descritto nei dettagli qui di seguito, si basa sull'osservazione che ogni matrice di Toeplitz può essere *immersa* in una matrice circolante.

Si consideri una generica matrice di Toeplitz T 4×4 ed un vettore $\mathbf{v}$ 4×1. Allora T può essere vista come la sottomatrice in alto a sinistra di una matrice circolante C 8×8, e per il vettore $T\mathbf{v}$ vale la seguente rappresentazione:

$$T\mathbf{v}=\begin{bmatrix} t_0 & t_{-1} & t_{-2} & t_{-3}\\ t_1 & t_0 & t_{-1} & t_{-2}\\ t_2 & t_1 & t_0 & t_{-1}\\ t_3 & t_2 & t_1 & t_0 \end{bmatrix}\begin{bmatrix} v_0\\ v_1\\ v_2\\ v_3 \end{bmatrix}=\left\{\begin{bmatrix} t_0 & t_{-1} & t_{-2} & t_{-3} & 0 & t_3 & t_2 & t_1\\ t_1 & t_0 & t_{-1} & t_{-2} & t_{-3} & 0 & t_3 & t_2\\ t_2 & t_1 & t_0 & t_{-1} & t_{-2} & t_{-3} & 0 & t_3\\ t_3 & t_2 & t_1 & t_0 & t_{-1} & t_{-2} & t_{-3} & 0\\ 0 & t_3 & t_2 & t_1 & t_0 & t_{-1} & t_{-2} & t_{-3}\\ t_{-3} & 0 & t_3 & t_2 & t_1 & t_0 & t_{-1} & t_{-2}\\ t_{-2} & t_{-3} & 0 & t_3 & t_2 & t_1 & t_0 & t_{-1}\\ t_{-1} & t_{-2} & t_{-3} & 0 & t_3 & t_2 & t_1 & t_0 \end{bmatrix}\begin{bmatrix} v_0\\ v_1\\ v_2\\ v_3\\ 0\\ 0\\ 0\\ 0 \end{bmatrix}\right\}_4=\left\{C\begin{bmatrix}\mathbf{v}\\ \mathbf{0}\end{bmatrix}\right\}_4$$

dove col simbolo $\{\mathbf{z}\}_4$ si intende il vettore 4×1 le cui componenti sono le prime quattro componenti del vettore $\mathbf{z}$.

Se T è $n\times n$ e $\mathbf{v}$ è $n\times 1$, allora l'osservazione vale ancora e può essere generalizzata:

$$T\mathbf{v}=\left\{C\begin{bmatrix}\mathbf{v}\\ \mathbf{0}_{(b-1)n}\end{bmatrix}\right\}_n,\ C=\mathcal{C}(\mathbf{a})=\sqrt{bn}F_{bn}d(F_{bn}\mathbf{a})F_{bn}^H,\ \mathbf{a}=\begin{bmatrix} t_0\\ t_{-1}\\ \cdot\\ t_{-n+1}\\ \mathbf{0}_{(b-2)n+1}\\ t_{n-1}\\ \cdot\\ t_1 \end{bmatrix}. \tag{9}$$

Se n è una potenza di b ($b=2,3,\ldots$), da questa formula si deduce immediatamente una procedura di costo $O(n\log_b n)$ per il calcolo del prodotto di una matrice di Toeplitz $n\times n$ per un vettore. Tale procedura ha come sotto-procedura l'algoritmo FFT considerato nella Proposizione 2.3.

Nella prossima sezione vedremo che la risoluzione di un sistema lineare triangolare inferiore di Toeplitz di n equazioni può ricondursi al calcolo di $O(\log n)$ prodotti matrice-vettore dove la matrice è sempre triangolare inferiore di Toeplitz ed è di dimensione variabile, che si dimezza ogni volta. Ne segue che è opportuno avere a disposizione un metodo che effettui tali prodotti il più efficientemente possibile. Un metodo abbastanza efficiente si ottiene ponendo $t_{-i}=0$, $i=1,\ldots,n-1$, nella procedura sopra illustrata. Sarebbero benvenuti metodi che sfruttino meglio la triangolarità delle nostre matrici di Toeplitz.

Un algoritmo per la risoluzione di sistemi triangolari di Toeplitz

In questa sezione si illustrerà un algoritmo di costo $O(n\log_2 n)$ per il calcolo di $\mathbf{x}$ tale che $A\mathbf{x}=\mathbf{f}$, essendo A una matrice triangolare inferiore di Toeplitz $n\times n$ con n potenza di 2 e $[A]_{11}=1$.

Lemmi preliminari

Dato un vettore $\mathbf{v}=[v_0\ v_1\ v_2\ \cdots]^T$, $v_i\in\mathbb{C}$ (in breve $\mathbf{v}\in\mathbb{C}^{\mathbb{N}}$), sia $L(\mathbf{v})$ la matrice semi-infinita

triangolare inferiore di Toeplitz con prima colonna $\mathbf{v}$, i.e.

$$L(\mathbf{v}) = \sum_{k=0}^{+\infty} v_k Z^k, \quad Z = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ & 1 & 0 & \\ & & \cdot & \cdot \end{bmatrix}.$$

Lemma 2.4 (Lemma 1) Siano $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ vettori di $\mathbb{C}^{\mathbb{N}}$. Allora $L(\mathbf{a})L(\mathbf{b}) = L(\mathbf{c})$ se e soltanto se $L(\mathbf{a})\mathbf{b} = \mathbf{c}$.

Dimostrazione. Se $L(\mathbf{a})L(\mathbf{b}) = L(\mathbf{c})$, allora la prima colonna della matrice $L(\mathbf{a})L(\mathbf{b})$ deve essere uguale alla prima colonna della matrice $L(\mathbf{c})$, e queste sono rispettivamente i vettori $L(\mathbf{a})\mathbf{b}$ e $\mathbf{c}$. Viceversa, supponiamo che $L(\mathbf{a})\mathbf{b} = \mathbf{c}$. Consideriamo la matrice $L(\mathbf{a})L(\mathbf{b})$. Questa, in quanto prodotto di matrici triangolari inferiori di Toeplitz, è una matrice triangolare inferiore di Toeplitz, e, per ipotesi, la sua prima colonna, $L(\mathbf{a})\mathbf{b}$, coincide con il vettore $\mathbf{c}$, che è la prima colonna della matrice triangolare inferiore di Toeplitz $L(\mathbf{c})$. La tesi segue dal fatto che le matrici triangolari inferiori di Toeplitz sono univocamente definite dalla loro prima colonna. □

Dato un vettore $\mathbf{v} = [v_0\, v_1\, v_2\, \cdots]^T \in \mathbb{C}^{\mathbb{N}}$, sia E la matrice semi-infinita di 0 e 1 che manda $\mathbf{v}$ nel vettore $E\mathbf{v} = [v_0\, 0\, v_1\, 0\, v_2\, 0\, \cdots]^T$:

$$E = \begin{bmatrix} 1 & & & \\ 0 & & & \\ 0 & 1 & & \\ 0 & 0 & & \\ 0 & 0 & 1 & \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix}.$$

In altre parole, l'azione di E su $\mathbf{v}$ ha l'effetto di inserire uno zero tra due successive componenti di $\mathbf{v}$. Si osserva facilmente che

$$E^2 = \begin{bmatrix} 1 & & & \\ 0 & & & \\ 0 & & & \\ 0 & & & \\ 0 & 1 & & \\ 0 & 0 & & \\ 0 & 0 & & \\ 0 & 0 & & \\ 0 & 0 & 1 & \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix}, \quad E^s = \begin{bmatrix} 1 & & & \\ \mathbf{0} & & & \\ 0 & 1 & & \\ \mathbf{0} & \mathbf{0} & & \\ 0 & 0 & 1 & \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix}, \ \mathbf{0} = \mathbf{0}_{2^s-1},$$

cioè l'azione di E^s su $\mathbf{v}$ ha l'effetto di inserire $2^s - 1$ zeri tra due successive componenti di $\mathbf{v}$.

Lemma 2.5 (Lemma 2) Siano $\mathbf{u}$, $\mathbf{v}$ vettori di $\mathbb{C}^{\mathbb{N}}$ con $u_0 = v_0 = 1$. Allora $L(E\mathbf{u})E\mathbf{v} = EL(\mathbf{u})\mathbf{v}$, e, più in generale, per ogni $s \in \mathbb{N}$ si ha $L(E^s\mathbf{u})E^s\mathbf{v} = E^sL(\mathbf{u})\mathbf{v}$.

Dimostrazione. Scrivendo i vettori $L(E\mathbf{u})E\mathbf{v}$ e $EL(\mathbf{u})\mathbf{v}$ si osserva che sono uguali. Moltiplicando a sinistra per E l'identità $L(E\mathbf{u})E\mathbf{v} = EL(\mathbf{u})\mathbf{v}$ ed utilizzando tale stessa identità con i vettori $E\mathbf{u}$ ed $E\mathbf{v}$ al posto, rispettivamente, di $\mathbf{u}$ e $\mathbf{v}$, si osserva che vale anche l'uguaglianza $L(E^2\mathbf{u})E^2\mathbf{v} = E^2L(\mathbf{u})\mathbf{v}$. ... □

L'algoritmo

Sia A una matrice t.i.T. $n \times n$ con $[A]_{11} = 1$. Si vuole risolvere il sistema $A\mathbf{x} = \mathbf{f}$. L'algoritmo seguente sfrutta l'osservazione che A^{-1} è ancora una matrice t.i.T. $n \times n$.

1 Si calcola la prima colonna della matrire t.i.T. A^{-1}, ovvero si risolve il sistema lineare particolare $A\mathbf{x} = \mathbf{e}_1$ utilizzando l'algoritmo di costo $O(n \log_2 n)$ illustrato nella sezione seguente, basato sulla ripetuta applicazione dei Lemmi 1 e 2.

2 Si calcola il prodotto matrice-vettore $A^{-1}\mathbf{f}$, ad esempio tramite la formula (9), effettuando non più di $O(n \log_2 n)$ operazioni aritmetiche.

Prima di procedere è importante osservare che è possibile definire una procedura più generale ma di costo $O(n \log_b n)$, conveniente quando n, l'ordine della matrice t.i.T., è uguale a una potenza di b, riformulando il Lemma 2 di cui sopra nel caso in cui E è definita come la matrice di 0 e 1 la cui azione su $\mathbf{v}$ ha l'effetto di inserire $b-1$ zeri tra due componenti successive di $\mathbf{v}$. Alla base di tale procedura c'è l'algoritmo FFT di costo $O(n \log_b n)$ per il calcolo veloce della DFT di un vettore di dimensione n potenza di b. Si veda [8].

Il calcolo della prima colonna dell'inversa di una matrice t.i.T.

Per semplicità illustriamo l'algoritmo per il calcolo di $\mathbf{x}$ tale che $A\mathbf{x} = \mathbf{e}_1$ nel caso $n = 8$. Indicheremo, a volte, cos'è che cambia nel caso generale $n = 2^s$, $s \in \mathbb{N}$; comunque tale caso è facilmente deducibile da quello considerato. L'algoritmo si divide in due parti. Nella prima parte si introducono e si calcolano matrici triangolari inferiori di Toeplitz che moltiplicate, una dopo l'altra, a sinistra per la matrice A, hanno l'effetto di trasformarla nella matrice identica. Nella seconda parte si moltiplicano tali matrici, di nuovo una dopo l'altra, per il vettore $\mathbf{e}_1$. Come si vedrà, non si fa altro che applicare una specie di eliminazione di Gauss, ma, invece di annullare colonne, si annullano diagonali. Il costo finale $O(n \log_2 n)$ dell'algoritmo deriva dal fatto che ad ogni passo della prima parte si annullano metà delle diagonali rimaste non nulle, e dal fatto che la seconda parte è semplificabile sfruttando il fatto che il vettore $\mathbf{e}_1$ ha solo una componente non nulla.

Per prima cosa osserviamo che la matrice A 8×8 può essere vista come la sottomatrice in alto a sinistra di una matrice $L(\mathbf{a})$ semi-infinita triangolare inferiore di Toeplitz con prima colonna $[1\, a_1\, a_2\, \cdot\, a_7\, a_8\, \cdot]^T$.

Passo 1. Trovare $\hat{\mathbf{a}}$ tale che

$$L(\mathbf{a})\hat{\mathbf{a}} = \begin{bmatrix} 1 & & & & & & & & \\ a_1 & 1 & & & & & & & \\ a_2 & a_1 & 1 & & & & & & \\ a_3 & a_2 & a_1 & 1 & & & & & \\ a_4 & a_3 & a_2 & a_1 & 1 & & & & \\ a_5 & a_4 & a_3 & a_2 & a_1 & 1 & & & \\ a_6 & a_5 & a_4 & a_3 & a_2 & a_1 & 1 & & \\ a_7 & a_6 & a_5 & a_4 & a_3 & a_2 & a_1 & 1 & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} 1 \\ \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \hat{a}_4 \\ \hat{a}_5 \\ \hat{a}_6 \\ \hat{a}_7 \\ \cdot \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ a_1^{(1)} \\ 0 \\ a_2^{(1)} \\ 0 \\ a_3^{(1)} \\ 0 \\ \cdot \end{bmatrix} = E\mathbf{a}^{(1)}$$

per certi $a_i^{(1)} \in \mathbb{C}$ e calcolare tali $a_i^{(1)}$. Il calcolo degli $a_i^{(1)}$ richiede, una volta noto $\hat{\mathbf{a}}$, il prodotto di una matrice t.i.T. 8×8 ($2^s \times 2^s$) per un vettore − o, più precisamente, due prodotti t.i.T. 4×4 ($2^{s-1} \times 2^{s-1}$) per vettore (perché?). Vedremo che $\hat{\mathbf{a}}$ è disponibile a costo zero.

Notiamo che allora, per il Lemma 1, si ha $L(\hat{\mathbf{a}})L(\mathbf{a}) = L(E\mathbf{a}^{(1)})$, cioè la matrice t.i.T. $L(\mathbf{a})$ è trasformata in una matrice t.i.T. che alterna ciascuna diagonale non nulla con una nulla.

Passo 2. Trovare $\hat{\mathbf{a}}^{(1)}$ tale che

$$L(E\mathbf{a}^{(1)})E\hat{\mathbf{a}}^{(1)} = \begin{bmatrix} 1 & & & & & & & & \\ 0 & 1 & & & & & & & \\ a_1^{(1)} & 0 & 1 & & & & & & \\ 0 & a_1^{(1)} & 0 & 1 & & & & & \\ a_2^{(1)} & 0 & a_1^{(1)} & 0 & 1 & & & & \\ 0 & a_2^{(1)} & 0 & a_1^{(1)} & 0 & 1 & & & \\ a_3^{(1)} & 0 & a_2^{(1)} & 0 & a_1^{(1)} & 0 & 1 & & \\ 0 & a_3^{(1)} & 0 & a_2^{(1)} & 0 & a_1^{(1)} & 0 & 1 & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ \hat{a}_1^{(1)} \\ 0 \\ \hat{a}_2^{(1)} \\ 0 \\ \hat{a}_3^{(1)} \\ 0 \\ \cdot \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ a_1^{(2)} \\ 0 \\ 0 \\ 0 \\ \cdot \end{bmatrix} = E^2\mathbf{a}^{(2)}$$

per certi $a_i^{(2)} \in \mathbb{C}$ e calcolare tali $a_i^{(2)}$. Il calcolo degli $a_i^{(2)}$ richiede, una volta noto $\hat{\mathbf{a}}^{(1)}$, il prodotto di una matrice t.i.T. 4×4 ($2^{s-1} \times 2^{s-1}$) per un vettore − o, più precisamente, due prodotti t.i.T. 2×2 ($2^{s-2} \times 2^{s-2}$) per vettore.

Notiamo che allora, per il Lemma 1, si ha $L(E\hat{\mathbf{a}}^{(1)})L(E\mathbf{a}^{(1)}) = L(E^2\mathbf{a}^{(2)})$, cioè la matrice t.i.T. $L(\mathbf{a})$ è trasformata in una matrice t.i.T. che alterna ciascuna diagonale non nulla con tre nulle.

Si noti anche che, per il Lemma 2, se $L(\mathbf{a}^{(1)})\hat{\mathbf{a}}^{(1)} = E\mathbf{a}^{(2)}$ allora $L(E\mathbf{a}^{(1)})E\hat{\mathbf{a}}^{(1)} = E^2\mathbf{a}^{(2)}$. Vedremo che $\hat{\mathbf{a}}^{(1)}$ tale che $L(\mathbf{a}^{(1)})\hat{\mathbf{a}}^{(1)} = E\mathbf{a}^{(2)}$ è disponibile a costo zero.

Passo 3. Trovare $\hat{\mathbf{a}}^{(2)}$ tale che

$$L(E^2\mathbf{a}^{(2)})E^2\hat{\mathbf{a}}^{(2)} = \begin{bmatrix} 1 & & & & & & & & \\ 0 & 1 & & & & & & & \\ 0 & 0 & 1 & & & & & & \\ 0 & 0 & 0 & 1 & & & & & \\ a_1^{(2)} & 0 & 0 & 0 & 1 & & & & \\ 0 & a_1^{(2)} & 0 & 0 & 0 & 1 & & & \\ 0 & 0 & a_1^{(2)} & 0 & 0 & 0 & 1 & & \\ 0 & 0 & 0 & a_1^{(2)} & 0 & 0 & 0 & 1 & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \hat{a}_1^{(2)} \\ 0 \\ 0 \\ 0 \\ \cdot \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \cdot \end{bmatrix} = E^3\mathbf{a}^{(3)}$$

per certi $a_i^{(3)} \in \mathbb{C}$ e calcolare tali $a_i^{(3)}$. Il calcolo degli $a_i^{(3)}$ richiede, una volta noto $\hat{\mathbf{a}}^{(2)}$, il prodotto di una matrice t.i.T. 2×2 ($2^{s-2} \times 2^{s-2}$) per un vettore − o, più precisamente, due prodotti t.i.T. 1×1 ($2^{s-3} \times 2^{s-3}$) per vettore. Cioè, nessuna operazione nel nostro caso $n = 8$, in cui nessun $a_i^{(3)}$, $i \geq 1$, è richiesto.

Notiamo che allora per il Lemma 1 si ha $L(E^2\hat{\mathbf{a}}^{(2)})L(E^2\mathbf{a}^{(2)}) = L(E^3\mathbf{a}^{(3)})$, cioè la matrice t.i.T. $L(\mathbf{a})$ è trasformata in una matrice t.i.T. che alterna ciascuna diagonale non nulla con sette nulle.

Si noti anche che per il Lemma 2, se $L(\mathbf{a}^{(2)})\hat{\mathbf{a}}^{(2)} = E\mathbf{a}^{(3)}$ allora $L(E^2\mathbf{a}^{(2)})E^2\hat{\mathbf{a}}^{(2)} = E^3\mathbf{a}^{(3)}$. Vedremo che $\hat{\mathbf{a}}^{(2)}$ tale che $L(\mathbf{a}^{(2)})\hat{\mathbf{a}}^{(2)} = E\mathbf{a}^{(3)}$ è disponibile a costo zero.

Se $n = 2^s > 8$, allora si continua. Altrimenti è completata la prima parte dell'algoritmo.

Riassumendo, abbiamo provato che

$$L(E^2\hat{\mathbf{a}}^{(2)})L(E\hat{\mathbf{a}}^{(1)})L(\hat{\mathbf{a}})L(\mathbf{a}) = L(E^3\mathbf{a}^{(3)}) \tag{10}$$

dove, osserviamo, le sottomatrici 8×8 in alto a sinistra di $L(\mathbf{a})$ e $L(E^3\mathbf{a}^{(3)})$ sono, rispettivamente, la matrice triangolare inferiore di Toeplitz data inizialmente A e la matrice identica I,

$$L(\mathbf{a}) = \begin{bmatrix} 1 & & & & & \\ a_1 & 1 & & & & \\ \cdot & \cdot & \cdot & & & \\ a_7 & \cdot & a_1 & 1 & & \\ a_8 & a_7 & \cdot & a_1 & 1 & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}, \quad L(E^3\mathbf{a}^{(3)}) = \begin{bmatrix} 1 & & & & & \\ 0 & 1 & & & & \\ \cdot & \cdot & \cdot & & & \\ 0 & \cdot & 0 & 1 & & \\ a_1^{(3)} & 0 & \cdot & 0 & 1 & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}.$$

Le operazioni che abbiamo dovuto fare finora sono state: 8×8 t.i.T. $\cdot$ vettore $+ 4 \times 4$ t.i.T. $\cdot$ vettore (se A è $n \times n$ con $n = 2^s$ le operazioni da farsi sarebbero state: $2^s \times 2^s$ t.i.T. $\cdot$ vettore $+ \ldots + 4 \times 4$ t.i.T. $\cdot$ vettore).

Veniamo ora al nostro scopo, calcolare la prima colonna di A^{-1}, e illustriamo quindi la seconda parte dell'algoritmo. Consideriamo il seguente sistema lineare semi-infinito:

$$L(\mathbf{a})\mathbf{z} = E^2\mathbf{v} \tag{11}$$

dove $\mathbf{v}$ è un generico vettore semi-infinito di $\mathbb{C}^{\mathbb{N}}$ (se A è $n \times n$ con $n = 2^s$, allora la matrice E in (11) va elevata a $s-1$ e non a 2). Tale sistema può essere riscritto come segue

$$\begin{bmatrix} A & O \\ \vdots & \ddots \end{bmatrix} \begin{bmatrix} \{\mathbf{z}\}_8 \\ z_8 \\ \cdot \end{bmatrix} = \begin{bmatrix} v_0 \\ 0 \\ 0 \\ 0 \\ v_1 \\ 0 \\ 0 \\ 0 \\ v_2 \\ \cdot \end{bmatrix}$$

cioè evidenziando la parte superiore del sistema, di sole 8 equazioni.

Prima di procedere, si noti che $\{\mathbf{z}\}_8$ è tale che $A\{\mathbf{z}\}_8 = [v_0\ 0\ 0\ 0\ v_1\ 0\ 0\ 0]^T$, $v_0, v_1 \in \mathbb{C}$. Quindi la scelta $v_0 = 1$ e $v_1 = 0$, renderebbe $\{\mathbf{z}\}_8$ uguale al vettore da noi cercato, $A^{-1}\mathbf{e}_1$.

Usando l'uguaglianza (10) si dimostra immediatamente che il sistema $L(\mathbf{a})\mathbf{z} = E^2\mathbf{v}$ è equivalente al seguente sistema:

$$\begin{bmatrix} I_8 & O \\ \vdots & \ddots \end{bmatrix} \begin{bmatrix} \{\mathbf{z}\}_8 \\ \vdots \end{bmatrix} = L(E^3\mathbf{a}^{(3)})\mathbf{z} = L(\hat{\mathbf{a}})L(E\hat{\mathbf{a}}^{(1)})L(E^2\hat{\mathbf{a}}^{(2)})E^2\mathbf{v}.$$

Per il Lemma 2 il secondo membro di quest'ultima uguaglianza può essere riscritto più convenientemente:

$$L(\hat{\mathbf{a}})L(E\hat{\mathbf{a}}^{(1)})L(E^2\hat{\mathbf{a}}^{(2)})E^2\mathbf{v} = L(\hat{\mathbf{a}})L(E\hat{\mathbf{a}}^{(1)})E^2L(\hat{\mathbf{a}}^{(2)})\mathbf{v} = L(\hat{\mathbf{a}})EL(\hat{\mathbf{a}}^{(1)})EL(\hat{\mathbf{a}}^{(2)})\mathbf{v}.$$

Quindi, vale la seguente identità:

$$\begin{bmatrix} I_8 & O \\ \vdots & \ddots \end{bmatrix} \begin{bmatrix} \{\mathbf{z}\}_8 \\ \vdots \end{bmatrix} = L(\hat{\mathbf{a}})EL(\hat{\mathbf{a}}^{(1)})EL(\hat{\mathbf{a}}^{(2)})\mathbf{v}.$$

Le matrici coinvolte nella rappresentazione a secondo membro sono triangolari inferiori e le sottomatrici quadrate di E in alto a sinistra, 8×8, 4×4, hanno le colonne sul lato destro nulle,

$$\{E\}_4 = \left[\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right], \quad \{E\}_8 = \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

Queste due osservazioni ci permettono di ottenere una efficiente rappresentazione di $\{\mathbf{z}\}_8$:

$$\{\mathbf{z}\}_8 = \{L(\hat{\mathbf{a}})\}_8\{E\}_8\{L(\hat{\mathbf{a}}^{(1)})\}_8\{E\}_8\{L(\hat{\mathbf{a}}^{(2)})\}_8\{\mathbf{v}\}_8 = \{L(\hat{\mathbf{a}})\}_8\{E\}_{8,4}\{L(\hat{\mathbf{a}}^{(1)})\}_4\{E\}_{4,2}\{L(\hat{\mathbf{a}}^{(2)})\}_2\{\mathbf{v}\}_2.$$

Usando tale formula, quando $v_0 = 1$, $v_1 = 0$, il vettore $\{\mathbf{z}\}_8$ può essere calcolato effettuando 4×4 t.i.T. $\cdot$ vettore $+$ 8×8 t.i.T. $\cdot$ vettore (se A è $n \times n$ con $n = 2^s$ le operazioni da farsi sarebbero state: 4×4 t.i.T. $\cdot$ vettore $+ \ldots +$ $2^s \times 2^s$ t.i.T.$\cdot$ vettore), ovvero allo stesso costo dell'*eliminazione Gaussiana*, la prima parte dell'algoritmo.

In conclusione, se $cj2^j$ è un limite superiore per il costo dell'operazione $2^j \times 2^j$ t.i.T. $\cdot$ vettore, allora il costo dell'algoritmo illustrato, nel caso di A $n \times n$ con $n = 2^s$, è $\tilde{c}\sum_{j=2}^{s} j2^j = O(s2^s) = O(n \log_2 n)$.

Ci resta da provare che, effettivamente, si ha $\hat{\mathbf{a}}$ tale che $L(\mathbf{a})\hat{\mathbf{a}} = E\mathbf{a}^{(1)}$ a costo zero. A tal fine è sufficiente osservare con un calcolo diretto che

$$\begin{bmatrix} 1 & & & & & & & & & & \\ a_1 & 1 & & & & & & & & & \\ a_2 & a_1 & 1 & & & & & & & & \\ a_3 & a_2 & a_1 & 1 & & & & & & & \\ a_4 & a_3 & a_2 & a_1 & 1 & & & & & & \\ a_5 & a_4 & a_3 & a_2 & a_1 & 1 & & & & & \\ a_6 & a_5 & a_4 & a_3 & a_2 & a_1 & 1 & & & & \\ a_7 & a_6 & a_5 & a_4 & a_3 & a_2 & a_1 & 1 & & & \\ a_8 & a_7 & a_6 & a_5 & a_4 & a_3 & a_2 & a_1 & 1 & & \\ a_9 & a_8 & a_7 & a_6 & a_5 & a_4 & a_3 & a_2 & a_1 & 1 & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} 1 \\ -a_1 \\ a_2 \\ -a_3 \\ a_4 \\ -a_5 \\ a_6 \\ -a_7 \\ a_8 \\ -a_9 \\ \cdot \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 2a_2 - a_1^2 \\ 0 \\ 2a_4 - 2a_1a_3 + a_2^2 \\ 0 \\ 2a_6 - 2a_1a_5 + 2a_2a_4 - a_3^2 \\ 0 \\ 2a_8 - 2a_1a_7 + 2a_2a_6 - 2a_3a_5 + a_4^2 \\ 0 \\ \cdot \end{bmatrix}, \tag{12}$$

$$L(\mathbf{a})\big(\mathbf{e}_1 + \sum_{i=1}^{+\infty}(-1)^i a_i \mathbf{e}_{i+1}\big) = \mathbf{e}_1 + \sum_{i=1}^{+\infty}\delta_{i=0 \bmod 2}\big(2a_i + \sum_{j=1}^{i-1}(-1)^j a_j a_{i-j}\big)\mathbf{e}_{i+1}.$$

Osservazioni finali

Dato il vettore $\mathbf{a}$ il problema del calcolo di un vettore $\hat{\mathbf{a}}$ tale che $L(\mathbf{a})\hat{\mathbf{a}} = E\mathbf{a}^{(1)}$, per qualche $\mathbf{a}^{(1)}$, è in realtà un problema di aritmetica polinomiale. Infatti, per il Lemma 1, l'identità $L(\mathbf{a})\hat{\mathbf{a}} = E\mathbf{a}^{(1)}$ è equivalente all'uguaglianza $L(\mathbf{a})L(\hat{\mathbf{a}}) = L(E\mathbf{a}^{(1)})$, i.e.

$$(\sum_{k=0}^{+\infty} a_k Z^k)(\sum_{k=0}^{+\infty} \hat{a}_k Z^k) = \sum_{k=0}^{+\infty} a_k^{(1)} Z^{2k}.$$

Quindi il problema di aritmetica polinomiale in questione è il seguente:

dato il polinomio $a(z) = \sum_{k=0}^{+\infty} a_k z^k$, trovare un polinomio $\hat{a}(z) = \sum_{k=0}^{+\infty} \hat{a}_k z^k$ tale che

$$\hat{a}(z)a(z) = a_0^{(1)} + a_1^{(1)} z^2 + a_2^{(1)} z^4 + \ldots =: a^{(1)}(z)$$

per certi $a_i^{(1)}$.

Tale problema è un caso particolare di un problema più generale: trasformare un polinomio *pieno* $a(z)$ in un polinomio *sparso* $a^{(1)}(z) = \sum_{k=0}^{+\infty} a_k^{(1)} z^{rk}$. È possibile scrivere esplicitamente un polinomio $\hat{a}(z)$ che realizza questa trasformazione, infatti vale il seguente teorema.

Teorema 2.6 [16]. Dato $a(z) = \sum_{k=0}^{+\infty} a_k z^k$, posto $\hat{a}(z) = a(zt)a(zt^2)\cdots a(zt^{r-1})$ dove t è una radice r-esima principale dell'unità ($t \in \mathbb{C}$, $t^r = 1$, $t^i \neq 1$ per $0 < i < r$), si ha che

$$\hat{a}(z)a(z) = a_0^{(1)} + a_1^{(1)} z^r + a_2^{(1)} z^{2r} + \ldots =: a^{(1)}(z)$$

per certi $a_i^{(1)}$. Inoltre, se i coefficienti di a sono reali allora anche i coefficienti di $\hat{a}$ sono reali.

Il risultato (12) si ottiene immediatamente dal Teorema 2.6, ponendo $r = 2$: $\hat{a}(z) = a(-z)$. È evidente che $a(-z)a(z) = a_0^{(1)} + a_1^{(1)} z^2 + a_2^{(1)} z^4 + \ldots$ [17], e che in tal caso i coefficienti di $\hat{a}$ sono disponibili a costo zero, occorre calcolare solo gli $a_i^{(1)}$.

Esercizio. Scrivere una formula per $\hat{\mathbf{a}}$ tale che $L(\mathbf{a})\hat{\mathbf{a}} = E\mathbf{a}^{(1)}$, dove E è la matrice di 0 e 1 la cui azione su $\mathbf{v} \in \mathbb{C}^{\mathbb{N}}$ ha l'effetto di inserire due zeri tra due componenti successive di $\mathbf{v}$, e valutare il costo del calcolo di $\hat{\mathbf{a}}$.

2.3 Una applicazione: il calcolo dei numeri di Bernoulli

Il j-esimo numero di Bernoulli, $B_{2j}(0)$, è un numero razionale definito per ogni $j \in \mathbb{N}$, positivo se j è dispari e negativo se j è pari, il cui denominatore è noto, nel senso che è il prodotto di tutti i numeri primi p tali che $p - 1$ divide $2j$ [18], e, invece, si hanno solo informazioni parziali sul numeratore [19], [20], [21]. In breve, $B_{2j}(0)$, $j \geq 1$, potrebbe essere definito in termini della funzione Zeta-Riemann, attraverso la ben nota formula di Eulero $B_{2j}(0) = (-1)^{j+1} \frac{2(2j)!}{(2\pi)^{2j}} \sum_{k=1}^{+\infty} \frac{1}{k^{2j}}$ [22], [23]. Quest'ultima formula da sola forse è sufficiente per giustificare l'interesse passato e presente nello studio dei numeri di Bernoulli. Si noti che come immediata conseguenza della formula di Eulero si ha che i numeri di Bernoulli $B_{2j}(0)$ vanno a infinito per $j \to +\infty$.

In questa sezione si osserva che i numeri di Bernoulli, a meno di fattori noti $x_j \in \mathbb{R}$, risolvono un sistema triangolare inferiore di Toeplitz. Ne segue che il calcolo dei loro numeratori può essere effettuato (1) utilizzando l'algoritmo descritto nella sezione precedente, per ottenere buone approssimazioni z_j^* dei numeri reali $z_j = x_{j-1} B_{2j-2}(0)$, $j = 1, \ldots, 2^s$, e (2) estraendoli dalle approssimazioni z_j^*.

I polinomi e numeri di Bernoulli

Le condizioni

$$B(x+1) - B(x) = nx^{n-1}, \quad \int_0^1 B(x)\,dx = 0, \quad B(x) \text{ polinomio}$$

definiscono univocamente la funzione $B(x)$. Essa è un particolare polinomio monico di grado n chiamato *n-esimo polinomio di Bernoulli* e indicato con il simbolo $B_n(x)$. È semplice calcolare i primi polinomi di Bernoulli:

$$B_1(x) = x - \frac{1}{2},\ B_2(x) = x^2 - x + \frac{1}{6},\ B_3(x) = x(x - \frac{1}{2})(x-1),\ \ldots\ .$$

Per convenzione $B_0(x) = 1$.

Si dimostra che i polinomi di Bernoulli definiscono i coefficienti dello sviluppo in serie di potenze di diverse funzioni; ad esempio, per ciò che segue, è opportuno ricordare che vale il seguente sviluppo:

$$\frac{te^{xt}}{e^t-1}=\sum_{n=0}^{+\infty}\frac{B_n(x)}{n!}t^n. \tag{13}$$

Inoltre, i polinomi di Bernoulli soddisfano diverse identità. Due delle più importanti sono quelle concernenti il valore della loro derivata e le loro proprietà di simmetria/antisimmetria rispetto all'asse $x=\frac{1}{2}$:

$$B_n'(x)=nB_{n-1}(x),\quad B_n(1-x)=(-1)^nB_n(x).$$

In particolare, come conseguenza di quest'ultima identità e della loro definizione, si vede facilmente che tutti i polinomi di Bernoulli di grado dispari eccetto il primo si annullano in zero. Al contrario, il valore in zero dei polinomi di Bernoulli di grado pari è, oltre che diverso da zero, particolarmente significativo. In particolare, vale la seguente formula di Eulero

$$\zeta(2j)=\frac{|B_{2j}(0)|(2\pi)^{2j}}{2(2j)!},\quad \zeta(s)=\sum_{k=1}^{+\infty}\frac{1}{k^s},$$

che mette in stretta relazione i valori $B_{2j}(0)$ con i valori della funzione Zeta di Riemann ζ nei numeri interi positivi pari $2j$. Ad esempio, da questa relazione e dal fatto che $\zeta(2j)\to 1$ se $j\to+\infty$, si deduce che $|B_{2j}(0)|$ ha lo stesso andamento di $2(2j)!/(2\pi)^{2j}$ per grandi valori di j. Un'altra formula importante coinvolgente i valori $B_{2j}(0)$ è quella di Eulero-Maclaurin, utile per il calcolo di somme: se f è una funzione sufficientemente regolare in $[m,n]$, $m,n\in\mathbb{Z}$, allora

$$\sum_{r=m}^{n}f(r)=\frac{1}{2}[f(m)+f(n)]+\int_m^n f(x)\,dx+\sum_{j=1}^{k}\frac{B_{2j}(0)}{(2j)!}[f^{(2j-1)}(n)-f^{(2j-1)}(m)]+u_{k+1}, \tag{14}$$

dove

$$\begin{aligned}u_{k+1}&=\frac{1}{(2k+1)!}\int_m^n f^{(2k+1)}(x)\overline{B}_{2k+1}(x)\,dx\\&=-\frac{1}{(2k)!}\int_m^n f^{(2k)}(x)\overline{B}_{2k}(x)\,dx\\&=\frac{1}{(2k+2)!}\int_m^n f^{(2k+2)}(x)[B_{2k+2}(0)-\overline{B}_{2k+2}(x)]\,dx\end{aligned}$$

e $\overline{B}_n$ è l'estensione periodica su $\mathbb{R}$ di $B_n|_{[0,1)}$. Ricordiamo che la formula di Eulero-Maclaurin conduce anche a una rappresentazione importante dell'errore commesso dalla formula dei trapezi $\mathcal{I}_h=h[\frac{1}{2}g(a)+\sum_{r=1}^{n-1}g(a+rh)+\frac{1}{2}g(b)]$, $h=\frac{b-a}{n}$, nell'approssimazione dell'integrale definito $\mathcal{I}=\int_a^b g(x)\,dx$. Tale rappresentazione, valida per g sufficientemente regolare in $[a,b]$, si ottiene ponendo $m=0$ e $f(t)=g(a+th)$ in (14):

$$\mathcal{I}_h=\mathcal{I}+\sum_{j=1}^{k}\frac{h^{2j}B_{2j}(0)}{(2j)!}[g^{(2j-1)}(b)-g^{(2j-1)}(a)]+r_{k+1},\ \ r_{k+1}=\frac{g^{(2k+2)}(\xi)h^{2k+2}(b-a)B_{2k+2}(0)}{(2k+2)!}, \tag{15}$$

$\xi \in (a,b)$. Tale rappresentazione dell'errore, in termini di potenze pari di h, giustifica l'efficienza del metodo di estrapolazione di Romberg per la stima di integrali, quando tale metodo è applicato in combinazione con la formula dei trapezi. È evidente infatti da (15) che $\tilde{\mathcal{I}}_{h/2} := (2^2\mathcal{I}_{h/2} - \mathcal{I}_h)/(2^2-1)$ approssima $\mathcal{I}$ con un errore dell'ordine di $O(h^4)$ mentre l'errore di $\mathcal{I}_h$ e $\mathcal{I}_{h/2}$, nell'approssimazione di $\mathcal{I}$, è dell'ordine di $O(h^2)$.

Esercizio. Mostrare che la conoscenza di una formula esplicita per la somma $\sum_{x=1}^{n} x^j$ equivale alla conoscenza di $B_{j+1}(n+1)$ e del numero di Bernoulli $B_{j+1}(0)$.

Per questi e tanti altri motivi (si veda ad esempio [25]) i valori $B_{2j}(0)$ sono ben noti in letteratura, col nome di *numeri di Bernoulli*.

I numeri di Bernoulli risolvono sistemi triangolari di Toeplitz

Dall'identità (13) della precedente sezione segue che i numeri di Bernoulli soddisfano la seguente uguaglianza:

$$\frac{t}{e^t-1} = -\frac{1}{2}t + \sum_{k=0}^{+\infty} \frac{B_{2k}(0)}{(2k)!} t^{2k}.$$

Moltiplicando quest'ultima per e^t-1, sviluppando e^t in potenze di t, ed imponendo che i coefficienti di t^i, $i = 2,3,4,\ldots$, del secondo membro siano uguali a zero, si ottengono le equazioni:

$$-\frac{1}{2}j + \sum_{k=0}^{[\frac{j-1}{2}]} \binom{j}{2k} B_{2k}(0) = 0, \quad j = 2,3,4,\ldots \ . \tag{16}$$

Ora, scegliendo j pari o j dispari, si ottengono due sistemi lineari triangolari inferiori che definiscono univocamente i numeri di Bernoulli. Esaminiamo qui nei dettagli solo la scelta j pari. La scelta j dispari, comunque, porta ad analoghi risultati e osservazioni (si veda [8]).

Dunque, mettendo insieme le equazioni (16) per j pari, si ottiene il seguente sistema lineare triangolare inferiore:

$$\begin{bmatrix} \binom{2}{0} & & & & \\ \binom{4}{0} & \binom{4}{2} & & & \\ \binom{6}{0} & \binom{6}{2} & \binom{6}{4} & & \\ \binom{8}{0} & \binom{8}{2} & \binom{8}{4} & \binom{8}{6} & \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} B_0(0) \\ B_2(0) \\ B_4(0) \\ B_6(0) \\ \cdot \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ \cdot \end{bmatrix}.$$

Da questo possiamo ricavare i primi numeri di Bernoulli:

$$1, \ , \frac{1}{6}, \ -\frac{1}{30}, \ \frac{1}{42}, \ -\frac{1}{30}, \ \frac{5}{66}, \ -\frac{691}{2730}, \ \frac{7}{6}, \ -\frac{3617}{510}. \tag{17}$$

Vogliamo dare una forma analitica alla matrice dei coefficienti W di tale sistema lineare. Per farlo è sufficiente osservare che tale matrice è una particolare sottomatrice della matrice di Tartaglia, che

può essere rappresentata come una serie di potenze. Più precisamente, poniamo

$$Y = \begin{bmatrix} 0 & & & & \\ 1 & 0 & & & \\ & 2 & 0 & & \\ & & 3 & 0 & \\ & & & \cdot & \cdot \end{bmatrix}, \quad \phi = \begin{bmatrix} 0 & & & & & \\ 2 & 0 & & & & \\ & 12 & 0 & & & \\ & & 30 & 0 & & \\ & & & 56 & 0 & \\ & & & & \cdot & \cdot \end{bmatrix}, \quad 2 = 1*2,\ 12 = 3*4,\ 30 = 5*6,\ \ldots,$$

e notiamo che dall'uguaglianza

$$X := \begin{bmatrix} \binom{0}{0} & & & & & & & \\ \binom{1}{0} & \binom{1}{1} & & & & & & \\ \binom{2}{0} & \binom{2}{1} & \binom{2}{2} & & & & & \\ \binom{3}{0} & \binom{3}{1} & \binom{3}{2} & \binom{3}{3} & & & & \\ \binom{4}{0} & \binom{4}{1} & \binom{4}{2} & \binom{4}{3} & \binom{4}{4} & & & \\ \binom{5}{0} & \binom{5}{1} & \binom{5}{2} & \binom{5}{3} & \binom{5}{4} & \binom{5}{5} & & \\ \binom{6}{0} & \binom{6}{1} & \binom{6}{2} & \binom{6}{3} & \binom{6}{4} & \binom{6}{5} & \binom{6}{6} & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} = \begin{bmatrix} 1 & & & & & & \\ 1 & 1 & & & & & \\ 1 & 2 & 1 & & & & \\ 1 & 3 & 3 & 1 & & & \\ 1 & 4 & 6 & 4 & 1 & & \\ 1 & 5 & 10 & 10 & 5 & 1 & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} = \sum_{k=0}^{+\infty} \frac{1}{k!} Y^k,$$

che vale perché $[X]_{ij} = \frac{1}{(i-j)!}[Y^{i-j}]_{ij} = \frac{1}{(i-j)!} j \cdots (i-2)(i-1) = \binom{i-1}{j-1}$, $1 \le j \le i \le n$, segue che

$$W = Z^T \phi \sum_{k=0}^{+\infty} \frac{1}{(2k+2)!} \phi^k.$$

Possiamo dunque riscrivere il sistema lineare risolto dai numeri di Bernoulli come segue:

$$\sum_{k=0}^{+\infty} \frac{2}{(2k+2)!} \phi^k \mathbf{b} = \mathbf{q}, \quad \mathbf{b} = \begin{bmatrix} B_0(0) \\ B_2(0) \\ B_4(0) \\ B_6(0) \\ \cdot \end{bmatrix}, \ \mathbf{q} = \begin{bmatrix} 1 \\ 1/3 \\ 1/5 \\ 1/7 \\ \cdot \end{bmatrix}. \tag{18}$$

Ora mostriamo che tale sistema è equivalente a un sistema lineare triangolare inferiore di Toeplitz. Il nostro scopo è sostituire ϕ, una matrice sulla cui sottodiagonale ci sono elementi tutti diversi tra loro, con una matrice sulla cui sottodiagonale ci sono elementi tutti uguali tra loro.

Sia $D = \operatorname{diag}(d_1, d_2, d_3, \ldots)$, $d_i \neq 0$. Se si scrive la matrice $D\phi D^{-1}$ ci si accorge che si può imporre che essa sia uguale a una matrice del tipo xZ, è infatti sufficiente porre $d_k = x^{k-1} d_1/(2k -$

2)!, $k = 1, 2, 3, \ldots$. Sia dunque

$$D = \begin{bmatrix} 1 & & & & & \\ & \frac{x}{2!} & & & & \\ & & \frac{x^2}{4!} & & & \\ & & & . & & \\ & & & & \frac{x^{n-1}}{(2n-2)!} & \\ & & & & & . \end{bmatrix}.$$

(si è scelto $d_1 = 1$). Allora (18) è equivalente ai seguenti sistemi lineari $\sum_{k=0}^{+\infty} \frac{2}{(2k+2)!} D\phi^k D^{-1} D\mathbf{b} = D\mathbf{q}$, $\sum_{k=0}^{+\infty} \frac{2}{(2k+2)!} (D\phi D^{-1})^k D\mathbf{b} = D\mathbf{q}$, $\sum_{k=0}^{+\infty} \frac{2x^k}{(2k+2)!} Z^k D\mathbf{b} = D\mathbf{q}$.

Riassumendo, sia $\mathbf{z}$ il vettore $D\mathbf{b}$. Allora il vettore $\{\mathbf{b}\}_n$ (ovvero, il vettore con i primi n numeri di Bernoulli) può essere ottenuto:

1 calcolando le prime n componenti della soluzione del seguente sistema lineare triangolare inferiore di Toeplitz:

$$\left(\sum_{k=0}^{+\infty} \frac{2x^k}{(2k+2)!} Z^k\right) \mathbf{z} = D\mathbf{q} \tag{19}$$

(ovvero, $\{\mathbf{z}\}_n$ tale che $\{\sum_{k=0}^{+\infty} \frac{2x^k}{(2k+2)!} Z^k\}_n \{\mathbf{z}\}_n = \{D\mathbf{q}\}_n$);

2 risolvendo il sistema lineare $\{D\}_n \{\mathbf{b}\}_n = \{\mathbf{z}\}_n$ nel campo razionale.

Osserviamo che il calcolo in 1 può essere effettuato con l'algoritmo descritto nella sezione precedente ad un costo $O(n \log_2 n)$ e che tale algoritmo può essere reso stabile numericamente mediante una scelta opportuna del parametro x. Ad esempio, osservando che la scelta $x = (2\pi)^2$ renderebbe la successione $z_n = \frac{x^{n-1}}{(2n-2)!} B_{2n-2}(0)$, $n \in \mathbb{N}$, limitata; infatti in tal caso $|z_n| \to 2$ se $n \to +\infty$, per la formula di Eulero. Effettuato il calcolo in 1 si ottengono n numeri macchina che costituiscono una ottima approssimazione in $\mathbb{R}$ delle quantità $x^s B_{2s}(0)/(2s)!$, $s = 0, 1, \ldots, n-1$. Poi, nella fase 2, da questi numeri macchina occorre ricavare i numeri *razionali* di Bernoulli $B_{2s}(0)$, $s = 0, 1, \ldots, n-1$.

Concludiamo osservando che i numeri di Bernoulli, di nuovo a meno di fattori noti, risolvono un sistema triangolare inferiore di Toeplitz dove 2/3 delle diagonali della matrice dei coefficienti sono nulle [24], [8]. Questo ovviamente comporta una ulteriore riduzione del costo computazionale, soprattutto se si definisce un algoritmo ad hoc per tali sistemi sparsi di Toeplitz. Per maggiori dettagli si veda [8].

2.4 Algebre di bassa complessità computazionale

Finora abbiamo studiato più o meno approfonditamente, le algebre di matrici triangolari di Toeplitz, le circolanti e ϕ-circolanti, le algebre di tipo Hartley e l'algebra τ, come esempio delle algebre trigonometriche (o di tipo Jacobi). In particolare si è visto che, per ognuna di queste algebre $\mathcal{L}$, se A è una generica matrice di $\mathcal{L}$ ed $\mathbf{f}$ è un generico vettore di $\mathbb{C}^n$, allora valgono le seguenti affermazioni:

1 lo spettro di A è calcolabile dagli elementi di A con al più $O(n \log n)$ operazioni aritmetiche

2 il prodotto matrice-vettore $A\mathbf{f}$ è calcolabile con al più $O(n \log n)$ operazioni aritmetiche

3 se $\det A \neq 0$, il vettore $\mathbf{x}$ tale che $A\mathbf{x} = \mathbf{f}$ è calcolabile con al più $O(n \log n)$ operazioni aritmetiche

In altre parole, i principali problemi dell'algebra lineare numerica associati ad una matrice $n \times n$ A, ovvero la risoluzione di un sistema lineare con A come matrice dei coefficienti, il prodotto di A per un vettore, e il calcolo degli autovalori di A, sono facilmente risolvibili quando A è una matrice di una delle algebre $\mathcal{L}$ sopra elencate. Per questo si può dire che tali algebre sono *di bassa complessità computazionale.*

DEFINIZIONE 2.7 *Un sottoinsieme $\mathcal{L}$ di matrici di $\mathbb{C}^{n\times n}$, ben definito per ogni n (per infiniti valori di n), si dice di bassa complessità computazionale se per le matrici di $\mathcal{L}$ valgono le proprietà 1, 2 e 3 di cui sopra.*

Ovviamente, oltre le algebre di Jacobi, Hartley e ϕ-circolanti, ci sono molti altri esempi di algebre $\mathcal{L}$ di bassa complessità computazionale, in particolare ci sono algebre significative di complessità minima, $O(n)$, ad esempio quelle di Householder e di Haar (ma Haar non e' $O(n \log n)$!?).

Esercizio. Sia $\mathbf{u} \in \mathbb{C}^n$, $\mathbf{u} \neq \mathbf{0}$, e $M = I - \frac{2}{\|\mathbf{u}\|^2}\mathbf{u}\mathbf{u}^H$. La matrice M, nota come matrice di Householder, è unitaria ed hermitiana, e i suoi autovalori sono noti esplicitamente (calcolarli!). Sia $\mathcal{L} = \operatorname{sd} M$. Valutare il costo dei tre problemi basilari di algebra lineare numerica associati ad $A \in \mathcal{L}$. Trovare delle condizioni su $\mathbf{u}$ che rendano l'algebra $\operatorname{sd} M$ rappresentabile nella forma $\operatorname{sd} M = \{Md(M^T\mathbf{z})d(M^T\mathbf{e}_1)^{-1}M^{-1} : \mathbf{z} \in \mathbb{C}^n\}$. Trovare delle condizioni su $\mathbf{u}$ per cui la matrice $\mathbf{e}\mathbf{e}^T = (1)_{i,j=1}^n$ appartiene a $\operatorname{sd} M$. 8È possibile scrivere in forma esplicita gli autovalori di $A \in \operatorname{sd} M$?

Esercizio. Poniamo

$$Q_1 = 1,\ D_1 = 1,$$

$$Q_{2m} = \begin{bmatrix} \mathbf{e}\mathbf{e}_1^T & Q_m \\ Q_m & -\mathbf{e}\mathbf{e}_1^T \end{bmatrix},\ S_m = \operatorname{diag}(\frac{1}{\sqrt{2}}, 1, \ldots, 1),\ D_{2m} = \begin{bmatrix} D_m S_m & 0 \\ 0 & D_m S_m \end{bmatrix},$$

$$U_{2m} = Q_{2m}D_{2m} = \begin{bmatrix} \mathbf{e}\mathbf{e}_1^T & Q_m \\ Q_m & -\mathbf{e}\mathbf{e}_1^T \end{bmatrix} \begin{bmatrix} D_m S_m & \\ & D_m S_m \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2m}}\mathbf{e}\mathbf{e}_1^T & U_m S_m \\ U_m S_m & -\frac{1}{\sqrt{2m}}\mathbf{e}\mathbf{e}_1^T \end{bmatrix},\ m = 1, 2, 4, \ldots, 2^{s-1}.$$

Per esempio

$$U_2 = Q_2 D_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \\ & \frac{1}{\sqrt{2}} \end{bmatrix};$$

$$U_8 = Q_8 D_8 = \left[\begin{array}{cccc|cccc} 1 & & & & 1 & & 1 & 1 \\ 1 & & & & 1 & & 1 & -1 \\ 1 & & & & 1 & 1 & -1 & \\ 1 & & & & 1 & -1 & -1 & \\ \hline 1 & & 1 & 1 & -1 & & & \\ 1 & & 1 & -1 & -1 & & & \\ 1 & 1 & -1 & & -1 & & & \\ 1 & -1 & -1 & & -1 & & & \end{array}\right] \left[\begin{array}{cccc|cccc} \frac{1}{\sqrt{8}} & & & & & & & \\ & \frac{1}{\sqrt{2}} & & & & & & \\ & & \frac{1}{\sqrt{4}} & & & & & \\ & & & \frac{1}{\sqrt{2}} & & & & \\ \hline & & & & \frac{1}{\sqrt{8}} & & & \\ & & & & & \frac{1}{\sqrt{2}} & & \\ & & & & & & \frac{1}{\sqrt{4}} & \\ & & & & & & & \frac{1}{\sqrt{2}} \end{array}\right].$$

Per costruzione le matrici $U := U_n$, $n = 2^s$, e U^T sono reali unitarie e definiscono trasformate discrete veloci, note come trasformate di Haar [26]. È vero che ogni matrice dell'algebra $\mathcal{L} = \mathrm{sd}\, U$ è univocamente determinata dalla sua prima riga? E per $\mathcal{L}' = \mathrm{sd}\, U^T$? Trovare matrici J_k, $k = 1, \ldots, n$, i cui elementi siano -1, 0, oppure 1 per cui $\mathcal{L} = \mathrm{Span}\{J_k\}$ e $(J_k, J_i)_F = 0$ se $k \neq i$, essendo $(X, Y)_F$ il prodotto scalare $\sum_{i,j=1}^{n} \overline{x}_{ij} y_{ij}$ in $\mathbb{C}^{n \times n}$.

A questo punto ci si può chiedere: come utilizzare la bassa complessità computazionale di certe algebre di matrici $\mathcal{L}$? Come vedremo meglio in seguito, ci sono due modi per farlo.

Il primo nasce dall'osservazione, maturata attraverso diversi risultati ottenuti in letteratura, che i problemi di algebra lineare numerica basilari su menzionati spesso possono essere risolti efficientemente anche quando la matrice $A \in \mathbb{C}^{n \times n}$ è solo *vicina strutturalmente* ad un'algebra di bassa complessità computazionale. Ad esempio, si è visto questo nella teoria delle formule di dislocamento per la rappresentazione di matrici e nel precondizionamento di sistemi lineari. Se infatti si sapesse che per una matrice A $n \times n$, definita per ogni n, vale una uguaglianza del tipo

$$A = \sum_{i=1}^{\alpha} M_i N_i, \quad M_i \in \mathcal{L},\ N_i \in \mathcal{L}',$$

con $\mathcal{L}$ ed $\mathcal{L}'$ entrambe algebre di bassa complessità computazionale, e α indipendente da n, allora si avrebbe immediatamente un algoritmo di costo $O(n \log n)$ per il calcolo di $A\mathbf{f}$. Vedi .... Oppure, se per una matrice A $n \times n$ definita positiva, definita per ogni n, si potesse trovare un'algebra di bassa complessità computazionale $\mathcal{L}$ ed una matrice $P \in \mathcal{L}$ definita positiva tali che gli autovalori di $P^{-1}A$ si *raggruppano su* 1 quando $n \to +\infty$, allora il metodo del Gradiente Coniugato applicato al sistema $(E^{-1}AE^{-H})(E^H\mathbf{x}) = E^{-1}\mathbf{b}$, $EE^H = P$, fornirebbe in pochi passi una buona approssimazione di $\mathbf{x} = A^{-1}\mathbf{f}$, $\forall \mathbf{f} \in \mathbb{C}^n$, e l'operazione in più per passo, data dalla risoluzione di un sistema del tipo $P\mathbf{z} = \mathbf{h}_k$, non rallenterebbe il metodo perché sarebbe di costo al più $O(n \log n)$, e quindi in generale inferiore o al più uguale al costo del prodotto matrice-vettore $A\mathbf{u}_k$. Vedi .... Inoltre, sono in corso studi sulla possibilità di ottenere informazioni sullo spettro di matrici non negative A utilizzando lo spettro, facilmente calcolabile, di matrici di algebre $\mathcal{L}$ vicine ad A [28], [29]. Si pensi in particolare a una matrice A tipo quella di Google, il cui autovettore dominante ha come componenti le importanze delle pagine del web [27]. Google calcola tale autovettore utilizzando il metodo delle potenze. Ora, se fosse disponibile a poco costo una stima del secondo autovalore della matrice A, prima di "partire" con il metodo delle potenze, si potrebbe pensare di apportare opportune modifiche ad A, che rendano la convergenza del metodo più rapida senza distorcere troppo dalla realtà il (grafo) modello del web. Vedi ....

Per quanto riguarda il secondo modo di utilizzo delle algebre di matrici di bassa complessità $\mathcal{L}$, osserviamo che esse, grazie alle loro proprietà 1,2,3, possono risultare estremamente utili nell'ideare algoritmi efficienti per la risoluzione di problemi non lineari di grandi dimensioni, come il calcolo del minimo delle funzioni errore associate a) a reti neurali per l'apprendimento [30], [31], [32], [9] oppure b) alla ricostruzione di immagini affette da rumore [33]. Per maggiori dettagli si rimanda a ....

2.5 Algebre ben condizionate spettralmente

Matrici con autovalori ottimamente condizionati

Sia $A \in \mathbb{C}^n$. Un numero complesso λ si dice autovalore di A se esiste $\mathbf{v} \in \mathbb{C}^n$ non nullo tale che $A\mathbf{v} = \lambda\mathbf{v}$, ovvero se $\det(\lambda I - A) = 0$. Un tale vettore $\mathbf{v}$ è detto autovettore di A corrispondente all'autovalore λ.

L'insieme degli autovalori di A, $\sigma(A)$, è quindi dato dall'insieme delle radici dell'equazione algebrica $p_A(t) = \det(tI - A) = 0$. Essendo p_A un polinomio monico di grado esattamente n a coefficienti in $\mathbb{C}$, l'equazione, $p_A(t) = 0$ ha in $\mathbb{C}$ esattamente n radici. In altre parole, gli autovalori di una matrice A $n\times n$ sono gli n zeri del polinomio caratteristico p_A di A. Dall'identità $p_{S^{-1}AS} = p_A$, vera $\forall\, S$ invertibile, segue che gli autovalori di A sono uguali agli autovalori di ogni matrice *simile ad* A, ovvero del tipo $S^{-1}AS$ con S generica non singolare. Se A ha elementi reali, allora p_A ha coefficienti reali e, quindi, $p_A(\lambda) = 0$ implica $p_A(\overline{\lambda}) = 0$; cioè $\sigma(A)$ è chiuso rispetto alla trasposizione coniugata.

Dato λ un autovalore di A, l'insieme degli autovettori di A corrispondenti a λ forma un sottospazio di $\mathbb{C}^n$ di dimensione $m_g(\lambda) := n - \mathrm{rank}(\lambda I - A)$. Se $\mathbf{v}_1$, ..., $\mathbf{v}_{m_g(\lambda)}$ è una base per tale sottospazio e V è una matrice di $\mathbb{C}^{n\times n}$ invertibile tale che $V\mathbf{e}_j = \mathbf{v}_j$, $j = 1, \ldots, m_g(\lambda)$, allora

$$V^{-1}AV = \begin{bmatrix} \lambda I_{m_g(\lambda)} & \\ O & B \end{bmatrix}.$$

Ne segue che λ è zero di p_A con molteplicità algebrica, $m_a(\lambda)$, maggiore o uguale di $m_g(\lambda)$. Inoltre, se $\hat{\lambda}$ è un autovalore di A diverso da λ, ogni autovettore di A corrispondente a $\hat{\lambda}$ è linearmente indipendente da ogni autovettore di A corrispondente a λ. Queste osservazioni ci permettono di concludere che A è simile a una matrice diagonale, ovvero è diagonalizzabile, se e solo se $m_a(\lambda) = m_g(\lambda)$ per ogni autovalore λ di A; in questo caso, le colonne della matrice V che diagonalizza A sono autovettori di A linearmente indipendenti.

Tra le matrici A diagonalizzabili, quelle i cui autovalori sono meno sensibili a variazioni dei loro elementi sono del tipo $A = U\,\mathrm{diag}\,(z_i,\, i = 1, \ldots, n)U^H$, $z_i \in \mathbb{C}$, U unitaria, ovvero sono le matrici diagonalizzabili da trasformazioni per similitudine unitarie. Questo risultato segue dal Teorema di Bauer-Fike.

Teorema 2.8 Sia $A \in \mathbb{C}^{n\times n}$ diagonalizzabile, e sia V tale che $V^{-1}AV = \mathrm{diag}\,(\lambda_i,\, i = 1, \ldots, n)$. Sia λ un autovalore di $B \in \mathbb{C}^{n\times n}$ che non sia autovalore di A. Allora

$$\min_i |\lambda - \lambda_i| \le \mu(V)\|B - A\|, \quad \mu(V) = \|V\|\|V^{-1}\| \ge 1,$$

dove $\|\cdot\|$ è una norma matriciale compatibile con una norma vettoriale, cioè $\|M\mathbf{v}\| \le \|M\|\|\mathbf{v}\|\ \forall\, M \in \mathbb{C}^{n\times n}$ e $\forall\, \mathbf{v} \in \mathbb{C}^n$, e tale che $\|M\| = \rho(M)$ se M è diagonale. Ad esempio $\|\cdot\| = \|\cdot\|_2$, $\|\cdot\|_1$, $\|\cdot\|_\infty$.

Il Teorema ci dice in particolare che per ogni autovalore $\tilde{\lambda}$ di una perturbazione $A+\delta A$ di A esiste un autovalore λ di A che dista da $\tilde{\lambda}$ non più di $\mu(V)\|\delta A\|$, dove V è una matrice che diagonalizza A:

$$\tilde{\lambda} \in \sigma(A + \delta A) \ \Rightarrow\ \exists\, \lambda \in \sigma(A) \mid\ |\tilde{\lambda} - \lambda| \le \inf_{V:\, V^{-1}AV = \mathrm{diag}} \mu(V)\|\delta A\|. \tag{20}$$

Quindi, il problema degli autovalori di una matrice A diagonalizzabile è ben condizionato se tra le matrici V che diagonalizzano A ve n'è una per cui $\mu(V)$ non è troppo maggiore di 1; se poi, tra

queste, ve n'è una tale che $\mu(V) = 1$ per qualche norma, allora il problema si dice ottimamente condizionato.

Notiamo, prima di proseguire, che se A non è diagonalizzabile, allora il problema degli autovalori di A non si potrà mai definire ottimamente condizionato perchè A ha almeno un autovalore λ con molteplicità algebrica maggiore della geometrica e la misura della variazione $\delta\lambda$ che subisce tale autovalore quando si perturba A di una quantità δA può assumere anche valori uguali a o maggiori di $\sqrt{\|\delta A\|}$ (tali λ si dicono per questo motivo patologici). Ad esempio, se si perturba di ε un elemento di una matrice triangolare superiore A $n \times n$, se l'elemento ha posto (i, j) con $i \leq j$ allora è ovvio che gli autovalori di A non possono subire variazioni superiori a $|\varepsilon|$. Ma se l'elemento ha posto $(n, 1)$, allora gli autovalori possono variare di molto, anche di $|\varepsilon|^{1/n}$, come accade se si sceglie $A = Z^T$ dove Z è la matrice lower-shift in (2).

È un semplice esercizio dimostrare che, se $M \in \mathbb{C}^{n\times n}$, allora $\mu_2(M) = \|M\|_2\|M^{-1}\|_2 = 1$ se e solo se $M = \alpha U$, $\alpha \in \mathbb{C}$ $\alpha \neq 0$, U unitaria. Ne segue che $\inf_{V: V^{-1}AV = \text{diag}} \mu(V)$ in (20) è uguale a 1 per $\|\cdot\| = \|\cdot\|_2$ se e solo se tra le matrici V che diagonalizzano A ve n'è almeno una unitaria. Diamo una caratterizzazione delle matrici A diagonalizzabili con una trasformazione per similitudine unitaria, ovvero delle matrici A per cui il problema degli autovalori di A è ottimamente condizionato.

Teorema 2.9 Esiste una matrice unitaria U tale che $U^{-1}AU = D$ con D diagonale se e solo se A è normale, ovvero $AA^H = A^HA$.

Dimostrazione. Si usa il Teorema di Schur, che per ogni $A \in \mathbb{C}^{n\times n}$ stabilisce l'esistenza di una matrice unitaria U tale che $U^{-1}AU = T$ con T triangolare superiore [4], insieme all'osservazione che matrici contemporaneamente triangolari e normali devono essere necessariamente diagonali. □

Esercizio. Dimostrare che le seguenti matrici sono rispettivamente non diagonalizzabile, diagonalizzabile ma non con trasformazioni unitarie, diagonalizzabile con trasformazione unitaria:

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 1 \\ 2 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 1 \\ \phi & 0 \end{bmatrix}, \ |\phi| = 1.$$

Risoluzione. Nel primo caso $m_g(0) = 1 < m_a(0) = 2$. Nel secondo, la matrice ha due autovalori distinti, quindi due autovettori linearmente indipendenti. Nel terzo caso la matrice è normale.

Osserviamo che se una matrice A è diagonalizzabile da una matrice V, cioè $V^{-1}AV = \text{diag}$ (come avviene nel secondo e terzo esempio dell'esercizio), allora la stessa matrice V diagonalizza l'algebra $\mathcal{L} = \{p(A)\}$ dei polinomi in A, e si dice che le matrici di $\mathcal{L}$ sono simultaneamente diagonalizzate da V. Nel caso in cui V si può scegliere unitaria $V = U$, cioè quando A è normale, le matrici di $\mathcal{L}$ sono tutte normali, perchè sono tutte diagonalizzate da U (ad es. $A = \chi$, $\mathcal{L} = \mathcal{C}^S$, vedi Capitolo 2). Se poi A, oltre ad essere normale, è anche non derogatoria, allora A ha autovalori distinti, e l'algebra $\mathcal{L}$ ha dimensione n ed ammette le tre seguenti diverse rappresentazioni: $\mathcal{L} = \{p(A) : p \,\text{polinomi}\} = \{Ud(\mathbf{z})U^H : \mathbf{z} \in \mathbb{C}^n\} = \{M \in \mathbb{C} : MA = AM\}$, dove con $d(\mathbf{z})$ intendiamo la matrice diagonale con elementi diagonali le componenti z_i del vettore $\mathbf{z}$ (ad es. $A = \Pi$, $\mathcal{L} = \mathcal{C}$ oppure $A = Z + Z^T$, $\mathcal{L} = \tau$, vedi Capitolo 2).

È evidente che per ogni matrice M che sia un polinomio in A normale non derogatoria, il problema degli autovalori di M è ottimamente condizionato. Ma è vero anche il contrario: se il

problema degli autovalori di $M \in \mathbb{C}^{n\times n}$ è ottimamente condizionato, allora per il Teorema 2.9 M deve essere uguale a UDU^H per qualche U unitaria e D diagonale, e quindi $M = p(A)$ per qualche A normale non derogatoria e p polinomio ($A = U\tilde{D}U^H$ con $\tilde{D}_{ii}$ distinti e p tale che $p(\tilde{D}) = D$). Dunque si ha la seguente

Proposizione 2.10 Il problema degli autovalori di una matrice $M \in \mathbb{C}^{n\times n}$ è ottimamente condizionato se e solo se $M \in \mathcal{L}$, dove $\mathcal{L} = \{Ud(\mathbf{z})U^H : \mathbf{z} \in \mathbb{C}^n\}$, U unitaria. Ovviamente, anche il problema degli autovalori di una matrice $M \in \mathbb{C}^{n\times n}$ triangolare superiore (inferiore) è ottimamente condizionato, a condizione che la perturbazione δM di M interessi solo la parte superiore (inferiore) di M.

Esercizio [9]. Sia $\mathcal{L} = \mathrm{sd}\, U = \{Ud(\mathbf{z})U^H : \mathbf{z} \in \mathbb{C}^n\}$, U unitaria. Dimostrare che se $\mathbf{s}$ ed $\mathbf{y}$ sono due vettori di $\mathbb{R}^n$ tali che $\mathbf{s}^T\mathbf{y} > 0$, allora esiste ed è unica $A \in \mathrm{sd}\, U$ definita positiva tale che $A\mathbf{s} = \mathbf{y}$.

Abbiamo in pratica dimostrato che le algebre ottimali dal punto di vista del condizionamento degli autovalori delle matrici che le compongono, sono le algebre $\mathrm{sd}\, U$ con U unitaria. Ne segue che ogni algebra $\mathrm{sd}\, U$ è un sottospazio di $\mathbb{C}^{n\times n}$ indicato dove andare a scegliere, quando opportuno, "sostituti semplici" delle matrici A più generali che capitano nelle applicazioni; infatti tali sostituti saranno per lo meno ottimamente condizionati spettralmente. Nella prossima sezione, data una matrice A generica, studieremo un suo possibile sostituto, la migliore approssimazione di A in $\mathcal{L} \subset \mathbb{C}^{n\times n}$ in norma di Frobenius, che è univocamente determinata se $\mathcal{L}$ è un sottospazio di $\mathbb{C}^{n\times n}$.

2.6 Migliore approssimazione in algebre di matrici generiche

Migliore approssimazione di A in sottospazi $\mathcal{L}$ di $\mathbb{C}^{n\times n}$

Sia $\mathcal{L}$ un sottospazio di $\mathbb{C}^{n\times n}$ di dimensione m e siano J_k, $k = 1, \ldots, m$, m matrici di $\mathcal{L}$ linearmente indipendenti. Poiché $\mathbb{C}^{n\times n}$ è uno spazio di Hilbert rispetto al prodotto scalare $(X, Y) = \sum_{i,j=1}^n \overline{x}_{ij} y_{ij}$ e la norma indotta da tale prodotto scalare è la norma di Frobenius $\|X\|_F = (\sum_{i,j=1}^n |x_{i,j}|^2)^{1/2}$, per il Teorema della proiezione di Hilbert, comunque presa $A \in \mathbb{C}^{n\times n}$, esiste ed è unica la sua migliore approssimazione in $\mathcal{L}$ in norma di Frobenius, ovvero esiste ed è unica una matrice $\mathcal{L}_A$ tale che

$$\mathcal{L}_A \in \mathcal{L}, \quad \|\mathcal{L}_A - A\|_F = \min_{X\in\mathcal{L}} \|X - A\|_F$$

e la matrice $\mathcal{L}_A$ è anche caratterizzata dalla condizione

$$\mathcal{L}_A \in \mathcal{L}, \quad (X, A - \mathcal{L}_A) = 0, \quad X \in \mathcal{L}. \tag{21}$$

Sostituendo in (21) X con J_s e $\mathcal{L}_A$ con $\sum_{k=1}^m \alpha_k J_k$, si ottiene la seguente utile rappresentazione di A:

$$\mathcal{L}_A = \sum_{k=1}^m \alpha_k J_k, \ \alpha_k = [B^{-1}\mathbf{c}]_k, \ B_{s,k} = (J_s, J_k), \ c_s = (J_s, A), \ s, k = 1, \ldots, m. \tag{22}$$

Notiamo che B, la matrice dei coefficienti del sistema lineare che occorre risolvere per ottenere gli α_k in (22), è definita positiva. È interessante osservare che per una ampia classe di spazi $\mathcal{L}$ n-dimensionali, la matrice B risulta un elemeto di $\mathcal{L}$ [5].

Esercizio. Sia $\mathcal{L}$ un sottospazio di $\mathbb{C}^{n\times n}$ di dimensione n tale che $I \in \mathcal{L}$ e $\mathcal{L} = \text{Span}\{J_k\}$ dove le J_k soddisfano le identità $J_i^H J_j = \sum_{k=1}^n \overline{(J_k)_{i,j}} J_k$, $1 \le i,j \le n$ (si noti che tali ipotesi sono soddisfatte in particolare dalle algebre $\text{sd}\, U$ e dalle algebre di gruppo). Dimostrare che $\overline{B} \in \mathcal{L}$. Osservare inoltre che tale spazio $\mathcal{L}$ deve essere necessariamente un'algebra chiusa per trasposizione coniugata e che i v_k definiti dall'identità $I = \sum_{k=1}^n v_k J_k$ verificano le uguaglianze $[v_1 \cdots v_n] J_k = \mathbf{e}_k^T$, $k = 1, \ldots, n$, (cioè $\mathcal{L}$ è uno spazio di classe $\mathbb{V}$ [5]), quindi ogni matrice A di $\mathcal{L}$ è univocamente determinata dal vettore $[v_1 \cdots v_n]A$.

Nel caso particolare in cui $\mathcal{L} = \text{sd}\, U = \{Ud(\mathbf{z})U^H : \mathbf{z} \in \mathbb{C}^n\}$ ($\mathcal{L} = \mu\circ$), U unitaria, per $\mathcal{L}_A$ si può ottenere la seguente altra rappresentazione:

$$\mathcal{L}_A = U \,\text{diag}\,((U^H A U)_{ii}) U^H. \tag{23}$$

Questa si ottiene osservando che la norma di Frobenius è invariante per trasformazioni unitarie, quindi

$$\min_{\mathbf{z}\in\mathbb{C}^n} \|A - Ud(\mathbf{z})U^H\|_F^2 = \min_{\mathbf{z}\in\mathbb{C}^n} \|U^H A U - d(\mathbf{z})\|_F^2 = \|U^H A U - \text{diag}\,((U^H A U)_{ii})\|_F^2 = \sum_{i\neq j} |(U^H A U)_{ij}|^2.$$

Un modo per aumentare l'efficienza di algoritmi nella risoluzione di certi problemi applicativi ove siano coinvolte matrici generiche A, consiste nell'usare, in tali algoritmi, in modo opportuno, matrici di $\mathcal{L}$ che approssimino in qualche senso A. Ad esempio usarle per valutare lo spettro di A, per precondizionare oppure rappresentare A, o, perfino, al posto di A. Ma per poter fare questo, le matrici di $\mathcal{L}$ prese in considerazione devono sia conservare le proprietà di A importanti in tali problemi, che avere bassa complessità computazionale.

Quindi, se in particolare si vuole usare $\mathcal{L}_A$, allora sarà opportuno studiare quando (per quali $\mathcal{L}$) la matrice $\mathcal{L}_A$ mantiene certe proprietà di A. Ad esempio, è facile dimostrare che se lo spazio $\mathcal{L}$ ammette una base reale, allora la proiezione $\mathcal{L}_A$ di una matrice A reale è ancora reale. Se $\mathcal{L}$ è chiuso per trasposizione coniugata ($M \in \mathcal{L} \Rightarrow M^H \in \mathcal{L}$), allora la proiezione $\mathcal{L}_A$ di una matrice A hermitiana è ancora hermitiana. Un problema di ricerca interessante, affrontato preliminarmente in [5], [6], è caratterizzare gli spazi $\mathcal{L} \subset \mathbb{C}^{n\times n}$ per cui la proiezione $\mathcal{L}_A$ di una matrice A definita positiva è ancora definita positiva. Infatti, ogni qual volta $\mathcal{L}_A$ è definita positiva, la si potrebbe usare come precondizionatore di A, se si deve risolvere un sistema $A\mathbf{x} = \mathbf{b}$, (vedi ...), oppure al posto di A, se A rappresenta l'approssimazione dell'Hessiano definita nel generico passo del metodo di minimizzazione BFGS (vedi ...). Per altre applicazioni può invece essere utile studiare spazi $\mathcal{L}$ per cui la proiezione $\mathcal{L}_A$ di una matrice A non negativa e/o wstocastica per colonne è ancora non negativa e/o wstocastica per colonne [28] (vedi ...).

Come abbiamo già detto, le matrici dello spazio $\mathcal{L}$ che vogliamo utilizzare, dovrebbero avere anche bassa complessità computazionale. Un sottospazio $\mathcal{L}$ di $\mathbb{C}^{n\times n}$ si dice di bassa complessità se, comunque presa $M \in \mathcal{L}$, il calcolo degli autovalori di M, il prodotto di M per un vettore, e la risoluzione di un sistema lineare con M come matrice dei coefficienti, sono tutte operazioni eseguibili effettuando non più di $O(n \log n)$ operazioni aritmetiche. Ad esempio, se U è una matrice unitaria $n \times n$ per cui le trasformate $U\mathbf{z}$ e $U^H\mathbf{z}$ sono calcolabili con al più $O(n \log n)$ operazioni aritmetiche, allora l'algebra $\mathcal{L} = \{Ud(\mathbf{z})U^H : \mathbf{z} \in \mathbb{C}^n\}$ è un sottospazio di $\mathbb{C}^{n\times n}$ di bassa complessità. Un altro esempio di sottospazio di $\mathbb{C}^{n\times n}$ di bassa complessità è l'algebra $\mathcal{L}$ delle matrici triangolari

di Toeplitz (studiata approfonditamente nel Capitolo 2). Notiamo che per entrambi gli esempi, il problema degli autovalori di $M \in \mathcal{L}$ è ottimamente condizionato.

Un esempio: la migliore approssimazione di $A \in \mathbb{C}^{n\times n}$ nell'algebra $\mathcal{C}_\phi$ delle matrici ϕ-circolanti

Sia Π_ϕ la seguente matrice $n \times n$

$$\Pi_\phi = \begin{bmatrix} 0 & 1 & & & \\ \cdot & 0 & 1 & & \\ \cdot & \cdot & \cdot & \cdot & \\ 0 & \cdot & \cdot & 0 & 1 \\ \phi & 0 & \cdot & \cdot & 0 \end{bmatrix}, \ \phi \in \mathbb{C}.$$

Si osserva che il polinomio minimo di Π_ϕ coincide con il polinomio caratteristico di Π_ϕ, $p_{\Pi_\phi}(\lambda) = \lambda^n - \phi$. Quindi l'algebra $\{p(\Pi_\phi)\} = \{\sum_{j=1}^n \alpha_j \Pi_\phi^{j-1} : \alpha_j \in \mathbb{C}\}$ ha dimensine n e coincide con l'insieme $\{A : A\Pi_\phi = \Pi_\phi A\}$. Chiamiamo $\mathcal{C}_\phi$ tale algebra, o algebra delle matrici ϕ-circolanti. Notiamo che, comunque dati $a_i \in \mathbb{C}$, è sempre ben definita la matrice ϕ-circolante la cui prima riga è $\mathbf{a}^T = [a_1\, a_2 \cdots a_n]$:

$$\mathcal{C}_\phi(\mathbf{a}) := \sum_{j=1}^n a_j \Pi_\phi^{j-1} = \begin{bmatrix} a_1 & a_2 & a_3 & \cdot & a_n \\ \phi a_n & a_1 & a_2 & \cdot & a_{n-1} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \phi a_3 & \cdot & \cdot & \cdot & a_2 \\ \phi a_2 & \phi a_3 & \cdot & \phi a_n & a_1 \end{bmatrix}.$$

Osserviamo che $\mathcal{C}_0$ è lo spazio delle matrici triangolari superiori di Toeplitz. Le matrici di $\mathcal{C}_0$ ovviamente non sono in generale diagonalizzabili (vedi ad esempio $\mathcal{C}_0(\mathbf{e}_2)$).

Se invece $\phi \neq 0$, allora la matrice Π_ϕ ha n autovalori distinti ed è, quindi, diagonalizzabile, cioè esiste U invertibile tale che $U^{-1}\Pi_\phi U$ è diagonale. Poiché la stessa U diagonalizza anche ogni polinomio in Π_ϕ, l'algebra $\mathcal{C}_\phi$, se $\phi \neq 0$, ammette l'ulteriore rappresentazione $\mathcal{C}_\phi = \{Ud(\mathbf{z})U^{-1} : \mathbf{z} \in \mathbb{C}^n\}$. In particolare, si ha l'identità $\mathcal{C}_\phi(\mathbf{a}) = Ud(U^T\mathbf{a})d(U^T\mathbf{e}_1)^{-1}U^{-1}$. Si osserva inoltre che, per il Teorema 2.9, la matrice U può essere scelta unitaria se e soltanto se Π_ϕ è normale ($\Pi_\phi\Pi_\phi^H = \Pi_\phi^H\Pi_\phi$) ovvero se e solo se ϕ ha modulo uguale a 1.

Notiamo che, per $\phi \neq 0$, la matrice U tale che $U^{-1}\Pi_\phi U = \text{diag}$ è nota esplicitamente. Infatti, $U = DF$ dove F è la matrice di Fourier (vedi (3)) e D è una opportuna matrice diagonale, che è unitaria se e solo se $|\phi| = 1$. Ricordando le proprietà di F (vedi Proposizione 2.3), si può quindi dire immediatamente che, per $\phi \neq 0$, il calcolo degli autovalori di $\mathcal{C}_\phi(\mathbf{a})$, il prodotto di $\mathcal{C}_\phi(\mathbf{a})$ per un vettore, e la risoluzione di un sistema lineare con matrice dei coefficienti $\mathcal{C}_\phi(\mathbf{a})$, sono tutte operazioni di complessità non maggiore di $O(n \log n)$. Di fatto, le stesse affermazioni si dimostrano, per altre vie, anche nel caso $\phi = 0$ (basta "trasporre" i risultati della Sezione 2.2). Quindi $\mathcal{C}_\phi$ è per ogni valore di ϕ uno spazio di bassa complessità.

Calcoliamo la migliore approssimazione di $A \in \mathbb{C}^{n\times n}$ in $\mathcal{C}_\phi$, cioè la matrice $(\mathcal{C}_\phi)_A \in \mathcal{C}_\phi$ tale che

$$\|A - (\mathcal{C}_\phi)_A\|_F = \min_{X\in\mathcal{C}_\phi} \|A - X\|_F, \quad (X, A - (\mathcal{C}_\phi)_A) = 0, \ \forall\, X \in \mathcal{C}_\phi.$$

Usiamo a tal fine quest'ultima condizione e il fatto che le matrici Π^{j-1}, $j = 1, \ldots, n$, costituiscono una base per $\mathcal{C}_\phi$. Per semplicità si usa il simbolo Π anziché Π_ϕ. Si ha:

$$(\Pi^{s-1}, A - \sum_{j=1}^{n} \alpha_j \Pi^{j-1}) = 0, \ s = 1, \ldots, n,$$

$$(\Pi^{s-1}, A) = \sum_{j=1}^{n} \alpha_j (\Pi^{s-1}, \Pi^{j-1}) = \alpha_s (\Pi^{s-1}, \Pi^{s-1}) = \alpha_s (n - s + 1 + |\phi|^2 (s-1)), \ s = 1, \ldots, n.$$

Caso $\phi = 0$:

$$(\mathcal{C}_0)_A = \sum_{j=1}^{n} \frac{1}{n-j+1} (\Pi^{j-1}, A) \Pi^{j-1}.$$

Caso $\phi \neq 0$:

$$(\mathcal{C}_\phi)_A = \sum_{j=1}^{n} \alpha_j \Pi^{j-1} = U d(U^T \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdot \\ \alpha_n \end{bmatrix}) d(U^T \mathbf{e}_1)^{-1} U^{-1}, \quad \alpha_j = \frac{1}{n - j + 1 + |\phi|^2 (j-1)} (\Pi^{j-1}, A).$$

Caso $|\phi| = 1$:

$$(\mathcal{C}_\phi)_A = \frac{1}{n} \sum_{j=1}^{n} (\Pi^{j-1}, A) \Pi^{j-1} = U d(U^T \frac{1}{n} \begin{bmatrix} (I, A) \\ (\Pi, A) \\ \cdot \\ (\Pi^{n-1}, A) \end{bmatrix}) d(U^T \mathbf{e}_1)^{-1} U^H = U \, \mathrm{diag}\,((U^H A U)_{ii}) U^H.$$

Quali proprietà di A sono ereditate da $(\mathcal{C}_\phi)_A$? Ad esempio, A hermitiana implica $(\mathcal{C}_\phi)_A$ hermitiana? A definita positiva implica $(\mathcal{C}_\phi)_A$ definita positiva? Dall'ultima formula per $(\mathcal{C}_\phi)_A$ è evidente che la risposta a queste due domande è si, a patto che $|\phi| = 1$. Un'altra domanda può essere: A reale implica $(\mathcal{C}_\phi)_A$ reale? Risposta: si, se $\mathcal{C}_\phi$ è generata da matrici reali, ovvero se $\phi \in \mathbb{R}$. Altre domande possono essere: eventuali proprietà di nonnegatività, irriducibilità, o contemporanea nonnegatività e irriducibilità di A quando (per quali valori di ϕ) sono ereditate $(\mathcal{C}_\phi)_A$? quando A wstocastica implica $(\mathcal{C}_\phi)_A$ wstocastica? lo spettro di $(\mathcal{C}_\phi)_A$ è in qualche relazione con lo spettro di A?

È chiaro che ha interesse il seguente problema più generale: per quali spazi $\mathcal{L}$ le proprietà di cui sopra sono ereditate da $\mathcal{L}_A$?

Esempio. Scriviamo $(\mathcal{C}_1)_A = \mathcal{C}_A$, la matrice circolante che meglio approssima A, per alcune scelte di A, e confrontiamo lo spettro di $\mathcal{C}_A$ con quello di A. Col simbolo $\overline{\sigma(M)}$, $\sigma(M) = \{\lambda_1(M), \ldots, \lambda_n(M)\}$, intendiamo il più piccolo insieme convesso in $\mathbb{C}$ contenente lo spettro $\sigma(M)$ di M.

$$A = \begin{bmatrix} 0 & 1 \\ 2 & 0 \end{bmatrix}, \ \mathcal{C}_A = \begin{bmatrix} 0 & 3/2 \\ 3/2 & 0 \end{bmatrix}, \ \overline{\sigma(A)} = [-\sqrt{2}, \sqrt{2}] \subset \overline{\sigma(\mathcal{C}_A)} = [-3/2, 3/2],$$

$$A = \begin{bmatrix} 0 & 1 \\ -4 & 0 \end{bmatrix}, \ \mathcal{C}_A = \begin{bmatrix} 0 & -3/2 \\ -3/2 & 0 \end{bmatrix}, \ \overline{\sigma(A)} = \{\mathbf{i}b : \ -2 \leq b \leq 2\}, \ \overline{\sigma(\mathcal{C}_A)} = [-3/2, 3/2],$$

$$A = \begin{bmatrix} 0 & 1 \\ \mathbf{i} & 0 \end{bmatrix}, \quad \mathcal{C}_A = \begin{bmatrix} 0 & \frac{1}{2}(1+\mathbf{i}) \\ \frac{1}{2}(1+\mathbf{i}) & 0 \end{bmatrix},$$

$$\overline{\sigma(\mathcal{C}_A)} = \{be^{\mathbf{i}\pi/4} : -\frac{1}{\sqrt{2}} \le b \le \frac{1}{\sqrt{2}}\} \subset \overline{\sigma(A)} = \{be^{\mathbf{i}\pi/4} : -1 \le b \le 1\}.$$

Osserviamo che nel terzo caso si ha $\overline{\sigma(\mathcal{C}_A)} \subset \overline{\sigma(A)}$. Come è provato poco più avanti è vera la seguente affermazione piu generale: se A è una matrice normale e $|\phi| = 1$, allora $\overline{\sigma((\mathcal{C}_\phi)_A)} \subset \overline{\sigma(A)}$.

Esempio. Sia A $n \times n$ non negativa ed $\mathcal{L}$ un sottospazio di $\mathbb{C}^{n\times n}$ generico con base $\{J_k\}$ ortogonale a elementi non negativi. Allora anche $\mathcal{L}_A$ è ovviamente non negativa (si usi la rappresentazione (22) di $\mathcal{L}_A$). In [52] è dimostrato il viceversa, cioè condizione necessaria affinché la proiezione di A in un sottospazio $\mathcal{L}$ sia non negativa per ogni A non negativa è che $\mathcal{L}$ ammetta una base ortogonale non negativa. Quindi, ad esempio si può dire che tra tutte le algebre $\mathcal{C}_\phi$ le uniche che conservano una eventuale non negatività di A sono $\mathcal{C}_1$ (circolanti) e $\mathcal{C}_0$ (triangolari di Toeplitz).

Lo studio su $\mathcal{C}_\phi$ e $(\mathcal{C}_\phi)_A$ svolto in questa sezione può essere facilmente ripetuto per altri spazi $\mathcal{L} = \{Ud(\mathbf{z})U^H : \mathbf{z} \in \mathbb{C}^n\}$ di bassa complessità, come le algebre di tipo Jacobi e di tipo Hartley. Cerchiamo ad esempio l'espressione esplicita di τ_A essendo τ l'algebra di bassa complessità descritta nella Sezione 2.1. Siano J_k le matrici di τ con prima riga $\mathbf{e}_k^T$, $k = 1, \ldots, n$. Per (22) si ha che $\tau_A = \sum_{k=1}^n (B^{-1}\mathbf{c})_k J_k$, con $B_{k,s} = (J_k, J_s)$, $c_k = (J_k, A)$. Scriviamo B^{-1} esplicitamente. Si ha

$$B = \begin{bmatrix} 4 & 0 & 2 & 0 \\ 0 & 6 & 0 & 2 \\ 2 & 0 & 6 & 0 \\ 0 & 2 & 0 & 4 \end{bmatrix}, \quad B = nJ_1 + (n-2)J_3 + \ldots + \begin{cases} J_n & n \text{ dispari} \\ 2J_{n-1} & n \text{ pari} \end{cases}.$$

Poiché l'inversa di B deve essere un elemento di τ (perché τ appartiene alla classe delle *algebre, vedi [6], [5]), per conoscerla è sufficiente conoscere la sua prima riga, cioè il vettore $\mathbf{z}$ tale che $\mathbf{z}^T B = \mathbf{e}_1^T$. Calcolando $\mathbf{z}$ per i primi valori di n si intuisce l'espressione esplicita di B^{-1} per n generico: $B^{-1} = \frac{1}{2n+2}(3J_1 - J_3)$. Quindi

$$\begin{aligned} \tau_A \;=\; & \frac{1}{2n+2}\Big((3c_1 - c_3)J_1 + (2c_2 - c_4)J_2 + (-c_1 + 2c_3 - c_5)J_3 + \ldots \\ & +(-c_{k-2} + 2c_k - c_{k+2})J_k + \ldots (-c_{n-4} + 2c_{n-2} - c_n)J_{n-2} \\ & +(-c_{n-3} + 2c_{n-1})J_{n-1} + (-c_{n-2} + 3c_n)J_n\Big), \quad c_k = (J_k, A). \end{aligned}$$

Usando questa rappresentazione di τ_A si dimostra in particolare che τ_A in generale non eredita da A una sua eventuale non negatività. Ad esempio, essendo $[\tau_A]_{11} = \frac{1}{2n+2}(3c_1 - c_3)[J_1]_{11} = \frac{1}{2n+2}(3c_1 - c_3)$, si può prendere una A non negativa per cui $3c_1 - c_3$ sia minore di zero; per $n = 3$ una tale A è la seguente

$$A = \begin{bmatrix} 1 & 0 & 5 \\ 0 & 1 & 0 \\ 5 & 0 & 1 \end{bmatrix}, \quad 3c_1 - c_3 = 3(J_1, A) - (J_3, A) = 3 * 3 - 11 = -2.$$

Esercizio. Sia $\mathcal{L}$ l'algebra di Haar considerata nella Sezione 2.4. Si dia una rappresentazione di $\mathcal{L}_A$ utilizzando la base di $\mathcal{L}$ costituita da matrici i cui elementi sono -1, 0, oppure 1.

Localizzazione e confronto degli spettri di A ed $\mathcal{L}_A$ dove $\mathcal{L} = \{Ud(\mathbf{z})U^H\}$, $\mathcal{L} = \ldots$?

Riportiamo alcuni risultati sulla localizzazione degli autovalori di A e della sua migliore approssimazione in uno spazio $\mathcal{L}$ di matrici simultaneamente diagonalizzate da U unitaria (ad esempio $\mathcal{L} = \tau$, $\mathcal{C}_\phi$, $|\phi| = 1$, $\mathcal{L}$ =Hartley-type). Per $M \in \mathbb{C}^{n\times n}$, con il simbolo $\overline{\sigma(M)}$ intendiamo il più piccolo insieme convesso in $\mathbb{C}$ contenente gli autovalori di M, e con il simbolo $\mathcal{F}(M)$ intendiamo il campo dei valori di M, cioè l'insieme

$$\mathcal{F}(M) = \Big\{ \frac{\mathbf{x}^H M \mathbf{x}}{\mathbf{x}^H \mathbf{x}} : \mathbf{x} \in \mathbb{C}^n \Big\}.$$

Se λ è un autovalore di M allora esso ammette la seguente rappresentazione

$$\lambda = \frac{\mathbf{x}^H M \mathbf{x}}{\mathbf{x}^H \mathbf{x}}$$

dove $\mathbf{x}$ è un qualsiasi autovettore di M corrispondente a λ. Da questa osservazione e da altri argomenti, seguono i risultati elencati qui di seguito, sugli spettri di $A \in \mathbb{C}^{n\times n}$ e di $\mathcal{L}_A$. Lasciamo al lettore la dimostrazione non difficile di tali risultati.

1 Se $A = A^H$ $[A = -A^H$ $(A = (A^H)^{-1})]$ e λ è un qualsiasi autovalore di A, allora $\lambda \in \mathbb{R}$ $[\mathbf{i}\lambda \in \mathbb{R}$ $(|\lambda| = 1)]$

2 $\mathcal{F}(A)$ contiene tutti gli autovalori di A. $\mathcal{F}(A)$ è un sottoinsieme convesso di $\mathbb{C}$ [34], quindi $\overline{\sigma(A)} \subset \mathcal{F}(A)$. Se A ha elementi reali, allora $\mathcal{F}(A)$, come $\overline{\sigma(A)}$, è simmetrico rispetto all'asse reale.

3 Sia $\mathcal{L} = \{Ud(\mathbf{z})U^H : \mathbf{z} \in \mathbb{C}^n\}$, dove U è una matrice unitaria (ad esempio $\mathcal{L} = \mathcal{C}_\phi$ con $|\phi| = 1$). Sia $\mathcal{L}_A = U\,\mathrm{diag}\,((U^H A U)_{ii})U^H$ la migliore approssimazione di $A \in \mathbb{C}^{n\times n}$ in $\mathcal{L}$. Allora $\overline{\sigma(\mathcal{L}_A)} \subset \mathcal{F}(A)$

4 Se A è normale, allora $\mathcal{F}(A)$ è il più sottoinsieme convesso di $\mathbb{C}$ contenente gli autovalori di A, cioè $\mathcal{F}(A) = \overline{\sigma(A)}$ e, per il punto precedente, $\overline{\sigma(\mathcal{L}_A)} \subset \overline{\sigma(A)}$. Inoltre, per ogni $\mathbf{x} \in \mathbb{C}^n$ $\mathbf{x} \neq \mathbf{0}$, nel cerchio chiuso in $\mathbb{C}$ di centro $\frac{\mathbf{x}^H A \mathbf{x}}{\mathbf{x}^H \mathbf{x}}$ e raggio il numero non negativo $\big(\frac{\|A\mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2} - \big|\frac{\mathbf{x}^H A \mathbf{x}}{\mathbf{x}^H \mathbf{x}}\big|^2\big)^{1/2}$ deve esserci almeno un autovalore di A (vedi Teorema 3.2).

5 Se A è hermitiana, $\mathcal{F}(A) = \overline{\sigma(A)}$ è il più piccolo intervallo della retta reale contenente gli autovalori di A, $\mathcal{L}_A$ è hermitiana e $\overline{\sigma(\mathcal{L}_A)} \subset \overline{\sigma(A)}$. Se A è definita positiva, allora $\mathcal{L}_A$ è definita positiva e $\overline{\sigma(\mathcal{L}_A)} \subset \overline{\sigma(A)}$

6 Sia $A \in \mathbb{C}^{n\times n}$. Ad A possiamo associare le seguenti due matrici normali $A_h = \frac{1}{2}(A + A^H)$ e $A_{ah} = \frac{1}{2}(A - A^H)$, tali che $A = A_h + A_{ah}$, che chiamiamo rispettivamente parte hermitiana e parte anti hermitiana di A. Allora

$$\overline{\sigma(A)} \subset \mathcal{F}(A) \subset \mathcal{F}(A_h) \times \mathcal{F}(A_{ah}) = \overline{\sigma(A_h)} \times \overline{\sigma(A_{ah})}.$$

$$\mathcal{F}(\mathcal{L}_A) = \overline{\sigma(\mathcal{L}_A)} \subset \mathcal{F}(\mathcal{L}_{A_h}) \times \mathcal{F}(\mathcal{L}_{A_{ah}}) = \overline{\sigma(\mathcal{L}_{A_h})} \times \overline{\sigma(\mathcal{L}_{A_{ah}})} \subset \mathcal{F}(A_h) \times \mathcal{F}(A_{ah}) = \overline{\sigma(A_h)} \times \overline{\sigma(A_{ah})}.$$

In particolare, se A ha parte hermitiana definita positiva (negativa), allora gli autovalori di A hanno parte reale positiva (negativa).

Le prime inclusioni del punto 6 seguono dall'osservazione che per ogni $\mathbf{z} \in \mathbb{C}^n$

$$\frac{\mathbf{z}^H A\mathbf{z}}{\mathbf{z}^H\mathbf{z}} = a + b = a + \mathbf{i}\frac{b}{\mathbf{i}},\ a = \frac{\mathbf{z}^H A_h\mathbf{z}}{\mathbf{z}^H\mathbf{z}} \in \mathbb{R},\ b = \frac{\mathbf{z}^H A_{ah}\mathbf{z}}{\mathbf{z}^H\mathbf{z}} \in \mathbf{i}\mathbb{R}.$$

Domanda: anche per $\mathcal{L} = \mu\circ$ si puo' stabilire qualcosa ?

$(\mathcal{C}_\phi)_A$ **(**$\mathcal{L}_A$**,** $\mathcal{L}$ **=) come precondizionatore di** A **Toeplitz**

$(\mathcal{C}_\phi)_A$ **(**$\mathcal{L}_A$**,** $\mathcal{L}$ **=) al posto di** A **approssimazione dell'Hessiano**

3 Preliminari, Teoria di Perron-Frobenius, Page-Rank

Introduciamo innanzitutto il concetto di irriducibilità di una matrice e stabiliamo due risultati interessanti sulle matrici irriducibili. Uno riguarda la localizzazione degli autovalori di una matrice irriducibile; è una sorta di teorema di Gershgorin rafforzato, in genere ignorato nella letteratura di carattere didattico. Nel secondo si suppone la matrice anche non negativa e si stabilisce che le potenze di questa, perturbata di I, diventano a un certo punto positive. Il secondo risultato è alla base della tecnica di dimostrazione stile Gantmacher-Varga della teoria di Perron-Frobenius, che seguiremo anche in questo testo. Poi, vi saranno tre sezioni. Nella prima, nella risoluzione del problema calcolo dell'autovettore dominante di una matrice non negativa wstocastica per colonne, si illustrano le principali affermazioni della teoria di Perron-Frobenius (PF), ricordando e utilizzando risultati e metodi classici sull'analisi numerica degli autovalori. Nella seconda si enunciano e dimostrano i risultati della teoria di PF. Nella terza si considera l'applicazione della teoria sviluppata nella formulazione e risoluzione del problema pagerank (vertexrank) del web (di un grafo).

Irriducibilità e normalità in teoremi per la localizzazione degli autovalori. Irriducibilità nel Lemma alla base della teoria di Perron-Frobenius

Sia $A \in \mathbb{C}^{n\times n}$. Se esiste $\mathcal{I} \subset \{1,\dots,n\}$ tale che $0 < |\mathcal{I}| < n$ e $a_{ik} = 0$ per ogni $i \in \mathcal{I}$, $k \notin \mathcal{I}$, allora A si dice riducibile. Ciò equivale a dire che esistono matrici di permutazione P e Q tali che

$$P^T AP = \begin{bmatrix} M & W \\ O & N \end{bmatrix}, \quad Q^T AQ = \begin{bmatrix} M' & O \\ W' & N' \end{bmatrix},$$

con M, M' ed N, N' matrici quadrate entrambe almeno 1×1. Un esempio:

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 \\ & & & \\ 0 & 0 & 0 & 0 \\ & & & \\ 0 & 0 & 0 & 0 \\ & & & \end{bmatrix}, \; Q = \begin{bmatrix} 1 & & & & & & \\ & & & 1 & & & \\ & & & & & 1 & \\ & 1 & & & & & \\ & & & & 1 & & \\ & & 1 & & & & \\ & & & & & & 1 \end{bmatrix}, \quad Q^T AQ = \begin{bmatrix} 0 & 0 & 0 & 0 & & & \\ 0 & 0 & 0 & 0 & & & \\ 0 & 0 & 0 & 0 & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \end{bmatrix}$$

(gli elementi non specificati in A possono essere numeri arbitrari, qualcuno può essere anche zero).

Ricordiamo il ben noto risultato che caratterizza le matrici A irriducibili come quelle associate a grafi fortemente connessi. Data $A \in \mathbb{C}^{n\times n}$, il grafo orientato $\mathcal{G}$ associato ad A ha come nodi $1, 2, \dots, n$ e $\mathcal{G}$ contiene l'arco che unisce il nodo i al nodo j se e solo se $a_{ij} \neq 0$. Tale $\mathcal{G}$ si dice fortemente connesso se comunque presi due nodi i, j in $\mathcal{G}$ esiste un cammino in $\mathcal{G}$ che li unisce.

Ad ogni matrice $A \in \mathbb{C}^{n\times n}$ si possono associare i seguenti n sottoinsiemi di $\mathbb{C}$, detti *cerchi di Gershgorin*:

$$K_i = \{z \in \mathbb{C} : \; |z - a_{ii}| \leq \sum_{j:\, j\neq i} |a_{ij}|\}, \; i = 1, \dots, n\,.$$

Notiamo che i cerchi K_i sono chiusi, quindi indichiamo con $\dot{K}_i = K_i \backslash (\partial K_i)$ la loro parte interna.

Teorema 3.1 (Gershgorin) Sia $A \in \mathbb{C}^{n\times n}$ e λ un generico autovalore di A. Sia $\mathbf{x}$ un autovettore di A corrispondente all'autovalore λ, e $\mathcal{I} \subset \{1, 2, \dots, n\}$ l'insieme degli indici i per cui $|x_i| = \|\mathbf{x}\|_\infty$. Allora valgono le seguenti affermazioni:

1. $\lambda \in K_i$ per ogni $i \in \mathcal{I}$.

2. Se A è irriducibile e $\mathcal{I} \neq \{1, 2, \ldots, n\}$, allora $\exists\, r \in \mathcal{I}$ tale che $\lambda \in \dot{K}_r$.

3. Se A è irriducibile e $\mathcal{I} = \{1, 2, \ldots, n\}$, allora $\lambda \in K_i$ per ogni i, quindi può accadere che $\lambda \in \partial K_i$ per ogni i (ma non e' che accade sempre ? ...) o che λ è interno ad uno dei cerchi.

Quindi, si ha sempre che $\lambda \in \cup_i K_i$, e, nel caso particolare in cui A è irriducibile, si ha più precisamente che $\lambda \in \big(\cup_i \dot{K}_i)\big) \cup (\cap_i \partial K_i)$.

Dimostrazione. Dall'identità $A\mathbf{x} = \lambda\mathbf{x}$ segue che $\sum_j a_{ij}x_j = \lambda x_i$, $\sum_{j:\, j\neq i} a_{ij}x_j = (\lambda - a_{ii})x_i$,

$$|\lambda - a_{ii}||x_i| \leq \sum_{j:\, j\neq i} |a_{ij}||x_j|,\ i = 1, 2, \ldots, n\,. \tag{24}$$

Per dimostrare il punto 1 basta osservare che scegliendo $i \in \mathcal{I}$ in (24) si ottengono le disuguaglianze $|\lambda - a_{ii}| \leq \sum_{j:\, j\neq i} |a_{ij}|$, $i \in \mathcal{I}$. Per 2, si nota che, poiché A è irriducibile, deve necessariamente esistere almeno una coppia di indici $r \in \mathcal{I}$ e $k \in \{1, \ldots, n\}\backslash\mathcal{I}$ per cui $a_{rk} \neq 0$, quindi (24) con $i = r$ implica

$$\begin{array}{rcl} |\lambda - a_{rr}||x_r| & \leq & \sum_{j:\, j\neq r} |a_{rj}||x_j| = \sum_{j\in\mathcal{I}:\, j\neq r} |a_{rj}||x_j| + \sum_{j\notin\mathcal{I}:\, j\neq k} |a_{rj}||x_j| + |a_{rk}||x_k| \\ & < & \sum_{j\in\mathcal{I}:\, j\neq r} |a_{rj}||x_r| + \sum_{j\notin\mathcal{I}:\, j\neq k} |a_{rj}||x_r| + |a_{rk}||x_r| = |x_r| \sum_{j:\, j\neq r} |a_{rj}|\,, \end{array}$$

dove la diseguaglianza stretta segue dal fatto che $|x_k| < |x_r|$ e $a_{rk} \neq 0$. Infine, il punto 3 segue dal punto 1 (non serve nemmeno l'ipotesi A irriducibile). □

È ben noto che $\lambda \in \mathbb{C}$ è autovalore di A se e solo se è autovalore di A^T. Quindi, oltre i cerchi $K_i = K_i(A)$ è utile considerare i cerchi $K_i(A^T) = \{z \in \mathbb{C} :\ |z - a_{ii}| \leq \sum_{j:\, j\neq i} |a_{ji}|\}$, $i = 1, \ldots, n$, e, ogni volta che $K_i(A^T) \neq K_i(A)$ per almeno qualche i, applicare il risultato di Gershgorin, oltre che ad A, anche ad A^T (tenendo presente che il vettore $\mathbf{x}$ in tal caso sarebbe un vettore tale che $A^T\mathbf{x} = \lambda\mathbf{x}$). In questo modo si ottengono più informazioni sulla localizzazione degli autovalori di A in $\mathbb{C}$. Ad esempio, si conclude subito che

$$\lambda \text{ autovalore di } A \ \Rightarrow\ \lambda \in \big(\cup_i K_i(A)\big) \cap \big(\cup_i K_i(A^T)\big),$$

$$\lambda \text{ autovalore di } A \text{ irriducibile } \Rightarrow\ \lambda \in \Big((\cup_i \dot{K}_i(A)) \cup (\cap_i \partial K_i(A))\Big) \cap \Big((\cup_i \dot{K}_i(A^T)) \cup (\cap_i \partial K_i(A^T))\Big).$$

Esercizio. ...

Una matrice $A \in \mathbb{C}^{n\times n}$ si dice normale se $AA^H = A^HA$, ovvero se e solo se A è diagonalizzabile da una matrice unitaria (Teorema 2.9). Ad ogni matrice $A \in \mathbb{C}^{n\times n}$ normale si possono associare i seguenti sottoinsiemi di $\mathbb{C}$, detti *cerchi di Weinstein*:

$$\mathbf{w} \in \mathbb{C}^n \quad \rightarrow \quad K_{\mathbf{w}}(A) = \Big\{z \in \mathbb{C} :\ \big|z - \frac{\mathbf{w}^HA\mathbf{w}}{\mathbf{w}^H\mathbf{w}}\big| \leq \big(\frac{\|A\mathbf{w}\|^2}{\|\mathbf{w}\|^2} - \big|\frac{\mathbf{w}^HA\mathbf{w}}{\mathbf{w}^H\mathbf{w}}\big|^2\big)^{1/2}\Big\}.$$

In particolare, possiamo associare ad A gli n cerchi:

$$K_{\mathbf{e}_i}(A) = \Big\{z \in \mathbb{C} :\ |z - a_{ii}| \leq \Big(\sum_{j:\, j\neq i} |a_{ji}|^2\Big)^{1/2}\Big\},\ i = 1, \ldots, n,$$

centrati anch'essi negli elementi diagonali di A, come i cerchi di Gershgorin. Osserviamo che $K_{\mathbf{e}_i}(A) \subset K_i(A^T)$ e $K_{\mathbf{e}_i}(A^T) \subset K_i(A)$.

Teorema 3.2 (Weinstein) Sia $A \in \mathbb{C}^{n\times n}$ normale. Allora

i) per ogni $\mathbf{w} \in \mathbb{C}^n$, sia il cerchio $K_{\mathbf{w}}(A)$ che il cerchio $K_{\mathbf{w}}(A^T)$ contengono almeno un autovalore di A

ii) $\forall\, i \in \{1, \ldots, n\}$, sia il cerchio $K_{\mathbf{e}_i}(A)$ che il cerchio $K_{\mathbf{e}_i}(A^T)$ contengono almeno un autovalore di A

iii) Si ha l'identità $K_{\mathbf{w}}(A) = \frac{\mathbf{w}^H A\mathbf{w}}{\mathbf{w}^H\mathbf{w}}$ se e solo se $\mathbf{w}$ è un autovettore di A. In tal caso, $K_{\mathbf{w}}(A)$ è l'autovalore di A con autovettore $\mathbf{w}$.

Dimostrazione. Per ogni $\mathbf{w} \in \mathbb{C}^n$ si ha $\min_i |\lambda_i - \mu|^2 \leq \frac{\|(A-\mu I)\mathbf{w}\|_2^2}{\|\mathbf{w}\|_2^2}$. Scegliendo $\mu = \frac{\mathbf{w}^H A\mathbf{w}}{\mathbf{w}^H\mathbf{w}}$ si ha la tesi. ... Aggiungere detagli ... $\square$

Esempio. Sia A la seguente matrice 3×3:

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}.$$

I cerchi di Weinstein $K_{\mathbf{e}_i}(A^T) = K_{\mathbf{e}_i}(A) = \{z \in \mathbb{C} : |z-2| \leq \sqrt{2}\}$ contengono effettivamente almeno un autovalore di A, 1. Osserviamo anche che c'è un autovalore di A, 4, che non è contenuto in nessuno di tali cerchi. L'autovalore 4 è invece un punto dell'insieme $\cap_i \partial K_i$. Sia A la matrice

$$A = \begin{bmatrix} 0 & 1 \\ -\frac{1}{4} & 1 \end{bmatrix}.$$

Uno dei cerchi di Gershgorin di A (o di A^T) non contiene nessun autovalore di A. Ne segue, in particolare, che A non può essere normale.

Cambiamo argomento, considerando un altro risultato ove l'ipotesi di irriducibilità di A interviene pesantemente. Tale risultato, che presuppone A anche non negativa ($A \geq O$), è alla base delle tecniche di dimostrazione (prese essenzialmente dal Varga [39]) che useremo per introdurre più avanti la teoria di Perron-Frobenius per matrici non negative irriducibili.

Teorema 3.3 Sia $A \in \mathbb{R}^{n\times n}$ non negativa e irriducibile. Allora $(I + A)^{n-1} > O$, cioè gli elementi della matrice $(I + A)^{n-1}$ devono necessariamente essere tutti positivi.

Dimostrazione. Si ha che $(I + A)^{n-1} > O$ se e solo se le componenti del vettore $(I + A)^{n-1}\mathbf{x}$ sono tutte positive per ogni vettore $\mathbf{x} \in \mathbb{R}^n$ non negativo non nullo (provare tale affermazione!). Quindi è sufficiente mostrare che $(I + A)^{n-1}\mathbf{x} > \mathbf{0}\ \forall\ \mathbf{x} \geq \mathbf{0}\ \mathbf{x} \neq \mathbf{0}$. Sia $\mathbf{x} \geq \mathbf{0}$, $\mathbf{x} \neq \mathbf{0}$. Sia $\mathbf{x}_0 = \mathbf{x}$ e $\mathbf{x}_{k+1} = (I + A)\mathbf{x}_k = \mathbf{x}_k + A\mathbf{x}_k$, $k = 0, 1, 2, \ldots$. Osserviamo che $\mathbf{x}_k = (I + A)^k\mathbf{x}$. Si vuole dimostrare che $\mathbf{x}_{n-1} > \mathbf{0}$. Come prima cosa notiamo che gli $\mathbf{x}_k$ sono tutti vettori non negativi (per induzione su k). Come seconda cosa, meno immediata, notiamo che $\mathbf{x}_{k+1}$ deve avere meno zeri di $\mathbf{x}_k$. Infatti, di più non ne può avere perché $\mathbf{x}_{k+1}$ si ottiene da $\mathbf{x}_k$ aggiungendo un vettore non negativo. Inoltre, se $\mathbf{x}_{k+1}$ e $\mathbf{x}_k$ avessero lo stesso numero di zeri, questi dovrebbero occupare le stesse posizioni (se $(\mathbf{x}_{k+1})_i = 0$, la componente i di $\mathbf{x}_k$ non può essere positiva perché $(A\mathbf{x}_k)_i \geq 0$!), quindi esisterebbe

P matrice di permutazione tale che $P\mathbf{x}_{k+1} = \begin{bmatrix} \mathbf{a} \\ \mathbf{0} \end{bmatrix}$, $P\mathbf{x}_k = \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix}$, con $\mathbf{a}$ e $\mathbf{b}$ positivi e della stessa dimensione m, $1 \leq m \leq n-1$. Ne segue che

$$P\mathbf{x}_{k+1} = P\mathbf{x}_k + PA\mathbf{x}_k = P\mathbf{x}_k + PAP^T P\mathbf{x}_k,$$

$$\begin{bmatrix} \mathbf{a} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix}, \ M_{11} \ m \times m,$$

da cui l'identità $M_{21}\mathbf{b} = \mathbf{0}$, che, essendo $M_{21} \geq O$ e $\mathbf{b} > \mathbf{0}$, implicherebbe $M_{21} = O$. In altre parole, supporre che $\mathbf{x}_{k+1}$ e $\mathbf{x}_k$ abbiano lo stesso numero di zeri ci farebbe concludere che A è riducibile, ma A non lo è! Dunque $\mathbf{x}_{k+1}$ deve avere meno zeri di $\mathbf{x}_k$ e, di conseguenza, $\mathbf{x}_{n-1}$ deve essere necessariamente un vettore positivo. Si noti che non è escluso che $\mathbf{x}_k$ ($(I+A)^k$) sia positivo (positiva) già per $k < n-1$. □

Esercizio. Si può dire qualcosa sulla struttura delle matrici normali riducibili (irriducibili)?

3.1 Calcolo dell'autovalore dominante di una matrice A e di un suo corrispondente autovettore: studio del caso R^T wstocastica per colonne non negativa usando la teoria di Perron-Frobenius e risultati/metodi basilari dell'analisi numerica degli autovalori di una matrice

Un noto risultato di Perron e Frobenius (vedi la Sezione 3.2) afferma che se $A \in \mathbb{C}^{n\times n}$ è non negativa irriducibile, allora $\rho(A)$ è autovalore positivo semplice di A e a $\rho(A)$ corrisponde un autovettore con componenti tutte positive. Prima di enunciare in maniera più completa e dimostrare la teoria di Perron-Frobenius, cominciamo a illustrarla ed usarla, studiando nei dettagli il seguente problema: determinare un vettore $\mathbf{p} \neq \mathbf{0}$ con elementi non negativi tale che $R^T\mathbf{p} = \mathbf{p}$ essendo R una matrice wstocastica per righe non negativa. Studiando questo problema particolare avremo bisogno di enunciare e dimostrare risultati importanti riguardanti l'analisi numerica degli autovalori di una matrice generica: i metodi delle potenze e delle potenze inverse, e le tecniche di deflazione.

Sia R una matrice $n \times n$ non negativa wstocastica per righe. Allora $R\mathbf{e} = \mathbf{e}$, $\mathbf{e} = [1\,1\,1\,\cdots\,1]^T$, quindi 1 è autovalore di R e un corrispondente autovettore è $\mathbf{e}$. Poiché per ogni λ autovalore di R esiste almeno un i per cui $|\lambda| = |\lambda - r_{ii} + r_{ii}| \leq |\lambda - r_{ii}| + |r_{ii}| \leq \sum_{j\neq i} |r_{ij}| + |r_{ii}| = \sum_j |r_{ij}| = \sum_j r_{ij} = 1$, possiamo anche dire che R non può avere autovalori di modulo maggiore di 1. La stessa affermazione segue anche dall'osservazione che i cerchi di Gershgorin di una matrice non negativa wstocastica per righe sono tutti centrati in punti dell'intervallo $[0,1] \subset \mathbb{C}$, passano tutti per il punto di $\mathbb{C}$ 1 e, quindi, sono tutti contenuti nel cerchio $\{z : |z| \leq 1\} \subset \mathbb{C}$.

In generale 1, come autovalore di R, può avere molteplicità algebrica e geometrica maggiore di 1 (si prenda ad esempio $R = I$). Però, non può mai verificarsi che $m_a(1) > m_g(1)$.

Proposizione 3.4 Se $R \in \mathbb{C}^{n\times n}$ è non negativa wstocastica per righe, allora $1 = \rho(R)$ è autovalore di R con molteplicità algebrica e geometrica coincidenti, e il vettore $\mathbf{e}$ è un autovettore di R relativo all'autovalore 1.

Dimostrazione. Occorre solo dimostrare che $m_a(1) = m_g(1)$. Se R è irriducibile, si ha che $m_a(1) = m_g(1) = 1$ per il teorema di Perron-Frobenius. Sia allora R riducibile. Possiamo supporre R già in

forma ridotta (perché?), cioè

$$R = \begin{bmatrix} R_{11} & & & \\ R_{21} & R_{22} & & \\ \cdot & \cdot & \cdot & \\ R_{k1} & R_{k2} & \cdot & R_{kk} \end{bmatrix}, \ k \geq 2, \tag{25}$$

dove R_{11} è irriducibile e il generico blocco diagonale R_{ii}, $i = 2, \ldots, k$, è irriducibile oppure nullo. Sia $\mathcal{I}$ il sottoinsieme di $\{1, \ldots, k\}$, $\mathcal{I} = \{1\} \cup \{i : R_{ij} = O, j = 1, \ldots, i-1\}$. Si noti che $i \in \mathcal{I}$ implica R_{ii} irriducibile wstocastica per righe, e $i \notin \mathcal{I}$ implica $\rho(R_{ii}) < 1$.

Se $i \in \mathcal{I}$, allora per il Teorema di Perron-Frobenius, 1 è autovalore semplice di R_{ii} (infatti, R_{ii} è non negativa, wstocastica per righe, irriducibile). Se $i \notin \mathcal{I}$ ed R_{ii} è irriducibile, allora 1 non è autovalore di R_{ii} e $\rho(R_{ii}) < 1$. Infatti, R_{ii} è non negativa irriducibile, i cerchi di Gershgorin di R_{ii} sono tutti centrati in $[0, 1]$ e contenuti nell'insieme $\{z : |z| \leq 1\} \subset \mathbb{C}$, e almeno uno di tali cerchi è contenuto in $\{z : |z| < 1\}$. Quindi, un numero complesso μ di modulo 1 sta fuori da almeno uno di tali cerchi, e se sta in uno degli altri cerchi necessariamente deve stare sulla sua frontiera. In altre parole μ non può essere interno a nessun cerchio e, nello stesso tempo, non può essere nell'intersezione delle frontiere di tutti i cerchi. Ne segue che, per il Teorema 3.1, μ non può essere autovalore di R_{ii}. Se $i \notin \mathcal{I}$ ed $R_{ii} = O$, allora ovviamente 1 non è autovalore di R_{ii}. Da queste tre osservazioni segue che $m_a(1) = |\mathcal{I}|$.

Valutiamo ora $m_g(1)$. Sempre per il Teorema di Perron-Frobenius, se $i \in \mathcal{I}$, allora $\dim\{\mathbf{x} : R_{ii}\mathbf{x} = \mathbf{x}\} = 1$ e $\{\mathbf{x} : R_{ii}\mathbf{x} = \mathbf{x}\} = \{\alpha_i \mathbf{e}^i : \alpha_i \in \mathbb{C}\}$ ove $\mathbf{e}^i$ è il vettore di uni $\mathbf{e}$ con dimensione uguale a quella di R_{ii}. Studiamo la dimensione dello spazio degli autovettori $\mathbf{x} = [\mathbf{x}_1^T \, \mathbf{x}_2^T \, \cdots \, \mathbf{x}_k^T]^T$ di R relativi all'autovalore 1. L'uguaglianza $R\mathbf{x} = \mathbf{x}$ è verificata se e solo se

$$\begin{aligned} &R_{ii}\mathbf{x}_i = \mathbf{x}_i, \ i \in \mathcal{I}, \ \ (I - R_{ii})\mathbf{x}_i = R_{i1}\mathbf{x}_1 + \ldots + R_{i,i-1}\mathbf{x}_{i-1}, \ i \notin \mathcal{I} \ \& \ R_{ii} \text{ irriducibile}, \\ &\mathbf{x}_i = R_{i1}\mathbf{x}_1 + \ldots + R_{i,i-1}\mathbf{x}_{i-1}, \ i \notin \mathcal{I} \ \& \ R_{ii} = O\,, \end{aligned}$$

se e solo se $\mathbf{x}_i = \alpha_i \mathbf{e}^i$, per $i \in \mathcal{I}$, $\mathbf{x}_i = (I - R_{ii})^{-1}(R_{i1}\mathbf{x}_1 + \ldots + R_{ii-1}\mathbf{x}_{i-1})$, per $i \notin \mathcal{I}$ & R_{ii} irriducibile, e $\mathbf{x}_i = R_{i1}\mathbf{x}_1 + \ldots + R_{i,i-1}\mathbf{x}_{i-1}$, per $i \notin \mathcal{I}$ & $R_{ii} = O$. Ne segue facilmente che $m_g(1) = |\mathcal{I}|$.

Quindi $m_a(1) = |\mathcal{I}| = m_g(1)$. □

Esercizio. Osservare che per la seguente matrice

$$M = \begin{bmatrix} 1 + b & -b \\ b & 1 - b \end{bmatrix}, \ b \in \mathbb{R},$$

si ha $2 = m_a(1) > m_g(1) = 1$, se $b \neq 0$. Ma, quando $b \neq 0$, M, pur essendo wstocastica per righe, non è non negativa.

Può accadere che $1 = \rho(R)$ domini gli altri autovalori, cioè $|\lambda| < 1$ per ogni λ autovalore di R, $\lambda \neq 1$ (si prenda ad esempio $R = \mathbf{e}\mathbf{e}_1^T$), come può accadere che 1 non sia l'unico numero di modulo 1 autovalore di R. Ad esempio, la matrice

$$R = \begin{bmatrix} 0 & 1 & 0 \\ b & 0 & 1 - b \\ 0 & 1 & 0 \end{bmatrix}, \ b \in (0, 1),$$

ha -1 come autovalore. Un altro esempio è fornito dalla matrice di permutazione Π che genera l'algebra delle matrici circolanti, infatti gli autovalori di Π sono distinti e hanno tutti modulo 1.

Se consideriamo ora la matrice trasposta di R, R^T, possiamo dire che essa ha gli stessi autovalori λ di R, con le stesse molteplicità algebriche e geometriche, ma gli autovettori di R^T relativi a tali λ, pur generando sempre uno spazio di dimensione $m_g(\lambda)$, non hanno in generale nulla a che fare con gli autovettori di R. In altre parole, la conoscenza degli autovettori di R non implica in generale la conoscenza degli autovettori di R^T.

Proposizione 3.5 Se R è non negativa wstocastica per righe, allora $1 = \rho(R)$ è autovalore di R^T con molteplicità algebrica e geometrica coincidenti ed esiste $\mathbf{p}$ non negativo non nullo tale che $R^T\mathbf{p} = \mathbf{p}$. In generale per l'autovettore $\mathbf{p}$ non è nota una formula esplicita.
($\exists$ un metodo diretto per il calcolo di $\mathbf{p}$, ovvero per il calcolo di soluzioni del sistema $(I-R^T)\mathbf{x} = \mathbf{0}$?)

Dimostrazione. È sufficiente dimostrate che l'uguaglianza $R^T\mathbf{x} = \mathbf{x}$ è verificata da un vettore $\mathbf{x}$ non negativo non nullo quando R è la matrice in (25). Sia $\mathbf{x} = [\mathbf{x}_1^T\, \mathbf{x}_2^T\, \cdots\, \mathbf{x}_k^T]^T$. Allora $R^T\mathbf{x} = \mathbf{x}$, con $\mathbf{x} \neq \mathbf{0}$, $\mathbf{x} \geq \mathbf{0}$, se e solo se $\sum_{j=i}^{k} R_{ji}^T\mathbf{x}_j = \mathbf{x}_i$, $i = 1, 2, \ldots, k$, con $\mathbf{x}_i \geq \mathbf{0}$, $\forall\, i$, e $\mathbf{x}_i \neq \mathbf{0}$ per almeno un i. Sia

$$\mathbf{x}_j = \begin{cases} \mathbf{0} & j \notin \mathcal{I} \\ \mathbf{p}_j & j \in \mathcal{I} \end{cases},$$

dove $\mathbf{p}_j$ è il vettore positivo tale che $\mathbf{p}_j = R_{jj}^T\mathbf{p}_j$, ben definito quando $j \in \mathcal{I}$ per il Teorema di Perron-Frobenius (perché per $j \in \mathcal{I}$ la matrice R_{jj}^T è non negativa irriducibile wstocastica per colonne). Per tale scelta degli $\mathbf{x}_j$ si ha

$$\sum_{j=i}^{k} R_{ji}^T\mathbf{x}_j = \sum_{i \leq j \leq k,\, j \in \mathcal{I}} R_{ji}^T\mathbf{p}_j = \begin{cases} \mathbf{0} & i \notin \mathcal{I} \\ R_{ii}^T\mathbf{p}_i = \mathbf{p}_i & i \in \mathcal{I} \end{cases} = \mathbf{x}_i$$

(infatti, per $j \in \mathcal{I}$ si ha $R_{ji}^T = O\ \forall\, i < j$). Dunque il vettore $\mathbf{x} = [\mathbf{x}_1^T\, \mathbf{x}_2^T\, \cdots\, \mathbf{x}_k^T]^T$ è ben definito, non negativo, non nullo e tale che $R^T\mathbf{x} = \mathbf{x}$. Si osservi che $\mathbf{x}$ non è noto esplicitamente, perché non sono noti esplicitamente i vettori $\mathbf{p}_i = R_{ii}^T\mathbf{p}_i$, $i \in \mathcal{I}$. $\square$

A meno che non siamo nel caso banale in cui R è anche wstocastica per colonne, non è in generale semplice avere una formula esplicita per il vettore $\mathbf{p}$ definito nella Proposizione 3.5. Una approssimazione buona quanto si vuole del vettore $\mathbf{p}$ si può calcolare con il *metodo delle potenze* nell'ipotesi in cui 1 domini gli altri autovalori di R. Prima di proseguire il nostro discorso su R ricordiamo dunque due argomenti classici sugli autovalori: il metodo delle potenze e le tecniche di deflazione.

Per una matrice generica A, sotto opportune ipotesi, il metodo delle potenze genera due successioni, una di scalari e una di vettori, convergenti rispettivamente a λ tale che $|\lambda| = \rho(A)$ e ad un autovettore $\mathbf{x}$ di A relativo a λ. L'ipotesi cruciale per la convergenza è che per ogni μ autovalore di A diverso da λ si abbia $|\mu| < |\lambda| = \rho(A)$, cioè che λ sia autovalore di A dominante (allora, anche ogni suo autovettore verrà detto dominante). Un'altra ipotesi importante è che λ abbia molteplicità algebrica e geometrica coincidenti. Senza quest'ultima ipotesi il limite della successione di vettori, pur esistendo, sarebbe in generale una combinazione lineare dei vettori principali di A relativi a λ e, quindi, potrebbe non essere autovettore.

Teorema 3.6 Sia $A \in \mathbb{C}^{n\times n}$ e λ_j, $j = 1, \ldots, n$, i suoi autovalori. Supponiamo che $\lambda_1 = \ldots = \lambda_t$, $t \geq 1$, sia tale che $m_a(\lambda_1) = m_g(\lambda_1)$ e $|\lambda_1| > |\lambda_j|$, $j = t+1, \ldots, n$. Sia $X = [\mathbf{x}_1 \cdots \mathbf{x}_n] \in \mathbb{C}^{n\times n}$ invertibile tale che $X^{-1}AX = J$ essendo J la forma canonica di Jordan di A e supponiamo che la sottomatrice $t \times t$ in alto a sinistra di J sia $\lambda_1 I$. Sia $\mathbf{v} \in \mathbb{C}^n$ generico tale che nell'espressione $\mathbf{v} = \sum_{j\leq t} \alpha_j \mathbf{x}_j + \sum_{j>t} \alpha_j \mathbf{x}_j$ il vettore $\mathbf{x} = \sum_{j\leq t} \alpha_j \mathbf{x}_j$ sia non nullo. Si noti che $\mathbf{x}$ è un autovettore di A relativo a λ_1. Posto $\mathbf{v}_0 = \mathbf{v}$, si considerino le seguenti successioni, di vettori e scalari:

$$\mathbf{a}_k = A\mathbf{v}_{k-1}, \quad \varphi_k = \frac{\mathbf{u}^H \mathbf{a}_k}{\mathbf{u}^H \mathbf{v}_{k-1}}, \quad \mathbf{v}_k = \frac{1}{\|\mathbf{a}_k\|}\mathbf{a}_k, \quad k = 1, 2, 3, \ldots \ ,$$

dove si suppone $\mathbf{u}$ scelto in modo che $\mathbf{u}^H\mathbf{x} \neq 0$ e gli scalari $\mathbf{u}^H\mathbf{v}_{k-1}$ siano (almeno definitivamente) diversi da zero. Allora si ha

$$\mathbf{v}_k = \frac{1}{\|A^k\mathbf{v}\|}A^k\mathbf{v} \quad \to \quad \frac{1}{\|\mathbf{x}\|}\mathbf{x}, \qquad \varphi_k = \frac{\mathbf{u}^H A^k \mathbf{v}}{\mathbf{u}^H A^{k-1}\mathbf{v}} \quad \to \quad \lambda_1 \, .$$

La velocità di convergenza è:

$$\max_{j:\, \lambda_j \neq \lambda_1} \max_{s_{\lambda_j}} O(|p_{s_{\lambda_j}-1}(k)|\big|\frac{\lambda_j}{\lambda_1}\big|^k), \tag{26}$$

dove per ogni $\lambda_j \neq \lambda_1$, il numero s_{λ_j} indica la dimensione del generico blocco di Jordan associato a λ_j, e $p_{s_{\lambda_j}-1}(k)$ è un polinomio di grado $s_{\lambda_j} - 1$, i cui coefficienti dipendono da $\frac{1}{\lambda_j^r}$, $r = 1, \ldots, s_{\lambda_j} - 1$, e dai coefficientti α_j, $j > t$, di $\mathbf{v}$ rispetto ai vettori colonna di X (vettori principali) corrispondenti al blocco di Jordan in considerazione. Se A è diagonalizzabile, allora la velocità di convergenza è $\max_{j:\, \lambda_j \neq \lambda_1} O((\big|\frac{\lambda_j}{\lambda_1}\big|)^k)$.

Dimostrazione (caso A diagonalizzabile). Si ha $A\mathbf{x}_i = \lambda_i \mathbf{x}_i$, $i = 1, \ldots, n$, e $\mathbf{v} = \sum_{j\leq t} \alpha_j \mathbf{x}_j + \sum_{j>t} \alpha_j \mathbf{x}_j = \mathbf{x} + \sum_{j>t} \alpha_j \mathbf{x}_j$. Quindi,

$$\frac{1}{\lambda_1^k} A^k \mathbf{v} = \mathbf{x} + \sum_{j>t} \alpha_j \Big(\frac{\lambda_j}{\lambda_1}\Big)^k \mathbf{x}_j,$$

$$\frac{1}{\lambda_1^k} \mathbf{u}^H A^k \mathbf{v} = \mathbf{u}^H \mathbf{x} + \sum_{j>t} \alpha_j \Big(\frac{\lambda_j}{\lambda_1}\Big)^k \mathbf{u}^H \mathbf{x}_j$$

Ne segue che, se $\mathbf{x} \neq \mathbf{0}$ e $\mathbf{u}^H\mathbf{x} \neq 0$, allora la sucessione $(1/\lambda_1^k)A^k\mathbf{v}$ converge a $\mathbf{x}$, autovettore di A relativo al'autovalore λ_1, e la successione $(\mathbf{u}^H A^k \mathbf{v})/(\mathbf{u}^H A^{k-1}\mathbf{v})$ converge a λ_1.

Dimostrazione del caso A non diagonalizzabile: ... $\square$

Per i nostri propositi è utile ricordare anche il seguente risultato classico sulla deflazione di matrici, una tecnica per introdurre una matrice i cui autovalori sono tutti uguali agli autovalori di A eccetto uno, che, invece, è zero. Ad esempio, tale metodo può essere utile per "eliminare" l'autovalore dominante di una matrice, dopo averlo calcolato, insieme ad un suo autovettore, con il metodo delle potenze.

Teorema 3.7 Sia $A \in \mathbb{C}^{n\times n}$. Sia μ un autovalore non nullo di A e $\mathbf{x}$ un corrispondente autovettore, i.e. $A\mathbf{x} = \mu\mathbf{x}$. Siano $\mu_2, \mu_3, \ldots, \mu_n$ i rimanenti autovalori di A. Allora per ogni vettore $\mathbf{w}$ tale che $\mathbf{w}^H\mathbf{x} \neq 0$, la matrice $W = A - \frac{\mu}{\mathbf{w}^H\mathbf{x}}\mathbf{x}\mathbf{w}^H$ ha autovalori $0, \mu_2, \mu_3, \ldots, \mu_n$.

Dimostrazione. Si introduce la matrice $S = [\,\mathbf{x}\ \mathbf{x}_2 \cdots \mathbf{x}_n\,]$ con $\mathbf{x}_j$ scelti in modo che $\det S \neq 0$, e si osserva che $p_A(\lambda) = p_{S^{-1}AS}(\lambda) = (\lambda - \mu)q(\lambda)$, $p_W(\lambda) = p_{S^{-1}WS}(\lambda) = \lambda q(\lambda)$. □

Esercizio. Le seguenti due matrici

$$A = \begin{bmatrix} a & b & 1-a-b \\ 1-a-b & a & b \\ b & 1-a-b & a \end{bmatrix},\ a, b \in \mathbb{C}, \quad A = \begin{bmatrix} -12 & 4 & 0 \\ -10 & -5 & 5 \\ -12 & 0 & -6 \end{bmatrix},$$

hanno rispettivamente autovalori 1 e $-12 = -3(n+1)$. Per ognuna, calcolare gli altri due autovalori utilizzando il teorema di deflazione.

Proseguiamo ora il nostro discorso su R^T. Il fatto che R^T sia una matrice wstocastica per colonne ci permette di concludere che per ogni vettore $\mathbf{v} \in \mathbb{C}^n$ vale la seguente identità: $\mathbf{e}^T R^T \mathbf{v} = \sum_i (R^T\mathbf{v})_i = \sum_i v_i = \mathbf{e}^T\mathbf{v}$. Questa semplice osservazione ha tre importanti conseguenze:

1 Se $\mathbf{v} \in \mathbb{C}^n$ è tale che $v_i \geq 0$, $\forall\, i$, e $\|\mathbf{v}\|_1 = \mathbf{e}^T\mathbf{v} = \sum_i v_i = 1$ ($\mathbf{v}$ =wdistribuzione discreta di probabilità = wddp), allora anche $R^T\mathbf{v}$ è una wddp.

2 L'identità $R^T\mathbf{v} = \lambda\mathbf{v}$, $\mathbf{v} \in \mathbb{C}^n$, implica $\mathbf{e}^T\mathbf{v} = \sum_i v_i = 0$ se $\lambda \neq 1$ (la somma degli elementi degli autovettori di R^T relativi agli autovalori diversi da 1 deve essere nulla).

3 Sia $X = [\mathbf{x}_1 \cdots \mathbf{x}_n] \in \mathbb{C}^{n\times n}$ invertibile tale che $X^{-1}R^TX = J$ essendo J la forma canonica di Jordan di R^T, e supponiamo che la sottomatrice $t \times t$ in alto a sinistra di J sia I, dove t è la molteplicita di 1 come autovalore di R (vedi la Proposizione 3.4). Allora $\mathbf{e}^T\mathbf{x}_j = 0$ per ogni $j > t$. Di conseguenza, per ogni wddp $\mathbf{v}$, il vettore $\mathbf{x}$ nella rappresentazione $\mathbf{v} = \mathbf{x} + \mathbf{y}$, $\mathbf{x} \in \mathrm{Span}\,\{\mathbf{x}_j : j \leq t\}$, $\mathbf{y} \in \mathrm{Span}\,\{\mathbf{x}_j : j > t\}$, è tale che $\mathbf{e}^T\mathbf{x} = 1$ e quindi è non nullo. Si noti che, per la Proposizione 3.5, nello spazio $\mathrm{Span}\,\{\mathbf{x}_j : j \leq t\}$ (degli autovettori di R^T relativi all'autovalore 1) c'è almeno un vettore a componenti non negative.

Dimostrazione. Occorre provare solo una parte del punto 3). Sia λ un autovalore di R^T diverso da 1. Se $\mathbf{x}_i$, $i > t$, è tale che $R^T\mathbf{x}_i = \lambda\mathbf{x}_i$ allora, per il punto 2), deve essere $\mathbf{e}^T\mathbf{x}_i = 0$. Se invece è tale che $R^T\mathbf{x}_i = \lambda\mathbf{x}_i + \mathbf{x}_{i-1}$, allora $\mathbf{e}^T\mathbf{x}_i = \mathbf{e}^TR^T\mathbf{x}_i = \lambda\mathbf{e}^T\mathbf{x}_i + \mathbf{e}^T\mathbf{x}_{i-1}$, cioè $\mathbf{e}^T\mathbf{x}_i(1-\lambda) = \mathbf{e}^T\mathbf{x}_{i-1}$, ed esiste $k \geq 1$ per cui $\mathbf{e}^T\mathbf{x}_{i-j+1}(1-\lambda) = \mathbf{e}^T\mathbf{x}_{i-j}$, $j = 1, \ldots, k$, e $\mathbf{e}^T\mathbf{x}_{i-k}(1-\lambda) = 0$. Dunque $\mathbf{e}^T\mathbf{x}_{i-j} = 0$, $j = k, \ldots, 1, 0$. In altre parole, si hanno le identità $\mathbf{e}^T\mathbf{x}_i = 0$, per ogni $i > t$. Infine, dalla rappresentazione $\mathbf{v} = \mathbf{x} + \mathbf{y}$ di cui all'enunciato, segue che $1 = \mathbf{e}^T\mathbf{v} = \mathbf{e}^T\mathbf{x} + \mathbf{e}^T\mathbf{y} = \mathbf{e}^T\mathbf{x}$. □

Dai risultati appena osservati segue che, nell'applicare il metodo delle potenze alla matrice R^T, con lo scopo di calcolare un autovettore $\mathbf{p}$ non negativo di R^T corrispondente all'autovalore 1, conviene scegliere, nel Teorema 3.6, $\mathbf{v}_0$ uguale a una wdistribuzione discreta di probabilità, $\|\cdot\|$ uguale alla norma 1, e $\mathbf{u} = \mathbf{e}$. Otteniamo così il seguente corollario del Teorema 3.6:

Proposizione 3.8 Sia R $n \times n$ non negativa wstocastica per righe. Si consideri la successione di vettori

$$\mathbf{v}_0 = \text{wddp}, \quad \mathbf{v}_k = R^T\mathbf{v}_{k-1}, \quad k = 1, 2, \ldots\ . \tag{27}$$

I vettori $\mathbf{v}_k$ sono tutti wddp, quindi $\varphi_k = 1$, $\forall\, k$, e se $|\lambda| < 1$ per ogni λ autovalore di R^T diverso da 1, allora i $\mathbf{v}_k$ convergono a $\mathbf{x}$ tale che $R^T\mathbf{x} = \mathbf{x}$, $\|\mathbf{x}\|_1 = 1$, $\mathbf{x} \geq \mathbf{0}$, con velocità di convergenza (26) dove i λ_j sono gli autovalori di R^T e $\lambda_1 = 1$. Si noti che $\mathbf{x}$ è un vettore $\mathbf{p}$ di cui alla Proposizione 3.5.

Se R è anche irriducibile, allora, per il Teorema di Perron-Frobenius, il vettore $\mathbf{p} \geq \mathbf{0}$, $\|\mathbf{p}\|_1 = 1$, tale che $R^T\mathbf{p} = \mathbf{p}$ è unico ed ha componenti positive, e conviene scegliere $\mathbf{v}_0$ =ddp, ovvero richiedere che $(\mathbf{v}_0)_i > 0$, $\forall\, i$, in modo che tutti i vettori della successione (27) (convergente a $\mathbf{x} = \mathbf{p}$) siano ddp.

Esempio (R irriducibile).
Sia R la seguente matrice non negativa wstocastica per righe

$$R = \begin{bmatrix} 1/4 & 1/4 & 1/2 \\ 3/4 & 1/8 & 1/8 \\ 11/16 & 1/4 & 1/16 \end{bmatrix}, \tag{28}$$

e 1, μ_2 e μ_3 i suoi autovalori. Poiché $R\mathbf{e} = \mathbf{e}$, per il Teorema 3.7 la matrice

$$W = R - \frac{1}{\mathbf{w}^H\mathbf{e}}\mathbf{e}\mathbf{w}^H,\ \mathbf{w}^H\mathbf{e} \neq 0,$$

ha autovalori 0, μ_2 e μ_3. Scegliendo $\mathbf{w}^H = \mathbf{e}_1^T R$, la matrice W diventa

$$W = \begin{bmatrix} 0 & 0 & 0 \\ 1/2 & -1/8 & -3/8 \\ 7/16 & 0 & -7/16 \end{bmatrix}.$$

È quindi evidente che $\mu_2 = -1/8$, $\mu_3 = -7/16$. Notiamo che, per la Proposizione 3.8, la successione (27) con R in (28) converge a $\mathbf{p}$ tale che $\mathbf{p} \geq \mathbf{0}$, $\|\mathbf{p}\|_1 = 1$, $R^T\mathbf{p} = \mathbf{p}$, con velocità di convergenza $\|\mathbf{v}_k - \mathbf{p}\| = O((7/16)^k)$.

Esempio (R riducibile).
Sia R la seguente matrice non negativa wstocastica per righe

$$R = \frac{1}{4}B, \quad B = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 3 \\ 4a & 1 & 2 & 1-4a \\ 0 & 1 & 0 & 3 \end{bmatrix},\ a \in [0, 1/4], \tag{29}$$

e 4, μ_2, μ_3, μ_4 gli autovalori di B. Poiché $B\mathbf{e} = 4\mathbf{e}$, per il Teorema 3.7 la matrice

$$W = B - \frac{4}{\mathbf{w}^H\mathbf{e}}\mathbf{e}\mathbf{w}^H,\ \mathbf{w}^H\mathbf{e} \neq 0,$$

ha autovalori 0, μ_2, μ_3 e μ_4. Scegliendo $\mathbf{w}^H = \mathbf{e}_1^T B$, la matrice W diventa

$$W = \begin{bmatrix} 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 2 \\ 4a-1 & 0 & 1 & -4a \\ -1 & 0 & -1 & 2 \end{bmatrix}.$$

È quindi evidente che $\mu_2 = (3+\sqrt{1+16a})/2$, $\mu_3 = (3-\sqrt{1+16a})/2$, $\mu_4 = 0$. Abbiamo dunque gli autovalori della matrice R: 0, $(3-\sqrt{1+16a})/8$, $(3+\sqrt{1+16a})/8$ e 1. Per la Proposizione 3.8, la successione (27) con R in (29) converge a $\mathbf{p}$ tale che $\mathbf{p} \geq \mathbf{0}$, $\|\mathbf{p}\|_1 = 1$, $R^T\mathbf{p} = \mathbf{p}$, con velocità di convergenza $\|\mathbf{v}_k - \mathbf{p}\| = O(((3+\sqrt{1+16a})/8)^k)$. Si osserva infine che

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 3 \\ 4a & 1 & 2 & 1-4a \\ 0 & 1 & 0 & 3 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} = \frac{1}{4} \begin{bmatrix} 1 & 3 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 1 & 1-4a & 2 & 4a \\ 1 & 1 & 1 & 1 \end{bmatrix},$$

e che $\mathbf{p} = [0\ 1/4\ 0\ 3/4]^T$.

Come risulta dalla Proposizione 3.8, il metodo delle potenze per il calcolo dell'autovettore $\mathbf{p} = R^T\mathbf{p} \geq \mathbf{0}$ potrebbe essere molto lento, o addirittura, potrebbe non convergere. Per capire se il calcolo di $\mathbf{p}$ può essere reso in qualche modo più efficiente, c'è bisogno di ricordare un altro risultato classico, che si può formulare per una matrice A generica: il metodo delle potenze inverse. Il vantaggio del metodo delle potenze consiste nel suo basso costo computazionale. Infatti, ad ogni passo occorre semplicemente effettuare il prodotto $A\mathbf{v}_{k-1}$ (vedi il Teorema 3.6) e l'implementazione del metodo richiede un numero minimo di celle di memoria. Il problema del metodo delle potenze è nel fatto che la sua convergenza può essere molto lenta se l'autovalore dominante, domina di poco gli altri autovalori. Illustriamo ora il metodo delle potenze inverse, il quale, vedremo, può avere una rapidità di convergenza molto maggiore, anche se il costo per passo si alza. Occorre infatti risolvere un sistema lineare ad ogni iterazione e la matrice dei coefficienti di tale sistema, definita in termini di A, può non essere ben condizionata. Il metodo delle potenze inverse sarà quindi improponibile per certe applicazioni dove l'ordine di A è molto elevato, per la proibitiva quantità di calcoli necessari per passo e per la troppa memoria necessaria per la sua implementazione.

Teorema 3.9 Sia $A \in \mathbb{C}^{n\times n}$ e λ_j, $j = 1, \ldots, n$, i suoi autovalori. Sia λ_i^* una approssimazione dell'autovalore λ_i, i.e. $|\lambda_i - \lambda_i^*|$ è più piccolo di $|\lambda_j - \lambda_i^*|$, $\forall\, \lambda_j \neq \lambda_i$, e supponiamo $m_a(\lambda_i) = m_g(\lambda_i) = t$. Sia $X = [\mathbf{x}_1 \cdots \mathbf{x}_n] \in \mathbb{C}^{n\times n}$ invertibile tale che $X^{-1}AX = J$ essendo J la forma canonica di Jordan di A e supponiamo che la sottomatrice $t \times t$ in alto a sinistra di J sia $\lambda_i I$. Sia $\mathbf{v} \in \mathbb{C}^n$ generico tale che nell'espressione $\mathbf{v} = \sum_{j\leq t} \alpha_j \mathbf{x}_j + \sum_{j>t} \alpha_j \mathbf{x}_j$ il vettore $\mathbf{x} = \sum_{j\leq t} \alpha_j \mathbf{x}_j$ sia non nullo. Si noti che $\mathbf{x}$ è un autovettore di A relativo a λ_i. Posto $\mathbf{v}_0 = \mathbf{v}$, si considerino le seguenti successioni di vettori:

$$(\lambda_i^* I - A)\mathbf{a}_k = \mathbf{v}_{k-1}, \quad \mathbf{v}_k = \frac{1}{\|\mathbf{a}_k\|}\mathbf{a}_k, \quad k = 1, 2, \ldots.$$

Allora i vettori $\mathbf{v}_k$ convergono al vettore $\mathbf{x}/\|\mathbf{x}\|$. La velocità di convergenza è:

$$\max_{j:\,\lambda_j\neq\lambda_i} \max_{s_{\lambda_j}} O(|p_{s_{\lambda_j}-1}(k)|\,|\frac{\lambda_i - \lambda_i^*}{\lambda_j - \lambda_i^*}|^k),$$

dove per ogni $\lambda_j \neq \lambda_i$, il numero s_{λ_j} indica la dimensione del generico blocco di Jordan associato a λ_j, e $p_{s_{\lambda_j}-1}(k)$ è un polinomio di grado $s_{\lambda_j}-1$, i cui coefficienti dipendono da $\frac{1}{(\lambda_j-\lambda_i^*)^r}$, $r = 1, \ldots, s_{\lambda_j}-1$, e dai coefficienti α_j, $j > t$, di $\mathbf{v}$ rispetto ai vettori colonna di X (vettori principali) corrispondenti

al blocco di Jordan in considerazione. Se A è diagonalizzabile, allora la velocità di convergenza è $\max_{j:\lambda_j\neq\lambda_i} O(\left|\frac{\lambda_i-\lambda_i^*}{\lambda_j-\lambda_i^*}\right|^k)$. Si osserva infine che se $|\lambda_i-\lambda_i^*| << |\lambda_j-\lambda_i^*|$, $\forall\,\lambda_j\neq\lambda_i$, allora il metodo delle potenze inverse converge molto rapidamente.

Dimostrazione. Usare il Teorema 3.6 applicato alla matrice $(\lambda_i^*I-A)^{-1}$, osservando che $1/(\lambda_i^*-\lambda_i)$ è autovalore dominante di $(\lambda_i^*I-A)^{-1}$ con corrispondente autovettore $\mathbf{x}$. □

Tornando alla nostra R^T osserviamo che $1+\varepsilon$, $\varepsilon>0$, è una approssimazione dell'autovalore 1 di R^T, infatti $|(1+\varepsilon)-1|<|(1+\varepsilon)-\lambda|$ per ogni $\lambda\in\sigma(R^T)$, $\lambda\neq 1$. Quindi possiamo pensare di ottenere un autovettore $\mathbf{p}=R^T\mathbf{p}\geq\mathbf{0}$ di cui alla Proposizione 3.5 applicando il Teorema 3.9 ad R^T.

Proposizione 3.10 Sia $R\in\mathbb{R}^{n\times n}$ non negativa wstocastica per righe, ed $\varepsilon>0$. Allora la successione $\mathbf{v}_k$,

$$\mathbf{v}_0=\text{wddp},\quad ((1+\varepsilon)I-R^T)\mathbf{a}_k=\mathbf{v}_{k-1},\quad \mathbf{v}_k=\frac{1}{\|\mathbf{a}_k\|_1}\mathbf{a}_k,\quad k=1,2,\ldots, \tag{30}$$

è ben definita, è formata da vettori wddp ($((1+\varepsilon)I-R^T)^{-1}\geq O!$) e converge a $\mathbf{p}=R^T\mathbf{p}\geq\mathbf{0}$, $\|\mathbf{p}\|_1=1$. Scegliendo ε abbastanza piccolo si ha $|(1+\varepsilon)-1|<<|(1+\varepsilon)-\lambda|$ per ogni $\lambda\in\sigma(R^T)$, $\lambda\neq 1$ (cioè $1+\varepsilon$ è una ottima approssimazione di 1), quindi dopo pochi passi $\mathbf{v}_k$ approssima molto bene $\mathbf{p}$. Si noti tuttavia che scegliere ε troppo piccolo rende la matrice $(1+\varepsilon)I-R^T$ quasi singolare.

Se R è anche irriducibile, allora, per il Teorema di Perron-Frobenius, il vettore $\mathbf{p}\geq\mathbf{0}$, $\|\mathbf{p}\|_1=1$, tale che $R^T\mathbf{p}=\mathbf{p}$ è unico ed ha componenti positive, e conviene scegliere $\mathbf{v}_0=$ddp, ovvero richiedere che $(\mathbf{v}_0)_i>0$, $\forall\,i$, in modo che tutti i vettori della successione (30) (convergente a $\mathbf{p}$) siano ddp ($((1+\varepsilon)I-R^T)^{-1}$ è non negativa e irriducibile!).

Esercizio. Sia $M\in\mathbb{R}^{n\times n}$ invertibile tale che $M_{ii}>0$, $\forall\,i$, e $M_{ij}\leq 0$, $\forall\,i\neq j$. Allora $M^{-1}\geq O$.

Nei due esercizi che seguono si discutono due possibili procedimenti per la risoluzione del sistema lineare $((1+\varepsilon)I-R^T)\mathbf{a}_k=\mathbf{v}_{k-1}$, operazione da fare ad ogni passo del metodo delle potenze inverse (30) per costruire $\mathbf{v}_k$, il nuovo vettore approssimante $\mathbf{p}$, a partire dal vecchio $\mathbf{v}_{k-1}$.

Esercizio (il metodo diretto LU oppure QU). Si introducono due matrici L, $L_{ii}=1$, ed U, rispettivamente triangolare inferiore e triangolare superiore, e una matrice di permutazione P tali che $(1+\varepsilon)I-R^T=PLU$, e ad ogni passo si risolvono i due sistemi triangolari $L\mathbf{z}=P^T\mathbf{v}_{k-1}$, $U\mathbf{a}_k=\mathbf{z}$, oppure si introducono due matrici Q unitaria ed U triangolare superiore tali che $(1+\varepsilon)I-R^T=QU$, e ad ogni passo si risolvono i due sistemi $Q\mathbf{z}=\mathbf{v}_{k-1}$ ($\mathbf{z}=Q^H\mathbf{v}_{k-1}$), $U\mathbf{a}_k=\mathbf{z}$. In generale non è consigliabile usare il primo metodo, infatti la matrice dei coefficienti $(1+\varepsilon)I-R^T$ potrebbe essere mal condizionata. Entrambi i metodi sono sconsigliabili quando la matrice R è molto grande e sparsa, perchè la loro implementazione richiederebbe, oltre che molte operazioni ($O(n^3)$), troppe celle di memoria ($O(n^2)$).

Provare ad applicare i due metodi quando R è la matrice di cui agli esempi (28), (29).

Esercizio (il metodo iterativo di Richardson-Eulero). Sia $A\in\mathbb{C}^{n\times n}$ non singolare. Dato $\mathbf{f}\in\mathbb{C}^n$, un metodo per trovare $\mathbf{x}\in\mathbb{C}^n$ tale che $A\mathbf{x}=\mathbf{f}$ è quello di Richardson-Eulero [40]. Sia $\mathbf{x}_0\in\mathbb{C}^n$, e $\{\mathbf{x}_k\}$ la successione generata a partire da $\mathbf{x}_0$ con la regola $\mathbf{x}_{k+1}=\mathbf{x}_k+\omega(\mathbf{f}-A\mathbf{x}_k)$, $k=0,1,2,\ldots$ (i valori ammessi del parametro ω, per semplicità, sono solo quelli reali). Allora $\mathbf{x}-\mathbf{x}_k=(I-\omega A)^k(\mathbf{x}-\mathbf{x}_0)\ \to\ 0$, $\forall\,\mathbf{x}_0$, se e solo se $\rho(I-\omega A)<1$, se e solo se $|1-\omega\lambda|^2=(1-\omega\Re(\lambda))^2+\omega^2(\Im(\lambda))^2<1$

per ogni autovalore λ di A. Ne segue che esistono valori reali di ω per cui il metodo converge se e solo se $\Re(\lambda) > 0$ (< 0) per ogni $\lambda \in \sigma(A)$, e i valori di ω per cui si ha la convergenza sono $(0, \omega^*)$ $((-\omega^*, 0))$ per un $\omega^* > 0$.

Una matrice A può non essere definita positiva (negativa), ma avere parte hermitiana definita positiva (negativa). In tal caso gli autovalori di A, pur non essendo in generale reali e positivi (negativi), devono avere, per quanto osservato alla fine del Capitolo 2, parte reale positiva (negativa). Quindi si può usare il metodo di Richardson-Eulero per risolvere un sistema lineare con matrice dei coefficienti con parte hermitiana definita positiva (negativa).

Inoltre, si può usare Richardson-Eulero per risolvere il nostro sistema $((1+\varepsilon)I - R^T)\mathbf{a}_k = \mathbf{v}_{k-1}$, infatti gli autovalori della matrice reale $(1+\varepsilon)I - R^T$ sono nel cerchio di centro $1+\varepsilon$ e raggio 1, quindi hanno tutti parte reale positiva. Dunque, se $\omega > 0$ è abbastanza piccolo, la successione $\mathbf{a}_k^0 =?$, $\mathbf{a}_k^{j+1} = \mathbf{a}_k^j + \omega(\mathbf{v}_{k-1} - ((1+\varepsilon)I - R^T)\mathbf{a}_k^j) = \big((1-\omega(1+\varepsilon))I + \omega R^T\big)\mathbf{a}_k^j + \omega\mathbf{v}_{k-1}$, $j = 0, 1, 2, \ldots$, converge ad $\mathbf{a}_k$. D'altro canto, nel caso particolare in cui 1 domina gli autovalori di R, cioè $1 > |\lambda|$ per ogni λ autovalore di R diverso da 1, il raggio spettrale della matrice $-\varepsilon I + R^T$ è minore di 1 se ε è abbastanza piccolo. In altre parole, se 1 domina gli autovalori di R, è consentita la scelta $\omega = 1$ e l'equazione del metodo di Richardson-Eulero diventa semplicemente $\mathbf{a}_k^{j+1} = (-\varepsilon I + R^T)\mathbf{a}_k^j + \mathbf{v}_{k-1}$. Notiamo che, per ottenere una buona approssimazione del vettore $\mathbf{a}_k$ con il metodo di Richardson-Eulero occorre effettuare molte iterazioni dello stesso, essendo in generale il raggio spettrale della matrice di iterazione $(1 - \omega(1+\varepsilon))I + \omega R^T$ poco minore di 1. Tuttavia con Richardson-Eulero è possibile trattare i casi in cui R è molto grande e sparsa, visto che la sua implementazione richiede un numero minimo di celle di memoria.

Provare ad applicare il metodo di Richardson-Eulero quando R è la matrice di cui agli esempi (28), (29).

3.2 La teoria di Perron-Frobenius

Sia A $n \times n$ non negativa ($A \geq O$) e irriducibile. Ad ogni $\mathbf{x} \neq \mathbf{0}$, $\mathbf{x} \geq \mathbf{0}$ associamo il numero reale $r_{\mathbf{x}} = \min_{i:\, x_i>0}(A\mathbf{x})_i/x_i$. È semplice verificare che $r_{\mathbf{x}} \geq 0$, $A\mathbf{x} \geq r_{\mathbf{x}}\mathbf{x}$, $r_{\mathbf{x}} = \sup\{\rho \in \mathbb{R} :\, A\mathbf{x} \geq \rho\mathbf{x}\}$, e $r_{\alpha\mathbf{x}} = r_{\mathbf{x}}$ se $\alpha > 0$. Ad esempio,

$$\begin{aligned}
\sup\{\rho \in \mathbb{R} :\, A\mathbf{x} \geq \rho\mathbf{x}\} &= \sup\{\rho \in \mathbb{R} :\, (A\mathbf{x})_i \geq \rho x_i,\, \forall\, i\} \\
&= \sup\{\rho \in \mathbb{R} :\, (A\mathbf{x})_i \geq \rho x_i,\, \forall\, i \,|\, x_i > 0\} \\
&= \sup\{\rho \in \mathbb{R} :\, (A\mathbf{x})_i/x_i \geq \rho,\, \forall\, i \,|\, x_i > 0\} \\
&= \min_{i:\, x_i>0} (A\mathbf{x})_i/x_i = r_{\mathbf{x}}\,.
\end{aligned}$$

Esercizio. Dimostrare che $r_{\mathbf{x}}$ è una funzione continua di $\mathbf{x}$.

Ad A associamo il numero $r = \sup_{\mathbf{x}\neq\mathbf{0},\, \mathbf{x}\geq\mathbf{0}} r_{\mathbf{x}}$.

Proposizione 3.11 Il numero $r = \sup_{\mathbf{x}\neq\mathbf{0},\, \mathbf{x}\geq\mathbf{0}} \min_{i:\, x_i>0}(A\mathbf{x})_i/x_i > 0$ è positivo.
(Nota: l'ipotesi A irriducibile non è necessaria; il risultato vale nell'ipotesi più debole che A non abbia righe nulle o che A abbia sulla diagonale almeno un elemento A_{kk} positivo ($\mathbf{x} = \mathbf{e}_k \Rightarrow r_{\mathbf{x}} = \min_{i:\, x_i>0}(A\mathbf{x})_i/x_i = (A\mathbf{x})_k/x_k = A_{kk} > 0$))

Dimostrazione. È sufficiente mostrare che $r_{\mathbf{x}} > 0$ per qualche $\mathbf{x} \geq \mathbf{0}$, $\mathbf{x} \neq \mathbf{0}$. Sia $\mathbf{e} = [\,1\ 1\ \cdots\ 1\,]^T$, e dunque $r_{\mathbf{e}} = \min_i \sum_j a_{ij}$. Se fosse $r_{\mathbf{e}} = 0$, la matrice A dovrebbe avere una riga nulla. Poiché A è irriducibile ciò non è possibile (una matrice con una riga o una colonna nulle è ovviamente riducibile). Ne segue che $0 < r_{\mathbf{e}} \leq r$. □

Sia p un polinomio positivo su $[0, +\infty)$ tale $p(A) > O$, ovvero tale che $p(A)\mathbf{x} > \mathbf{0}$, $\forall\, \mathbf{x} \geq \mathbf{0}$ $\mathbf{x} \neq \mathbf{0}$ (ad esempio $p(t) = (1+t)^{n-1}$, vedi il Teorema 3.3). Si noti che affinché tale p possa esistere è necessario che A sia irriducibile.

Proposizione 3.12 Se $\mathbf{w} \neq \mathbf{0}$, $\mathbf{w} \geq \mathbf{0}$ è tale che $A\mathbf{w} \geq r\mathbf{w}$, allora $A\mathbf{w} = r\mathbf{w}$ e $\mathbf{w} > \mathbf{0}$.

Dimostrazione. Sia $\eta = A\mathbf{w} - r\mathbf{w}$. Sappiamo che $\eta \geq \mathbf{0}$. Supponiamo che $\eta \neq \mathbf{0}$. Allora, per l'irriducibilità di A, si avrebbe $p(A)\eta > \mathbf{0}$, ovvero $A\mathbf{y} > r\mathbf{y}$, con $\mathbf{y} = p(A)\mathbf{w} > \mathbf{0}$. Quindi, $(A\mathbf{y})_i/y_i > r$, $\forall\, i$, cioè $r_{\mathbf{y}} = \min_i (A\mathbf{y})_i/y_i > r$, ma questo è un assurdo, vista la definizione di r. Dunque deve valere l'identità $A\mathbf{w} = r\mathbf{w}$. Inoltre, da questa segue l'uguaglianza $\mathbf{y} = p(A)\mathbf{w} = p(r)\mathbf{w}$, che, essendo $\mathbf{y} > \mathbf{0}$, implica la positività del vettore $\mathbf{w}$. □

Proposizione 3.13 Esiste $\mathbf{z} \geq \mathbf{0}$, $\mathbf{z} \neq \mathbf{0}$, tale che $r = r_{\mathbf{z}}$ e, quindi, tale che $A\mathbf{z} \geq r_{\mathbf{z}}\mathbf{z} = r\mathbf{z}$.

Dimostrazione.

$$r = \sup_{\mathbf{x}\neq\mathbf{0},\, \mathbf{x}\geq\mathbf{0}} r_{\mathbf{x}} = \sup_{\mathbf{x}\neq\mathbf{0},\, \mathbf{x}\geq\mathbf{0}} r_{\mathbf{x}/\|\mathbf{x}\|} = \sup_{\|\mathbf{y}\|=1,\, \mathbf{y}\geq\mathbf{0}} r_{\mathbf{y}}.$$

Si osserva che $r_{\mathbf{y}}$ è tale che $A\mathbf{y} \geq r_{\mathbf{y}}\mathbf{y}$, $Ap(A)\mathbf{y} \geq r_{\mathbf{y}}p(A)\mathbf{y}$ e, quindi, per definizione di $r_{p(A)\mathbf{y}}$, tale che $r_{\mathbf{y}} \leq r_{p(A)\mathbf{y}}$ (Nota: $\mathbf{y} \geq \mathbf{0}$, $\mathbf{y} \neq \mathbf{0}$, mentre $p(A)\mathbf{y} > \mathbf{0}$). Ne segue che, posto $Q = \{p(A)\mathbf{y} :\, \mathbf{y} \in P\}$, $P = \{\mathbf{y} :\, \|\mathbf{y}\| = 1,\, \mathbf{y} \geq \mathbf{0}\}$, si ha

$$r \leq \sup_{\|\mathbf{y}\|=1,\, \mathbf{y}\geq\mathbf{0}} r_{p(A)\mathbf{y}} = \sup_{\tilde{\mathbf{z}}\in Q} r_{\tilde{\mathbf{z}}} \leq r.$$

Poiché Q è un compatto, si può allora dire che

$$\exists\, \mathbf{z} \in Q \mid r \leq \max_{\tilde{\mathbf{z}} \in Q} r_{\tilde{\mathbf{z}}} = r_{\mathbf{z}} \leq r,$$

da cui la tesi, visto che i vettori di Q sono tutti positivi. □

Proposizione 3.14 *r è un autovalore di A ed $r = \rho(A)$.*

Dimostrazione. Sia $\lambda \in \mathbb{C}$ un autovalore di A, quindi $A\mathbf{x} = \lambda\mathbf{x}$ per un vettore $\mathbf{x} \neq \mathbf{0}$. Per $\mathbf{v} \in \mathbb{C}^n$ sia $|\mathbf{v}|$ il vettore le cui componenti sono i moduli delle componenti di $\mathbf{v}$. Allora $|\lambda||\mathbf{x}| = |\lambda\mathbf{x}| = |A\mathbf{x}| \leq |A||\mathbf{x}| = A|\mathbf{x}|$. Poiché $|\mathbf{x}| \geq \mathbf{0}$ e $|\mathbf{x}| \neq \mathbf{0}$, la disuguaglianza $|\lambda||\mathbf{x}| \leq A|\mathbf{x}|$ implica $|\lambda| \leq r_{|\mathbf{x}|} \leq r$. Quindi $\rho(A) \leq r$. Ma r, per le Proposizioni 3.12 e 3.13, è un autovalore di A, quindi $\rho(A) = r$. □

Enunciamo ora il risultato principale della teoria di Perron-Frobenius [59], [60], [39], come corollario delle Proposizioni precedenti 3.11, 3.12, 3.13, 3.14.

Teorema 3.15 *Esiste $\mathbf{z} > \mathbf{0}$ tale che $A\mathbf{z} = r\mathbf{z}$ con $r = \sup_{\mathbf{x}\neq\mathbf{0}, \mathbf{x}\geq\mathbf{0}} \min_{i:\, x_i>0} (A\mathbf{x})_i/x_i = \rho(A)$ positivo. Inoltre, r è radice semplice del polinomio caratteristico di A.*

Dimostrazione. È da dimostrare solo il fatto che r, come autovalore di A, è semplice. Vedi [39]. □

Esercizio. Scrivere una matrice A almeno 2×2 che non abbia il raggio spettrale come autovalore.

Il Teorema 3.15 giustifica la seguente

DEFINIZIONE 3.16 *La coppia $(r, \mathbf{z})$ con $r = \rho(A)$, $\mathbf{z} > \mathbf{0}$ tale che $A\mathbf{z} = r\mathbf{z}$, $\|\mathbf{z}\|_1 = \sum_i z_i = 1$, è detta coppia di Perron. Gli altri $n-1$ autovalori di A, $\lambda_2, \ldots, \lambda_n$, sono diversi da r.*

Calcolo della coppia di Perron

Per calcolare la coppia di Perron $(\rho(A), \mathbf{z})$ si può usare il metodo delle potenze (vedi il Teorema 3.6) che, riscritto nel nostro caso in cui A è non negativa irriducibile, diventa:

Teorema 3.17 (metodo delle potenze per matrici A non negative irriducibili). Siano

$$\mathbf{v}_0 \in \mathbb{R}^n,\ \mathbf{v}_0 > \mathbf{0},\ \|\mathbf{v}_0\|_1 = 1,\ \mathbf{a}_k = A\mathbf{v}_{k-1},\ \mathbf{v}_k = \mathbf{a}_k/\|\mathbf{a}_k\|_1,\ k = 1, 2, \ldots\ .$$

Si noti che i vettori $\mathbf{a}_k$ e $\mathbf{v}_k$ sono ben definiti e positivi per ogni k (dimostrarlo!). Supponiamo che $\lambda_1 = r = \rho(A)$ domina gli autovalori di A, cioè $\max_{j\neq 1} |\lambda_j| < \rho(A)$. Sia $X = [\mathbf{z}\ \mathbf{x}_2\ \cdot\ \mathbf{x}_n]$ tale che $X^{-1}AX = J$ dove J è la forma di Jordan di A. Se $\mathbf{v}_0$ è tale che nell'espressione $\mathbf{v}_0 = \alpha\mathbf{z} + \sum_{j=2}^n \alpha_j \mathbf{x}_j$ il coefficiente di $\mathbf{z}$, α, è diverso da zero, allora, per $k \to +\infty$,

$$\mathbf{v}_k\ \to\ \mathbf{z},\quad \|\mathbf{a}_k\|_1 = \mathbf{e}^T\mathbf{a}_k\ \to\ \rho(A).$$

L'ordine di convergenza è $\max_{j\neq 1} \max_{s_{\lambda_j}} O(|p_{s_{\lambda_j}-1}(k)||\lambda_j/\lambda_1|^k)$, dove $p_{s_{\lambda_j}-1}$ è un polinomio di grado $s_{\lambda_j} - 1$ con s_{λ_j} = dimensione del generico blocco di Jordan relativo all'autovalore λ_j. Se A è diagonalizzabile, allora l'ordine di convergenza è $\max_{j\neq 1} O(|\lambda_j/\lambda_1|^k)$.

Quindi, il metodo delle potenze ci permette di calcolare $(\rho(A), \mathbf{z})$ quando $r = \rho(A)$ domina gli altri $n-1$ autovalori di A, ovvero se è soddisfatta la seguente condizione

$$\max_{j=2,\ldots,n} |\lambda_j| < \rho(A). \tag{31}$$

Viceversa, se tale condizione non è soddisfatta il metodo delle potenze in generale non può convergere, e quindi è inutilizzabile nel calcolo della coppia di Perron.

Esempi. Le seguenti matrici

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} 0 & a & 0 \\ 1 & 0 & 1 \\ 0 & 1-a & 0 \end{bmatrix}, \ a \in (0,1),$$

sono irriducibili e non negative, quindi $\rho(A) = 1$ è autovalore semplice per tali A. Tutte e tre hanno hanno però come autovalori numeri $\lambda \in \mathbb{C}$ tali che $|\lambda| = \rho(A)$, $\lambda \neq \rho(A)$. Quindi, ad esempio per la terza A, non possiamo usare il metodo delle potenze per calcolare il vettore ben definito, non banale $\mathbf{z} = A\mathbf{z} > \mathbf{0}$, $\|\mathbf{z}\|_1 = 1$. Anche le matrici

$$A = \begin{bmatrix} 0 & 0 & 1 \\ a & 0 & 0 \\ 1-a & 1 & 0 \end{bmatrix}, \ a \in (0,1), \quad A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}, \quad A = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix},$$

sono non negative irriducibili, ma oltre $\rho(A)$, autovalore semplice, non hanno altri autovalori di modulo uguale a $\rho(A)$. Dunque, per la prima e per la terza A, che hanno vettori di Perron non banali, si può usare con successo il metodo delle potenze.

È noto che, per una matrice A non negativa irriducibile, la condizione (31) è soddisfatta se e solo se esiste k intero positivo tale che $A^k > O$ [39]. Quindi, in particolare, ogni matrice A $n \times n$ i cui elementi sono tutti positivi ha $\rho(A)$ come autovalore dominante. Dimostriamo nei dettagli solo quest'ultimo risultato, di per se importante essendo stato il primo ad essere ottenuto in letteratura nell'ambito della teoria di Perron-Frobenius (vedi [59]). Notiamo che il carattere "semplice" di $\rho(A)$ come autovalore di A segue immediatamente da (31) (per $A > O$).

Premettiamo due risultati ulteriori della teoria di Perron-Frobenius per le matrici A non negative irriducibili.

Proposizione 3.18 Sia $B \in \mathbb{C}^{n\times n}$ tale che $|B| \leq A$. Allora $\rho(B) \leq \rho(A)$.

Dimostrazione. Sia $\beta \in \mathbb{C}$ un autovalore di B, quindi $B\mathbf{x} = \beta\mathbf{x}$ per un vettore $\mathbf{x} \neq \mathbf{0}$. Allora $|\beta||\mathbf{x}| = |\beta\mathbf{x}| = |B\mathbf{x}| \leq |B||\mathbf{x}| \leq A|\mathbf{x}|$. Poiché $|\mathbf{x}| \geq \mathbf{0}$ e $|\mathbf{x}| \neq \mathbf{0}$, la disuguaglianza $|\beta||\mathbf{x}| \leq A|\mathbf{x}|$ implica $|\beta| \leq r_{|\mathbf{x}|} \leq r = \rho(A)$, quindi $\rho(B) \leq \rho(A)$. □

Proposizione 3.19 Sia $B \in \mathbb{C}^{n\times n}$ tale che $|B| \leq A$, $|B| \neq A$. Allora $\rho(B) < \rho(A)$.

Dimostrazione. Per la Proposizione 3.18 deve essere $\rho(B) \leq \rho(A)$. Supponiamo $\rho(B) = \rho(A)$. Sia $\beta \in \mathbb{C}$ un autovalore di B di modulo uguale ad r. Quindi si ha $B\mathbf{x} = \beta\mathbf{x}$ per un β tale che $|\beta| = r = \rho(A)$ e per un vettore $\mathbf{x} \neq \mathbf{0}$. Allora $r|\mathbf{x}| = |\beta||\mathbf{x}| = |\beta\mathbf{x}| = |B\mathbf{x}| \leq |B||\mathbf{x}| \leq A|\mathbf{x}|$, con $|\mathbf{x}| \geq \mathbf{0}$ e $|\mathbf{x}| \neq \mathbf{0}$. Ma per la Proposizione 3.12 la disuguaglianza $r|\mathbf{x}| \leq A|\mathbf{x}|$ implica $|\mathbf{x}| > \mathbf{0}$ e $r|\mathbf{x}| = A|\mathbf{x}|$. Ne segue che $r|\mathbf{x}| = |\beta||\mathbf{x}| = |\beta\mathbf{x}| = |B\mathbf{x}| \leq |B||\mathbf{x}| \leq A|\mathbf{x}| = r|\mathbf{x}|$, e quindi che $|B||\mathbf{x}| = A|\mathbf{x}|$ con $|\mathbf{x}| > \mathbf{0}$ e $|B| \leq A$. Ciò può essere possibile solo se $|B| = A$. □

Sia ora $W = A - s\mathbf{z}\mathbf{e}^T$ dove $\mathbf{z}$ è l'autovettore di Perron di A (sempre supposta non negativa

irriducibile) e s è un parametro reale. Sia T una matrice invertibile la cui prima colonna è $\mathbf{z}$. Allora

$$\begin{aligned} T^{-1}AT &= T^{-1}\left[\begin{array}{cccc} \rho(A)\mathbf{z} & * & \cdot & * \end{array}\right] = \left[\begin{array}{cc} \rho(A) & * \\ \mathbf{0} & B \end{array}\right], \\ T^{-1}WT &= T^{-1}AT - sT^{-1}\mathbf{z}\mathbf{e}^TT = \left[\begin{array}{cc} \rho(A) & * \\ \mathbf{0} & B \end{array}\right] - s\mathbf{e}_1\left[\begin{array}{cccc} 1 & * & \cdot & * \end{array}\right], \end{aligned}$$

quindi $p_A(\lambda) = (\lambda - r)p_B(\lambda)$, e $p_W(\lambda) = (\lambda - (r-s))p_B(\lambda)$. Sia ora $s = \min_{i,j} a_{ij}$. Se $A > O$ allora $|W| = W < A$, quindi, per la Proposizione 3.19, si ha $\max\{|\lambda| : \lambda \text{ autovalori di } B\} \leq \rho(W) < \rho(A)$. Poiché gli autovalori di B non sono altro che gli autovalori $\lambda_2, \ldots, \lambda_n$ di A, si ha la disuguaglianza in (31). Abbiamo quindi ottenuto il seguente

Teorema 3.20 Se A è positiva allora gli autovalori di A diversi da $r = \rho(A)$, autovalore semplice di A (con autovettore positivo), hanno modulo minore di $r = \rho(A)$.

Teorema 3.21 Se A è non negativa e irriducibile, e almeno un elemento diagonale di A è positivo, allora gli autovalori di A diversi da $r = \rho(A)$, autovalore semplice di A (con autovettore positivo), hanno modulo minore di $r = \rho(A)$.

Dimostrazione. Vedi gli esercizi in fondo alla sezione. □

Citare, enunciare solamente il teorema in base al grafo, tesi Cardinali...ALTRA CONDIZIONE CHE, AGGIUNTA A NON NEG E IRRIDUC, ASSICURI (31)

Il caso in cui A è anche wstocastica per colonne

Se la matrice A, oltre che non negativa e irriducibile, è anche wstocastica per colonne, cioè $A^T\mathbf{e} = \mathbf{e}$, allora 1 è autovalore di A, e $\rho(A) \leq \|A\|_1 = 1$. Quindi, $1 = \rho(A)$ è l'autovalore semplice di Perron. L'autovettore $\mathbf{z}$ di Perron, cioè il vettore $\mathbf{z} > \mathbf{0}$, tale che $\mathbf{z} = A\mathbf{z}$, $\|\mathbf{z}\|_1 = 1$, non è noto in generale (lo è, ovviamente, se A è anche wstocastica per righe). Lo si può allora calcolare utilizzando il metodo delle potenze di cui al Teorema 3.17. Come vedremo subito, tale metodo quando A è wstocastica per colonne si semplifica e converge nella sola ipotesi che $1 = \rho(A)$ domini gli autovalori di A e già sappiamo che quest'ultima ipotesi è ad esempio soddisfatta se A ha almeno un elemento diagonale positivo.

Lemma 3.22 Sia $A \in \mathbb{C}^{n\times n}$ tale che $\sum_{i=1}^n a_{ij} = 1$ per ogni $j = 1, \ldots, n$. Per ogni $\mathbf{v} \in \mathbb{C}^n$ vale l'identità $\mathbf{e}^TA\mathbf{v} = \sum_i (A\mathbf{v})_i = \sum_i v_i = \mathbf{e}^T\mathbf{v}$, e quindi

(i) $\sum_i v_i = 1 \Rightarrow \sum_i (A\mathbf{v})_i = 1$; se in particolare A è non negativa e irriducibile, e $\mathbf{v}$ è una distribuzione discreta di probabilità (ddp), cioè $v_i > 0\ \forall\, i$, $\sum_i v_i = 1$, allora anche $A\mathbf{v}$ è una ddp.

(ii) $A\mathbf{v} = \lambda\mathbf{v}$, $\mathbf{v} \in \mathbb{C}^n$, $\mathbf{v} \neq \mathbf{0}$, $\lambda \neq 1 \Rightarrow \sum_i v_i = 0$.

(iii) $A \geq O$, irriducibile, $X^{-1}AX = J$, J forma di Jordan di A, $J\mathbf{e}_1 = r\mathbf{e}_1$, $X\mathbf{e}_1 = \mathbf{z} \Rightarrow \sum_i x_{ij} = 0$ per ogni $j \geq 2$.

Dimostrazione. È lasciata al lettore. Notare che sono stati ottenuti risultati molto simili nella Sezione dove si studia il problema del calcolo del vettore $\mathbf{p} = R^T\mathbf{p} \geq \mathbf{0}$, essendo $R \geq O$, wstocastica per righe. □

Teorema 3.23 (metodo delle potenze per matrici A non negative irriducibili wstocastiche per colonne) Siano

$$\mathbf{v}_0 \in \mathbb{R}^n,\ \mathbf{v}_0 > \mathbf{0},\ \|\mathbf{v}_0\|_1 = 1,\ \mathbf{v}_k = A\mathbf{v}_{k-1},\ k = 1, 2, \dots\ .$$

Si noti che i vettori $\mathbf{v}_k$ sono ben definiti, positivi e ddp per ogni k. Supponiamo che $1 = r = \rho(A)$ domina gli autovalori di A, cioè $\max_{j\neq 1} |\lambda_j| < 1 = r = \rho(A)$. Sia $X = [\mathbf{z}\ \mathbf{x}_2\ \cdot\ \mathbf{x}_n]$ tale che $X^{-1}AX = J$ dove J è la forma di Jordan di A. Allora, nell'espressione $\mathbf{v}_0 = \alpha\mathbf{z} + \sum_{j=2}^{n} \alpha_j \mathbf{x}_j$, il coefficiente di $\mathbf{z}$, α, è diverso da zero (è uguale a 1). Quindi, per $k \to +\infty$,

$$\mathbf{v}_k\ \to\ \mathbf{z},\ \mathbf{z} > \mathbf{0},\ \|\mathbf{z}\|_1 = 1,\ A\mathbf{z} = \mathbf{z},\ \text{ e }\ \|\mathbf{a}_k\|_1 = 1 \to\ 1 = \rho(A).$$

L'ordine di convergenza è $\max_{j\neq 1} \max_{s_{\lambda_j}} O(|p_{s_{\lambda_j}-1}(k)||\lambda_j|^k)$, dove $p_{s_{\lambda_j}-1}$ è un polinomio di grado $s_{\lambda_j} - 1$ con s_{λ_j} = dimensione del generico blocco di Jordan relativo all'autovalore λ_j. Se A è diagonalizzabile, allora l'ordine di convergenza è $\max_{j\neq 1} O(|\lambda_j|^k)$.

Dimostrazione. Nel caso in cui $A^T\mathbf{e} = \mathbf{e}$, nel metodo delle potenze per matrici A non negative irriducibili (vedi il Teorema 3.17), per il vettore $\mathbf{a}_k = A\mathbf{v}_{k-1}$ si ha $\|\mathbf{a}_k\|_1 = \|\mathbf{v}_{k-1}\|_1 = 1$, quindi $\mathbf{v}_k = \mathbf{a}_k/\|\mathbf{a}_k\|_1 = \mathbf{a}_k$. Da qui la sparizione dal metodo della succesione $\mathbf{a}_k$. Inoltre, l'espressione $\mathbf{v}_0 = \alpha\mathbf{z} + \sum \alpha_j \mathbf{x}_j$ implica $1 = \mathbf{e}^T\mathbf{v}_0 = \alpha\mathbf{e}^T\mathbf{z} + \sum \alpha_j \mathbf{e}^T\mathbf{x}_j = \alpha$. Il fatto che $\mathbf{e}^T\mathbf{x}_j = 0$, $j = 2, \dots, n$, segue nel caso A diagonalizzabile dal punto (ii) del Lemma e, nel caso generico, dal punto (iii) del Lemma. □

Esercizio. Sia $A \geq O$ irriducibile. Dimostrare che esiste una matrice diagonale D, con $D_{ii} > 0$, tale che $\sum_i (D^{-1}AD)_{ij} = \rho(A)$, $\forall\, j$ ($\sum_j (D^{-1}AD)_{ij} = \rho(A)$, $\forall\, i$).

Esercizio (dimostrazione del Teorema 3.21). Se A è non negativa e irriducibile ed ha almeno un elemento diagonale positivo, allora $n - 1$ autovalori di A hanno modulo strettamente minore dell'autovalore $\rho(A)$. Si noti che questo risultato è molto piu forte del risultato di cui al Teorema 3.20. (Suggerimento: usare l'affermazione dell'esercizio precedente).

Esercizio. Sia $A \geq O$ irriducibile. Dimostrare che esiste una matrice $D + \mathbf{u}\mathbf{v}^T$, con D diagonale, $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, non singolare tale che per $M = (D + \mathbf{u}\mathbf{v}^T)^{-1} A (D + \mathbf{u}\mathbf{v}^T)$ si ha $\sum_i (M)_{ij} = \sum_j (M)_{ij} = \rho(A)$, $\forall\, i, j$.

Esercizio. Dimostrare che $A \in \mathbb{C}^{n\times n}$ wstocastica per colonne con una riga costante deve essere singolare.

Esercizio (matrice B che non soddisfa entrambe le ipotesi $B \geq O$, B irriducibile, ma che ammette lo stesso una coppia di Perron $(\rho(B), \mathbf{w})$ con $\rho(B)$ autovalore semplice che domina gli altri autovalori e $\mathbf{w} > \mathbf{0}$). La matrice di Hessenberg superiore

$$B = \begin{bmatrix} 2 & 1/3 & 1 \\ 3 & -5/3 & 1 \\ 0 & 11/9 & 5/3 \end{bmatrix}$$

ha gli autovalori 3, 1, -2. Si nota che $3 = \rho(B)$ è autovalore semplice di B e un autovettore corrispondente può essere scelto a componenti positive con somma uguale a 1, infatti $B\mathbf{w} = 3\mathbf{w}$, se $w_1 = 15/38$, $w_2 = 12/38$, $w_3 = 11/38$. Quindi, per essere verificate le conclusioni della teoria di

Perron-Frobenius non è necessario che la matrice sia non negativa. Possibili generalizzazioni della teoria al caso $A \in \mathbb{R}^{n\times n}$ e anche al caso $A \in \mathbb{C}^{n\times n}$ sono considerate in [61], [62], [63].

Continuando l'esercizio, notiamo che $\rho(B)$ e $\mathbf{w}$ potrebbero essere calcolati con il metodo delle potenze (Teorema 3.6). A partire da un vettore $\mathbf{v}_0$, le successioni $\mathbf{v}_k$ e φ_k generate dal seguente algoritmo $\mathbf{v}_0 \in \mathbb{R}^n$, $\mathbf{v}_0 > \mathbf{0}$, $\|\mathbf{v}_0\|_1 = 1$, $\mathbf{a}_k = B\mathbf{v}_{k-1}$, $\varphi_k = \mathbf{u}^H\mathbf{a}_k/\mathbf{u}^H\mathbf{v}_{k-1} = \mathbf{u}^H B^k\mathbf{v}_0/\mathbf{u}^H B^{k-1}\mathbf{v}_0$, $\mathbf{v}_k = \mathbf{a}_k/\|\mathbf{a}_k\|_1 = B^k\mathbf{v}_0/\|B^k\mathbf{v}_0\|_1$, $k = 1, 2, \ldots$, se ben definite (ad esempio, se $\mathbf{u}^H\mathbf{v}_{k-1} \neq 0$) dovrebbero convergere rispettivamente a $\mathbf{w}$ e a $\rho(B) = 3$. Tuttavia, con una semplice osservazione si può evitare l'incertezza se le equazioni dell'algoritmo siano ben definite oppure no.

Si nota infatti che la matrice $A = B + (5/3)I$ è non negativa irriducibile, ed ha, sulla diagonale un elemento positivo. Quindi $\lambda_1 = r = \rho(A) > 0$ è autovalore semplice di A ed esiste ed è unico $\mathbf{z}$, $\mathbf{z} > \mathbf{0}$, $\|\mathbf{z}\|_1 = 1$, $A\mathbf{z} = \rho(A)\mathbf{z}$. Inoltre, per il Teorema 3.21, i rimanenti autovalori λ_2 e λ_3 hanno modulo minore di r, quindi il metodo delle potenze $\mathbf{v}_0 \in \mathbb{R}^n$, $\mathbf{v}_0 > \mathbf{0}$, $\|\mathbf{v}_0\|_1 = 1$, $\mathbf{a}_k = A\mathbf{v}_{k-1}$, $\mathbf{v}_k = \mathbf{a}_k/\|\mathbf{a}_k\|_1$, $k = 1, 2, \ldots$, genera due successioni di vettori positivi tali che $\mathbf{v}_k \to \mathbf{z}$, $\|\mathbf{a}_k\|_1 \to \rho(A)$, se $k \to \infty$. Per essere precisi, questo è vero nell'ipotesi che il coefficiente α nella rappresentazione $\mathbf{v}_0 = \alpha\mathbf{z} + \beta\mathbf{x} + \gamma\mathbf{y}$, $A\mathbf{x} = \mathbf{x}$, $A\mathbf{y} = -2\mathbf{y}$, sia non nullo. Trovata la coppia di Perron $(\rho(A), \mathbf{z})$ di A si può risalire immediatamente alla coppia di Perron $(\rho(B), \mathbf{w})$ di B, infatti $\rho(B) = \rho(A) - 5/3$ e $\mathbf{w} = \mathbf{z}$.

In questi ragionamenti, oltre l'osservazione che B differisce da una matrice non negativa irriducibile per un multiplo della matrice identica, si è usato il fatto che $\rho(B)$ è autovalore di B, cosa in generale a priori ignota per una matrice generica $B \in \mathbb{R}^{n\times n}$.

Esercizio. Sia

$$A = \begin{bmatrix} 0 & 1-a & a \\ a & 0 & 1-a \\ 1-a & a & 0 \end{bmatrix}, \ a \in \mathbb{R}.$$

Dimostrare che A ha almeno un autovalore in modulo minore di 1 se $a \in (0, 1)$. Studiare esistenza e unicità della coppia di Perron-Frobenius, e discutere la convergenza del metodo delle potenze nel calcolo della stessa.

Esercizio. Sia

$$A = \begin{bmatrix} 0 & b_1 & & & & \\ 1 & 0 & b_2 & & & \\ & 1-b_1 & 0 & \ddots & & \\ & & 1-b_2 & \ddots & b_{n-2} & \\ & & & \ddots & 0 & 1 \\ & & & & 1-b_{n-2} & 0 \end{bmatrix}, \ b_i \in (0, 1).$$

Studiare esistenza e unicità della coppia di Perron-Frobenius, e discutere la convergenza del metodo delle potenze nel calcolo della stessa.

3.3 Il calcolo dell'importanza dei nodi di un grafo: pagerank

Illustriamo ora una interessante applicazione della teoria sull'esistenza e unicità della coppia di Perron di una matrice e dei metodi illustrati per calcolarla. Consideriamo un insieme di entità

interconnesse. Data una questione è opportuno poter individuare nell'insieme il più rapidamente possibile, ovvero facendo pochi salti da una entità all'altra, l'entità in grado di rispondere esaurientemente alla questione. Individuare il nodo del grafo che soddisfa la questione. Ci potrebbero essere entità che si mostrano in grado di risolvere tante questioni di ogni tipo, che effettivamente possono essere in grado di farlo oppure possono esserlo per certe ma non per altre; ci potrebbero essere entità che si dichiarano specializzate in un ristretto numero di questioni, e, di nuovo, questo può più o meno corrispondere al vero. La capacità dichiarata di rispondere alle questioni piano piano assume un altro valore, maggiore o minore, a seconda di quanto, nel tempo, questa entità viene interpellata, a seconda se e quante volte i visitatori, coloro che sottopongono le questioni, tornano a interpellarla. Piano piano a queste entità vengono associati dei valori oggettivi che le collocano al di sopra o al di sotto di altre entità in grado di rispondere come loro a certe questioni. Quando questi valori vengono resi noti, dal più alto al più basso, il visitatore potrà individuare presto le entità che con più probabilità sono in grado di soddisfare le sue questioni, e potrà visitarle, standoci tanto e tornandoci più volte, perché sembrano rispondere alle sue questioni o, se insoddisfatto, abbandonandole subito in favore delle successive. E se questo secondo caso capiterà sempre più spesso, le "seconde scelte" acquisteranno più valore e cominceranno ad essere le "prime scelte" delle nuove visite. Come stabilire (aggiornare) periodicamente i valori oggettivi di tutte le entità dell'insieme? Visto che l'insieme delle entità e delle loro interconnessioni sono rappresentabili in modo naturale dai vertici e dagli archi di un grafo orientato, la domanda diventa: come associare al generico vertice di un grafo orientato un valore oggettivo, stabilito semplicemente sulla base della conoscenza di tutte le connessioni del grafo? Nel seguito vedremo uno dei possibili modi, seguendo brani del survey di Berkhin [27] sull'argomento.

Ad un grafo orientato con vertici $1, 2, \ldots, n$, si associ la sua matrice di transizione $\tilde{R}$. Se $\deg(i)$ denota il numero di archi uscenti dal vertice i, allora $\tilde{R}$ è la matrice $n \times n$ definita come segue

$$\tilde{R}_{ij} = \begin{cases} \frac{1}{\deg(i)} & \text{se c'è un arco da } i \text{ a } j \\ 0 & \text{altrimenti} \end{cases} .$$

Si noti che $\sum_j \tilde{R}_{ij} = 1$ se $\deg(i) > 0$ e $\sum_j \tilde{R}_{ij} = 0$ altrimenti. Quindi, la matrice $\tilde{R}$ è una matrice non negativa *quasi*-wstocastica per righe. Si noti, inoltre, che la riga i di $\tilde{R}$ è nulla se e solo se nessun arco esce da i, e la colonna j di $\tilde{R}$ è nulla se e solo se nessun arco punta a j.

Sia $\mathbf{p} \in \mathbb{R}^n$ il vettore il cui elemento j, p_j, è l'importanza (authority) del vertice j. Allora si può dire che p_j è proporzionale all'importanza del generico vertice i che punta a j ed inversamente proporzionale al numero dei vertici ai quali i punta, cioè in formule

$$p_j = \sum_{i:\, i \to j} \frac{p_i}{\deg(i)} = \sum_{i=1}^{n} \tilde{R}_{ij} p_i = \sum_{i=1}^{n} \tilde{R}^T_{ji} p_i = (\tilde{R}^T \mathbf{p})_j, \quad \mathbf{p} = \tilde{R}^T \mathbf{p}.$$

La condizione $\mathbf{p} = \tilde{R}^T \mathbf{p}$ sul vettore delle importanze $\mathbf{p}$ esprime un buon modello della realtà, ma, dal punto di vista matematico, può essere in contrasto con la ragionevole richiesta che ogni vertice j abbia una porzione di importanza p_j, univocamente definita e positiva, di una importanza totale $\sum_j p_j = 1$. Infatti, come vedremo, un vettore $\mathbf{p}$ tale che $\mathbf{p} = \tilde{R}^T \mathbf{p}$ può non esistere, oppure, se esiste, può non essere unico o avere elementi nulli.

Sia $p_j^{(k+1)}$ la probabilità che un certo utente al passo $k+1$ della visita del grafo (navigazione sul web) sia sul vertice (sulla pagina) j. È chiaro che $p_j^{(k+1)}$ sarà proporzionale alla probabilità che tale utente al passo k sia stato su un generico vertice i che punta a j e inversamente proporzionale al numero degli archi uscenti da i. Dunque,

$$p_j^{(k+1)} = \sum_{i:\, i\to j} \frac{p_i^{(k)}}{\deg(i)} = \sum_{i=1}^{n} \tilde{R}_{ij} p_i^{(k)} = \sum_{i=1}^{n} P_{ji}^T p_i^{(k)} = (\tilde{R}^T \mathbf{p}^{(k)})_j, \quad \mathbf{p}^{(k+1)} = \tilde{R}^T \mathbf{p}^{(k)}.$$

Anche qui sarebbe ragionevole richiedere che $p_j^{(k)}$ sia positiva per ogni j, ovvero che ci siano possibilità che l'utente al passo k possa trovarsi nell'arbitrario vertice j del grafo, e che la somma su j dei $p_j^{(k)}$ sia 1, ovvero che sia certo che al passo k l'utente stia su un qualche vertice del grafo, cioè richiedere che $\mathbf{p}^{(k)}$ sia una distribuzione discreta di probabilità (ddp). Invece, una volta dato $\mathbf{p}^{(0)} > \mathbf{0}$, $\sum_j p_j^{(0)} = 1$, i vettori generati dallo schema iterativo $\mathbf{p}^{(k+1)} = \tilde{R}^T \mathbf{p}^{(k)}$, pur esistendo ed essendo univocamente definiti, potrebbero non conservare la caratteristica di $\mathbf{p}^{(0)}$, di essere ddp. Vedi più avanti.

Riassumendo, vorremmo che valgano le seguenti affermazioni:

a) $\mathbf{p}$ tale che $\mathbf{p} = \tilde{R}^T \mathbf{p}$, $\mathbf{p} > \mathbf{0}$, $\|\mathbf{p}\|_1 = 1$, esiste ed è univocamente definito,

b) il metodo $\mathbf{p}^{(0)}$ =ddp, $\mathbf{p}^{(k+1)} = \tilde{R}^T \mathbf{p}^{(k)}$, genera una successione di ddp convergente a $\mathbf{p}$.

Adesso mostriamo che affinché a) e b) possano essere verificate, è necessario che la matrice $\tilde{R}^T$ sia wstocastica per colonne e irriducibile.

Teorema 3.24 Se i fatti a) e b) sono veri, allora $\tilde{R}^T$ deve essere wstocastica per colonne.

Dimostrazione. Sappiamo che $\exists!\ \mathbf{p}$ tale che $\mathbf{p} = \tilde{R}^T \mathbf{p}$, $\mathbf{p} > \mathbf{0}$, $\|\mathbf{p}\|_1 = 1$. Supponiamo che $\tilde{R}^T$ sia quasi-wstocastica ma non wstocastica per colonne. Allora

$$\begin{aligned} \|\tilde{R}^T \mathbf{p}\|_1 &= \sum_i (\tilde{R}^T \mathbf{p})_i = \sum_i \sum_j (\tilde{R}^T)_{ij} p_j = \sum_j \sum_i \tilde{R}_{ji} p_j \\ &= \sum_{j:\, \deg(j)>0} 1 \cdot p_j + \sum_{j:\, \deg(j)=0} 0 \cdot p_j < \sum_j p_j = \|\mathbf{p}\|_1, \end{aligned}$$

e, analogamente,

$$\|\mathbf{p}^{(k+1)}\|_1 = \|\tilde{R}^T \mathbf{p}^{(k)}\|_1 \le \|\mathbf{p}^{(k)}\|_1 \le \|\mathbf{p}^{(1)}\|_1 < \|\mathbf{p}^{(0)}\|_1 = 1.$$

Quindi $\mathbf{p}$ non può essere uguale a $\tilde{R}^T \mathbf{p}$, e $\mathbf{p}^{(k)}$ non può convergere a una ddp. □

Teorema 3.25 Se i fatti a) e b) sono veri, allora $\tilde{R}^T$ deve essere irriducibile.

Dimostrazione. Possiamo supporre, oltre che i fatti a) e b) siano veri, anche che la matrice $\tilde{R}^T$ sia wstocastica per colonne. Supponiamo allora per assurdo che $\tilde{R}^T$ sia riducibile. Allora esiste una matrice di permutazione Q tale che

$$Q^T \tilde{R}^T Q = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix},$$

con A_{11} e A_{22} matrici quadrate almeno 1×1. Si noti che $Q^T \tilde{R}^T Q$ è wstocastica per colonne, come $\tilde{R}^T$. Sappiamo che $\exists!$ $\mathbf{p}$ tale che $\mathbf{p} = \tilde{R}^T\mathbf{p}$, $\mathbf{p} > \mathbf{0}$, $\|\mathbf{p}\|_1 = 1$, ma ciò è equivalente a dire che $\exists!$ $Q^T\mathbf{p}$ tale che

$$Q^T\mathbf{p} = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} Q^T\mathbf{p}, \quad Q^T\mathbf{p} > \mathbf{0}, \ \|Q^T\mathbf{p}\|_1 = 1. \tag{32}$$

Caso 1: Supponiamo $A_{12} = 0$. Allora A_{11} e A_{22} sono matrici non negative wstocastiche per colonne. Possiamo supporre che siano anche irriducibili (perché?). Dunque, per la teoria di Perron-Frobenius,

$$\exists!\ \mathbf{y}_1, \mathbf{y}_2 > \mathbf{0}, \ \|\mathbf{y}_1\|_1 = \|\mathbf{y}_2\|_1 = 1, \ \mathbf{y}_1 = A_{11}\mathbf{y}_1, \ \mathbf{y}_2 = A_{22}\mathbf{y}_2,$$

e, di conseguenza, per tutti gli $\alpha \in (0,1)$ abbiamo

$$\begin{bmatrix} \alpha\mathbf{y}_1 \\ (1-\alpha)\mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} \alpha\mathbf{y}_1 \\ (1-\alpha)\mathbf{y}_2 \end{bmatrix},$$

$$\begin{bmatrix} \alpha\mathbf{y}_1 \\ (1-\alpha)\mathbf{y}_2 \end{bmatrix} > \mathbf{0}, \ \| \begin{bmatrix} \alpha\mathbf{y}_1 \\ (1-\alpha)\mathbf{y}_2 \end{bmatrix} \|_1 = 1.$$

Quindi, tutti i vettori $\mathbf{p}$ tali che $\mathbf{p}^T Q = [\, \alpha\mathbf{y}_1^T \ \ (1-\alpha)\mathbf{y}_2^T \,]$, $\alpha \in (0,1)$, soddisfano le proprietà $\mathbf{p} > \mathbf{0}$, $\|\mathbf{p}\|_1 = 1$, $\mathbf{p} = \tilde{R}^T\mathbf{p}$, contro l'ipotesi di unicità.

Caso 2: Supponiamo $A_{12} \neq 0$. Allora A_{22} è non negativa quasi-wstocastica ma non wstocastica per colonne, quindi A_{22} può non avere nessun autovalore uguale a 1 (non ce lo ha sicuramente se è irriducibile). Ne segue che le equazioni in (32) coinvolgenti A_{22} possono essere verificate solo se parte delle componenti del vettore $\mathbf{p}$ sono nulle, contro l'ipotesi di positività di $\mathbf{p}$. $\square$

Viceversa, per il Teorema 3.15, sappiamo che se la matrice non negativa $\tilde{R}^T$ è irriducibile e wstocastica per colonne, allora $1 = \rho(\tilde{R}^T)$ è autovalore semplice di $\tilde{R}^T$ ed esiste un unico vettore $\mathbf{p}$ tale che $\mathbf{p} = \tilde{R}^T\mathbf{p}$, $\mathbf{p} > \mathbf{0}$, $\|\mathbf{p}\|_1 = 1$. Sappiamo anche, però, che tali ipotesi non sono sufficienti per assicurare la convergenza (a $\mathbf{p}$) della successione $\mathbf{p}^{(k+1)} = \tilde{R}^T\mathbf{p}^{(k)}$, $\mathbf{p}^{(0)} > \mathbf{0}$, $\|\mathbf{p}_0\|_1 = 1$, o, equivalentemente, ad assicurare che i rimanenti autovalori di $\tilde{R}^T$ abbiano valore assoluto più piccolo di 1. Possiamo solo dire che la successione $\{\mathbf{p}^{(k)}\}$ è una successione ben definita di ddp.

La nostra matrice di transizione $\tilde{R}$, così com'è, fedele alla realtà, può non essere wstocastica; inoltre può essere riducibile e avere molte righe nulle. Dunque, per i risultati visti, dobbiamo modificarla in modo da rendere ben posto ($\exists!$) il problema matematico e rendere convergente l'algoritmo per la sua risoluzione.

Primo passo: rendere $\tilde{R}^T$ wstocastica per colonne. Consideriamo la seguente modifica R della matrice $\tilde{R}$,

$$R = \tilde{R} + \mathbf{d}\mathbf{v}^T, \quad \mathbf{d} = \begin{bmatrix} \delta_{\deg(1),0} \\ \vdots \\ \delta_{\deg(n),0} \end{bmatrix}, \quad \mathbf{v} = \begin{bmatrix} \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{bmatrix},$$

i.e, dove $\tilde{R}$ ha righe nulle R ha il vettore riga $\mathbf{v}^T$. Con questa modifica, il vertice i che aveva $\deg(i) = 0$ ora punta a tutti i vertici del grafo, con la stessa "minima" probabilità. (Discutiamo qui il caso uniforme, ma ciò che segue può essere ripetuto per il caso generale in cui $\mathbf{v}$ è una generica ddp).

Si osservi che R^T è wstocastica per colonne e non negativa. Quindi, 1 è autovalore di R^T con molteplicità algebrica e geometrica coincidenti (Proposizione 3.4), gli altri autovalori di R^T, $\lambda_2, \ldots, \lambda_n$, hanno modulo minore o uguale a 1, e almeno un autovettore $\mathbf{p}$ relativo all'autovalore 1 è non negativo (Proposizione 3.5). Un tale autovettore $\mathbf{p}$ può essere calcolato con il metodo delle potenze o delle potenze inverse (vedi le Proposizioni 3.8 e 3.10). Tuttavia, poiché la matrice R^T potrebbe essere anch'essa riducibile, $\mathbf{p}$ non è in generale univocamente definito e/o può avere componenti nulle, e questo può non essere accettabile.

Secondo passo: rendere R^T irriducibile. Consideriamo la seguente modifica R' della matrice R,

$$R' = cR + (1-c)\mathbf{e}\mathbf{v}^T, \quad \mathbf{e} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \ c \in (0,1).$$

Poiché

$$\mathbf{e}_i^T R' = \begin{cases} c\mathbf{v}^T + (1-c)\mathbf{v}^T = \mathbf{v}^T & \deg(i) = 0 \\ c[\cdots 0 \ \frac{1}{\deg(i)} \ 0 \cdots] + (1-c)\mathbf{v}^T & \deg(i) > 0 \end{cases},$$

stiamo supponendo che l'utente, visitando il grafo, possa andare dal vertice i ad uno dei suoi vicini con probabilità $c/\deg(i) + (1-c)/n$, e ad un arbitrario altro vertice del grafo con probabilità $(1-c)/n$. Questo è ammissibile, infatti, improvvisamente, l'utente potrebbe decidere di passare ad un'altra ricerca (ad un altro argomento). Naturalmente, il parametro c deve essere scelto vicino a 1, in modo che non ci si discosti troppo dal modello di visita del grafo (web) supposto inizialmente. (Il motore di ricerca Google pone $c = 0.85$ [27]).

Si osservi che $(R')^T$ è wstocastica per colonne e positiva, perciò, in particolare, è non negativa e irriducibile. Quindi abbiamo tutto ciò di cui abbiamo bisogno.

Teorema 3.26 $1 = \rho((R')^T)$ è un autovalore semplice di $(R')^T$, esiste un unico vettore $\mathbf{p}$ tale che $\mathbf{p} = (R')^T\mathbf{p}$, $\mathbf{p} > \mathbf{0}$, $\|\mathbf{p}\|_1 = 1$ (i.e. si ha il fatto (a)), e gli altri autovalori di $(R')^T$, $\lambda'_2, \ldots, \lambda'_n$, hanno modulo minore di 1 (per il Teorema 3.20). Quindi, $\mathbf{p}^{(k+1)} = (R')^T\mathbf{p}^{(k)}$, $k = 0, 1, \ldots$, $\mathbf{p}^{(0)} > \mathbf{0}$, $\|\mathbf{p}^{(0)}\|_1 = 1$, è una successione di ddp convergente a $\mathbf{p}$ con velocità di convergenza

$$\|\mathbf{p}^{(k)} - \mathbf{p}\| = \max_{j=2,\ldots,n} \max_{s_{\lambda'_j}} O(|p_{s_{\lambda'_j}-1}(k)||\lambda'_j|^k)$$

dove $s_{\lambda'_j}$ è l'ordine del generico blocco di Jordan di $(R')^T$ relativo a λ'_j (i.e. si ha il fatto (b)). Inoltre, per la particolare scelta di $(R')^T$, il costo di ciascun passo del metodo delle potenze è dominato dal costo del prodotto della matrice $\tilde{R}^T$ per un vettore, ed è quindi $O(n)$, e sulla la velocità di convergenza si ha che

$$\|\mathbf{p}^{(k)} - \mathbf{p}\| = \max_{s_{\lambda'_j}} O(|p_{s_{\lambda'_j}-1}(k)|c^k).$$

Dimostrazione: Segue quasi tutto dal Teorema 3.23. Si devono provare solo le ultime due affermazioni, che sfruttano la particolare struttura di $(R')^T$. Dimostriamo la prima. Poiché

$$(R')^T = c(\tilde{R}^T + \mathbf{v}\mathbf{d}^T) + (1-c)\mathbf{v}\mathbf{e}^T,$$

possiamo dire che, per ogni vettore $\mathbf{z}$ non negativo,

$$(R')^T\mathbf{z} = c\tilde{R}^T\mathbf{z} + \gamma\mathbf{v},$$
$$\gamma = c\mathbf{d}^T\mathbf{z} + (1-c)\mathbf{e}^T\mathbf{z} = \mathbf{e}^T\mathbf{z} - c[\mathbf{e}^T\mathbf{z} - \mathbf{d}^T\mathbf{z}] = \|\mathbf{z}\|_1 - c\|\tilde{R}^T\mathbf{z}\|_1$$

(dimostrare l'ultima uguaglianza!). Quindi, il vettore $\mathbf{p}^{(k+1)}$ può essere calcolato da $\mathbf{z} := \mathbf{p}^{(k)}$ nei seguenti tre passi:

$$\mathbf{y} = c\tilde{R}^T\mathbf{z}, \quad \gamma = \|\mathbf{z}\|_1 - \|\mathbf{y}\|_1, \quad \mathbf{p}^{(k+1)} := (R')^T\mathbf{z} = \mathbf{y} + \gamma\mathbf{v}\,,$$

dove l'operazione dominante è chiaramente il prodotto matrice-vettore $\tilde{R}^T\mathbf{z}$. Si noti che ciascuna riga j di $\tilde{R}^T$ ha un numero di elementi non nulli uguale al numero dei vertici che puntano a j, e quest'ultimo in media è molto piccolo. Quindi $\tilde{R}^T\mathbf{z}$ si può calcolare con $O(n)$ operazioni aritmetiche. Si noti anche che per implementare i tre passi di cui sopra, per il calcolo di $(R')^T\mathbf{z}$, sono sufficienti solo $2n + O(1)$ celle di memoria.

Dimostriamo ora la seconda affermazione. A tale scopo sarà sufficiente mostrare che gli autovalori di R' diversi da 1 hanno modulo minore o uguale a c. Per prima cosa si dimostra che se $\mathbf{e}^T\mathbf{v} = 1$, allora

$$p_R(\lambda) = (\lambda - 1)q(\lambda) \quad \Rightarrow \quad p_{R+\frac{1-c}{c}\mathbf{e}\mathbf{v}^T}(\lambda) = (\lambda - \frac{1}{c})q(\lambda), \tag{33}$$

dove $p_M(\lambda)$ indica il polinomio caratteristico della matrice M. Sia infatti S la matrice $\begin{bmatrix}\mathbf{e}\ \mathbf{y}_2\ \cdots\ \mathbf{y}_n\end{bmatrix}$ con $\mathbf{e} = [1\ 1\ \cdots\ 1]^T$ e $\mathbf{y}_j$ scelti in modo che $\det(S) \neq 0$. Allora

$$S^{-1}RS = \begin{bmatrix} 1 & \mathbf{u}^T \\ \mathbf{0} & M \end{bmatrix},\ p_R(\lambda) = (\lambda - 1)p_M(\lambda),$$

$$S^{-1}(R + \frac{1-c}{c}\mathbf{e}\mathbf{v}^T)S = \begin{bmatrix} 1 & \mathbf{u}^T \\ \mathbf{0} & M \end{bmatrix} + \frac{1-c}{c}S^{-1}\mathbf{e}\mathbf{v}^T S = \begin{bmatrix} \frac{1}{c} & \tilde{\mathbf{u}}^T \\ \mathbf{0} & M \end{bmatrix},$$

da cui segue l'identità $p_{R+((1-c)/c)\mathbf{e}\mathbf{v}^T}(\lambda) = (\lambda - 1/c)p_M(\lambda)$, e quindi la tesi. Come conseguenza di (33), se $1, \lambda_2, \ldots, \lambda_n$ sono gli autovalori di R (si ricorda che $|\lambda_j| \leq 1$, $j = 2, \ldots, n$), allora gli autovalori di $R + ((1-c)/c)\mathbf{e}\mathbf{v}^T$ sono $\frac{1}{c}, \lambda_2, \ldots, \lambda_n$, e quindi gli autovalori di $R' = cR + (1-c)\mathbf{e}\mathbf{v}^T$ sono $1, c\lambda_2, \ldots, c\lambda_n$. Ne segue che gli autovalori di R' diversi da 1 hanno modulo minore o uguale a c. □

Esercizio Si dimostra che $\mathbf{p} = (R')^T\mathbf{p}$, $\mathbf{p} > \mathbf{0}$, $\|\mathbf{p}\|_1$=1, soddisfa la condizione $\mathbf{p} = (1/\|\mathbf{x}\|)\mathbf{x}$ essendo $\mathbf{x}$ la soluzione del sistema lineare $(I - \alpha\tilde{R}^T)\mathbf{x} = \mathbf{v}$ (è vero ??). Poiché $\tilde{R}$ ha gli autovalori in modulo minore o uguale a 1 ($\tilde{R}$ è non negativa e quasi-wstocastica per righe), gli autovalori della matrice $I - \alpha\tilde{R}^T$ hanno parte reale positiva. Quindi si può usare il metodo di Richardson-Eulero per calcolare $\mathbf{x}$: per $\omega \in (0, \omega^*)$ la successione $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \omega(\mathbf{v} - (I - \alpha\tilde{R}^T)\mathbf{x}^{(k)})$, $k = 0, 1, \ldots$, converge a $\mathbf{x}$. D'altro canto, per la Proposizione 3.10 si ha che, per ε abbastanza piccolo, la successione $\mathbf{v}_k$

$$\mathbf{v}_0 = \text{wddp}, \quad ((1+\varepsilon)I - (R')^T)\mathbf{a}_k = \mathbf{v}_{k-1}, \quad \mathbf{v}_k = (1/\|\mathbf{a}_k\|_1)\mathbf{a}_k, \quad k = 1, 2, \ldots,$$

converge in poche iterazioni a $\mathbf{p}$. C'è una relazione tra i due risultati? (Nota: $\frac{1}{1+\varepsilon} < 1$ come α!)

Perchè calcolare $\mathbf{p}$? Illustriamo con un esempio come utilizza $\mathbf{p}$ il motore di ricerca Google [27]. L'utente sottopone una QUERY, ad esempio la QUERY: "Berkhin survey". Google come

prima cosa va nel "inverted terms document file". In tale file, a ciascun termine di un "collection's dictionary" è associata una lista di tutti i documenti che contengono tale termine:

$$\begin{array}{l}\vdots\\ \text{term} \quad \rightarrow \quad \text{LISTA}_{term} = \{i_1, i_2, \ldots, i_k\} \subset \{1, 2, \ldots, n\}\\ \vdots\\ \text{Berkhin} \quad \rightarrow \quad \text{LISTA}_{Berkhin} = \{1, 4, 6\}\\ \vdots\\ \text{survey} \quad \rightarrow \quad \text{LISTA}_{survey} = \{1, 3\}\\ \vdots\end{array}$$

Google individua le liste corrispondenti ai termini contenuti nella QUERY, nel nostro caso le liste corrispondenti ai termini "Berkhin" e "survey", e definisce l'insieme di rilevanza per la query:

$$\bigcup_{term \in \text{QUERY}} \text{LISTA}_{term} = \{1, 3, 4, 6\}.$$

Poi, leggendo $\mathbf{p}$ ($\mathbf{p}$ è aggiornato una volta al mese), considera il corrispondente insieme delle "authorities" $\{p_1, p_3, p_4, p_6\}$ e ordina i suoi elementi, per esempio $p_4 \geq p_6 \geq p_3 \geq p_1$. Infine, mostra all'utente i titoli dei documenti $1, 3, 4, 6$ nell'ordine $4, 6, 3, 1$, da quello con la più grande authority a quello con la più piccola.

Una delle critiche che si muovono a questa procedura è che la sua indipendenza dal tipo di QUERY, da un lato permette una veloce risposta, ma dall'altro lato non permette di distinguere le pagine con una certa authority in tanti argomenti, dalle pagine con la stessa authority su un argomento specifico.

Esercizio. Disegnare il grafo con la seguente matrice di transizione:

$$\tilde{R} = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

Si noti che $\tilde{R}$ è non negativa, riducibile e quasi-wstocastica, ma non wstocastica, per righe. Dimostrare che non c'è nessun vettore positivo $\mathbf{p}$ tale che $\mathbf{p} = \tilde{R}^T\mathbf{p}$. Partendo da $\tilde{R}$ e procedendo come indicato nella teoria, introdurre una matrice non negativa R wstocastica per righe. Si noti che R è riducibile, come $\tilde{R}$. Provare che non c'è nessun vettore positivo $\mathbf{p}$ tale che $\mathbf{p} = R^T\mathbf{p}$. Partendo da R e procedendo come indicato nella teoria, introdurre una matrice non negativa R' irriducibile e wstocastica per righe. Provare che esiste un unico vettore positivo $\mathbf{p}$ tale che $\mathbf{p} = (R')^T\mathbf{p}$, $\|\mathbf{p}\|_1 = 1$, e descrivere il metodo delle potenze per il calcolo di $\mathbf{p}$. Ecco qui di seguito una approssimazione del vettore $\mathbf{p}$ ottenuta con tale algoritmo:

$$\mathbf{p}^T = [0.03721\ 0.05396\ 0.04151\ 0.3751\ 0.206\ 0.2862].$$

Assumendo, per esempio, che l'insieme di rilevanza della query sia {1, 3, 4, 6}, i documenti $1, 3, 4, 6$ vengono presentati nell'ordine 4, 6, 3, 1, essendo $p_4 \geq p_6 \geq p_3 \geq p_1$.

Spettro di $(C_\phi)_A$ ($\mathcal{L}_A$, $\mathcal{L}$ =) per stimare lo spettro di A wstocastica ...?

4 Metodi iterativi e tecniche di precondizionamento per la risoluzione di sistemi lineari

4.1 ...Teoria sul Precondizionamento (Daniele)...

4.2 Precondizionamento, preliminari, precondizionamento con algebre (di sistemi di Toeplitz)

Come abbiamo visto nella Sezione 4.1, si può migliorare l'efficienza di un metodo iterativo nella risoluzione di un sistema lineare $A\mathbf{x} = \mathbf{b}$, $A \in \mathbb{C}^{n\times n}$, $\mathbf{b} \in \mathbb{C}^n$, applicandolo a un sistema equivalente $P^{-1}A\mathbf{x} = P^{-1}\mathbf{b}$, a due condizioni: (1) lo spettro di $P^{-1}A$ deve essere distribuito più convenientemente di quello di A (ai fini di una maggiore rapidità di convergenza del metodo); (2) la presenza della matrice P nelle equazioni del metodo non deve aumentare troppo il costo di ogni sua iterazione.

In particolare, se A è reale definita positiva, applicare il metodo del Gradiente Coniugato (GC) al sistema $P^{-1}A\mathbf{x} = P^{-1}\mathbf{b}$, $\mathbf{b} \in \mathbb{R}^n$, P reale definita positiva, può essere molto vantaggioso perché sotto certe ipotesi su A e per opportune scelte di P lo spettro di $P^{-1}A$ si raggruppa e ciò si traduce in un ordine di convergenza superlineare di GC. Mostreremo ciò nei dettagli quando A è una matrice di Toeplitz e la matrice P è scelta in una algebra $\mathcal{L}$, con $\mathcal{L} = C_\phi$, $\phi = \pm 1$ [45], [43], [44]. Molte delle cose che diremo valgono più in generale per altre algebre di bassa complessità di tipo $\mathcal{L} = \{Ud(\mathbf{z})U^H : \mathbf{z} \in \mathbb{C}^n\}$ dove U è unitaria, ed, in particolare, per l'algebra τ e le algebre di tipo Hartley (vedi [46], [36], [47], [5]). La risoluzione dei sistemi lineari di Toeplitz con metodi (diretti e) iterativi e lo studio di ogni possibile precondizionatore per tali sistemi sono argomenti trattati approfonditamente in [48].

È importante citare almeno un altro caso in cui un sistema reale definito positivo $A\mathbf{x} = \mathbf{b}$ può essere trasformato con poco costo in un sistema equivalente risolubile con molte meno iterazioni di GC. Dato un poligono $\Omega \subset \mathbb{R}^2$, la soluzione $u \in H_0^1(\Omega)$ della formulazione variazionale del problema differenziale di Poisson su Ω, $-\Delta u = f$, $x \in \Omega$, $u = 0$, $x \in \partial\Omega$, e lo spazio $V_h \subset H_0^1(\Omega)$ delle funzioni nulle su $\partial\Omega$, continue lineari a tratti su una triangolazione τ_h di Ω, i coefficienti della approssimazione interna $u_h \in V_h$ di u rispetto alla base standard di Lagrange risolvono un sistema $A\mathbf{x} = \mathbf{b}$ con A reale definita positiva e numero di condizionamento che va a infinito come $O(1/h^2)$. La semplice sostituzione della base standard con una base di tipo gerarchico porta alla definizione di un sistema $P^{-1}A\mathbf{x} = P^{-1}\mathbf{b}$, con P reale definita positiva, tale che lo spettro di $P^{-1}A$ al diminuire del parametro di discretizzazione h, pur non raggruppandosi da nessuna parte, occupa su $\mathbb{R}^+$ un intervallo molto più piccolo di quello occupato da $\sigma(A)$, infatti $\max\sigma(P^{-1}A)/\min\sigma(P^{-1}A)$ va a infinito come $O((\log 1/h)^2)$ [49]. Dunque anche in questo caso è vantaggioso applicare GC al sistema $P^{-1}A\mathbf{x} = P^{-1}\mathbf{b}$ anziché ad $A\mathbf{x} = \mathbf{b}$.

Caratteristiche importanti del metodo GCP

Data una matrice $A \in \mathbb{R}^{n\times n}$ definita positiva e $\mathbf{b} \in \mathbb{R}^n$, il seguente *metodo GCP*

$P \in \mathbb{R}^{n\times n}$ definita positiva, $\mathbf{x}_0 \in \mathbb{R}^n$, $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$, $\mathbf{d}_0 = \mathbf{h}_0 = P^{-1}\mathbf{r}_0$;
Per $k = 0, 1, 2, \ldots$:
$\tau_k = \frac{\mathbf{r}_k^T\mathbf{h}_k}{\mathbf{d}_k^T A\mathbf{d}_k}$, $\mathbf{x}_{k+1} = \mathbf{x}_k + \tau_k\mathbf{d}_k$, $\mathbf{r}_{k+1} = \mathbf{r}_k + \tau_k A\mathbf{d}_k$,
$\mathbf{h}_{k+1} = P^{-1}\mathbf{r}_{k+1}$, $\beta_k = \frac{\mathbf{r}_{k+1}^T\mathbf{h}_{k+1}}{\mathbf{r}_k^T\mathbf{h}_k}$, $\mathbf{d}_{k+1} = \mathbf{h}_{k+1} + \beta_k\mathbf{d}_k$,

(ottenuto applicando GC [41] al sistema lineare $\tilde{A}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$, dove $\tilde{A} = E^{-1}AE^{-T}$, $\tilde{\mathbf{b}} = E^{-1}\mathbf{b}$, $\tilde{\mathbf{x}} = E^T\mathbf{x}$, $EE^T = P$, e riscrivendo le equazioni del metodo in termini di $\mathbf{x}_k := E^{-T}\tilde{\mathbf{x}}_k$), fornisce la soluzione del sistema lineare $A\mathbf{x} = \mathbf{b}$, o, equivalentemente, il punto di minimo della funzione $f : \mathbb{R}^n \to \mathbb{R}$, $f(\mathbf{z}) = \frac{1}{2}\mathbf{z}^T A\mathbf{z} - \mathbf{z}^T\mathbf{b}$, in al più n passi.

Inoltre, valgono i seguenti risultati:

1 $\mathbf{d}_k^T A\mathbf{d}_j = 0$, $\mathbf{r}_k^T\mathbf{h}_j = 0$, $0 \le j < k$, finché $\mathbf{r}_k \neq \mathbf{0}$.

2 Esiste $\dot{k}$ minore o uguale del numero degli autovalori distinti di $P^{-1}A$ tale che $\mathbf{x}_{\dot{k}} = \mathbf{x} = A^{-1}\mathbf{b}$.

3 Per l'errore al passo k, misurato con la norma $\|\cdot\|_A = (\,\cdot^T A\,\cdot\,)^{\frac{1}{2}}$, vale la seguente maggiorazione

$$\|\mathbf{x} - \mathbf{x}_k\|_A \le \min_{p\in\Pi_k^1} \max_{t\in\sigma(P^{-1}A)} |p(t)| \|\mathbf{x} - \mathbf{x}_0\|_A,$$

essendo Π_k^1 l'insieme dei polinomi di grado esattamente k che in zero valgono 1 (si noti che gli autovalori di $P^{-1}A$ sono reali positivi perché coincidenti con quelli di $\tilde{A}$).

4 Siano $[a, b] \subset \mathbb{R}^+$ (si pensi $[a, b] \cap \sigma(P^{-1}A) \neq \emptyset$), r il numero degli autovalori di $P^{-1}A$ fuori dall'intervallo $[a, b]$, e $\hat{\lambda}$ gli autovalori di $P^{-1}A$ tali che $\hat{\lambda} < \frac{1}{2}b$ e $\hat{\lambda} < a$. Allora, per $k \ge r$,

$$\|\mathbf{x} - \mathbf{x}_k\|_A \le 2\left(\frac{\sqrt{\frac{b}{a}} - 1}{\sqrt{\frac{b}{a}} + 1}\right)^{k-r} s_{\hat{\lambda}} \, \|\mathbf{x} - \mathbf{x}_0\|_A, \quad s_{\hat{\lambda}} = \prod_{\hat{\lambda}} \left(\frac{b}{\hat{\lambda}} - 1\right).$$

4a) Scegliendo in 4) $a = \min\sigma(P^{-1}A)$, $b = \max\sigma(P^{-1}A)$, si ottiene la seguente maggiorazione per l'errore al passo k:

$$\|\mathbf{x} - \mathbf{x}_k\|_A \le 2\left(\frac{\sqrt{\mu_2(\tilde{A})} - 1}{\sqrt{\mu_2(\tilde{A})} + 1}\right)^{k} \|\mathbf{x} - \mathbf{x}_0\|_A, \quad \mu_2(\tilde{A}) = \frac{\max\sigma(P^{-1}A)}{\min\sigma(P^{-1}A)}.$$

4b) Scegliendo in 4) $a = 1 - \varepsilon$, $b = 1 + \varepsilon$, $0 < \varepsilon < 1$, si prova che il fattore di riduzione dell'errore iniziale $\|\mathbf{x} - \mathbf{x}_0\|_A$ è, per $k \ge r$, minore o uguale di

$$\begin{aligned}\prod_{\hat{\lambda}} \left(\frac{1+\varepsilon}{\hat{\lambda}} - 1\right) 2\left(\frac{\sqrt{\frac{1+\varepsilon}{1-\varepsilon}} - 1}{\sqrt{\frac{1+\varepsilon}{1-\varepsilon}} + 1}\right)^{k-r} &= \prod_{\hat{\lambda}} \left(\frac{1+\varepsilon}{\hat{\lambda}} - 1\right) 2\left(\frac{\varepsilon^k}{\varepsilon^r 2^{k-r}} + \ldots\right) \\ &\le \left(\frac{1+\varepsilon}{\min\hat{\lambda}} - 1\right)^{\#\hat{\lambda}} 2\left(\frac{\varepsilon^k}{\varepsilon^r 2^{k-r}} + \ldots\right),\end{aligned}$$

dove i $\hat{\lambda}$ sono gli autovalori di $P^{-1}A$ tali che $\hat{\lambda} < 1 - \varepsilon$ e $\hat{\lambda} < \frac{1}{2}(1 + \varepsilon)$. Quindi, se per ε tale che $(1+\varepsilon)/(1-\varepsilon) << \mu_2(\tilde{A})$ si ha $r << n$, allora GCP può essere molto più rapido di quanto si potrebbe dedurre dalla maggiorazione in 4a).

4c) Sia $A_n\mathbf{x} = \mathbf{b}_n$, $\mathbf{b}_n \in \mathbb{R}^n$, $A_n \in \mathbb{R}^{n\times n}$ reale definita positiva, una successione di sistemi lineari, e $P_n^{-1}A_n\mathbf{x} = P_n^{-1}\mathbf{b}_n$, $P_n \in \mathbb{R}^{n\times n}$ definita positiva, una corrispondente successione di sistemi equivalenti. Se

$$\mu_2(\tilde{A}_n) = \frac{\max\sigma(P_n^{-1}A_n)}{\min\sigma(P_n^{-1}A_n)} \leq M,\ \forall\, n, \tag{34}$$

allora (si dice che) GCP converge con ordine di convergenza lineare quando applicato a $\{P_n^{-1}A_n\mathbf{x} = P_n^{-1}\mathbf{b}_n\}_n$. Se invece, per $\varepsilon > 0$ comunque piccolo, esiste un ν_ε tale che per $n > \nu_\varepsilon$ il numero $r_{n,\varepsilon}$ degli autovalori di $P_n^{-1}A_n$ fuori dell'intervallo $[1-\varepsilon, 1+\varepsilon]$ è limitato da un numero r_ε indipendente da n e il numero $\min\sigma(P_n^{-1}A_n)$ è limitato dal basso da un numero positivo $\hat{\lambda}_{MIN}$ indipendente da n, ovvero

$$\forall\,\varepsilon > 0,\ \exists\,\nu_\varepsilon,\ r_\varepsilon \ \big|\ \forall\, n > \nu_\varepsilon \text{ si ha} \quad \#\{\lambda \in \sigma(P_n^{-1}A_n),\ |\lambda - 1| > \varepsilon\} \ =: r_{n,\varepsilon} < r_\varepsilon \tag{35}$$

$$\text{ed}\ \ \exists\,\hat{\lambda}_{MIN} > 0 \ \big|\ \min\sigma(P_n^{-1}A_n) > \hat{\lambda}_{MIN},\ \forall\, n\ , \tag{36}$$

allora (si dice che) GCP converge con ordine di convergenza superlineare quando applicato a $\{P_n^{-1}A_n\mathbf{x} = P_n^{-1}\mathbf{b}_n\}_n$.

Dimostrazione. Per i punti 1,2,3 si rimanda a [41], [42]. Per dimostrare il punto 4 si usa il punto 3, scegliendo come p il seguente polinomio in Π_k^1:

$$p(t) = \prod_{\lambda\in\sigma(P^{-1}A),\,\lambda\notin[a,b]} (1 - \frac{t}{\lambda})\tilde{q}(t),$$

dove $\tilde{q}(t)$ è il polinomio che rende la quantità $\max_{t\in[a,b]}|q(t)|$ minima al variare di q nell'insieme Π_{k-r}^1 dei polinomi di grado esattamente $k-r$ che in zero valgono 1, cioè $\tilde{q}(t) = T_{k-r}(\frac{b+a-2t}{b-a})\,/\,T_{k-r}(\frac{b+a}{b-a})$ ove $T_m(t)$ è l'm-esimo polinomio di Chebycev (vedi [41]). Quindi

$$\|\mathbf{x} - \mathbf{x}_k\|_A \leq \max_{t\in\sigma(P^{-1}A)} |p(t)|\|\mathbf{x} - \mathbf{x}_0\|_A,$$

con

$$\begin{aligned}
\max_{t\in\sigma(P^{-1}A)} |p(t)| &= \max_{t\in\sigma(P^{-1}A)} |\Pi_{\lambda\in\sigma(P^{-1}A),\,\lambda\notin[a,b]}(1 - \frac{t}{\lambda})\tilde{q}(t)| \\
&\leq \max_{t\in[a,b]} |\Pi_{\lambda\in\sigma(P^{-1}A),\,\lambda\notin[a,b]}(1 - \frac{t}{\lambda})\tilde{q}(t)| \\
&= \max_{t\in[a,b]} \Pi_{\lambda\in\sigma(P^{-1}A),\,\lambda\notin[a,b]}|1 - \frac{t}{\lambda}||\tilde{q}(t)| \\
&\leq \max_{t\in[a,b]} \Pi_{\lambda\in\sigma(P^{-1}A),\,\lambda\notin[a,b],\,\lambda<\frac{1}{2}b}|1 - \frac{t}{\lambda}| \max_{t\in[a,b]} |\tilde{q}(t)| \\
&= \frac{1}{T_{k-r}(\frac{b+a}{b-a})} \max_{t\in[a,b]} \Pi_{\hat{\lambda}}|1 - \frac{t}{\hat{\lambda}}|,
\end{aligned}$$

da cui la tesi, essendo $T_{k-r}(\frac{b/a+1}{b/a-1}) \geq \frac{1}{2}\big(\frac{\sqrt{b/a}+1}{\sqrt{b/a}-1}\big)^{k-r}$ [41]. L'affermazione 4b) segue dall'identità

$$f(\varepsilon) = \frac{\sqrt{\frac{1+\varepsilon}{1-\varepsilon}} - 1}{\sqrt{\frac{1+\varepsilon}{1-\varepsilon}} + 1} = \frac{\varepsilon}{2} + \frac{\varepsilon^2}{2}f''(0) + \dots\ .$$

Infine, i risultati 4a) e 4b) giustificano le affermazioni in 4c). □

Le equazioni del metodo GCP e le sue proprietà 1),2),3),4) sopra descritte, ci permettono di fare diverse considerazioni. È evidente che il costo di ogni passo dell'algoritmo GCP è determinato dal costo del prodotto matrice-vettore $A\mathbf{d}_k$ e dal costo della risoluzione del sistema lineare $P\mathbf{h}_{k+1} = \mathbf{r}_{k+1}$. Scegliendo $P = I$ si evita questa seconda operazione, ma se lo spettro di A non è favorevole, ovvero gli autovalori di A sono tutti diversi tra loro, sono distribuiti su un ampio intervallo di $\mathbb{R}^+$ e non sono per la maggior parte raggruppati, potrebbero essere necessari quasi tutti gli n passi per ottenere un vettore $\mathbf{x}_k \approx \mathbf{x}$. Invece, scegliendo P tale che lo spettro di $P^{-1}A$ sia più favorevole di quello di A, se l'operazione $P\mathbf{h}_{k+1} = \mathbf{r}_{k+1}$ non appesantisce troppo il generico passo, allora si potrebbe ottenere lo scopo, $\mathbf{x}_k \approx \mathbf{x}$, a un costo computazionale totale minore. Dai punti 2)-4a)-4b)-4c) di cui sopra, si deduce che per "più favorevole" si possono intendere tre cose: i) il rapporto tra il massimo e il minimo autovalore di $P^{-1}A$ (il numero di condizionamento di $\tilde{A} = E^{-1}AE^{-T}$, $P = EE^T$) è molto minore di $\mu_2(A)$; ii) $P^{-1}A$ ha molti meno autovalori distinti di quanti ne ha A; iii) quasi tutti gli autovalori di $P^{-1}A$ sono contenuti nell'intervallo $[a, b]$, con $a > 0$ non troppo vicino a zero e b/a molto minore di $\mu_2(A)$, e quelli fuori di $[a, b]$ non sono troppo vicini a zero. In ogni caso, è opportuno (vedi [42]) che il numero di condizionamento, nel passaggio da A ad $\tilde{A}$, se non diminuisce, almeno rimanga pressoché invariato.

È opportuno fare considerazioni particolari riguardanti il punto 4c). Il sistema lineare $A\mathbf{x} = \mathbf{b}$ dato, che occorre risolvere, fa parte quasi sempre di un insieme di sistemi lineari $\{A_n\mathbf{x} = \mathbf{b}_n\}_n$ di dimensione n variabile (ad es. crescente al rimpiccolirsi di un parametro di discretizzazione), e il voler ottenere una $\mathbf{x}$ "migliore" potrebbe richiedere la sostituzione di $A\mathbf{x} = \mathbf{b}$ con un altro sistema dell'insieme, di dimensione "maggiore". Sarebbe quindi opportuno poter associare alla sequenza $\{A_n\}$ una sequenza di matrici P_n che si comportino da buoni precondizionatori anche al crescere di n. Nel punto 4c) si sono formalizzate le condizioni ideali che la sequenza P_n dovrebbe soddisfare. Osserviamo che anche nel caso in cui la condizione (34) non è soddisfatta, cioè $\mu_2(\tilde{A}_n) \to +\infty$, possono essere soddisfatte le (35),(36) e il metodo GCP può quindi avere comunque una alta velocità di convergenza, anche per valori grandi di n. Tuttavia, in tale caso, la soluzione del sistema $P_n^{-1}A_n\mathbf{x} = P_n^{-1}\mathbf{b}_n$ potrebbe diventare (al crescere di n) molto sensibile a variazioni dei dati; in altre parole, il vettore $(A_n + \delta A_n)^{-1}(\mathbf{b}_n + \delta\mathbf{b}_n)$ e, quindi, il vettore che lo approssima calcolato da GCP, potrebbe essere molto diverso dal vettore $A_n^{-1}\mathbf{b}_n$, quello cui siamo interessati [42]. È da notare che, tranne che per poche classi di matrici strutturate A_n — vedi il caso A_n =Toeplitz ($A_n = (t_{|i-j|})_{i,j=1}^n$, $\sum_k |t_k| < +\infty$, $\sum_{k\in\mathbb{Z}} t_{|k|}e^{\mathbf{i}k\theta} > 0$) considerato più avanti —, in generale non si riescono a soddisfare le condizioni (34) o (35),(36), a meno che non si sceglie P_n troppo vicino ad A_n ($P_n = A_n \Rightarrow \mu_2(\tilde{A}_n) = 1\ \forall n$!), "troppo" nel senso che la risoluzione del sistema $P_n\mathbf{z} = (\mathbf{r}_n)_{k+1}$, da effettuarsi in GCP ad ogni passo k, verrebbe a costare quasi quanto la risoluzione del sistema $A_n\mathbf{x} = \mathbf{b}_n$! Quindi, in generale è già molto se si riesce a *scegliere P_n in modo che* 1) *i numeri $\mu_2(\tilde{A}_n)$ (e/o $\#\{\lambda \in \sigma(P_n^{-1}A_n),\ |\lambda - 1| > \varepsilon\}$ e/o $1/\min\sigma(P_n^{-1}A_n)$) vadano a $+\infty$ meno velocemente dei numeri $\mu_2(A_n)$ (e/o $\#\{\lambda \in \sigma(A_n),\ |\lambda - 1| > \varepsilon\}$ e/o $1/\min\sigma(A_n)$), quando $n \to +\infty$, e* 2) *il costo per passo di* GCP *con $P = P_n$ sia dell'ordine del costo per passo del metodo* GCP *con $P = I_n$, ovvero dell'ordine del costo del prodotto matrice $A_n \times$ vettore.* Si riesce a far ciò ad esempio sostituendo una base di tipo standard con una base gerarchica, nella rappresentazione della soluzione [49], come abbiamo accennato all'inizio di questa sezione.

La costruzione del precondizionatore P

Data $A \in \mathbb{R}^{n\times n}$ definita positiva, si vuole quindi introdurre una matrice $P \in \mathbb{R}^{n\times n}$ definita positiva tale che (1) lo spettro di $P^{-1}A$ sia distribuito più convenientemente di quello di A, (2) la presenza della matrice P nelle equazioni del metodo GCP non aumenti troppo il costo di ogni sua iterazione. Inoltre, vorremmo che $\mu_2(\tilde{A}) \not> \mu_2(A)$, essendo $\tilde{A} = E^{-1}AE^{-T}$ e $EE^T = P$, ossia che il sistema $P^{-1}A\mathbf{x} = P^{-1}\mathbf{b}$ non sia peggio condizionato del sistema $A\mathbf{x} = \mathbf{b}$ [42].

Esempio 1

Sia $A = 2I - X$ con

$$X = \begin{bmatrix} & 1 & & & \\ 1 & & 1 & & \\ & 1 & & \cdot & \\ & & \cdot & & 1 \\ & & & 1 & \end{bmatrix},$$

dove gli elementi non specificati sono zeri. Poiché $X(\sin\frac{ij\pi}{n+1})_{i=1}^n = 2\cos\frac{j\pi}{n+1}(\sin\frac{ij\pi}{n+1})_{i=1}^n$, $j = 1,\ldots,n$, gli autovalori di A sono: $2 - 2\cos\frac{j\pi}{n+1}$, $j = 1,\ldots,n$. Quindi A è definita positiva, gli autovalori di A sono distribuiti uniformemente nell'intervallo $(0,4)$ e il numero di condizionamento di A in norma 2 va a infinito come $O(n^2)$ (dimostrarlo!). Dunque, lo spettro di A non è favorevole e, di conseguenza, il calcolo di una buona approssimazione di $A^{-1}\mathbf{b}$ con GCP, $P = I$, potrebbe richiedere un gran numero di iterazioni.

Precondizionatore di Axelsson. In [41], definita la matrice E come

$$E = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \cdot & \cdot & \\ & & -1 & 1 \end{bmatrix},$$

si osserva che $\tilde{A} = E^{-1}AE^{-T}$,

$$\tilde{A} = \begin{bmatrix} 1 & & & \\ 1 & 1 & & \\ \cdot & \cdot & \cdot & \\ 1 & \cdot & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \cdot & \\ & \cdot & \cdot & -1 \\ & & -1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdot & 1 \\ & 1 & \cdot & \cdot \\ & & \cdot & 1 \\ & & & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 & \cdot & 1 \\ 1 & 2 & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 \\ 1 & \cdot & 1 & 2 \end{bmatrix} = I + \mathbf{e}\mathbf{e}^T,$$

è definita positiva e ha gli autovalori tutti uguali a 1 eccetto uno, $n+1$. Quindi, il numero di condizionamento di $\tilde{A}$ in norma 2 è $n+1$, cioè lineare in n, e il numero dei passi da effettuare con GCP, $P = EE^T$, per calcolare $A^{-1}\mathbf{b}$ è due. È inoltre semplice verificare che le operazioni in più per passo, dovute alla presenza di $P = EE^T$ nel metodo, sono $2(n-1)$ addizioni. Osserviamo infine che per $P = EE^T$ le condizioni (35), (36) sono soddisfatte (mentre non lo sono per $P = I$).

Diamo una possibile giustificazione della scelta di E in [41]. Siano

$$P = EE^T = \begin{bmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & \cdot & \cdot & \\ & & \cdot & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}, \quad X = \begin{bmatrix} 0 & 1 & & & \\ 1 & -1 & 1 & & \\ & 1 & \cdot & \cdot & \\ & & \cdot & -1 & 1 \\ & & & 1 & -1 \end{bmatrix}. \tag{37}$$

Si osserva che P è una matrice dello spazio vettoriale $\mathcal{L} = \{\alpha I + \beta X : \alpha, \beta \in \mathbb{R}\}$. Calcoliamo la matrice in $\mathcal{L}$ che meglio approssima A in norma di Frobenius:

$$\mathcal{L}_A = \alpha I + \beta X, \quad \alpha = 1 + 3\frac{1}{2n+1}, \ \beta = -1 + \frac{1}{2n+1}.$$

Ora notiamo che P si ottiene da $\mathcal{L}_A$ sostituendo il termine $\frac{1}{2n+1}$, in α e β, con 0. Tale sostituzione è giustificata pensando che $\mathcal{L}_A$ tende, in qualche senso, a P quando n tende a $+\infty$. Quindi P è l'approssimazione migliore possibile di A in $\mathcal{L}$ quando $n \to +\infty$, ovvero quando A ha dimensione semi-infinita.

Esercizio. Sia X la matrice in (37). Per quali α, β la quantità $\|\alpha I + \beta X - A\|_\infty = \max\{|\alpha - 2| + |\beta + 1|, |\alpha - \beta - 2| + 2|\beta + 1|\}$ è minima? Si noti che $\|\alpha I + \beta X - A\|_\infty = 1$ se $\alpha = 1$ e $\beta = -1$. Vale la disuguaglianza $\|\alpha I + \beta X - A\|_\infty \geq 1$, $\forall\, \alpha, \beta \in \mathbb{R}$?

Esercizio. Sia X la matrice in (37) e $\mathcal{L}$ l'insieme dei polinomi in X. Calcolare la matrice $\mathcal{L}_A$. $\mathcal{L}_A$ tende, in qualche senso, a P quando n tende a $+\infty$?

Precondizionatori circolanti di Strang e T. Chan. La seguente matrice circolante

$$P = \begin{bmatrix} 2 & c & & & c \\ c & 2 & c & & \\ & c & \cdot & \cdot & \\ & & \cdot & 2 & c \\ c & & & c & 2 \end{bmatrix}, \ c \in \mathbb{R}$$

(gli elementi non specificati sono zeri), potrebbe essere proposta come precondizionatore di A. Vedremo che per $c = -1$ essa non è altro che il precondizionatore circolante di A di Strang [51], [43], che ricalca la parte diagonale centrale della matrice A. Per $c = -1 + \frac{1}{n}$ essa coincide con il precondizionatore circolante di A di T. Chan [45], che minimizza $\|X - A\|_F$ al variare di X circolante. Osserviamo che P è definita positiva solo per $|c| < 1$, quindi il precondizionatore di Strang, in questo caso, a meno che non sia modificato opportunamente, è inutilizzabile perché non è una matrice definita positiva. Poiché il precondizionatore di T. Chan per $n \to +\infty$ viene a coincidere con quello di Strang, si può utilizzarlo solo per valori moderati di n, e, come è facilmente intuibile, anche per questi n non potrà essere molto efficiente. E infatti la matrice A non verifica le ipotesi del teorema di clustering sugli autovalori di $P^{-1}A$, A =Toeplitz e P =TChan, riportato più avanti, quindi non è detto che valga la proprietà (35) (e le (36), (34)?).

Esempio 2
Sia $A = (t^{|i-j|})_{i,j=1}^n$, con $t \in \mathbb{C}$. Vale la seguente identità:

$$A = \begin{bmatrix} 1 & t & \cdot & t^{n-1} \\ t & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & t \\ t^{n-1} & \cdot & t & 1 \end{bmatrix} = \begin{bmatrix} 1 & & & \\ t & 1 & & \\ \cdot & \cdot & \cdot & \\ t^{n-1} & \cdot & t & 1 \end{bmatrix} \begin{bmatrix} 1 & & & \\ & 1-t^2 & & \\ & & \cdot & \\ & & & 1-t^2 \end{bmatrix} \begin{bmatrix} 1 & t & \cdot & t^{n-1} \\ & 1 & \cdot & \cdot \\ & & \cdot & t \\ & & & 1 \end{bmatrix}.$$

Infatti

$$\begin{bmatrix} 1 & t & \cdot & t^{n-1} \\ t & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & t \\ t^{n-1} & \cdot & t & 1 \end{bmatrix} = \begin{bmatrix} 1 & & & \\ t & 1 & & \\ \cdot & \cdot & \cdot & \\ t^{n-1} & \cdot & t & 1 \end{bmatrix} + \begin{bmatrix} 1 & t & \cdot & t^{n-1} \\ & 1 & \cdot & \cdot \\ & & \cdot & t \\ & & & 1 \end{bmatrix} - I,$$

$$\begin{bmatrix} 1 & -t & & \\ & 1 & \cdot & \\ & & \cdot & -t \\ & & & 1 \end{bmatrix} \begin{bmatrix} 1 & t & \cdot & t^{n-1} \\ & 1 & \cdot & \cdot \\ & & \cdot & t \\ & & & 1 \end{bmatrix} = I.$$

Ne segue che

Se $t^2 = 1$, allora

$$A = \begin{bmatrix} 1 \\ (\pm 1) \\ \cdot \\ (\pm 1)^{n-1} \end{bmatrix} [1 \;\; (\pm 1) \;\; \cdot \;\; (\pm 1)^{n-1}]$$

è una matrice di rango 1 i cui autovalori sono 0 con molteplicità $n-1$, ed n.

Se $t^2 \neq 1$ allora A è non singolare e l'inversa di A è tridiagonale non Toeplitz (calcolarla!).

Se t è reale e $t^2 < 1$, allora A è definita positiva.

Supponiamo dunque t reale e $t^2 < 1$, in modo che A sia definita positiva, e studiamo l'efficienza dei precondizionatori P circolanti di Strang e di T. Chan di A, che denotiamo con i simboli P_{st}, P_{tc}. Entrambi verificano la proprietà (35). Nel caso del precondizionatore di Strang si riportano esplicitamente gli autovalori di $P_{st}^{-1}A$ $(\forall\, n)$ da cui sarà evidente la tesi (in effetti, sarà evidente che sono soddisfatte tutte e tre le condizioni (35), (36), (34)). Nel caso del precondizionatore di T. Chan il risultato (35) verrà dimostrato per una classe di matrici di Toeplitz di cui la nostra A è un esempio particolare (e le (34), (36)?). Si può inoltre affermare che ogni passo di GCP, con $P = P_{st}$, P_{tc}, ha un costo dell'ordine del costo di GCP, $P = I$, infatti, come sappiamo, la risoluzione di un sistema lineare con matrice dei coefficienti circolante non costa più del prodotto di una matrice di Toeplitz per un vettore.

Precondizionatore circolante di Strang. Sia n pari. Alla matrice A, $-1 < t < 1$, si può associare il precondizionatore circolante di Strang, ovvero la matrice P_{st} circolante con prima riga $[1\ t\ t^2 \cdots\ t^{\frac{n}{2}-1}\ t^{\frac{n}{2}}\ t^{\frac{n}{2}-1}\ \cdots t^2\ t]$. Ad esempio, per $n = 4$

$$A = \begin{bmatrix} 1 & t & t^2 & t^3 \\ t & 1 & t & t^2 \\ t^2 & t & 1 & t \\ t^3 & t^2 & t & 1 \end{bmatrix}, \quad P_{st} = \begin{bmatrix} 1 & t & t^2 & t \\ t & 1 & t & t^2 \\ t^2 & t & 1 & t \\ t & t^2 & t & 1 \end{bmatrix}.$$

Sorprendentemente tale matrice P_{st}, oltre ad essere definita positiva (condizione essenziale affinché P_{st} possa essere usata come precondizionatore), rende lo spettro di $P_{st}^{-1}A$ particolarmente favorevole. Si dimostra infatti che

$$\sigma(P_{st}^{-1}A) = \{1,\ \frac{1}{1 \pm t^{n/2}},\ \frac{1}{1 \pm t}\},\ m_a(1) = 2,\ m_a(\frac{1}{1 \pm t^{n/2}}) = \frac{n}{2} - 2,\ m_a(\frac{1}{1 \pm t}) = 1 \qquad (38)$$

(vedi [50] e la Proposizione 4.1 qui sotto). Quindi, lo spettro di $P_{st}^{-1}A$ si raggruppa su 1 ed è limitato, al variare di n, sia dal basso che dall'alto, cioè le condizioni (34), (35), (36) sono tutte e tre

soddisfatte. Inoltre, $\sigma(P_{st}^{-1}A)$ è costituito da soli cinque autovalori distinti. Questo equivale a dire che GCP, $P = P_{st}$, converge in al più cinque passi ($\forall\, n$), quando applicato per risolvere il sistema $A\mathbf{x} = \mathbf{b}$.

Proposizione 4.1 Se $t^2 < 1$, allora esistono vettori $\mathbf{x}_i^{\pm}$, $i = 1, \ldots, \frac{n}{2} - 2$, linearmente indipendenti tali che $A\mathbf{x}_i^{\pm} = \frac{1}{1\pm t^{n/2}} P_{st}\mathbf{x}_i^{\pm}$ e vettori $\mathbf{y}^{\pm}$ tali che $A\mathbf{y}^{\pm} = \frac{1}{1\pm t} P_{st}\mathbf{y}^{\pm}$ [50]. Inoltre, $A\mathbf{e}_{n/2} = P_{st}\mathbf{e}_{n/2}$ e $A\mathbf{e}_{n/2+1} = P_{st}\mathbf{e}_{n/2+1}$. I vettori $\mathbf{x}_i^{\pm}$, $\mathbf{y}^{\pm}$, $\mathbf{e}_{n/2}$, $\mathbf{e}_{n/2+1}$ costituiscono un insieme di n vettori linearmente indipendenti. Quindi, la matrice P_{st} è non singolare, vale il risultato (38), e P_{st} è definita positiva.

Dimostrazione. Se $A\mathbf{z} = \lambda B\mathbf{z}$, $\mathbf{z} \neq \mathbf{0}$, e $A\mathbf{w} = \mu B\mathbf{w}$, $\mathbf{w} \neq \mathbf{0}$, con A non singolare, $\lambda \neq 0$, $\mu \neq 0$ e $\lambda \neq \mu$, allora necessariamente $\mathbf{z}$ e $\mathbf{w}$ devono essere linearmente indipendenti (infatti si avrebbe $\frac{1}{\lambda}\mathbf{z} = A^{-1}B\mathbf{z}$ e $\frac{1}{\mu}\mathbf{w} = A^{-1}B\mathbf{w}$, quindi $\mathbf{z}$ e $\mathbf{w}$ sarebbero autovettori della stessa matrice corrispondenti ad autovalori distinti!). Quindi X, la matrice con colonne i vettori $\mathbf{x}_i^{\pm}$, $\mathbf{y}^{\pm}$, $\mathbf{e}_{n/2}$, $\mathbf{e}_{n/2+1}$, è non singolare. Inoltre vale l'identità $AX = P_{st}XD$, con $D = \operatorname{diag}(\lambda_i)$ dove i λ_i sono scelti opportunamente nell'insieme $\{1,\ 1/(1\pm t^{n/2}),\ 1/(1\pm t)\}$. Ne segue che $\det(P_{st}) \neq 0$, $P_{st}^{-1}AX = XD$, e, di conseguenza, i λ_i sono gli autovalori di $P_{st}^{-1}A$ e le colonne di X sono i corrispondenti autovettori. Inoltre, poiché $(P_{st}^{-1}A)^{-1} = A^{-1}P_{st} = L^{-T}L^{-1}P_{st} = L^{-T}(L^{-1}P_{st}L^{-T})L^T$, gli autovalori di $L^{-1}P_{st}L^{-T}$ sono semplicemente gli inversi dei λ_i e dunque sono positivi. Quindi $L^{-1}P_{st}L^{-T}$ è definita positiva e anche P_{st} è definita positiva. In particolare, $P_{st}^{-1}A$ ed $\tilde{A} = E^{-1}AE^{-T}$, $EE^T = P_{st}$, hanno gli stessi autovalori. □

Precondizionatore circolante di T. Chan. La matrice circolante P_{tc} che meglio approssima $A = (t^{|i-j|})_{i,j=1}^n$ ha la seguente espressione

$$P_{tc} = \frac{1}{n}\sum_{k=1}^{n}((n-k+1)t^{k-1} + (k-1)t^{n-k+1})\Pi^{k-1},$$

dove Π è la matrice circolante con prima riga $\mathbf{e}_2^T$. Ad esempio, per $n = 4$

$$A = \begin{bmatrix} 1 & t & t^2 & t^3 \\ t & 1 & t & t^2 \\ t^2 & t & 1 & t \\ t^3 & t^2 & t & 1 \end{bmatrix}, \quad P_{st} = \begin{bmatrix} 1 & \frac{3t+t^3}{4} & t^2 & \frac{3t+t^3}{4} \\ \frac{3t+t^3}{4} & 1 & \frac{3t+t^3}{4} & t^2 \\ t^2 & \frac{3t+t^3}{4} & 1 & \frac{3t+t^3}{4} \\ \frac{3t+t^3}{4} & t^2 & \frac{3t+t^3}{4} & 1 \end{bmatrix}.$$

La matrice P_{tc} è, come P_{st}, un buon precondizionatore di A. Per i dettagli si rimanda all'esempio seguente dove si considera una generica matrice di Toeplitz simmetrica $A = (t_{|i-j|})_{i,j=1}^n$ e, sotto opportune ipotesi su A — soddisfatte quando $A = (t^{|i-j|})_{i,j=1}^n$ —, si dimostra che $P_{tc}^{-1}A$ soddisfa la proprietà (35), e quindi si prova l'efficienza di GCP, $P = P_{tc}$, nella risoluzione del sistema $A\mathbf{x} = \mathbf{b}$.

Esempio 3 (A = Toeplitz generica)

Sia $A = (t_{|i-j|})_{i,j=1}^n$, dove $t_0, \ldots, t_{n-1}$ sono n parametri reali tali che A è definita positiva. Sia $\phi = \pm 1$. Alla matrice A possiamo associare la matrice ϕ-circolante di Strang (Huckle), P_{st}, definita come la matrice ϕ-circolante la cui prima riga è, per n pari,

$$[t_0\ t_1\ t_2\ \cdots\ t_{\frac{n}{2}-1}\ \ t\ \ \frac{1}{\phi}t_{\frac{n}{2}-1}\ \cdots \frac{1}{\phi}t_2\ \frac{1}{\phi}t_1], \quad t = \begin{cases} 0 & \phi = -1 \\ t_{\frac{n}{2}} & \phi = 1 \end{cases},$$

e, per n dispari,

$$[t_0\ t_1\ t_2\ \cdots\ t_{\frac{n-1}{2}}\ \frac{1}{\phi}t_{\frac{n-1}{2}}\ \cdots \frac{1}{\phi}t_2\ \frac{1}{\phi}t_1].$$

La matrice P_{st} è reale simmetrica come A. Si è visto che P_{st} può non essere definita positiva, nonostante lo sia A (vedi l'esempio 1, dove P_{st} è semidefinita positiva), come può invece essere, oltre che definita positiva, particolarmente efficiente come precondizionatore di A, per certe A (per $P = P_{st}$ le proprietà (34), (35), (36) sono tutte e tre soddisfatte nell'esempio 2). Qui si omette lo studio di P_{st} come precondizionatore di matrici di Toeplitz A simmetriche generiche (vedi [51], [43], [50], [48]). Ci si limita a dire che vale un risultato analogo a quello enunciato nel Teorema clustering per il precondizionatore di T. Chan.

Ad A possiamo anche associare la matrice ϕ-circolante di T. Chan, P_{tc}, definita come la matrice ϕ-circolante che realizza il minimo:

$$\|P_{tc} - A\|_F = \min\big\{\|X - A\|_F : X\ \phi\text{-circolanti}\big\}.$$

Tale matrice per $\phi = 1$ è stata proposta come precondizionatore di matrici di Toeplitz simmetriche da T. Chan in [45]. Successivamente per $\phi = -1$ è stata studiata da Huckle [44].

Utilizzando l'espressione (22) si ottiene una formula per P_{tc}, valida per ogni scelta di ϕ e per A più generale, $A = (t_{i-j})_{i,j=1}^n$, $t_k \in \mathbb{C}$. Sia Π_ϕ la matrice ϕ-circolante con prima riga $\mathbf{e}_2^T$. Allora

$$\begin{aligned} P_{tc} = (\mathcal{C}_\phi)_A &= \sum_{j=1}^{n} \frac{1}{n-j+1+|\phi|^2(j-1)}((n-j+1)t_{-j+1} + (j-1)\overline{\phi}t_{n-j+1})\Pi_\phi^{j-1} \\ &= \begin{bmatrix} t_0 & \frac{(n-1)t_{-1}+\overline{\phi}t_{n-1}}{n-1+|\phi|^2} & \cdot & \frac{t_{-n+1}+\overline{\phi}(n-1)t_1}{1+|\phi|^2(n-1)} \\ \phi\frac{t_{-n+1}+\overline{\phi}(n-1)t_1}{1+|\phi|^2(n-1)} & & & \\ \cdot & & & \\ \phi\frac{(n-1)t_{-1}+\overline{\phi}t_{n-1}}{n-1+|\phi|^2} & & & \end{bmatrix}. \end{aligned}$$

È evidente da tale formula che se A è simmetrica, allora anche $P_{tc} = (\mathcal{C}_\phi)_A$ è simmetrica, a condizione che $\phi \in \mathbb{R}$, $|\phi| = 1$, ovvero che $\phi = \pm 1$. Analogamente, se A è hermitiana (definita positiva), allora, se ϕ è tale che $|\phi| = 1$, anche $P_{tc} = (\mathcal{C}_\phi)_A$ è hermitiana (definita positiva) e, da quanto visto nella Sezione 2.6, vale l'inclusione $\overline{\sigma((\mathcal{C}_\phi)_A)} \subset \overline{\sigma(A)}$.

La nostra matrice A di Toeplitz è reale, simmetrica e definita positiva, quindi, per quanto detto, se $\phi = \pm 1$, anche la matrice $P_{tc} = (\mathcal{C}_\phi)_A$,

$$\begin{aligned} P_{tc} = (\mathcal{C}_{\pm 1})_A &= \sum_{j=1}^{n} \frac{1}{n}((n-j+1)t_{j-1} + (j-1)(\pm 1)t_{n-j+1})\Pi_{\pm 1}^{j-1} \\ &= \begin{bmatrix} t_0 & \frac{(n-1)t_1+(\pm 1)t_{n-1}}{n} & \cdot & \frac{t_{n-1}+(\pm 1)(n-1)t_1}{n} \\ (\pm 1)\frac{t_{n-1}+(\pm 1)(n-1)t_1}{n} & & & \\ \cdot & & & \\ (\pm 1)\frac{(n-1)t_1+(\pm 1)t_{n-1}}{n} & & & \end{bmatrix} \end{aligned}$$

è reale, simmetrica e definita positiva.

Esercizio. Trovare la matrice circolante più vicina in norma di Frobenius a una matrice 4×4 simmetrica di Toeplitz $A = (t_{|i-j|})_{i,j=1}^4$, e chiamarla $P_{tc} = \mathcal{L}_A$, $\mathcal{L} = \mathcal{C}$. Scrivere $P_{tc} = \mathcal{L}_A$ nel caso particolare $t_0 = 2$, $t_1 = -1$, $t_2 = t_3 = 0$. Estendere il risultato al caso $n \times n$, notando che la prima riga di $P_{tc} = \mathcal{L}_A$ può essere calcolata con $O(n)$ operazioni aritmetiche. Ripetere tutto per $\mathcal{L} = \mathcal{C}_{-1}$.
Risoluzione. La matrice $P_{tc} = \mathcal{L}_A$ è la matrice circolante la cui prima riga è

$$[t_0 \quad (3t_1 + t_3)/4 \quad t_2 \quad (3t_1 + t_3)/4].$$

Quindi, nel caso particolare,

$$P_{tc} = \mathcal{L}_A = \frac{3}{4} \begin{bmatrix} 8/3 & -1 & 0 & -1 \\ -1 & 8/3 & -1 & 0 \\ 0 & -1 & 8/3 & -1 \\ -1 & 0 & -1 & 8/3 \end{bmatrix}.$$

L'ultima affermazione è evidente dalla formula generale per $\mathcal{C}_A$, $A = (t_{|i-j|})_{i,j=1}^n$, trovata prima dell'esercizio.

Esercizio. Trovare la matrice dell'algebra τ più vicina (in norma di Frobenius) a una matrice 4×4 simmetrica di Toeplitz $A = (t_{|i-j|})_{i,j=1}^4$ e chiamarla $P_\tau = \tau_A$. Scrivere τ_A nel caso particolare $t_0 = 2$, $t_1 = -1$, $t_2 = t_3 = 0$. Estendere il risultato al caso $n \times n$, notando che la prima riga di τ_A può essere calcolata con $O(n)$ operazioni aritmetiche.

Esercizio. Trovare le matrici $\mathcal{L}_A$ delle algebre di tipo Hartley $\mathcal{L} = \mathcal{C}^S + J\Pi\mathcal{C}^{SK}$, $\mathcal{C}^s + J\mathcal{C}^S$ più vicine (in norma di Frobenius) a una matrice 4×4 simmetrica di Toeplitz $A = (t_{|i-j|})_{i,j=1}^4$. Estendere il risultato al caso $n \times n$, notando che la prima riga di $\mathcal{L}_A$ può essere calcolata con $O(n)$ operazioni aritmetiche.

Prima di proseguire con lo studio della matrice di T. Chan (Huckle) $(\mathcal{C}_{\pm 1})_A$ nel caso di una particolare classe di matrici A di Toeplitz simmetriche reali, osserviamo che essa e, più in generale, ogni matrice $\mathcal{L}_A$, con $\mathcal{L} = \{Ud(\mathbf{z})U^H : \mathbf{z} \in \mathbb{C}^n\}$, U unitaria, potrebbe in principio essere proposta come precondizionatore del sistema $A\mathbf{x} = \mathbf{b}$ per ogni A definita positiva (quindi non necessariamente di Toeplitz). Ad esempio, è sempre vero che $\mathcal{L}_A$ è definita positiva e che lo spettro di $\mathcal{L}_A$ è contenuto in quello di A, cioè $\overline{\sigma(\mathcal{L}_A)} \subset \overline{\sigma(A)}$; quindi in particolare è sempre ben definito GCP con $P = \mathcal{L}_A$, come metodo per la risoluzione del sistema $A\mathbf{x} = \mathbf{b}$. Unica accortezza, quando A è reale, scegliere una $\mathcal{L}$ generata da matrici reali in modo che anche $\mathcal{L}_A$ sia reale (ad es. $\mathcal{C}_\phi$ con $|\phi| = 1$, $\phi \neq \pm 1$ produrrebbe una $(\mathcal{C}_\phi)_A$ non reale). Questa applicabilità della matrice $\mathcal{L}_A$ come possibile precondizionatore di A anche per A non Toeplitz, ha permesso ad esempio di risolvere iterativamente i sistemi lineari di Toeplitz non simmetrici $M\mathbf{x} = \mathbf{f}$, $M_{ij} = t_{i-j}$, applicando GCP con $P = \mathcal{L}_{M^H M}$ ai corrispondenti sistemi normali $M^H M\mathbf{x} = M^H\mathbf{f}$, ovvero ponendo $A = M^H M$, $\mathbf{b} = M^H\mathbf{f}$, $P = \mathcal{L}_A$ in GCP. Si vedano i lavori [53] per $\mathcal{L} = \mathcal{C}_\phi$, [54] per $\mathcal{L}$ di tipo Jacobi, e [6] per $\mathcal{L}$ di tipo Hartley.

Sullo spettro di una classe di matrici reali simmetriche di Toeplitz; precondizionamento di tipo T.Chan di tali matrici nelle algebre $\mathcal{C}_{\pm 1}$; il Teorema di clustering

Siano $\{t_k\}_{k=0}^{+\infty}$, $t_k \in \mathbb{R}$, una successione di numeri reali tale che $\sum_{k=0}^{+\infty} |t_k| < +\infty$, e $t(\theta)$ la

corrispondente funzione simbolo,

$$t(\theta) = \sum_{k=-\infty}^{+\infty} t_{|k|} e^{\mathbf{i}k\theta} = t_0 + 2\sum_{k=1}^{+\infty} t_k \cos(k\theta).$$

Osserviamo che $t(\theta)$ è una funzione continua, pari e di periodo 2π. Quindi sono ben definiti $t_{\min}$ e $t_{\max}$, rispettivamente il minimo e il massimo valore assunto da $t(\theta)$ in $\mathbb{R}$ (ovvero in $[-\pi, \pi]$). Il generico elemento della successione $\{t_k\}_{k=0}^{+\infty}$ può essere rappresentato in termini di $t(\theta)$ tramite la seguente formula:

$$t_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} t(\theta) e^{-\mathbf{i}k\theta} d\theta.$$

Infatti,

$$t(\theta)e^{-\mathbf{i}j\theta} = \sum_{k\in\mathbb{Z}} t_{|k|} e^{\mathbf{i}(k-j)\theta} \quad \Rightarrow \quad \int_{-\pi}^{\pi} t(\theta)e^{-\mathbf{i}j\theta} d\theta = \sum_{k\in\mathbb{Z}} t_{|k|} \int_{-\pi}^{\pi} e^{\mathbf{i}(k-j)\theta} d\theta \ = 2\pi t_j.$$

Per ogni n, sia ora A_n la matrice $n \times n$ reale simmetrica di Toeplitz $A_n = (t_{|i-j|})_{i,j=1}^n$. Si possono avere informazioni sullo spettro di A_n studiando la funzione $t(\theta)$, infatti vale la seguente

Proposizione 4.2 Sia data la successione $\{t_k\}_{k=0}^{+\infty}$, $\sum_{k=0}^{+\infty} |t_k| < +\infty$, e la corrispondente successione di matrici $\{A_n\}_{n\in\mathbb{N}}$, $A_n = (t_{|i-j|})_{i,j=1}^n$. L'intervallo $[t_{\min}, t_{\max}]$ contiene gli autovalori di A_n, cioè $\sigma(A_n) \subset [t_{\min}, t_{\max}]$ per ogni valore di n, e l'insieme $\bigcup_n \sigma(A_n)$ è denso in $[t_{\min}, t_{\max}]$. Se $t_{\min} > 0$, allora le matrici A_n sono definite positive per ogni n, e $\mu_2(A_n) \le t_{\max}/t_{\min}$. Se $t_{\min} = 0$, sotto opportune ipotesi sull'insieme $\{\theta : t(\theta) = 0\}$ si ha ancora che le matrici A_n sono definite positive per ogni n, ma $\mu_2(A_n)$ va a $+\infty$ (più o meno velocemente a seconda di come è fatto l'insieme $\{\theta : t(\theta) = 0\}$).

Dimostrazione. Per ogni $\mathbf{z} \in \mathbb{C}^n$,

$$\begin{aligned}
\mathbf{z}^H A_n \mathbf{z} &= \sum_{k,j} \overline{z}_k \big[\frac{1}{2\pi} \int_{-\pi}^{\pi} t(\theta) e^{-\mathbf{i}(k-j)\theta} d\theta\big] z_j = \frac{1}{2\pi} \int_{-\pi}^{\pi} (\sum_k \overline{z}_k e^{-\mathbf{i}k\theta}) t(\theta) (\sum_j z_j e^{\mathbf{i}j\theta}) d\theta \\
&\le t_{\max} \frac{1}{2\pi} \int_{-\pi}^{\pi} (\sum_k \overline{z}_k e^{-\mathbf{i}k\theta})(\sum_j z_j e^{\mathbf{i}j\theta}) d\theta = t_{\max} \frac{1}{2\pi} \sum_{k,j} \overline{z}_k z_j \int_{-\pi}^{\pi} e^{\mathbf{i}(j-k)\theta} d\theta = t_{\max} \mathbf{z}^H \mathbf{z}.
\end{aligned}$$

Quindi, $t_{\min}\mathbf{z}^H\mathbf{z} \le \mathbf{z}^H A\mathbf{z} \le t_{\max}\mathbf{z}^H\mathbf{z}$, $\forall \mathbf{z} \in \mathbb{C}^n$, e, in particolare, si ha $\overline{\sigma(A)} \in [t_{\min}, t_{\max}]$. □

Applichiamo la Proposizione 4.2, calcolando la funzione $t(\theta)$, nel caso degli esempi 1 e 2.

Nell'esempio 1 si ha $t_0 = 2$, $t_1 = -1$, e $t_k = 0$ per ogni $k > 1$. Allora $t(\theta) = 2 - 2\cos\theta$, e quindi $t(\theta)$ è minima per $\theta = 0$ e $t(0) = 0$, ed è massima per $\theta = \pm\pi$ e $t(\pm\pi) = 4$. Per la Proposizione 4.2, per ogni n si ha $\sigma(A_n) \subset [0, 4]$ e, per come è fatto l'insieme $\{\theta : t(\theta) = 0\}$, si ha $\sigma(A_n) \subset (0, 4)$ (tali informazioni ci venivano anche dal teorema di Gershgorin), inoltre $\mu_2(A_n) \to +\infty$. Effettivamente $\sigma(A_n) = \{2 - 2\cos\frac{j\pi}{n+1} : j = 1, \ldots, n\}$, e, quindi, $\mu_2(A_n) = O(n^2)$.

Invece nell'esempio 2 si ha $t_k = t^k$ con t parametro tale che $-1 < t < 1$. Quindi la funzione $t(\theta)$ è data da

$$\begin{aligned} t(\theta) &= \sum_{k\in\mathbb{Z}} t^{|k|} e^{\mathbf{i}k\theta} = \sum_{k=0}^{+\infty} t^k e^{\mathbf{i}k\theta} + \sum_{k=1}^{+\infty} t^k e^{-\mathbf{i}k\theta} \\ &= \frac{1}{1-te^{\mathbf{i}\theta}} + \frac{1}{1-te^{-\mathbf{i}\theta}} - 1 = \frac{1-t^2}{1+t^2-2t\cos\theta}. \end{aligned}$$

È semplice individuare i valori massimo e minimo di $t(\theta)$, infatti

$$0 < \frac{1-|t|}{1+|t|} \le t(\theta) \le \frac{1+|t|}{1-|t|}.$$

Ne segue che, per ogni n, la matrice $A_n = (t^{|i-j|})_{ij=1}^n$ è definita positiva, il suo spettro è contenuto nell'intervallo $\big[(1-|t|)/(1+|t|), (1+|t|)/(1-|t|)\big]$, e quindi $\mu_2(A_n) \le (1+|t|)^2/(1-|t|)^2$. Si noti che, stavolta, con il teorema di Gershgorin non avremmo potuto ottenere tali informazioni sullo spettro delle matrici A_n.

Enunciamo ora il seguente risultato che, in pratica, afferma l'esistenza di buoni precondizionatori $\mathcal{A}_n$ per le matrici A_n (quando queste sono definite positive), ovvero l'esistenza di matrici $\mathcal{A}_n$ che verificano la condizione (35) su P_n e quindi tali che GCP con $A = A_n$, $P = \mathcal{A}_n$ ha convergenza pressoché superlineare nella risoluzione del sistema $A_n\mathbf{x} = \mathbf{b}_n$.

Teorema 4.3 Sia data la successione $\{t_k\}_{k=0}^{+\infty}$, $\sum_{k=0}^{+\infty} |t_k| < +\infty$, e la corrispondente successione di matrici $\{A_n\}_{n\in\mathbb{N}}$, $A_n = (t_{|i-j|})_{i,j=1}^n$. Allora esiste una successione di matrici reali simmetriche $\{\mathcal{A}_n\}_{n\in\mathbb{N}}$ tali che

i) Se A_n e $\mathcal{A}_n$ sono non singolari allora risolvere un sistema lineare con $\mathcal{A}_n$ come matrice dei coefficienti è molto meno costoso che risolvere un sistema lineare con A_n come matrice dei coefficienti;

ii) Per ogni $\varepsilon > 0$ esistono ν_ε, k_ε tali che $\forall\, n > \nu_\varepsilon$ almeno $n - k_\varepsilon$ autovalori di $A_n - \mathcal{A}_n$ non distano da 0 più di ε.

iii) Se $t(\theta) = \sum_{k=-\infty}^{+\infty} t_{|k|} e^{\mathbf{i}k\theta} > 0$ (in $[-\pi, \pi]$), allora $\mathcal{A}_n$ è definita positiva e l'affermazione ii) vale anche per $I - \mathcal{A}_n^{-1} A_n$.

Per dimostrare il Teorema esibiremo semplicemente una particolare sequenza $\mathcal{A}_n$ che verifica i risultati enunciati, la sequenza $\mathcal{A}_n = \mathcal{C}_{A_n}$ di T. Chan. È però prima opportuno osservare che le conclusioni del Teorema sono vere anche per diverse altre particolari sequenze $\mathcal{A}_n$ e che alcune di tali conclusioni, in realtà, sono verificate per classi molto generali di sequenze $\mathcal{A}_n$.

In particolare, in [51], [43], [57] si mostrano le tesi del Teorema 4.3 per $\mathcal{A}_n$ definita $(\forall\, n)$ come la matrice circolante la cui parte diagonale centrale è identica alla parte diagonale centrale di A_n (vedi la matrice P_{st} definita in precedenza, associata ad ogni matrice reale simmetrica di Toeplitz). Osserviamo che nel caso dell'esempio 1 si ha che $\sum_{k=0}^{+\infty} |t_k| = 3 < +\infty$, ma $t(\theta) = 2 - 2\cos\theta$ non è positiva in $[-\pi, \pi]$, quindi l'ipotesi $t(\theta) > 0$ non è soddisfatta, ecco perché, pur essendo A_n definita positiva, la matrice $\mathcal{A}_n$ non viene definita positiva (gli autovalori di $\mathcal{A}_n$ in tal caso sono $2 - 2\cos\frac{2\pi(j-1)}{n}$, $j = 1, \ldots, n$, quindi $\mathcal{A}_n$ è solo semi definita positiva).

Le tesi del Teorema si possono poi mostrare per $\mathcal{A}_n$ scelta come la migliore approssimazione $\mathcal{L}_{A_n}$ di A_n in $\mathcal{L} = \mathcal{L}^{(n)} = \mathrm{sd}\, U_n = \{U_n d(\mathbf{z}) U_n : \mathbf{z} \in \mathbb{C}^n\}$, con U_n unitaria, a condizione che le matrici $\{U_n\}_{n\in\mathbb{N}}$ siano scelte in modo opportuno.

Si noti che, indipendentemente dalla scelta di U_n e, quindi, dell'algebra $\mathcal{L} = \mathrm{sd}\, U_n$, la matrice $\mathcal{L}_{A_n} = U_n \,\mathrm{diag}\,((U_n^H A_n U_n)_{ii}) U_n^H$ è hermitiana. Quindi, nell'ipotesi che $\mathcal{L}$ sia generata da matrici reali, $\mathcal{L}_{A_n}$ è reale simmetrica, come A_n. Inoltre $\overline{\sigma(\mathcal{L}_{A_n})} \subset \overline{\sigma(A_n)} \subset [t_{\min}, t_{\max}]$ per ogni valore di n. Quindi, nell'ipotesi $t_{\min} > 0$, le matrici A_n e dunque le $\mathcal{L}_{A_n}$ sono definite positive per ogni n, e $\mu_2(\mathcal{L}_{A_n}) \le \mu_2(A_n) \le t_{\max}/t_{\min}$, $\forall\, n$. Nel caso in cui $t_{\min} = 0$ e sotto opportune ipotesi su $\{\theta : t(\theta) = 0\}$ si ha ancora che le matrici A_n e quindi le $\mathcal{L}_{A_n}$ sono definite positive per ogni n e che $\mu_2(\mathcal{L}_{A_n}) \le \mu_2(A_n)$, ma $\mu_2(A_n)$ va a $+\infty$.

Una scelta opportuna di U_n, e, quindi, di $\mathcal{L} = \mathrm{sd}\, U_n$, è invece necessaria per ottenere le affermazioni centrali i) e ii) del teorema. Per ottenere i) occorre scegliere come $\mathcal{L}$ una algebra di bassa complessità computazionale, in modo che il costo della risoluzione di un sistema $M\mathbf{z} = \mathbf{f}$, $M \in \mathcal{L}$ – ad esempio $M = \mathcal{L}_{A_n}$ –, sia $O(n \log n)$. Per ottenere ii) occorre che la struttura delle matrici di $\mathcal{L}$ sia in qualche modo vicina alla struttura di Toeplitz delle matrici A_n. Il risultato ii) è ottenuto in [47], [5] nel caso di algebre $\mathcal{L}$ di tipo Hartley, e in [36], [46], [56] per $\mathcal{L}$ di tipo Jacobi. Tuttavia, è per $\mathcal{L} = \mathcal{C}_\phi$ che si è notato per la prima volta che poteva valere un risultato tipo ii) [43], [57], [58]. Si vedano anche i survey [55], [48].

Notiamo infine che una volta dimostrata la ii) per $\mathcal{A}_n = \mathcal{L}_{A_n}$, l'affermazione iii) per $\mathcal{A}_n = \mathcal{L}_{A_n}$ segue immediatamente dalla disuguaglianza ottenuta qui di seguito. Sia E_n tale che $E_n E_n^T = \mathcal{L}_{A_n}$, e $\alpha_j^{(n)}$ e $\beta_j^{(n)}$ rispettivamente gli autovalori di $I - E_n^{-1} A_n E_n^{-T}$ e $\mathcal{L}_{A_n} - A_n$ in ordine non decrescente. Allora

$$\frac{1}{t_{\max}}|\beta_j^{(n)}| \le \frac{1}{\max\lambda(\mathcal{L}_{A_n})}|\beta_j^{(n)}| \le |\alpha_j^{(n)}| \le \frac{1}{\min\lambda(\mathcal{L}_{A_n})}|\beta_j^{(n)}| \le \frac{1}{t_{\min}}|\beta_j^{(n)}| \tag{39}$$

(si applichi la caratterizzazione minimax di Courant-Fisher degli autovalori di una matrice reale simmetrica [42] alla matrice $I - E_n^{-1} A_n E_n^{-T}$).

Dimostrazione. Proviamo le affermazioni ii) e iii) del Teorema 4.3 per $\mathcal{A}_n = \mathcal{C}_{A_n}$. Per semplicità poniamo $A = A_n$. Si prenda un numero N, $n > 2N$, e siano $W^{(N)}$ e $E^{(N)}$ le matrici $n \times n$ definite come segue

$$[W^{(N)}]_{ij} = \begin{cases} [\mathcal{C}_A - A]_{ij} & i, j \le n - N \\ 0 & \text{altrimenti} \end{cases}$$

e

$$\mathcal{C}_A - A = E^{(N)} + W^{(N)}. \tag{40}$$

Si noti che $[\mathcal{C}_A]_{1j} = ((n-j+1)t_{j-1} + (j-1)t_{n-j+1})/n$, $j = 1, \ldots, n$; quindi, per $i, j = 1, \ldots, n$, abbiamo

$$[\mathcal{C}_A - A]_{ij} = -\frac{s_{|i-j|}|i-j|}{n},\ s_k = t_k - t_{n-k}.$$

Ora si osservi che il rango di $E^{(N)}$ è minore o uguale di $2N$, quindi $E^{(N)}$ ha almeno $n-2N$ autovalori nulli. Si osservi anche che $\mathcal{C}_A - A$, $E^{(N)}$ e $W^{(N)}$ sono tutte matrici reali simmetriche. Nel seguito dimostriamo che comunque fissato $\varepsilon > 0$, esistono N_ε e $\nu_\varepsilon \ge 2N_\varepsilon$ tali che

$$\|W^{(N_\varepsilon)}\|_1 < \varepsilon \quad \forall\, n > \nu_\varepsilon. \tag{41}$$

Come consequenza di questo fatto e dell'uguaglianza (40) per $N = N_\varepsilon$, avremo che per tutti gli $n > \nu_\varepsilon$ almeno $n - 2N_\varepsilon$ autovalori di $\mathcal{C}_A - A$ sono in $(-\varepsilon, \varepsilon)$. Inoltre, se $t_{\min} > 0$, allora, per (39), otterremo anche il cluster su 0 degli autovalori di $I - \mathcal{C}_A^{-1} A$.

Quindi dimostriamo (41). Prima di tutto si ha che

$$\|W^{(N)}\|_1 \leq \frac{2}{n} \sum_{j=1}^{n-N-1} j|s_j| \leq 2 \sum_{j=N+1}^{n-1} |t_j| + \frac{2}{n} \sum_{j=1}^{N} j|t_j|. \tag{42}$$

Poi, per ogni $\varepsilon > 0$ si scelga N_ε tale che $2 \sum_{j=N_\varepsilon+1}^{+\infty} |t_j| < \frac{\varepsilon}{2}$ e si ponga $N = N_\varepsilon$ in (42) e nei precedenti argomenti. Se ν_ε, $\nu_\varepsilon \geq 2N_\varepsilon$, è tale che, $\forall\, n > \nu_\varepsilon$, $\frac{2}{n} \sum_{j=1}^{N_\varepsilon} j|t_j| < \frac{\varepsilon}{2}$ (la successione $\frac{1}{n} \sum_{j=1}^{n-1} j|t_j|$ tende a 0 nell'ipotesi $\sum_{k=0}^{+\infty} |t_k| < +\infty$), allora per (42) abbiamo la tesi (41). □

Esercizio. Nelle ipotesi del Teorema 4.3 provare l'affermazione ii) per $A_n = (\mathcal{C}_{-1})_{A_n}$.

Esercizio [47], [5]. Nelle ipotesi del Teorema 4.3 provare l'affermazione ii) per $A_n = \mathcal{L}_{A_n}$, con $\mathcal{L} = \mathcal{C}^S + J\Pi\mathcal{C}^{SK}$, $\mathcal{C}^S + J\mathcal{C}^S$.

References

[1] P. J. Davis, *Circulant matrices*, Wiley, New York, 1979

[2] P. D. Gader, Displacement operator based decompositions of matrices using circulants or other group matrices, *Linear Algebra Appl.*, 139 (1990), pp.111–131

[3] E. Bozzo, Algebras of higher dimension for displacement decompositions and computations with Toeplitz plus Hankel matrices, *Linear Algebra Appl.*, 230 (1995), pp.127–150

[4] P. Lancaster, M. Tismenetsky, *The Theory of Matrices, second edition with applications*, Computer Science and Applied Mathematics, Academic Press, Orlando, Florida, 1985 (pp. 416–420)

[5] C. Di Fiore, P. Zellini, Matrix algebras in optimal preconditioning, *Linear Algebra Appl.*, 335 (2001), pp.1–54

[6] C. Di Fiore, S. Fanelli, P. Zellini, On the best least squares fit to a matrix and its applications, in *Algebra and Algebraic Topology*, ed. Ched E. Stedman, Nova Science Publishers Inc., pp.73–109, 2006

[7] R. Bevilacqua, C. Di Fiore, P. Zellini, h-Space structure in matrix displacement formulas, *Calcolo*, 33 (1996), pp.11–35

[8] C. Di Fiore, F. Tudisco, P. Zellini, Bernoulli, Ramanujan, Toeplitz and the triangular matrices, submitted, December 2012

[9] C. Di Fiore, S. Fanelli, P. Zellini, Low-complexity minimization algorithms, *Numerical Linear Algebra with Applications*, 12 (2005), pp.755–768

[10] G. S. Ammar, W. B. Gragg, Superfast solution of real positive definite Toeplitz systems, *SIAM J. Matrix Anal. Appl.*, 9 (1988), pp.61–76

[11] A. Bortoletti, C. Di Fiore, On a set of matrix algebras related to discrete Hartley-type transforms, *Linear Algebra and its Aplications*, 366 (2003), pp.65–85

[12] E. Bozzo, C. Di Fiore, On the use of certain matrix algebras associated with discrete trigonometric transforms in matrix displacement decomposition, *SIAM J. Matrix Anal. Appl.*, 16 (1995), pp.312–326

[13] C. Di Fiore, P. Zellini, Matrix decompositions using displacement rank and classes of commutative matrix algebras, *Linear Algebra and its Aplications*, 229 (1995), pp.49–99

[14] C. Di Fiore, P. Zellini, Matrx displacement decompositions and applications to Toeplitz linear systems, *Linear Algebra Appl.*, 268 (1998), pp.197–225

[15] C. Di Fiore, Matrix algebras and displacement decompositions, *SIAM J. Matrix Anal. Appl.*, 21 (2000), pp.646–667

[16] Dario Bini, Private comunication, February 2012

[17] D. H. Lehmer, The Graeffe process as applied to power series, *Math. Comp.*, 1 (1945), pp.377–383; Corrigendum: *Math. Comp.*, 3 (1948), 227

[18] T. S. Caley, *A review of the von Staudt Clausen theorem*, Master in Science thesis, Dalhousie University, Halifax, Nova Scotia, 2007

[19] S. S. Wagstaff Jr., Prime divisors of the Bernoulli and Euler numbers, in *Number theory for the millennium*, III (Urbana, IL, 2000), pages 357-374. A K Peters, Natick, MA, 2002

[20] B. C. Kellner, On a conjecture about numerators of the Bernoulli numbers, http://arxiv.org/pdf/math/0410297.pdf, 2004

[21] W. Stein, K. McGown, Computing Bernoulli numbers,
http://modular.math.washington.edu/talks/bernoulli/current.pdf, 2006

[22] T. M. Apostol, Another elementary proof of Euler's formula for $\zeta(2n)$, *The American Mathematical Monthly*, Vol. 80, No. 4 (Apr., 1973), pp.425–431

[23] Edwards H. M., *Riemann's Zeta Function*, Pure and Applied Mathematics (series), Academic Press, New York, 1974

[24] S. Ramanujan, Some properties of Bernoulli numbers (J. Indian Math. Soc. 3 (1911), 219–234), in *Collected Papers of Srinivasa Ramanujan*, (Edited by G. H. Hardy, P. V. Seshu Aiyar, and B. M. Wilson), New York, Chelsea, 1962

[25] B. Mazur, Bernoulli numbers and the unity of mathematics, March 2008, very rough notes for the Bartlett lecture
http://wiki.wstein.org/2008/480a?action=AttachFile&do=get&target=Bernoulli.pdf

[26] riferimento per matrice di Haar

[27] P. Berkhin, A survey of pagerank computing, *Internet Mathematics*, 2 (2005), pp.73–120

[28] B. Venanzoni, Approssimazioni di Matrici in Spazi di Bassa Complessità, Master thesis in Mathematics, Univ. of Rome "Tor Vergata", October 2012

[29] V. Cardinali, *La Teoria di Perron-Frobenius*, Master thesis in Mathematics, Univ. of Rome "Tor Vergata", December 2012

[30] C. Di Fiore, S. Fanelli, F. Lepore, P. Zellini, Matrix algebras in quasi-Newton methods for unconstrained minimization, *Numerishe Mathematik*, 94 (2003), p.479–500

[31] A. Bortoletti, C. Di Fiore, S. Fanelli, P. Zellini, A new class of quasi-newtonian methods for optimal learning in MLP-networks, IEEE Transactions on Neural Networks, 14 (2003), pp.263–273

[32] C. Di Fiore, Structured matrices in unconstrained minimization methods, in *Contemporary Mathematics*, 323, pp.205–219, 2003

[33] J. F. Cai, R. H. Chan, C. Di Fiore, Minimization of a detail-preserving regularization functional for impulse noise removal, *J. Math. Imaging Vis.*, 29 (2007), pp.79–91

[34] F. Hausdorff, Der Wertvorrat einer Bilinearform, *Math. Zeits.*, 3 (1919), pp.314–316

[35] F. Di Benedetto, Gram matrices of fast algebras have a rank structure, *SIAM J. Matrix Anal. Appl.*, 31 (2009), pp.526–545

[36] F. Di Benedetto, Analysis of preconditioning techniques for ill-conditioned Toeplitz matrices, SIAM J. Sci. Comput. 16 (1995), pp.682–697

[37] T. Ceccherini, F. Scarabotti, F. Tolli, Appendix of *Harmonic Analysis on Finite Groups, Representation Theory, Gelfand Pairs and Markov Chains*, Cambridge Studies in Advanced Mathematics, Cambridge Univ. Press, 2008

[38] F. Scarabotti, The discrete sine transform and the spectrum of the finite q-ary tree, SIAM J. Discrete Math., 19 (2005), pp.1004–1010

[39] R. S. Varga, *Matrix Iterative Analysis*, Springer Series in Computational Mathematics, vol. 27, Springer, 2nd edition, 1999

[40] riferimento per Richardson-Eulero

[41] O. Axelsson, V. A. Barker, *Finite Element Solution of Boundary Value Problems – Theory and Computation*, Academic Press, Orlando, FL, 1984

[42] D. Bini, M. Capovani, O. Menchi, Metodi Numerici per l'Algebra Lineare, Zanichelli, ??

[43] R. H. Chan, G. Strang, Toeplitz equations by conjugate gradients with circulant preconditioner, SIAM J. Sci. Statist. Comput., 10 (1989), pp.104–119

[44] T. Huckle, Circulant and skewcirculant matrices for solving Toeplitz matrix problems, *SIAM J. Matrix Anal. Appl.*, 13 (1992), pp.767–777

[45] T. F. Chan, An optimal circulant preconditioner for Toeplitz systems, *SIAM J. Sci. Statist. Comput.*, 9 (1988), pp.766–771

[46] D. Bini, F. Di Benedetto, A new preconditioner for the parallel solution of positive definite Toeplitz systems, in: *Proceedings of the 2nd ACM Symposium on Parallel Algorithms and Architectures*, Crete, Greece, 1990, pp.220–223

[47] D. Bini, P. Favati, On a matrix algebra related to the discrete Hartley transform, *SIAM J. Matrix Anal. Appl.*, 14 (1993), pp.500–507

[48] . M. K. Ng, *Iterative Methods for Toeplitz Systems*, Oxford University Press, 2004

[49] C. Canuto

[50] G. Strang, A. Edelman, The Toeplitz-circulant eigenvalue problem $A\mathbf{x} = \lambda\mathcal{C}\mathbf{x}$, in: L. Bragg, J. Dettman (Eds.), *Oakland Conference on PDE's*, Longman, London, 1987

[51] G. Strang, A proposal for Toeplitz matrix calculations, *Stud. Appl. Math.*, 74 (1986), pp.171–176

[52] P. Sentinelli, F. Tudisco, Right cones in euclidean spaces and the orthogonal projections leaving them invariant, presentato per la publicazione, Gennaio 2013

[53] R. H. Chan, X. Jin, M. Yeung, The circulant operator in the Banach algebra of matrices, *Linear Algebra Appl.*, 149 (1991), pp.41-53

[54] D. Potts, G. Steidl, Optimal trigonometric preconditioners for nonsymmetric Toeplitz systems, *Linear Algebra Appl.*, 281 (1998), pp.265–292

[55] R. H. Chan, M. K. Ng, Conjugate gradient methods for Toeplitz systems, *SIAM Review*, 38 (1996), pp.427–482

[56] R. H. Chan, M. K. Ng, C. K. Wong, Sine transform based preconditioners for symmetric Toeplitz systems, *Linear Algebra Appl.*, 232 (1996), pp.237–259

[57] R. H. Chan, Toeplitz preconditioners for Toeplitz systems with nonnegative generating functions, *IMA J. Numer. Anal.*, 11 (1991), pp.333–345

[58] R. H. Chan, Circulant preconditioners for Hermitian Toeplitz systems, *SIAM J. Matrix Anal. Appl.*, 10 (1989), pp.542–550

[59] Perron

[60] Frobenius

[61] S. Friedland, On an inverse problem for nonnegative and eventually nonnegative matrices, Israel Journal of Mathematics, 29 (1978), pp.43–60

[62] D. Noutsos, On Perron-Frobenius property of matrices having some negative entries, Linear Algebra Appl., 412 (2006), pp.132–153

[63] F. Tudisco, V. Cardinali, C. Di Fiore, On power nonnegative matrices and certain Rothblum spectral bounds, presentato per la pubblicazione

CONTINUAZIONE DELL'ESERCIZIO SULL'ALGEBRA DI BASSA COMPLESSITA' DI HAAR: BASE ORTOGONALE, STRUTTURA DELLE MATRICI DELL'ALGEBRA

Consider the matrix algebra $\mathcal{L}$ associated to the transform U: $\mathcal{L} = \{Ud(\mathbf{z})U^H\} = \{QDd(\mathbf{z})DQ^T\}$.

Exercise: Find an orthogonal basis $\{J_1, J_2, \ldots, J_n\}$, $n = 2^k$, for $\mathcal{L}$ made up with matrices whose entries are 0, 1 or -1. Solution: the following rank-one $0, 1, -1$-matrices form a basis for $\mathcal{L}$ ($n = 4$, $n = 8$, $n = 2^k$):

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} = Q\mathbf{e}_1\mathbf{e}_1^T Q^T, \quad \begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix} = Q\mathbf{e}_3\mathbf{e}_3^T Q^T,$$

$$\begin{bmatrix} 1 & -1 & & \\ -1 & 1 & & \\ & & & \\ & & & \end{bmatrix} = Q\mathbf{e}_4\mathbf{e}_4^T Q^T, \quad \begin{bmatrix} & & & \\ & & & \\ & & 1 & -1 \\ & & -1 & 1 \end{bmatrix} = Q\mathbf{e}_2\mathbf{e}_2^T Q^T;$$

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} = Q\mathbf{e}_1\mathbf{e}_1^T Q^T, \quad \begin{bmatrix} 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \end{bmatrix} = Q\mathbf{e}_5\mathbf{e}_5^T Q^T,$$

$$\begin{bmatrix} 1 & 1 & -1 & -1 & & & & \\ 1 & 1 & -1 & -1 & & & & \\ -1 & -1 & 1 & 1 & & & & \\ -1 & -1 & 1 & 1 & & & & \\ & & & & & & & \\ & & & & & & & \\ & & & & & & & \\ & & & & & & & \end{bmatrix} = Q\mathbf{e}_7\mathbf{e}_7^T Q^T, \quad \begin{bmatrix} & & & & & & & \\ & & & & & & & \\ & & & & 1 & 1 & -1 & -1 \\ & & & & 1 & 1 & -1 & -1 \\ & & & & -1 & -1 & 1 & 1 \\ & & & & -1 & -1 & 1 & 1 \\ & & & & & & & \\ & & & & & & & \end{bmatrix} = Q\mathbf{e}_3\mathbf{e}_3^T Q^T,$$

$$\begin{bmatrix} 1 & -1 & & & & & & \\ -1 & 1 & & & & & & \\ & & & & & & & \\ & & & & & & & \\ & & & & & & & \\ & & & & & & & \\ & & & & & & & \\ & & & & & & & \end{bmatrix} = Q\mathbf{e}_8\mathbf{e}_8^T Q^T, \quad \begin{bmatrix} & & & & & & & \\ & & & & & & & \\ & & 1 & -1 & & & & \\ & & -1 & 1 & & & & \\ & & & & & & & \\ & & & & & & & \\ & & & & & & & \\ & & & & & & & \end{bmatrix} = Q\mathbf{e}_6\mathbf{e}_6^T Q^T,$$

$$\begin{bmatrix} & & & & & & & \\ & & & & & & & \\ & & 1 & -1 & & & & \\ & & -1 & 1 & & & & \\ & & & & & & & \\ & & & & & & & \\ & & & & & & & \\ & & & & & & & \end{bmatrix} = Q\mathbf{e}_4\mathbf{e}_4^T Q^T, \quad \begin{bmatrix} & & & & & & & \\ & & & & & & & \\ & & & & & & & \\ & & & & & & & \\ & & & & & & & \\ & & & & & & & \\ & & & & & & 1 & -1 \\ & & & & & & -1 & 1 \end{bmatrix} = Q\mathbf{e}_2\mathbf{e}_2^T Q^T;$$

$$\mathbf{e}\mathbf{e}^T = Q\mathbf{e}_1\mathbf{e}_1^T Q^T,$$
$$J_{s,i} = \begin{bmatrix} 1_{2^{s-1}} & -1_{2^{s-1}} \\ -1_{2^{s-1}} & 1_{2^{s-1}} \end{bmatrix} \otimes \mu_i^{(\frac{n}{2^s})} = Q\mathbf{e}_*\mathbf{e}_*^T Q^T, \quad * = (n - 2^{s-1} + 1) - 2^s(i-1),$$
$$i = 1, \ldots, \tfrac{n}{2^s}, \ s = 1, 2, \ldots, \log_2 n = k$$

(1_r is the $r \times r$ matrix with ones everywhere; $\mu_i^{(r)}$ is the $r \times r$ matrix with 1 in position i, i and 0 elsewhere). Note that if X, Y are two matrices from such basis, then $(X, X) = n^2, \frac{n^2}{4}, \frac{n^2}{16}, \ldots, \frac{n^2}{2^{2k-2}} = 4$, and $(X, Y) = 0$ if $X \neq Y$. This implies, in particular, that we have immediately a formula for the best approximation of $A \in \mathbb{C}^{n\times n}$ in $\mathcal{L}$:

$$\mathcal{L}_A = \frac{(\mathbf{e}\mathbf{e}^T, A)}{n^2}\mathbf{e}\mathbf{e}^T + \sum_{s=1}^{\log_2 n} \frac{1}{2^{2s}} \sum_{i=1}^{\frac{n}{2^s}} (J_{s,i}, A) J_{s,i}.$$

Observe that for $n = 4$ and $n = 8$ the generic matrix in $\mathcal{L}$ has the following structure:

$$\begin{bmatrix} a+b & a-b & c & c \\ a-b & a+b & c & c \\ c & c & a+d & a-d \\ c & c & a-d & a+d \end{bmatrix},$$

$$\begin{bmatrix} (a+d)+c & (a+d)-c & a-d & a-d & b & b & b & b \\ (a+d)-c & (a+d)+c & a-d & a-d & b & b & b & b \\ a-d & a-d & (a+d)+e & (a+d)-e & b & b & b & b \\ a-d & a-d & (a+d)-e & (a+d)+e & b & b & b & b \\ b & b & b & b & (a+f)+g & (a+f)-g & a-f & a-f \\ b & b & b & b & (a+f)-g & (a+f)+g & a-f & a-f \\ b & b & b & b & a-f & a-f & (a+f)+h & (a+f)-h \\ b & b & b & b & a-f & a-f & (a+f)-h & (a+f)+h \end{bmatrix}.$$