Dispense di Calcolo Numerico
a.a. 2020-21*

1 Analisi dell'errore

Per questo argomento si fa riferimento alle slides della lezione sui numeri di macchina e al Capitolo 2, (paragrafi 1, 3, 4, 5 (prima parte), 6, 7, 8) del libro [1] ed al materiale (inclusi i relativi esercizi) presente in portineria.

Un riferimento in inglese è il Capitolo 1 di [2], Sezioni 1.1, 1.2 e 1.3.

Riferimenti bibliografici

[1] R. Bevilacqua, D. Bini, M. Capovani, O. Menchi. *Metodi Numerici.* Zanichelli, Bologna, 1992.
[2] J. Stoer, R. Bulirsch. *Introduction to Numerical Analysis.* Third Edition. Springer, 2002.

*Versione del 20 dicembre 2021. Queste dispense sono fornite in formato di bozza ed è vietata la pubblicazione su Internet (possono invece essere distribuite liberamente tra gli studenti).

2 Algoritmi elementari e loro accelerazione

Premessa

Descriveremo gli algoritmi per il calcolo di alcune quantità dell'analisi numerica e ne analizzeremo il costo computazionale, inteso come il numero di operazioni elementari richiesto per il calcolo. Le operazioni elementari sono la somma algebrica, il prodotto e la divisione tra scalari che indicheremo semplicemente con "ops".

La nostra trattazione sarà necessariamente più grossolana di quella che si fa in informatica, poiché l'interesse del calcolo numerico è più focalizzato sulle questioni analitiche che su quelle informatiche. Tuttavia, ci sembra necessario fissare le notazioni, che useremo spesso.

Sia Ω l'insieme di tutti i dati di un problema, il *costo computazionale assoluto* è una funzione $f : \Omega \to \mathbb{Z}_{\geqslant}$, dove $\mathbb{Z}_{\geqslant}$ è l'insieme dei numeri interi non negativi, che a un insieme di dati assegna il numero di operazioni da effettuare.

Ad esempio, se si vuole considerare il costo computazionale della somma tra due vettori v, w della stessa dimensione, occorrerà definire una funzione f sull'insieme $\Omega = \cup_{n=1}^{\infty}\{(v, w) : v, w \in \mathbb{K}^n\}$ di tutte le coppie di vettori della stessa dimensione e a valori in $\mathbb{Z}_{\geqslant}$. Chiaramente, il valore di questa funzione f sarà n, dove n è la dimensione di v e w, perché sono richieste esattamente n somme di scalari per calcolare $v + w$. Ci si accorge che il costo, in questo caso, dipende solo dalla dimensione dei vettori e non dal loro valore.

In molti casi, il costo dipende da semplici quantità legate ai dati, come per esempio la "dimensione dei dati". È quindi conveniente, dato un algoritmo, stabilire le quantità essenziali che caratterizzano il costo computazionale e definire un costo computazionale relativo a partire da esse. Le quantità essenziali sono definite (abbastanza arbitrariamente) tramite una funzione $\ell : \Omega \to \mathcal{U}$, dove $\mathcal{U}$ è un opportuno insieme e vale che se $\ell(x) = \ell(y)$ allora $f(x) = f(y)$. (Un altro modo di definire ℓ è tramite una relazione di equivalenza su Ω.)

Nel caso della somma tra vettori, si può definire la dimensione dei dati come $\ell : \Omega \to \mathbb{Z}_{\geqslant}$ tale che $\ell(v, w) = n$ e così possiamo definire il costo computazionale relativo come una funzione $\mathcal{C} : \mathbb{Z}_{\geqslant} \to \mathbb{Z}_{\geqslant}$, che varrà $\mathcal{C}(n) = n$. La funzione f sarà data da $f = \mathcal{C} \circ \ell$.

Nel seguito, per semplicità, considereremo esclusivamente il costo computazionale relativo $\mathcal{C} : \mathcal{U} \to \mathbb{Z}_{\geqslant}$, che chiameremo semplicemente *costo computazionale*, definito su un insieme $\mathcal{U}$ delle possibili "dimensioni dei dati", mentre la funzione $\ell : \Omega \to \mathcal{U}$ sarà chiara dal contesto e verrà omessa.

Ad esempio, se lavoriamo con due vettori reali di uguale dimensione, la grandezza dei dati può essere intesa come il numero di componenti dei vettori, e quindi in questo caso $\mathcal{U}$ è l'insieme dei numeri naturali. Se lavoriamo con matrici quadrate di dimensione $n \times n$, la grandezza dei dati è di solito intesa come il lato della matrice, cioè n, e quindi $\mathcal{U} = \mathbb{Z}_{\geqslant}$, mentre se la matrice è rettangolare $m \times n$, la grandezza dei dati è data dalle due dimensioni e $\mathcal{U} = \mathbb{Z}_{\geqslant} \times \mathbb{Z}_{\geqslant}$. Si noti l'ambiguità della notazione, che non generalizza il caso $n \times n$, tuttavia è abbastanza comoda.

Il calcolo del costo computazionale fornisce indicazioni sull'efficienza di un dato algoritmo. Nella maggior parte dei casi si è interessati al comportamento dell'algoritmo quando il costo computazionale $\mathcal{C}(n_1, n_2, \ldots, n_t)$ tende a infinito (si parla di comportamento *asintotico*). Questo accade normalmente quando le dimensioni dei dati tendono a infinito.

Se $\mathcal{C}$ tende a infinito, allora può essere sufficiente la parte principale del suo sviluppo asintotico che spesso è più facile da calcolare. Addirittura, spesso è sufficiente una stima usando le notazioni asintotiche del tipo O.

Il motivo per cui interessa l'andamento a infinito è che spesso lo sviluppo asintotico del costo computazionale fornisce indicazioni sul comportamento dell'algoritmo per valori grandi della dimensione dei dati, che sono quelli che creano maggiori problemi per il calcolo.

Nota. Nella nostra trattazione consideriamo unitario il costo della singola operazione aritmetica a prescindere dalla precisione di macchina usata o dal fatto che $\mathbb{K}$ sia l'insieme dei numeri reali o complessi, anche se in realtà i costi effettivi sono diversi (ma sono multipli l'uno dell'altro).

Valutazione di un polinomio

Un polinomio a coefficienti in $\mathbb{K}$,

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0, \qquad a_i \in \mathbb{K},$$

quando $\mathbb{K} = \mathbb{R}$ o $\mathbb{K} = \mathbb{C}$, può essere identificato con una funzione $p : \mathbb{K} \to \mathbb{K}$. Un problema computazionale è quindi quello di valutare p in un dato $x \in \mathbb{K}$, cioè calcolare $p(x)$.

Un algoritmo semplice che esegue questa operazione, consiste nel calcolare i termini del tipo $a_i x^i$ e sommarli. Se si parte dai termini di grado minore, quando si sta calcolando $a_i x^i$, per $i > 0$, al passo precedente si è calcolato $a_{i-1} x^{i-1}$, e quindi non conviene ricalcolare x^i ogni volta, ma si può ottenere $x^i = x^{i-1} x$ con una sola operazione. L'algoritmo è il seguente

```
1 s=0;
2 p=1;
3 for i=0:n
4    s=s+a(i)*p;
5    p=p*x;
6 end
```

Si osserva che l'algoritmo esegue $3(n+1)$ operazioni aritmetiche.

Con un piccolo trucco è possibile fare di meglio. Consideriamo dapprima il caso $n = 3$ e scriviamo

$$p(x) = ((a_3x + a_2)x + a_1)x + a_0;$$

a questo punto si osserva che sono richieste due operazioni per ogni coppia di parentesi. In generale si può scrivere

$$p(x) = (\cdots((a_nx + a_{n-1})x + a_{n-2})x + \cdots a_1)x + a_0,$$

e tradurre la formula in un algoritmo nel modo seguente

```
1 p=a(n);
2 for i=n-1:-1:0
3    p=p*x+a(i);
4 end
```

Si osserva che questa volta le operazioni richieste sono $2n$, quindi il metodo è più conveniente. Questo algoritmo di valutazione è detto *metodo di Ruffini-Horner*.

Dimostriamo ora più formalmente che il metodo di Ruffini-Horner fornisce il valore $p(x)$.

Proposizione 2.1. *Sia $p(x) = a_0 + a_1x + \cdots + a_nx^n \in \mathbb{K}[x]$, e per $y \in \mathbb{K}$, sia data la sequenza*

$$\begin{cases} s_n = a_n, \\ s_i = s_{i+1}y + a_i, \qquad i = n-1, \ldots, 0. \end{cases}$$

Si ha che $s_0 = p(y)$.

Dimostrazione. Si procede per induzione su n. Per $n = 0$, il polinomio è a_0 e quindi $p(x) = a_0$ per ogni x. La sequenza contiene il solo valore $s_0 = a_0 = p(y)$ e questo dimostra il passo base.

Dimostriamo ora che se la proposizione è vera per polinomi di grado $n-1$, allora lo è per polinomi di grado n. Scriviamo

$$p(x) = a_0 + x(a_1 + a_2x + \cdots + a_{n-1}x^{n-2} + a_nx^{n-1}) = a_0 + xq(x), \tag{1}$$

dove il polinomio q ha grado $n-1$. Sia $t_{n-1}, t_{n-2}, \cdots, t_0$ la successione ottenuta applicando il metodo di Ruffini-Horner per valutare il polinomio q nel punto y, per ipotesi induttiva si ha $t_0 = q(y)$. Inoltre per come è costruita la successione si osserva che $t_{n-1} = a_n = s_n$, e inoltre $t_{n-2} = a_ny + a_{n-1} = s_ny + a_{n-1} = s_{n-1}$ e così via si ha che $t_i = s_{i+1}$ per $i = n-1, \ldots, 1, 0$ e quindi $t_0 = s_1 = q(y)$. A questo punto $s_0 = s_1y + a_0 = q(y)y + a_0$, che da (**??**) è proprio $p(y)$. □

Il metodo di Ruffini-Horner può essere usato anche per valutare una funzione razionale del tipo $r(x) = \frac{p(x)}{q(x)}$, dove p è un polinomio di grado m e q è un polinomio di grado n. È sufficiente applicare l'algoritmo a $p(x)$ e $q(x)$ separatamente e poi dividere i risultati, e questo richiede $2(m+n)+1$ operazioni aritmetiche.

Algoritmi di base in algebra lineare

Dati due vettori $v, w \in \mathbb{K}^n$, che consideriamo vettori colonna, cioè $v, w \in \mathbb{K}^{n\times n}$, i loro trasposti $v^T, w^T \in \mathbb{K}^{1\times n}$ saranno considerati vettori riga. Possiamo definire il prodotto riga per colonna come

$$v^Tw = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} := v_1w_1 + \cdots + v_nw_n = \sum_{i=1}^{n} v_iw_i.$$

Si osservi che $w^Tv = v^Tw$.

L'operazione richiede il calcolo di n prodotti e $n-1$ somme e quindi il costo computazionale è $\mathcal{C}(n) = 2n-1$ o in termini asintotici $\mathcal{C}(n) \approx 2n$. Si può dire che il costo è lineare in n.

Per implementare il calcolo del prodotto scalare, è sufficiente scrivere un ciclo `for` che calcoli la somma $\sum_{i=1}^{n} v_iw_i$ e per questo è sufficiente un accumulatore s che conterrà la somma

```
1 s=0;
2 for i=1:n
3    s=s+v(i)*w(i);
4 end
```

Si noti che il codice esegue $2n$ ops e può essere leggermente migliorato inizializzando l'accumulatore con il primo termine della somma anziché con 0, ottenendo un algoritmo che richiede esattamente le $2n-1$ ops necessarie

```
1 s=v(1)*w(1);
2 for i=2:n
3    s=s+v(i)*w(i);
4 end
```

Il successivo algoritmo è il calcolo del prodotto di una matrice $A \in \mathbb{K}^{n\times n}$ per un vettore $b \in \mathbb{K}^n$, che può essere ottenuto a partire dal prodotti di una riga per una colonna. Si tratta di calcolare ciascuna delle componenti del vettore $c = Ab \in \mathbb{K}^n$, dove la componente c_i, per $i = 1, \ldots, n$, è data dal prodotto della riga i-esima di A, indicata con r_i^T per il vettore b, cioè

$$c_i := r_i^T b = a_{i1}b_1 + a_{i2}b_2 + \cdots + a_{in}b_n = \sum_{j=1}^{n} a_{ij}b_j, \quad i = 1, \ldots, n.$$

Il costo è dato quindi da n prodotti scalari, cioè $\mathcal{C}(n) = n(2n-1) = 2n^2 - n \approx 2n^2$. Siccome il costo è dato da un polinomio di grado due si parla di costo quadratico.

Se la matrice $A \in \mathbb{K}^{m\times n}$, essa dovrà essere moltiplicata per un vettore $b \in \mathbb{K}^n$ e darà un vettore $c \in \mathbb{K}^m$. In questo caso la moltiplicazione di una riga di A per il vettore b richiederà $2n-1$ operazioni, e siccome le righe di A sono m, il costo totale dell'algoritmo è $\mathcal{C}(m,n) = 2mn - m \approx 2mn$. Si noti che anche questa volta si è ottenuto un polinomio di grado due, anche se in due variabili.

L'implementazione di questo algoritmo è anch'essa abbastanza semplice, sarà necessario un ciclo `for` per calcolare ciascuna delle componenti di c all'interno del quale è innestato un ciclo `for` che calcola la somma

```
1 for i=1:m
2    s=a(i,1)*b(1);
3    for k=2:n
4       s=s+a(i,k)*b(k);
5    end
6    c(i)=s;
7 end
```

Consideriamo infine il prodotto righe per colonne di due matrici, partendo dal caso quadrato: $A, B \in \mathbb{K}^{n\times n}$. Si vuole calcolare $C = AB$, in questo caso l'elemento c_{ij} di C, per ogni i e j, viene calcolato come il prodotto della riga i-esima di A, indicata con r_i^T, per la colonna j-esima di B, indicata con c_j:

$$c_{ij} := r_i^T c_j = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj} = \sum_{\ell=1}^{n} a_{i\ell}b_{\ell j}, \quad , i, j = 1, \ldots, n,$$

dove r_i^T è la riga i-esima di A e c_j è la colonna j-esima di B. Siccome un prodotto tra una riga e una colonna costa $2n-1$ ops e questo va fatto per tutti gli i e j che sono n^2 allora si ha $\mathcal{C}(n) = 2n^3 - n^2 \approx 2n^3 = O(n^3)$ ops.

Se le matrici sono di dimensioni diverse, per esempio se $A \in \mathbb{K}^{m\times k}$ e $B \in \mathbb{K}^{k\times n}$ allora $C = AB \in \mathbb{K}^{m\times n}$ e quindi si devono fare mn prodotti di righe di A per colonne di B, essendo queste ultime lunghe k, il costo totale sarà $\mathcal{C}(n) = (2k-1)mn \approx 2mnk$.

Un altro modo di valutare il costo computazionale consiste nello scrivere dello pseudocodice per l'algoritmo e contare le operazioni elementari all'interno dei cicli a partire dal ciclo più interno. Il costo di un'istruzione del tipo `for k=k1:k2 f(k); end` è data dalla somma per k che va da k_1 a k_2 del costo della valutazione di $f(k)$.

Il prodotto righe per colonne si può scrivere in pseudocodice come

```
1 for i=1:m
2    for j=1:n
3       s=a(i,1)*b(1,j);
4       for t=2:k
5          s=s+a(i,t)*b(t,j);
6       end
```

```
7     c(i,j)=s;
8     end
9  end
```

I cicli innestati sono tre, quindi ci si aspetta un costo cubico. Il ciclo più interno, richiede 2 operazioni e va ripetuto $k-2+1$ volte, quindi costa $2k-2$ ops a cui va sommata l'operazione che si esegue durante l'assegnamento di s nella linea 4, per cui il costo della porzione di codice corrispondente alle righe 3-6 è di $2k-1$ operazioni e non dipende da j e da i. Per ottenere il costo complessivo è sufficiente sommare per ogni i e j il valore $2k-1$, ottenendo, come sopra, $\mathcal{C}(m,n,k) = mn(2k-1) = 2mnk - mn$.

Esercizio 2.2. Calcolare il costo computazionale della somma tra due matrici $A, B \in \mathbb{K}^{m\times n}$ e del prodotto αA, dove $\alpha \in \mathbb{K}$.

Soluzione. Semplicemente mn ops per ciascuna delle due operazioni. □

Esercizio 2.3. Dire che dimensione deve avere il vettore y affinché abbia senso il prodotto $y^T A$, con $A \in \mathbb{K}^{m\times n}$ e calcolare il costo computazionale di questa operazione.

Soluzione. Il vettore y deve avere 1 o m componenti. Nel primo caso si ha il prodotto di uno scalare per una matrice che richiede mn ops; nel secondo caso, detto $c = y^T A \in \mathbb{K}^n$, si ha $c_i = \sum_{k=1}^{m} y_k a_{ki}$, per $i = 1, \ldots, n$, quindi il costo è dato da $2nm - n$ ops. □

Esercizio 2.4. Dati due vettori v, w dire quali dimensioni devono avere affinché abbia senso il prodotto vw^T e dire qual è il costo computazionale del calcolo del prodotto. Mostrare che vw^T può essere diverso da wv^T.

Soluzione. Se $v \in \mathbb{K}^{m\times 1}$ e $w \in \mathbb{K}^{n\times 1}$, chiaramente $w^T \in \mathbb{K}^{1\times n}$ e quindi, immaginando v e w^T come matrici, il loro prodotto ha senso e $vw^T \in \mathbb{K}^{m\times n}$ è la matrice

$$\begin{bmatrix} v_1w_1 & v_1w_2 & \cdots & v_1w_n \\ v_2w_2 & v_2w_2 & \cdots & v_2w_n \\ \vdots & \vdots & \vdots & \vdots \\ v_mw_1 & v_mw_2 & \cdots & v_mw_n \end{bmatrix}.$$

Si vede immediatamente che sono richieste esattamente mn moltiplicazioni per la costruzione di vw^T.

Per mostrare che vw^T può essere diverso da wv^T basta osservare che se v e w hanno dimensioni diverse, allora vw^T e wv^T hanno dimensioni diverse e quindi non possono essere uguali. □

Accelerazione degli algoritmi

Gli algoritmi visti sopra sono standard e funzionano su qualsiasi insieme di dati, tuttavia esistono vari modi per accelerarli nei problemi concreti, ottenendo possibilmente risultati computazionali migliori.

Alcuni tra i metodi comunemente usati per accelerare gli algoritmi consistono nel

1. cercare un algoritmo che funzioni meglio di quello standard sul problema o i problemi che stiamo studiando anche se non funziona nel caso generale; in questo caso si parla di algoritmi strutturati;

2. utilizzare le risorse di calcolo (quasi sempre un elaboratore con la sua architettura) per ottenere prestazioni migliori;

3. trovare algoritmi basati su idee nuove e che risolvono i problemi per tutti i dati ma che abbiano un costo computazionale minore rispetto alle implementazioni standard.

Ci soffermeremo sul punto 1 nel seguito del capitolo, dando una menzione del punto 2 nella prossima sezione. Per quanto riguarda il punto 3 ci limiteremo a un breve cenno su come si è tentato di accelerare il prodotto righe per colonne.

Il prodotto righe per colonne non è l'algoritmo con la minore complessità per il calcolo del prodotto tra due matrici quadrate: nel 1969, V. Strassen ha mostrato che è possibile calcolare il prodotto tra due matrici con un costo asintotico pari a $O(n^{\log_2 7})$ ops. Non è ancora chiaro quale sia la complessità di questo problema, l'unica cosa che si è riusciti a ottenere è di abbassare il costo. L'algoritmo più veloce oggi richiede $O(n^{2.3727})$ ops. L'utilità pratica di questo tipo di algoritmi è abbastanza limitata poiché la O nasconde costanti a volte molto grosse, inoltre non è frequente eseguire moltiplicazioni di matrici molto grandi, a meno che queste matrici non siano strutturate. Nel resto della trattazione considereremo solo il prodotto righe per colonne.

Algoritmi paralleli

Il computo del numero di operazioni richieste da un algoritmo è normalmente una misura del tempo di esecuzione, nell'ipotesi che ogni operazione elementare, se eseguita su calcolatore, richieda un tempo finito, grosso modo costante.

Le moderne architetture permettono di effettuare operazioni in parallelo su più processori e quindi effettuare più operazioni contemporaneamente. Se si riescono ad eseguire n operazioni su n processori, le operazioni saranno effettuate in una sola unità di tempo, che chiameremo *ciclo*.

Nella pratica occorre tenere conto di vari problemi, tra cui la comunicazione (tra i processori) e l'accesso alla memoria, tuttavia in molti casi sfruttare i processori in parallelo accorcia i tempi effettivi di esecuzione.

Nel seguito considereremo un modello semplificato in cui i calcoli vengono effettuati in sequenza dai processori coinvolti, un'operazione per ogni ciclo trascurando i problemi di comunicazione. Inoltre, per semplicità, tratteremo solo il caso in cui si disponga di un numero sufficientemente alto di processori.

Se si vogliono sommare due vettori $v, w \in \mathbb{R}^n$, occorre calcolare le n somme $v_i + w_i$ per $i = 1, \ldots, n$. Si osserva che per calcolare $v_1 + w_1$ non serve il risultato di $v_2 + w_2$, quindi queste due operazioni possono essere eseguite da due processori diversi nello stesso istante, anziché su un solo processore in due istanti consecutivi. Più in generale, poiché nessuna delle operazioni richiede il risultato delle altre, le n operazioni $v_i + w_i$ possono essere effettuate in un solo ciclo di calcolo su n processori in parallelo.

La seguente rappresentazione grafica mostra a sinistra un'esecuzione su un solo processore e a destra l'esecuzione su n processori in parallelo, evidenziando il fatto che le stesse operazioni vengono eseguite in un minor numero di cicli di calcolo.

ciclo 1	$\boxed{v_1 + w_1}$
ciclo 2	$\boxed{v_2 + w_2}$
$\vdots$	$\vdots$
ciclo n	$\boxed{v_n + w_n}$

ciclo 1	$\boxed{v_1 + w_1}$	$\boxed{v_2 + w_2}$	$\cdots$	$\boxed{v_n + w_n}$
	processore 1	processore 2	$\cdots$	processore n

Un'implementazione dell'algoritmo parallelo è data da:

```
1 parallel for i=1:n
2   u(i)=v(i)+w(i)
3 end
```

Si noti l'uso dell'istruzione `parallel for` che indica che ogni ciclo deve essere eseguito, se possibile, in parallelo. Useremo anche l'istruzione `parallel` per racchiudere un blocco di comandi che devono essere eseguiti, se possibile, in parallelo.

Se si devono sommare tre numeri $x_0, x_1, x_2 \in \mathbb{R}$, il numero di operazioni da effettuare è 2. Si possono utilizzare tre algoritmi, per esempio, si può porre $y = x_0 + x_1$ e $s = y + x_2$ oppure si può porre $z = x_1 + x_2$ e $s = x_0 + z$ e infine si può porre $t = x_0 + x_2$ e $s = t + x_1$. Si noti che, poiché in aritmetica finita la proprietà associativa non vale, sono tre algoritmi diversi.

In tutti e tre i casi la seconda operazione richiede il risultato della prima e quindi, anche disponendo di due processori, le due operazioni non possono essere eseguite in parallelo. Il calcolo richiede sempre 2 cicli su qualsiasi numero di processori.

Consideriamo ora la somma di 4 numeri reali $s = x_0 + x_1 + x_2 + x_3$. Il calcolo richiede 3 operazioni, tuttavia se si dispone di almeno 2 processori è possibile sommare durante il primo ciclo $y_0 = x_0 + x_1$ e $y_2 = x_2 + x_3$ in parallelo poiché nessuna delle due operazioni richiede il risultato dell'altra. Al secondo ciclo si può calcolare $s = y_0 + y_2$. L'implementazione di questo algoritmo è la seguente

```
1 parallel
2   s=x(0)+x(1);
3   t=x(2)+x(3);
4 end
5 s=s+t;
```

da cui, poiché il blocco parallelo viene eseguito in un ciclo, si deduce facilmente che il numero di cicli richiesti è 2. In questo modo abbiamo risparmiato 1 ciclo di calcolo rispetto alla somma sequenziale che consiste nel calcolare $z_1 = x_0 + x_1$, poi $z_2 = z_1 + x_2$ e infine $s = z_2 + x_3$, senza possibilità di parallelizzare.

Nel caso di 8 numeri si può procedere come in figura calcolando la somma in soli $3 = \log_2 8$ cicli anziché 7 che è il numero di operazioni necessarie.

$$\underbrace{\underbrace{\underbrace{x_0 + x_1}_{v_0^{(1)}} + \underbrace{x_2 + x_3}_{v_1^{(1)}}}_{v_0^{(2)}} + \underbrace{\underbrace{x_4 + x_5}_{v_2^{(1)}} + \underbrace{x_6 + x_7}_{v_3^{(1)}}}_{v_1^{(2)}}}$$

Questo modo di procedere può essere generalizzato. Consideriamo il problema del calcolo della somma di n numeri reali, $x_i \in \mathbb{R}$ per $i = 0, \ldots, n-1$, cioè $s = \sum_{i=0}^{n-1} x_i$. Si può assumere che $n = 2^k$, altrimenti si possono aggiungere zeri quanto basta per arrivare a 2^k numeri.

L'algoritmo sequenziale per il calcolo della somma in pseudocodice ha la forma

```
1 s=0;
2 for i=0:2^k-1
3   s=s+x(i);
4 end
```

Posto $v_i^{(0)} = x_i$, per $i = 0, \ldots, n-1$, l'algoritmo parallelo ha invece questa forma

$$v_i^{(p)} = v_{2i}^{(p-1)} + v_{2i+1}^{(p-1)}, \qquad i = 0, \ldots, n/2^p - 1, \qquad p = 1, \ldots, k.$$

Un'implementazione in pseudocodice con sovrascrittura del vettore v è

```
1 for p=1:k
2    for i=0:n/2^p-1
3       v(i)=v(2*i)+v(2*i+1);
4    end
5 end
```

e la somma sarà contenuta in `v(0)`.

Si vede immediatamente che tutte le operazioni nel ciclo `for` interno sono indipendenti e si possono eseguire in un ciclo di calcolo e quindi in totale sono richiesti k cicli di calcolo, se si hanno a disposizione 2^{k-1} processori. Se n non è potenza di 2, il più piccolo k tale che $n \leqslant 2^k$ è $\lceil \log_2 n \rceil$. Si può concludere che con un numero di processori maggiore di $n/2$ è possibile effettuare la somma di n numeri in non più di $\lceil \log_2 n \rceil$ cicli di calcolo. L'implementazione è:

```
1 for p=1:k
2    n=n/2;
2    parallel for i=0:n-1
3       v(i)=v(2*i)+v(2*i+1);
4    end
5 end
```

Si noti che il parallel for interno va da 1 a $n/2^p - 1$, ma per evitare di calcolare ogni volta la potenza, viene utilizzata la variabile n che si dimezza a ogni passo.

Ora consideriamo gli algoritmi elementari di algebra lineare. Il prodotto scalare si può calcolare in $1+\lceil \log_2 n \rceil$ cicli su n processori, effettuando prima i prodotti e poi sommando gli n numeri così ottenuti con la somma parallela.

Il prodotto matrice-vettore richiede n prodotti scalari che si possono eseguire in parallelo, quindi su n^2 processori è possibile calcolare il prodotto matrice-vettore in $1 + \lceil \log_2 n \rceil$ cicli.

Il prodotto tra due matrici $n \times n$ righe per colonne, richiede n^2 prodotti scalari, tutti indipendenti, quindi su n^3 processori è possibile calcolare il prodotto matrice-matrice in $1 + \lceil \log_2 n \rceil$ cicli!

Anche se a prima vista non sembra, è possibile parallelizzare anche i due algoritmi per la valutazione di un polinomio visti sopra. Infatti il passo dell'algoritmo standard di valutazione può essere visto come il prodotto di una matrice 2×2 per un vettore

$$\begin{bmatrix} s \\ p \end{bmatrix} := \begin{bmatrix} 1 & a_i \\ 0 & x \end{bmatrix} \begin{bmatrix} s \\ p \end{bmatrix},$$

quindi l'intero procedimento può essere visto come il prodotto di $n+1$ matrici 2×2 per un vettore

$$\begin{bmatrix} 1 & a_n \\ 0 & x \end{bmatrix} \begin{bmatrix} 1 & a_{n-1} \\ 0 & x \end{bmatrix} \cdots \begin{bmatrix} 1 & a_2 \\ 0 & x \end{bmatrix} \begin{bmatrix} 1 & a_1 \\ 0 & x \end{bmatrix} \begin{bmatrix} 1 & a_0 \\ 0 & x \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Nell'algoritmo standard il prodotto sopra viene eseguito con l'ordine dato dalle parentesi nel seguente modo

$$\begin{bmatrix} 1 & a_n \\ 0 & x \end{bmatrix} \left(\begin{bmatrix} 1 & a_{n-1} \\ 0 & x \end{bmatrix} \cdots \left(\begin{bmatrix} 1 & a_2 \\ 0 & x \end{bmatrix} \left(\begin{bmatrix} 1 & a_1 \\ 0 & x \end{bmatrix} \left(\begin{bmatrix} 1 & a_0 \\ 0 & x \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right) \right) \right) \cdots \right).$$

Usando la proprietà associativa è possibile calcolare la stessa quantità, moltiplicando prima tra loro tutte le matrici e poi moltiplicando il risultato per il vettore $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

$$\left(\begin{bmatrix} 1 & a_n \\ 0 & x \end{bmatrix} \begin{bmatrix} 1 & a_{n-1} \\ 0 & x \end{bmatrix} \cdots \begin{bmatrix} 1 & a_2 \\ 0 & x \end{bmatrix} \begin{bmatrix} 1 & a_1 \\ 0 & x \end{bmatrix} \begin{bmatrix} 1 & a_0 \\ 0 & x \end{bmatrix} \right) \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Non sembra un grosso affare, visto che moltiplicare $n+1$ matrici 2×2 e poi moltiplicare il risultato per un vettore è più costoso che moltiplicarle consecutivamente per un vettore, ma facendo alcune osservazioni si scopre che può essere conveniente.

Per prima cosa, si osserva che tutte le matrici sono del tipo $\left[\begin{smallmatrix} 1 & \alpha \\ 0 & \beta \end{smallmatrix}\right]$, con $\alpha, \beta \in \mathbb{K}$, e che moltiplicando due matrici di questo tipo si ottiene un'altra matrice di questo tipo, infatti

$$\begin{bmatrix} 1 & \alpha \\ 0 & \beta \end{bmatrix} \begin{bmatrix} 1 & \gamma \\ 0 & \delta \end{bmatrix} = \begin{bmatrix} 1 & \gamma + \alpha\delta \\ 0 & \beta\delta \end{bmatrix}.$$

Avendo a disposizione due processori, un tale prodotto può essere eseguito in due cicli eseguendo prima i prodotti $\alpha\delta$ e $\beta\gamma$ contemporaneamente e, successivamente, la somma $\gamma + \alpha\delta$.

Come seconda osservazione, ci accorgiamo che lo stesso algoritmo usato per la somma parallela, può essere usato per moltiplicare 2^k matrici in $k\ell$ cicli, dove ℓ è il numero di cicli necessario per moltiplicare due matrici. Quindi, siccome per moltiplicare due matrici di quel tipo sono richiesti 2 cicli e la struttura si preserva, se $n = 2^k - 1$ allora è possibile calcolare il prodotto di tutte le matrici con $2k$ cicli. Se $n+1$ non è potenza di due, si possono anteporre al prodotto matrici del tipo $\left[\begin{smallmatrix} 1 & 0 \\ 0 & x \end{smallmatrix}\right]$ in numero sufficiente a raggiungere 2^k matrici, e in questo caso il costo è di $2\lceil \log_2(n+1) \rceil$ cicli.

Infine, detta $M = \left[\begin{smallmatrix} 1 & s \\ 0 & t \end{smallmatrix}\right]$, per trovare il valore del polinomio, occorre moltiplicare M per il vettore $\left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]$, e selezionare la prima componente del risultato. Ci si accorge facilmente che è proprio s e quindi non è necessario calcolare tale prodotto, ma la valutazione del polinomio si può leggere in m_{12}.

In definitiva, si riesce a valutare il polinomio in $2(1 + \lceil \log_2 n \rceil)$ cicli di calcolo utilizzando meno di $2n$ processori.

Procedendo in modo analogo per l'algoritmo di Ruffini-Horner si scopre che in questo caso il passo è dato dal prodotto

$$\begin{bmatrix} s \\ 1 \end{bmatrix} := \begin{bmatrix} x & a_i \\ 0 & 1 \end{bmatrix} \begin{bmatrix} s \\ 1 \end{bmatrix},$$

quindi l'intero procedimento può essere visto come il prodotto di $n+1$ matrici 2×2 per un vettore

$$\begin{bmatrix} x & a_0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x & a_1 \\ 0 & 1 \end{bmatrix} \cdots \begin{bmatrix} x & a_{n-1} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x & a_n \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Ancora una volta, si osserva che tutte le matrici hanno una struttura molto particolare, infatti sono del tipo $\left[\begin{smallmatrix} \alpha & \beta \\ 0 & 1 \end{smallmatrix}\right]$, con $\alpha, \beta \in \mathbb{K}$, e che moltiplicando due matrici di questo tipo si ottiene un'altra matrice di questo tipo, infatti

$$\begin{bmatrix} \alpha & \beta \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \gamma & \delta \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \alpha\gamma & \alpha\delta + \beta \\ 0 & 1 \end{bmatrix}.$$

Avendo a disposizione più processori è possibile calcolare prima il prodotto tra le matrici in modo parallelo e poi moltiplicare il risultato per il vettore $\left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]$ (quest'ultima operazione non è necessaria, in quanto, come prima, il risultato si trova nell'elemento m_{12} del prodotto delle matrici). Ancora una volta, si riesce a valutare il polinomio in $2\lceil \log_2(n+1) \rceil$ cicli di calcolo.

In ambito parallelo, quindi, non sembra che ci sia un vantaggio ad utilizzare un metodo ottenuto dall'algoritmo di Ruffini-Horner, invece che da quello tradizionale.

Esercizio 2.5. Si studi la possibilità di parallelizzare la somma tra matrici e tra polinomi.

Soluzione. Poiché le matrici e i polinomi di grado massimo fissato sono spazi vettoriali di dimensione finita, essi sono isomorfi ai vettori, e quindi per sommare due di essi basta sommare le coordinate (gli elementi della matrice e i coefficienti del polinomio). Vale quindi lo stesso discorso che vale per la somma di vettori. Si possono sommare due matrici $m \times n$ in un solo ciclo di calcolo con mn processori e due polinomi di grado h e k, rispettivamente, in un solo ciclo di calcolo con $\max\{h, k\}$ processori. □

Esercizio 2.6. Dire come è possibile calcolare la somma di tutti gli elementi di una matrice $m \times n$ in $\lceil \log_2(mn) \rceil$ cicli e specificare quanti processori sono necessari.

Soluzione. È sufficiente osservare che il problema è esattamente quello di sommare mn numeri e quindi utilizzando l'algoritmo di somma parallela, scoprima che con $mn/2$ processori si può eseguire il calcolo in $\lceil \log_2(mn) \rceil$ cicli. □

Abbiamo visto che è possibile sommare due vettori in un ciclo con n processori, quindi se $k \geqslant n$ basta un ciclo. Se i processori sono meno di n, il massimo numero di operazioni che si possono eseguire in parallelo in un ciclo sono k e quindi per eseguirne n sono necessari $\lceil n/k \rceil$ cicli (la formula vale anche per $k \geqslant n$). Un possibile pseudocodice è

```
for i=1:ceil(n/k)
  parallel for j=k*(i-1)+1:k*i
     u(j)=v(j)+w(j);
  end
end
```

Ad esempio se vogliamo sommare due vettori di dimensione 8 con tre processori, si può eseguire il seguente algoritmo

ciclo 1	$\boxed{v_1 + w_1}$	$\boxed{v_2 + w_2}$	$\boxed{v_3 + w_3}$
ciclo 2	$\boxed{v_4 + w_4}$	$\boxed{v_5 + w_5}$	$\boxed{v_6 + w_6}$
ciclo 3	$\boxed{v_7 + w_7}$	$\boxed{v_8 + w_8}$	

che in pseudocodice è

```
parallel
  u(1)=v(1)+w(1); u(2)=v(2)+w(2); u(3)=v(3)+w(3);
end
parallel
  u(4)=v(4)+w(4); u(5)=v(5)+w(5); u(6)=v(6)+w(6);
end
parallel
  u(7)=v(7)+w(7); u(8)=v(8)+w(8);
end
```

Si noti che all'ultimo passo, ci sono dei processori liberi che possono essere usati, in principio, per qualche altra operazione.

Strutture di matrici

Le matrici che provengono dalle applicazioni hanno, nella maggior parte dei casi, una qualche struttura più o meno facilmente identificabile. Possiamo definire *struttura di matrici* un qualsiasi sottoinsieme delle matrici, tuttavia le strutture interessanti sono quelle che ci permettono di progettare algoritmi *ad hoc* e che abbiano proprietà computazionali migliori rispetto a quelle date dall'algoritmo standard (per esempio un costo computazionale inferiore o maggiore stabilità numerica). In questo caso parleremo di *algoritmo strutturato*.

La struttura più facilmente identificabile è quella in cui un certo numero di elementi della matrice sono nulli. Ad esempio, una matrice A è detta diagonale se $a_{ij} = 0$ quando $i \neq j$, essa ha un numero di elementi non nulli che è al massimo n, inoltre le matrici diagonali formano una sottoalgebra delle matrici quadrate (cfr. Esercizio **??**). Nella pratica, è spesso sufficiente rappresentare le matrici strutturate dando una loro rappresentazione grafica utilizzando i "puntini", oppure indicando solo gli elementi eventualmente non nulli. Si può anche indicare con un asterisco l'elemento che può essere non nullo e identificare in questo modo la struttura. Una matrice D diagonale si potrà rappresentare come

$$D = \begin{bmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & d_{nn} \end{bmatrix} = \begin{bmatrix} d_{11} & & & \\ & d_{22} & & \\ & & \ddots & \\ & & & d_{nn} \end{bmatrix} = \begin{bmatrix} * & & & \\ & * & & \\ & & \ddots & \\ & & & * \end{bmatrix}.$$

Occorre prestare attenzione al fatto che nella definizione non è richiesto che gli elementi sulla diagonale siano non nulli.

Altri esempi di strutture sono le matrici triangolari superiori, cioè le matrici $T \in \mathbb{K}^{n\times n}$ tali che $t_{ij} = 0$ se $i > j$ e che si rappresentano come

$$T = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ 0 & t_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & t_{nn} \end{bmatrix} = \begin{bmatrix} * & * & \cdots & * \\ & * & \ddots & \vdots \\ & & \ddots & \vdots \\ & & & * \end{bmatrix}.$$

In modo analogo si possono definire le matrici triangolari inferiori (basta che $t_{ij} = 0$ per $i < j$).

Un altro esempio di struttura è dato dalle matrici tridiagonali, cioè le matrici $A \in \mathbb{K}^{n\times n}$ tali che $a_{ij} = 0$ se $|i - j| > 1$. Esse si rappresentano come

$$A = \begin{bmatrix} a_{11} & a_{12} & 0 & \cdots & 0 \\ a_{21} & a_{22} & a_{23} & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & a_{n-1,n} \\ 0 & \cdots & 0 & a_{n,n-1} & a_{nn} \end{bmatrix} = \begin{bmatrix} * & * & & & \\ * & * & * & & \\ & \ddots & \ddots & \ddots & \\ & & * & * & * \\ & & & * & * \end{bmatrix}.$$

L'insieme delle matrici tridiagonali formano uno spazio vettoriale di dimensione $3n-2$, tuttavia esse non formano un'algebra poiché il prodotto tra due matrici tridiagonali non è tridiagonale. Il termine tridiagonale è riferito al fatto che la matrice è nulla al di fuori di tre diagonali che sono la diagonale principale e la sua sopra- e sotto-diagonale. In modo analogo si può definire una matrice pentadiagonale.

In generale si può definire una matrice *A a banda di ampiezza k* come una matrice tale che $a_{ij} = 0$ per $|i - j| > k$ e quindi la matrice tridiagonale è a banda di ampiezza 1.

Se una struttura può essere definita per ogni m, n dove m e n sono le dimensioni della matrice, allora ha senso chiedersi qual è il massimo numero di elementi non nulli asintoticamente. Sia $S(m, n)$ il massimo numero di elementi non nulli di una matrice strutturata, se $S(m, n) \ll mn$ (nel senso che $\lim_{m,n\to\infty} S(m, n)/mn = 0$), allora si dice che la matrice è *sparsa* e la struttura si dice struttura di sparsità.

Nella pratica, si usa la definizione di matrice sparsa anche per matrici di dimensioni finite in cui il numero di elementi non nulli sia molto minore (in un qualche senso) delle dimensioni della matrice.

Per concludere, vorremmo sottolineare il fatto che una struttura utile dal punto di vista computazionale non deve essere per forza legata alla presenza di elementi nulli. Per esempio si può considerare la struttura

$$\mathcal{S}_n = \{A \in \mathbb{C}^{n\times n} \,:\, A = vw^T, v, w \in \mathbb{C}^n\}$$

in cui le matrici possono anche non avere alcun elemento nullo.

Esercizio 2.7. Provare che le matrici diagonali sono una sottoalgebra delle matrici $n \times n$ (con elementi in un campo $\mathbb{K}$).

Soluzione. Sia $\mathcal{D}_n$ l'insieme delle matrici diagonali di dimensione n. Occorre provare

1. $0 \in \mathcal{D}_n$ (dove 0 è intesa come la matrice nulla);
2. se $A, B \in \mathcal{D}_n$ allora $A + B \in \mathcal{D}_n$;
3. se $A \in \mathcal{D}_n$ e $\alpha \in \mathbb{K}$, allora $\alpha A \in \mathcal{D}_n$;
4. se $A, B \in \mathcal{D}_n$ allora $AB \in \mathcal{D}_n$.

Le prime tre proprietà ci dicono che $\mathcal{D}_n$ è un sottospazio vettoriale. Per mostrare la proprietà 1 è sufficiente osservare che la matrice nulla è una matrice diagonale. Per mostrare le proprietà 2 e 3 osserviamo che, se $\alpha \in \mathbb{K}$ e

$$A = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_{nn} \end{bmatrix}, \qquad B = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_{nn} \end{bmatrix}$$

allora

$$A + B = \begin{bmatrix} a_{11} + b_{11} & 0 & \cdots & 0 \\ 0 & a_{22} + b_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_{nn} + b_{nn} \end{bmatrix} \in \mathcal{D}_n, \qquad \alpha A = \begin{bmatrix} \alpha a_{11} & 0 & \cdots & 0 \\ 0 & \alpha a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \alpha a_{nn} \end{bmatrix} \in \mathcal{D}_n.$$

In modo più formale, si può dire che se $a_{ij} = b_{ij} = 0$ per $i \neq j$ allora $a_{ij} + b_{ij} = 0$ e $\alpha a_{ij} = 0$ per $i \neq j$ e questo mostra che la somma e il prodotto per scalare rimangono diagonali. Infine, per la proprietà 4, date $A, B \in \mathcal{D}_n$ e posto $C = AB$, si osserva che $c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}$ e la somma si riduce al solo addendo $a_{ii}b_{ij}$ (perché per $k \neq i$ il prodotto $a_{ik}b_{kj} = 0$). Ma $a_{ii}b_{ij} = 0$ se $i \neq 0$ (perché lo è b_{ij}) e quindi C è diagonale. In alternativa, si può osservare che

$$AB = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_{nn} \end{bmatrix} \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_{nn} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & 0 & \cdots & 0 \\ 0 & a_{22}b_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_{nn}b_{nn} \end{bmatrix} \in \mathcal{D}_n.$$

□

Esercizio 2.8. Dimostrare che la dimensione del sottospazio $\mathcal{D}_n$ delle matrici diagonali $n \times n$ è n.

Soluzione. La dimensione di un sottospazio vettoriale è il numero di elementi di una sua base. Occorre quindi trovare n matrici diagonali che sono linearmente indipendenti e generano $\mathcal{D}_n$. Tale insieme può essere, per esempio

$$D_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 0 \end{bmatrix}, \qquad D_2 = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 0 \end{bmatrix}, \qquad \cdots, \qquad D_n = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix}.$$

Si osserva che una combinazione lineare di esse è nulla, cioè

$$\alpha_1 D_1 + \alpha_2 D_2 + \cdots + \alpha_n D_n = \begin{bmatrix} \alpha_1 & 0 & \cdots & 0 \\ 0 & \alpha_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \alpha_n \end{bmatrix} = 0$$

se e solo se $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$ e quindi sono linearmente indipendenti. Inoltre generano $\mathcal{D}_n$, infatti data $A \in \mathcal{D}_n$ si ha

$$A = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_{nn} \end{bmatrix} = a_{11} D_1 + a_{22} D_2 + \cdots + a_{nn} D_n.$$

□

Esercizio 2.9. Provare che le matrici triangolari superiori o inferiori sono una sottoalgebra delle matrici quadrate e trovarne la dimensione (che è il massimo numero di elementi non nulli).

Soluzione. È abbastanza evidente che la somma di due matrici triangolari è ancora triangolare, così come il prodotto di una matrice triangolare per uno scalare. Per quanto riguarda il prodotto: siano $S, T \in \mathbb{K}^{n\times n}$ due matrici triangolari superiori e sia $C = ST$, si osserva che per $i > j$,

$$c_{ij} = \sum_{k=1}^{n} s_{ik}c_{kj} = \sum_{k=1}^{i-1} s_{ik}c_{kj} + \sum_{k=i}^{n} s_{ik}c_{kj} = \sum_{k=1}^{i-1} 0 \cdot c_{kj} + \sum_{k=i}^{n} s_{ik} \cdot 0 = 0.$$

La prima delle due somme è nulla poiché $s_{ik} = 0$ per $i > k$, in particolare per $k = 1, \ldots, i-1$, mentre la seconda è nulla poiché $c_{kj} = 0$ per $k > j$, in particolare per $k = i, \ldots, n$. □

Esercizio 2.10. Mostrare che il prodotto di due matrici tridiagonali non è tridiagonale, dire se esso ha qualche struttura.

Soluzione. Il fatto che non sia tridiagonale si può osservare prendendo una matrice tridiagonale generica 3×3 con elementi positivi sulle tre diagonali e calcolare il quadrato.

Si osserva che il prodotto di due matrici tridiagonali è pentadiagonale. Questo fatto si può dimostrare considerando due matrici tridiagonali $A, B \in \mathbb{K}^{n\times n}$ (con $n > 3$) e il loro prodotto $C = AB$. Per fissare le idee consideriamo gli elementi di indice $i - j > 2$

$$c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj} = \sum_{k=1}^{j+1} a_{ik}b_{kj} + \sum_{k=j+2}^{n} a_{ik}b_{kj} = \sum_{k=1}^{j+1} 0 \cdot b_{kj} + \sum_{k=j+2}^{n} a_{ik} \cdot 0 = 0,$$

la prima somma è nulla poiché $a_{ik} = 0$ per $k < i - 1$ in particolare per $k = 1, \ldots, j+1$, mentre la seconda somma è nulla poiché $b_{kj} = 0$ per $k > j + 1$ dalla definizione di matrice tridiagonale. In modo del tutto analogo si dimostra che $c_{ij} = 0$ anche quando $i - j < -2$ e quindi alla fine $c_{ij} = 0$ per $|i - j| > 2$ e la matrice è pentadiagonale. □

Esercizio 2.11. Dire se le matrici a banda di ampiezza t sono sparse.

Soluzione. Sono sparse, in quanto per n che tende a infinito gli elementi non nulli sono al più

$$n + \sum_{k=1}^{t} 2(n-k) = n + 2\left(tn - \frac{t(t+1)}{2}\right) = (2t+1)n - t^2 - t \approx (2t+1)n \ll n^2.$$

□

Esercizio 2.12. Sia data la struttura $\mathcal{S}_n = \{A \in \mathbb{C}^{n\times n} \,:\, A = vw^T, v, w \in \mathbb{C}^n\}$, si mostri che $\mathcal{S}_n$ è l'insieme delle matrici di rango al più 1 e che anche se è chiuso rispetto al prodotto, non è un'algebra.

Soluzione. Le matrici in $\mathcal{S}_n$ hanno rango minore o uguale a 1, infatti, scrivendo A per colonne come $A = [w_1 v | w_2 v | \cdots | w_n v]$, si osserva che tutte le colonne sono multiple dello stesso vettore e quindi A non può avere rango maggiore di 1. Viceversa, se una matrice A ha rango minore o uguale a 1, tutte le colonne sono multiple di un unico vettore per cui $A = vw^T$.

Siano $A = v_1 w_1^T$ e $B = v_2 w_2^T$ allora $AB = (v_1 w_1^T)(v_2 w_2^T) = v_1 (w_1^T v_2) w_2^T$, ma $\alpha = w_1^T v_2$ è uno scalare e quindi si può scrivere $AB = (\alpha v_1) w_2^T$ da cui $AB \in \mathcal{S}_n$. La struttura $\mathcal{S}_n$ non è un'algebra per $n \geqslant 2$, in quanto non è uno spazio vettoriale, poiché la somma di due matrici di $\mathcal{S}_n$ non è necessariamente una matrice di $\mathcal{S}_n$. Questo fatto si può vedere con un controesempio: siano

$$A = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix},$$

si osserva che $A, B \in \mathcal{S}_2$, ma $A + B = I$ che ha rango 2 e quindi $A + B \notin \mathcal{S}_2$. Analoghi controesempi si possono trovare per ogni $n \geqslant 2$, mentre $\mathcal{S}_1$ è un'algebra poiché $\mathcal{S}_1 = \mathbb{R}$.

In alternativa, si può osservare che $I \notin \mathcal{S}_n$ usando un metodo di forza bruta, cioè cercando soluzioni dell'equazione

$$I = \begin{bmatrix} x \\ y \end{bmatrix} \begin{bmatrix} z & t \end{bmatrix} \iff \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} xz & xt \\ yz & yt \end{bmatrix} \iff \begin{cases} xz = 1, \\ xt = 0, \\ yz = 0, \\ yt = 1 \end{cases}$$

dove le incognite sono x, y, z, t. Affinché l'identità sia vera, occorre che $xt = 0$ e quindi si danno due casi: 1) $x = 0$ e in questo caso $xz = 1$ non può essere verificata; 2) $t = 0$ e in questo caso $yt = 1$ non può essere verificata. Quindi I non si può scrivere come prodotto di due vettori. □

Algoritmi strutturati

Diamo alcuni esempi di algoritmi strutturati, facendo vedere come varia il costo computazionale.

Come primo esempio cerchiamo un algoritmo strutturato per il calcolo del prodotto di una matrice diagonale $A \in \mathbb{K}^{n\times n}$ per un vettore $b \in \mathbb{K}^n$. Detto $c = Ab$, per ogni $i = 1, \ldots, n$, si ha

$$c_i = a_{i1} b_1 + \cdots + a_{i,i-1} b_{i-1} + a_{ii} b_i + a_{i,i+1} b_{i+1} + \cdots + a_{in} b_n = a_{ii} b_i.$$

Per ottenere b_i è sufficiente calcolare il prodotto $a_{ii} b_i$ e quindi eseguire una sola operazione, anziché $2n - 1$. Quindi il costo di questo algoritmo strutturato è $\mathcal{C}(n) = n \ll 2n^2 - n$.

Consideriamo ora il caso in cui la matrice A sia tridiagonale. Ci accorgiamo che per $n > 2$ è possibile calcolare c nel seguente modo

$$\begin{cases} c_1 = a_{11} b_1 + a_{12} b_2, \\ c_i = a_{i,i-1} b_{i-1} + a_{ii} b_i + a_{i,i+1} b_{i+1}, \quad i = 2, \ldots, n-1, \\ c_n = a_{n,n-1} b_{n-1} + a_{nn} b_n, \end{cases}$$

e quindi abbiamo un algoritmo strutturato che esegue 3 operazioni per calcolare c_1 e c_n e 5 operazioni per il calcolo dei restanti $n - 2$ valori di c_i per un costo totale pari a

$$\mathcal{C}(n) = 5(n-2) + 3 + 3 = 5n - 4 \approx 5n.$$

La formula è valida per $n > 2$.

Un altro modo per calcolare il costo computazionale consiste nello scrivere l'algoritmo in pseudocodice

```
1 c(1)=a(1,1)*b(1)+a(1,2)*b(2);
2 for i=2:n-1
3    c(i)=a(i,i-1)*b(i-1)+a(i,i)*b(i)+a(i,i+1)*b(i+1);
4 end
5 c(n)=a(n,n-1)*b(n-1)+a(n,n)*b(n);
```

da cui si vede chiaramente che il costo di ogni istruzione del ciclo è 5 e quindi il costo del ciclo è $5(n-1-2+1) = 5(n-2)$, le altre due istruzioni costano 6 ops e alla fine si giunge allo stesso risultato.

Come ulteriore esempio consideriamo il prodotto di una matrice triangolare superiore $T \in \mathbb{K}^{n\times n}$ per un vettore $x \in \mathbb{K}^n$, anche in questo caso si può semplificare la formula che fornisce c_i tale che $c = Tb$.

Ci si accorge che, siccome $a_{ij} = 0$ per per $i > j$, si ha

$$c_i = a_{i1}b_1 + \cdots + a_{in}b_n = a_{ii}b_i + \cdots + a_{in}x_n,$$

cioè

$$c_i = \sum_{k=1}^{n} a_{ik}b_k = \sum_{k=i}^{n} a_{ij}b_k.$$

In altre parole, la somma che definisce c_i può partire da i senza che cambi il risultato e quindi c_i è la somma di $n - i + 1$ prodotti. Il costo per il calcolo di c_i è quindi di $2(n - i) + 1$ operazioni. Il costo totale si ottiene sommando i costi necessari per il calcolo di ciascun c_i, e insomma alla fine

$$\mathcal{C}(n) = \sum_{i=1}^{n} 2(n-i)+1 = 2\frac{n(n-1)}{2} + n = n^2.$$

Scrivendo l'algoritmo in pseudocodice

```
1 for i=1:n
2    c(i)=a(i,i)*b(i);
3    for k=i+1:n
4       c(i)=c(i)+a(i,k)*b(k);
5    end
6 end
```

si ottiene lo stesso risultato.

Per calcolare i costi computazionali possono essere utili le seguenti formule asintotiche

$$\sum_{k=1}^{n} k \approx \sum_{k=1}^{n}(n-k) \approx \frac{1}{2}n^2, \quad \sum_{k=1}^{n} k^2 \approx \frac{1}{3}n^3.$$

Naturalmente, il carattere asintotico delle precedenti formule, fa sì che esse continuino ad essere valide se, anziché sommare da 1 a n, si somma da un indice i_1 indipendente da n fino a $n - i_2$, dove i_2 è indipendente da n, ad esempio

$$\sum_{k=5}^{n+3} k^2 \approx \sum_{k=1000}^{n-5} (n-k)^2 \approx \frac{1}{3}n^3.$$

Nota. Lo sviluppo asintotico di $\sum_{k=1}^{n} k^\alpha$ con α intero positivo è dato dalla formula

$$\sum_{k=1}^{n} k^\alpha = \frac{k^{\alpha+1}}{\alpha+1},$$

la dimostrazione si può ottenere a partire dalla diseguaglianza $\int_0^n x^\alpha dx < \sum_{k=1}^{n} k^\alpha < \int_0^{n+1} x^\alpha dx$ che si verifica graficamente. Calcolando gli integrali si ottiene

$$\frac{n^{\alpha+1}}{\alpha+1} < \sum_{k=1}^{n} k^\alpha < \frac{(n+1)^{\alpha+1}}{\alpha+1} \approx \frac{n^{\alpha+1}}{\alpha+1},$$

da cui segue tutto.

Esercizio 2.13. Trovare un algoritmo strutturato per il calcolo del prodotto tra due matrici triangolari superiori e valutarne il costo computazionale.

Soluzione. Siano $A, B \in \mathbb{K}^{n\times n}$ e $C = AB$. Il prodotto di matrici triangolari è a sua volta triangolare, quindi basta calcolare c_{ij} per $i \leqslant j$,

$$c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj} = \sum_{k=1}^{i-1} a_{ik}b_{kj} + \sum_{k=i}^{j} a_{ik}b_{kj} + \sum_{k=j+1}^{n} a_{ik}b_{kj} = \sum_{k=i}^{j} a_{ik}b_{kj},$$

dove delle tre somme sopra la prima è nulla perché $a_{ik} = 0$ per $i > k$ e quindi per $k = 1, \ldots, j-1$, la terza è nulla perché $b_{kj} = 0$ per $k > j$ e quindi per $k = i+1, \ldots, n$.

Ora, per ogni $i \leqslant j$ si richiedono $2(j-i)+1$ ops e alla fine il costo è

$$\mathcal{C}(n) = \sum_{j=1}^{n} \sum_{i=1}^{j} 2(j-i)+1 \approx \sum_{j=1}^{n} 2(j^2 - \frac{j^2}{2}) = \sum_{j=1}^{n} j^2 \approx \frac{n^3}{3}.$$

L'esercizio può essere risolto anche in maniera meno formale facendo vedere con un disegno che per il calcolo di c_{ij} la sommatoria da calcolare ha indici estremi i e j. □

Prodotto di una matrice sparsa per un vettore

Finora abbiamo visto strutture più o meno facilmente identificabili dove l'implementazione di un algoritmo strutturato risultava semplice. Tuttavia, in altri casi è più difficile implementare un algoritmo strutturato.

Consideriamo il calcolo del prodotto di una matrice $A \in \mathbb{K}^{n \times n}$ che ha k elementi non nulli per un vettore $v \in \mathbb{K}^n$. Come abbiamo visto, se k è "piccolo" rispetto a n si dice comunemente che la matrice è sparsa.

Per prima cosa, ricordiamo l'implementazione per il calcolo del prodotto $c = Av$

```
1 for i=1:n
2   for j=1:n
3     c(i)=c(i)+a(i,j)*v(j);
4   end
5 end
```

dove si è assunto che `c` sia stato inizializzato come vettore nullo. Si osserva che nella formula, l'elemento a_{ij} compare solo una volta, moltiplicato per b_j e va a incrementare c_i. Quindi non è necessario calcolare il prodotto con l'ordine visto sopra, per esempio l'implementazione

```
1 for j=1:n
2   for i=1:n
3     c(i)=c(i)+a(i,j)*v(j);
4   end
5 end
```

fornisce lo stesso risultato. E la stessa cosa si ottiene se l'ordine con cui vengono selezionate le coppie (i, j) è arbitrario. A questo punto si può pensare di selezionare solo le coppie corrispondenti a elementi non nulli di A e fare un ciclo solo su esse.

Detto $\Omega \subset \{1, \ldots, n\} \times \{1, \ldots, n\}$ l'insieme delle coppie di indici tali che $a_{ij} \neq 0$, abbiamo che Ω contiene esattamente k elementi e il prodotto matrice-vettore $c = Av$ si può scrivere come

$$c_i = \sum_{(i,j) \in \Omega} a_{ik} v_k, \qquad i = 1, \ldots, n,$$

utilizzando la convenzione che la somma vuota è nulla (questo succede se non ci sono coppie del tipo $(i, k) \in \Omega$ e cioè se la riga i-esima di A è nulla). Notiamo che l'algoritmo scritto sopra richiede non più di $2k$ operazioni aritmetiche elementari anziché $2n^2$.

Un'implementazione in pseudocodice si può ottenere memorizzando l'insieme Ω in un array con k righe e due colonne che indicheremo con `ind`, in modo che `ind(t,1)` fornisca l'indice di riga e `ind(t,2)` fornisca l'indice di colonna dell'elemento t:

```
1 for t=1:k
2   i=ind(t,1);
3   j=ind(t,2);
4   c(i)=c(i)+a(i,j)*v(j);
5 end
```

dove si è assunto che il vettore `c` sia stato inizializzato a 0.

Questo algoritmo richiede solamente $2k$ ops, tuttavia è inefficiente dal punto di vista della gestione della memoria perché richiede che la matrice A venga mantenuta interamente in memoria. Ma, poiché gli elementi nulli non hanno nessuna funzione è possibile risparmiare memoria utilizzando un vettore di k strutture `mat`, in cui ogni struttura ha tre campi: uno per l'indice di riga i (ad es., `mat(t).i`), uno per l'indice di colonna j (ad es., `mat(t).j`) e uno per il valore a_{ij} (ad es., `mat(t).value`).

L'algoritmo relativo a questa implementazione, in pseudocodice è

```
1 for t=1:k
2   c(mat(t).i)=c(mat(t).i)+mat(t).value*v(mat(t).j);
3 end
```

Riferimenti bibliografici

Per quanto riguarda il metodo di Ruffini Horner, si faccia riferimento all'esempio 1.10 del testo [1]; per approfondire il metodo di Strassen, si faccia rifemento all'esempio 1.9 del testo [1] o all'articolo originale [2]; per gli algoritmi di somma sequenziale e parallela si faccia riferimento all'esempio 2.28 di [1], dove viene anche effettuata un'analisi dell'errore dei due algoritmi; per la parallelizzazione degli algoritmi di valutazione di polinomi si faccia riferimento a [3].

[1] R. Bevilacqua, D. Bini, M. Capovani, O. Menchi. *Metodi Numerici.* Zanichelli, Bologna, 1992.
[2] V. Strassen. *Gaussian Elimination is not Optimal.* Numer. Math., 13, 1969, pp. 354–356.
[3] D. Bertaccini, C. Di Fiore, P. Zellini. *Complessità e iterazione numerica. Percorsi, matrici e algoritmi veloci nel calcolo numerico.* Bollati Boringhieri, 2013.

3 Sistemi lineari e algoritmo di Gauss

Richiami sui sistemi lineari

Un sistema lineare di m equazioni in n incognite si può scrivere come

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2, \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m, \end{cases}$$

dove ogni equazione è lineare. Usando le sommatorie si può sintetizzare la scrittura come

$$\sum_{j=1}^{n} a_{ij}x_j = b_i, \quad i = 1, \ldots, m.$$

Tuttavia, c'è una notazione ancora più elegante. Definendo le matrici

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix},$$

il sistema può essere scritto come

$$Ax = b.$$

L'esistenza e la molteplicità delle soluzioni è caratterizzata dal seguente risultato dell'algebra lineare dove si usa la matrice $[A|b] \in \mathbb{K}^{m\times(n+1)}$ ottenuta incollando la colonna b in fondo alla matrice A.

Teorema 3.1 (Rouché-Capelli)**.** *Il sistema lineare $Ax = b$, con $A \in \mathbb{K}^{m\times n}$ e $b \in \mathbb{K}^m$, ammette soluzioni se e solo se il rango di A coincide con il rango di $[A|b]$. Nel caso in cui esistano soluzioni, esse formano un sottospazio affine di $\mathbb{K}^n$ di dimensione $n - \operatorname{rank}(A)$.*

Nel caso di matrici quadrate il teorema si semplifica molto e si ottiene il seguente corollario.

Teorema 3.2 (Cramer)**.** *Il sistema lineare quadrato $Ax = b$ ammette soluzione unica se e solo se la matrice A è invertibile.*

In quest'ultimo caso la soluzione simbolica si può scrivere semplicemente come $x = A^{-1}b$. Nell'ambito del calcolo numerico questa scrittura non è soddisfacente, al contrario, si vuole fornire un algoritmo che calcoli la soluzione.

Il problema della soluzione di un sistema lineare

Sia $A \in \mathbb{K}^{m\times n}$, essa definisce un'applicazione lineare $f : \mathbb{K}^n \to \mathbb{K}^m$, quella che al vettore $v \in \mathbb{K}^m$ associa il vettore $f(v) = Av \in \mathbb{K}^n$ (dove v e $f(v)$ sono espressi in opportune base). Si può definire l'insieme $\operatorname{Im}(A) = \{v \in \mathbb{K}^m : v = Aw \text{ per qualche } w \in \mathbb{K}^n\}$, detto l'*immagine* di A, che è un sottospazio vettoriale di $\mathbb{K}^m$ generato dalle colonne di A e la sua dimensione è $\operatorname{rank}(A)$ (cfr. Esercizio **??**).

Ci si può chiedere sotto quali ipotesi la soluzione di un sistema lineare è un problema ben posto per $\mathbb{K} = \mathbb{R}$ (il risultato è analogo per $\mathbb{K} = \mathbb{C}$).

Per definizione, il problema $Ax = b$ è *ben posto* se e solo se ammette soluzione unica x^* ed esiste $\varepsilon > 0$ tale che per ogni $\widetilde{A} = (\widetilde{a}_{ij})_{i=1,\ldots,m;j=1,\ldots,n}$ con $|\widetilde{a}_{ij} - a_{ij}| < \varepsilon$ e per ogni $\widetilde{b} = (\widetilde{b}_i)_{i=1,\ldots,m}$ con $|\widetilde{b}_i - b_i| < \varepsilon$ il sistema lineare $\widetilde{A}x = \widetilde{b}$ ammette soluzione unica $\widetilde{x}$ tale che $\lim_{\varepsilon\to 0} \widetilde{x} = x^*$.

Teorema 3.3. *Sia $A \in \mathbb{R}^{m\times n}$ e $b \in \mathbb{R}^m$, allora il sistema lineare $Ax = b$ è un problema ben posto se e solo se A è quadrata e invertibile.*

Dimostrazione. Distinguiamo quattro casi: $m < n$, $m > n$, $m = n$ con A non invertibile e $m = n$ con A invertibile. Mostreremo che nei primi tre casi il problema non è ben posto, mentre nel quarto lo è.

Osserviamo preliminarmente che una condizione necessaria affinché il problema sia ben posto è che $\operatorname{rank}(A) = n$, altrimenti, cioè se $\operatorname{rank}(A) < n$, il sistema non ammette soluzioni o esse sono un sottospazio affine di dimensione $n - \operatorname{rank}(A) > 0$ e quindi sono infinite. Quindi se $\operatorname{rank}(A) < n$, allora si possono avere nessuna o infinite soluzioni e il problema non è ben posto.

Caso $m < n$. Poiché il rango di A è sicuramente minore o uguale alla più piccola delle dimensioni di A, abbiamo che $\operatorname{rank}(A) \leqslant m < n$, e quindi $\operatorname{rank}(A) < n$ e il problema non è ben posto.

Caso $m > n$. Sappiamo che $\operatorname{rank}(A) \leqslant n < m$. Siccome il rango è la dimensione del sottospazio vettoriale generato dalle colonne di A, questo implica che le sue colonne $c_1, \dots, c_n$ non possono essere una base di $\mathbb{R}^m$, che è equivalente a dire che esiste $c_0 \in \mathbb{R}^m$ tale che c_0 non è combinazione lineare di $c_1, \dots, c_n$, che è come dire che il sistema $Ax = c_0$ non ammette soluzione. Se $Ax = b$ non ammette soluzione allora il problema non è ben posto. Se $Ax = b$ ammette soluzione x^*, dimostriamo che il sistema $Ax = b_\varepsilon = b + \varepsilon c_0$, con $\varepsilon > 0$, non ammette soluzione. Per contraddizione, se il esiste un vettore y tale che $Ay = b_\varepsilon$, allora

$$A\Big(\frac{y - x^*}{\varepsilon}\Big) = \frac{1}{\varepsilon}(Ay - Ax^*) = \frac{1}{\varepsilon}(b + \varepsilon c_0 - b) = c_0$$

e quindi avremmo trovato una soluzione del sistema $Ax = c_0$, che è una contraddizione, quindi il sistema il sistema $Ax = b_\varepsilon$ non ammette soluzione per ogni $\varepsilon > 0$. Ma $\lim_{\varepsilon \to 0} b_\varepsilon = b$ e quindi non si ha dipendenza continua dai dati e il problema non è ben posto.
Caso $m = n$ e A non invertibile. Se A non è invertibile, allora $\operatorname{rank}(A) < n$ e il problema non è ben posto.
Caso $m = n$ e A invertibile. Se A è invertibile allora la soluzione è unica, quindi è sufficiente mostrare che dipende in modo continuo dai dati. Siccome il determinante è una funzione continua, esiste $\varepsilon > 0$ tale che se $|\widetilde{a}_{ij} - a_{ij}| < \varepsilon$ per ogni i, j allora $\widetilde{A} = (\widetilde{a}_{ij})$ è invertibile e quindi il sistema $\widetilde{A}x = \widetilde{b}$ ammette soluzione unica $\widetilde{x} = \widetilde{A}^{-1}\widetilde{b}$, per ogni $\widetilde{b}$ (in particolare per ogni $\widetilde{b}$ tale che $|\widetilde{b}_i - b_i| < \varepsilon$). Ma dalle formule dell'inversa si ha che $\widetilde{A}^{-1}$ è una funzione razionale degli elementi di $\widetilde{A}$ e quindi $\widetilde{x}$ è una funzione razionale degli elementi di $\widetilde{A}$ e $\widetilde{b}$, e in particolare è continua. In conclusione, il problema è ben posto. □

In virtù del precedente risultato, ci preoccuperemo di fornire un algoritmo per la risoluzione numerica del sistema lineare $Ax = b$ solo non caso in cui A sia quadrata e invertibile.

Esercizio 3.4. Sia $A \in \mathbb{K}^{m \times n}$. Mostrare che $\operatorname{Im}(A)$ è un sottospazio vettoriale di $\mathbb{K}^m$ generato dalle colonne di A.

Idea degli algoritmi, sistemi triangolari

Nel seguito ci occuperemo di descrivere un'algoritmo per il calcolo della soluzione (unica) di un sistema quadrato con matrice dei coefficienti invertibile. Poiché sono note formule esplicite per la soluzione date da una funzione razionale dei dati, allora si può pensare a un algoritmo che, in aritmetica esatta, fornisca la soluzione con un numero finito di operazioni. Un tale metodo è detto *metodo diretto* per la soluzione di un sistema lineare (in contrapposizione con i metodi iterativi che invece costruiscono una successione che approssima la soluzione). Il metodo diretto di base, che con un'opportuna variante, è anche quello più usato, è il metodo di Gauss.

L'idea del metodo di Gauss è di eseguire una serie di operazioni che trasformino il sistema $Ax = b$ in un altro sistema $Tx = \widetilde{b}$ più facile da risolvere, ma che abbia la stessa soluzione. Per prima cosa cerchiamo di capire quali sono i sistemi facili da risolvere.

Consideriamo un sistema lineare $Tx = \widetilde{b}$ la cui matrice dei coefficienti è triangolare superiore. Ad esempio, nel caso 3×3, un sistema con matrice dei coefficienti triangolare è

$$\begin{cases} t_{11}x_1 + t_{12}x_2 + t_{13}x_3 = \widetilde{b}_1, \\ \qquad\quad t_{22}x_2 + t_{23}x_3 = \widetilde{b}_2, \\ \qquad\qquad\quad t_{33}x_3 = \widetilde{b}_3, \end{cases}$$

la cui soluzione è ottenuta nel seguente modo

- osservando che nell'ultima equazione compare la sola variabile x_3 si può ricavare direttamente $x_3 = \widetilde{b}_3/t_{33}$;
- una volta calcolato x_3, si osserva che la seconda equazione può essere scritta come $t_{22}x_2 = \widetilde{b}_2 - t_{23}x_3$ e si può ricavare $x_2 = (\widetilde{b}_2 - t_{23}x_3)/t_{22}$;
- infine, utilizzando i valori già calcolati di x_2 e x_3, si può calcolare x_1 dalla prima equazione riscritta come $t_{11}x_1 = \widetilde{b}_1 - t_{12}x_2 - t_{13}x_3$.

Questo procedimento si può generalizzare a un sistema di n equazioni con matrice dei coefficienti triangolari, diciamo $Tx = \widetilde{b}$, ottenedo il *metodo di sostituzione all'indietro*. Si parte dall'ultima equazione, si calcola x_n e si sostituisce nella penultima, dove si calcola x_{n-1}, e così via, finché non si arriva alla prima.

Per scrivere in modo più formale l'algoritmo consideriamo l'equazione k-esima

$$t_{kk}x_k + t_{k,k+1}x_{k+1} + \cdots + t_{kn}x_n = \widetilde{b}_k,$$

che verrà risolta, quando sono già noti $x_{k+1}, \dots, x_n$, e quindi può essere riscritta come

$$t_{kk}x_k = \widetilde{b}_k - t_{k,k+1}x_{k+1} + \cdots + t_{kn}x_n = \widetilde{b}_k - \sum_{j=k+1}^{n} t_{kj}x_j,$$

e alla fine

$$x_k = \frac{1}{t_{kk}}\left(\widetilde{b}_k - \sum_{j=k+1}^{n} t_{kj}x_j\right), \quad k = 1, \ldots, n, \tag{2}$$

con la convenzione che una somma da $n+1$ a n si intende uguale a zero.

L'equazione (**??**) può essere immediatamente trasformata in algoritmo, se la formula viene valutata a partire da x_n fino a x_1. Questo algoritmo prende il nome di metodo sostituzione all'indietro (o *backward (back) substitutiton*).

Diamo un'implementazione in pseudocodice del metodo di sostituzione all'indietro:

```
1 for k=n:-1:1
2    s=b(k);
3    for j=k+1:n
4       s=s-t(k,j)*x(j);
5    end
6    x(k)=s/t(k,k);
7 end
```

Analizziamo ora l'algoritmo di sostituzione all'indietro. Cercheremo di capire qual è il suo costo computazionale e per quali problemi l'algoritmo fornisce la soluzione senza *breakdown*, cioè senza interruzioni dovute a errori.

La valutazione della formula (**??**) (che coincide con le righe 2-6 del codice) richiede $n-k$ somme, $n-k$ prodotti e 1 divisione per un totale di $2(n-k)+1$ ops. Ma la formula va valutata per ogni k e quindi il costo totale è dato da

$$\mathcal{C}(n) = \sum_{k=1}^{n} 2(n-k)+1 = 2\frac{n(n-1)}{2} + n = n^2.$$

Si osservi che tale costo è identico al costo del prodotto di una matrice triangolare per un vettore, nonostante la risoluzione di un sistema lineare sia un problema appartentemente più difficile. Questo ci fa concludere che la risoluzione di un sistema lineare con matrice dei coefficienti triangolari è un "problema facile".

Ora cerchiamo di capire sotto quali ipotesi l'algoritmo termina. Si nota che l'unico caso in cui si ha un breakdown è quello in cui $t_{kk} = 0$ per qualche k, ma questo vale se e solo se la matrice T è singolare (si confronti l'esercizio **??**) che è come dire che il sistema non ha soluzione unica. Ne deduciamo che nelle nostre ipotesi l'algoritmo di sostituzione all'indietro è sempre applicabile.

In modo totalmente analogo si può definire l'algoritmo di *sostituzione in avanti* (o *forward substitution*) per risolvere un sistema lineare con matrice dei coefficienti triangolare inferiore, diciamo $Lx = c$.

La differenza consiste nel fatto che adesso si può calcolare x_1 dalla prima equazione, sostituirlo nella seconda ottenendo x_2 e ottenere via via $x_3, \ldots, x_n$.

La formula per il termine x_k è

$$x_k = \frac{1}{\ell_{kk}}\left(c_k - \sum_{j=1}^{k-1} \ell_{kj}x_j\right), \quad k = 1, \ldots, n,$$

mentre un'implementazione in pseudocodice del metodo di sostituzione in avanti è data da

```
1 for k=1:n
2    s=c(k);
3    for j=1:k-1
4       s=s-l(k,j)*x(j);
5    end
6    x(k)=s/l(k,k);
7 end
```

Come per il metodo di sostituzione in avanti il costo dell'algoritmo è di n^2 ops ed è sempre applicabile se la matrice L è invertibile.

Nota. I sistemi triangolari non sono gli unici sistemi che si sanno risolvere in $O(n^2)$ ops. Un altro caso interessante è quello del sistema $Qx = b$ dove $Q \in \mathbb{R}^{n\times n}$ è una matrice ortogonale, cioè una matrice tale che $Q^TQ = I$. In questo caso $Q^{-1} = Q^T$ e quindi $x = Q^{-1}b = Q^Tb$ si ottiene moltiplicando la trasposta di Q per b, con un costo di circa $2n^2$ ops. Questa idea è usata nel metodo di soluzione di un sistema lineare basato sulla fattorizzazione QR.

Esercizio 3.5. Mostrare che una matrice triangolare è invertibile se e solo se tutti gli elementi della diagonale sono diversi da 0.

Soluzione. Una matrice è invertibile se e solo se il suo determinante è diverso da 0. Si osserva che il determinante di una matrice triangolare è il prodotto degli elementi sulla diagonale (cfr. esercizio **??**) e questo prodotto è diverso da 0 se e solo se lo sono tutti i suoi fattori, cioè tutti gli elementi della diagonale. Si conclude che una matrice triangolare è invertibile se e solo se gli elementi sulla sua diagonale sono tutti nulli. □

Esercizio 3.6. Mostrare che il determinante di una matrice triangolare è il prodotto degli elementi sulla sua diagonale.

Soluzione. Sia $T \in \mathbb{K}^{n\times n}$. Si procede per induzione sulla dimensione della matrice. Per $n = 1$, la matrice T è sempre triangolare superiore. Supponiamo ora che $n > 1$ e che la proprietà sia vera per tutte le matrici triangolari di dimensione $n - 1$. Applicando la regola di Laplace alla prima colonna della matrice

$$T = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ 0 & t_{22} & \cdots & t_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & t_{nn} \end{bmatrix},$$

si ha che

$$\det(T) = (-1)^{1+1} t_{11} \det\left(\begin{bmatrix} t_{22} & t_{23} & \cdots & t_{2n} \\ 0 & t_{33} & \cdots & t_{3n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & t_{nn} \end{bmatrix}\right) \overset{\text{ipotesi induttiva}}{=} t_{11} t_{22} \cdots t_{nn}.$$

□

Il metodo di Gauss

Abbiamo visto che la soluzione di un sistema lineare la cui matrice dei coefficienti è triangolare superiore richiede "solo" n^2 ops. L'algoritmo di Gauss permette di trasformare un sistema lineare $Ax = b$ in cui A e invertibile (e sotto altre ipotesi che preciseremo in seguito) in un sistema lineare $Ux = \widetilde{b}$ la cui matrice dei coefficienti è triangolare superiore.

Ci sono vari modi di vedere l'algoritmo di Gauss:

- come una sequenza di operazioni sulla matrice A con l'obiettivo di ottenere U triangolare superiore;
- come una fattorizzazione $A = LU$, dove L è triangolare inferiore con uno sulla diagonale e U è triangolare superiore.

Inizialmente, considereremo il primo approccio che è più vicino a quello che si vede nei corsi di algebra lineare (matematica discreta) ed è anche più comprensibile dal punto di vista algoritmico. In seguito parleremo dell'approccio basato sulla fattorizzazione che è più moderno e versatile.

L'algoritmo si basa sul fatto che in un sistema lineare di n equazioni in n incognite con soluzione unica, scelti $i, j \in \{1, \ldots, n\}$, se si sostituisce l'equazione i-esima con la somma dell'equazione i-esima e un multiplo dell'equazione j-esima la soluzione non cambia. Nel nostro formalismo matriciale questa operazione corrisponde a sostituire alla riga i della matrice, la somma della riga i con un multiplo della riga j.

Vediamo com'è possibile, con questo tipo di operazioni, ottenere una sistema triangolare a partire dalla matrice $[A|b]$. Struttureremo l'algoritmo in $n - 1$ passi.

Al primo passo si pone $[A_1|b_1] = [A|b]$ e si cerca di ottenere una matrice che abbia la prima colonna nulla eccetto il primo elemento. Per far questo si eliminano gli elementi $a_{21}, a_{31}, \ldots, a_{n1}$ in sequenza, sostituendo la riga i, per $i = 2, \ldots, n$ con una opportuna combinazione di essa e della prima riga.

Per far questo si costruiscono i numeri $\ell_{21} = \frac{a_{21}}{a_{11}}$, $\ell_{31} = \frac{a_{31}}{a_{11}}$, e così via fino a $\ell_{n1} = \frac{a_{n1}}{a_{11}}$ e tramite essi si combinano le righe nel seguente modo $r_2 = r_2 - \dfrac{a_{21}}{a_{11}} r_1$, $r_3 = r_3 - \dfrac{a_{31}}{a_{11}} r_1$ e così via fino a $r_n = r_n - \dfrac{a_{n1}}{a_{11}} r_1$.

Per fissare le idee descriviamo in dettaglio il caso 4×4 (omettendo il vettore b_1 per semplicità):

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \quad \begin{array}{c} r_2 = r_2 - \dfrac{a_{21}}{a_{11}} r_1 \\ \longrightarrow \\ 7 \text{ ops} = 1\text{d}+3\text{m}+3\text{a} \end{array}$$

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & a_{24}^{(2)} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \quad \begin{array}{c} r_3 = r_3 - \dfrac{a_{31}}{a_{11}} r_1 \\ \longrightarrow \\ 7 \text{ ops} = 1\text{d}+3\text{m}+3\text{a} \end{array} \quad \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & a_{24}^{(2)} \\ 0 & a_{32}^{(2)} & a_{33}^{(2)} & a_{34}^{(2)} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

$$\begin{array}{c} r_4 = r_4 - \dfrac{a_{41}}{a_{11}} r_1 \\ \longrightarrow \\ 7 \text{ ops} = 1\text{d}+3\text{m}+3\text{a} \end{array} \quad \left[\begin{array}{c|ccc} a_{11} & a_{12} & a_{13} & a_{14} \\ \hline 0 & a_{22}^{(2)} & a_{23}^{(2)} & a_{24}^{(2)} \\ 0 & a_{32}^{(2)} & a_{33}^{(2)} & a_{34}^{(2)} \\ 0 & a_{42}^{(2)} & a_{43}^{(2)} & a_{44}^{(2)} \end{array}\right] = A_2,$$

$$L_2 = \begin{bmatrix} 1 & & & \\ \ell_{21} & 1 & & \\ \ell_{31} & 0 & 1 & \\ \ell_{41} & 0 & 0 & 1 \end{bmatrix}, \quad b_2 = \begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ b_3^{(2)} \\ b_4^{(2)} \end{bmatrix}.$$

Come si può osservare il costo del calcolo delle frazioni $\ell_{1i} := \frac{a_{1i}}{a_{11}}$ è di $n-1$ ops (3 ops nel caso 4×4), mentre la combinazione delle righe costa 2 operazioni per ogni elemento e quindi $2(n-1)$ per riga (6 nel caso particolare), per un totale $2(n-1)^2$ ops ($6 \cdot 3$ nel caso particolare). Il costo totale del calcolo di A_2 è quindi di $2(n-1)^2+(n-1)$ ops. La matrice L_2 accumula gli elementi ℓ_{ij} e non richiede ulteriori calcoli. Infine il calcolo di b_2 richiede 2 ops per ogni componente, quindi un totale di $2(n-1)$ ops, visto che la prima componente resta invariata (nel caso 4×4 le operazioni sono 6).

Al secondo passo si cerca di eliminare dalla matrice $[A_2|b_2]$ gli elementi della seconda colonna che stanno sotto l'elemento di posizione $(2,2)$ senza modificare la prima colonna. Per far questo è sufficiente ripetere lo stesso procedimento descritto sopra sulla matrice ottenuta eliminando la prima riga e la prima colonna da A_2 che è una matrice di dimensione $n-1$. Otterremo $[A_3|b_3]$ come desiderata, al passo tre occorrerà ripetere il procedimento per una matrice $n-2$; e così via fino ad arrivare al passo $n-1$ dove si lavora su una matrice 2×2.

Per fissare le idee riassumiamo gli altri passi nel caso $n = 4$.

$$A_2 = \left[\begin{array}{c|ccc} a_{11} & a_{12} & a_{13} & a_{14} \\ \hline 0 & a_{22}^{(2)} & a_{23}^{(2)} & a_{24}^{(2)} \\ 0 & a_{32}^{(2)} & a_{33}^{(2)} & a_{34}^{(2)} \\ 0 & a_{42}^{(2)} & a_{43}^{(2)} & a_{44}^{(2)} \end{array}\right] \quad \begin{array}{c} r_i = r_i - \dfrac{a_{i1}^{(2)}}{a_{22}^{(2)}} r_2, i = 3,4 \\ \longrightarrow \\ 6 \text{ ops} = 2\text{d}+4\text{m}+4\text{a} \end{array}$$

$$A_3 = \left[\begin{array}{cc|cc} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & a_{24}^{(2)} \\ \hline 0 & 0 & a_{33}^{(3)} & a_{34}^{(3)} \\ 0 & 0 & a_{43}^{(3)} & a_{44}^{(3)} \end{array}\right] \quad \begin{array}{c} r_4 = r_4 - \dfrac{a_{43}^{(3)}}{a_{33}^{(3)}} r_3, \\ \longrightarrow \\ 3 \text{ ops} = 1\text{d}+1\text{m}+1\text{a} \end{array}$$

$$U = A_4 = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & a_{24}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & a_{34}^{(3)} \\ 0 & 0 & 0 & a_{44}^{(4)} \end{bmatrix}$$

$$L = \begin{bmatrix} 1 & & & \\ \ell_{21} & 1 & & \\ \ell_{31} & \ell_{32} & 1 & \\ \ell_{41} & \ell_{42} & \ell_{43} & 1 \end{bmatrix}, \quad \widetilde{b} = b_4 = \begin{bmatrix} b_1 \\ b_2^{(2)} \\ b_3^{(3)} \\ b_4^{(4)} \end{bmatrix}.$$

Al secondo passo si devono calcolare $n-2$ frazioni del tipo $\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$ e si devono sommare gli ultimi $n-2$ elementi delle ultime $n-2$ righe con quelli della seconda riga, per un totale di $2(n-2)^2$ operazioni.

In definitiva al passo k si effettuano $(n-k)$ operazioni per calcolare i valori $\ell_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$ e $2(n-k)^2$ operazioni per fare le combinazioni delle righe. Il calcolo di $b^{(k)}$ richiede $2(n-k)$ operazioni.

Il costo totale della operazioni su A è dunque $\sum_{k=1}^{n-1} 2(n-k)^2 + (n-k)$ operazioni la cui parte principale è $\frac{2}{3}n^3$. Il calcolo di $\widetilde{b}$ richiede $\sum_{k=1}^{n-1} 2(n-k)$ operazioni la cui parte principale è n^2 (trascurabile rispetto a n^3). Una volta trovate U e $\widetilde{b}$, si può risolvere il sistema $Ux = \widetilde{b}$ con l'algoritmo di sostituzione all'indietro per cui sono necessarie altre $\sum_{k=1}^{n}(2k-1)$ operazioni, la cui parte principale è n^2 (trascurabile rispetto a n^3).

Si può concludere che la soluzione di un sistema lineare con il metodo di Gauss richiede $\frac{2}{3}n^3$ ops. Il metodo è detto anche metodo di eliminazione di Gauss o *Gaussian elimination*.

Per come è descritto però non è chiaro sotto quali condizioni l'algoritmo dovrebbe terminare, visto che vengono effettuate delle divisioni il cui denominatore può essere nullo. Prima di discutere di questi aspetti daremo alcuni differenti approcci dell'algoritmo di Gauss: prima riscriveremo l'algoritmo in termini delle singole componenti e poi faremo vedere che esso coincide con la fattorizzazione LU della matrice A.

Implementazione del metodo di Gauss

Riscriveremo la relazione di ricorrenza che fornisce gli elementi di A_{k+1} e b_{k+1}, per $k = 1, \ldots, n-1$, in termini degli elementi di A_k e b_k anziché delle righe delle matrici $[A_k|b_k]$, questo rende più facile l'implementazione.

Detti $a_{ij}^{(k)}$ gli elementi di A_k, posto $\ell_{ij}^{(k)} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$ e $b_i^{(k)}$ gli elementi di $b^{(k)}$, al passo k si ha

$$\begin{cases} a_{ij}^{(k+1)} = a_{ij}^{(k)}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n, \\ a_{ij}^{(k+1)} = 0, \quad i = k+1, \ldots, n, \quad j = 1, \ldots, k, \\ a_{ij}^{(k+1)} = a_{ij}^{(k)} - \dfrac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)}, \quad i = k+1, \ldots, n, \quad j = k+1, \ldots, n. \end{cases}$$

Da queste formule si deduce facilmente che il costo del passo k è di $2(n-k)^2$ ops e il costo totale per formare U è $\sum_{k=1}^{n-1} 2(n-k)^2$ ops la cui parte principale è $\frac{2}{3}n^3$ ops.

Il calcolo di $\widetilde{b} = b^{(n)}$ può essere ottenuto aggiungendo alla relazione sopra le righe

$$\begin{cases} b_i^{(k+1)} = b_i^{(k)}, \quad i = 1, \ldots, k \\ b_i^{(k+1)} = b_i^{(k+1)} - \dfrac{a_{ik}^{(k)}}{a_{kk}^{(k)}} b_k^{(k)}, \quad i = k+1, \ldots, n, \end{cases}$$

il cui costo è $2(n-k)$ (si ricordi $\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$ è già stato calcolato) che e quindi il costo totale per il calcolo di $\widetilde{b}$ è $\sum_{k=1}^{n-1} 2(n-k)$ ops la cui parte principale è n^2 ops. Questo costo è trascurabile rispetto al costo del calcolo degli elementi di A_k, tranne in rari casi in cui A è particolarmente strutturata.

Queste formule si traducono immediatamente in pseudocodice, dove, anziché definire le sequenze A_k e b_k per $k = 1, \ldots, n$, si sovrascrivono i dati, com'è uso in informatica.

```
for k=1:n-1
   % ciclo che scorre i passi del metodo di Gauss
   for i=k+1:n
      % ciclo che scorre le righe
      ell(i,k)=a(i,k)/a(k,k);
      a(i,k)=0;
      for j=k+1:n
        % ciclo che scorre gli elementi della riga i
        a(i,j)=a(i,j)-ell(i,k)*a(k,j);
      end
      b(i)=b(i)-ell(i,k)*b(k);
   end
end
```

A titolo di curiosità, osserviamo che se ci interessa solamente calcolare la matrice U, è possibile dare un'implementazione del metodo di Gauss in una sola istruzione

```
for k=1:n-1
  for i=k+1:n
    for j=n:-1:k
      a(i,j)=a(i,j)-a(i,k)*a(k,j)/a(k,k);
    end
  end
end
```

dove il terzo ciclo `for` prosegue con passo negativo perché il valore `a(i,k)`, che diventa nullo, va assegnato dopo che tutti gli altri elementi della riga (`a(i,j)` con $j > k$) sono stati calcolati.

Il metodo di Gauss come fattorizzazione LU

Mostreremo ora che il metodo di Gauss applicato ad una matrice A, se termina, permette di scrivere A come prodotto di una matrice triangolare inferiore con 1 sulla diagonale per una matrice triangolare superiore. In particolare $A = LU$ dove L è la matice degli elementi ℓ_{ij} e $U = A_n$. Tale fattorizzazione è detta *fattorizzazione* LU di A e, nell'ipotesi che la matrice A sia invertibile, se esiste è unica.

Teorema 3.7. *Sia $A \in \mathbb{K}^{n\times n}$ una matrice tale che il metodo di Gauss sia applicabile ad essa e siano L e U le matrici ottenute tramite il metodo di Gauss, allora $A = LU$.*

Dimostrazione. Ricordiamo che, detta $A_k = (a_{ij}^{(k)})_{i,j=1,\dots,n}$, la matrice ottenuta al passo k del metodo di Gauss, con $A_1 = A$ e $A_n = U$, si ha per definizione

$$u_{kj} = a_{kj}^{(k)}, \text{ per ogni } k \text{ e } j, \qquad a_{ij}^{(k+1)} = a_{ij}^{(k)} - \ell_{ik}a_{kj}^{(k)}, \text{ per } k = 1,\dots,n-1, \text{ e } i > k, j \geqslant k,$$

da cui segue che $\ell_{ik}a_{kj}^{(k)} = a_{ij}^{(k)} - a_{ij}^{(k+1)}$. Si osservi che $u_{kj} = a_{kj}^{(k)}$, per ogni k e j, segue dal fatto che la riga k-esima non viene mai modificata dal passo k in poi.

Dunque, per l'elemento di indici (i, j) di LU si ha

$$(LU)_{ij} = \sum_{k=1}^{n} \ell_{ik}u_{kj} \overset{\text{triangolarità}}{=} \sum_{k=1}^{\min\{i,j\}} \ell_{ik}u_{kj} \overset{[u_{kj}=a_{kj}^{(k)}]}{=} \sum_{k=1}^{\min\{i,j\}} \ell_{ik}a_{kj}^{(k)}$$

dove la penultima uguaglianza segue dal fatto che L è triangolare inferiore e quindi $\ell_{ik} = 0$ per $k > i$ e U è triangolare superiore e quindi $u_{kj} = 0$ per $k > j$.

Distinguiamo ora due casi: $i \leqslant j$ e $i > j$.

Se $i \leqslant j$, è $\min\{i, j\} = i$, per cui

$$\begin{aligned}(LU)_{ij} = \sum_{k=1}^{i} \ell_{ik}u_{kj} \overset{[\ell_{ii}=1]}{=} \sum_{k=1}^{i-1} \ell_{ik}a_{kj}^{(k)} + a_{ij}^{(i)} \overset{[a_{ij}^{(k)} - a_{ij}^{(k+1)} = \ell_{ik}a_{kj}^{(k)}]}{=} \sum_{k=1}^{i-1} \big(a_{ij}^{(k)} - a_{ij}^{(k+1)}\big) + a_{ij}^{(i)} \\ = a_{ij}^{(1)} - a_{ij}^{(2)} + a_{ij}^{(2)} - a_{ij}^{(3)} + \cdots + a_{ij}^{(i-1)} - a_{ij}^{(i)} + a_{ij}^{(i)} = a_{ij}^{(1)} = a_{ij}.\end{aligned}$$

Se $i > j$, è $\min\{i, j\} = j$, per cui

$$(LU)_{ij} = \sum_{k=1}^{j} \ell_{ik}u_{kj} = a_{ij}^{(1)} - a_{ij}^{(2)} + a_{ij}^{(2)} - a_{ij}^{(3)} + \cdots + a_{ij}^{(j)} - a_{ij}^{(j+1)} = a_{ij}^{(1)} - a_{ij}^{(j+1)} = a_{ij}$$

dove l'ultima uguaglianza segue dal fatto che $a_{ij}^{(j+1)} = 0$ poiché è un elemento sotto la diagonale e nella colonna j di A_{j+1} gli elementi sotto la diagonale sono nulli. □

Applicabilità dell'algoritmo di Gauss

L'algoritmo di Gauss applicato a una matrice $A \in \mathbb{K}^{n\times n}$ richiede che a ogni passo il pivot non sia nullo, cioè che $a_{kk}^{(k)} \neq 0$. Se esiste $k < n$ tale che $a_{kk}^{(k)} = 0$ mentre $a_{jj}^{(j)} \neq 0$ per $j = 1,\dots,k$, si dice che l'algoritmo ha *breakdown* al passo k, viceversa si dice che il metodo di Gauss è applicabile alla matrice A se non ha breakdown.

Diamo ora un teorema che fornisce una condizione necessaria e sufficiente affinché l'algoritmo sia applicabile ad una data matrice e cioè che i suoi minori principali di testa siano invertibili.

Teorema 3.8. *Sia $A \in \mathbb{K}^{n\times n}$ e siano M_k i suoi minori principali di testa di ordine k, per $k = 1, \dots, n-1$, allora l'algoritmo di Gauss è applicabile alla matrice A se e solo se* $\det(M_k) \neq 0$.

Dimostrazione. Supponiamo dapprima che il metodo sia applicabile, cioè che $a_{kk}^{(k)} \neq 0$ per ogni k, allora si può costruire la matrice $U = A_n$. Siano $N_1, \dots, N_{n-1}$, i suoi minori principali, essi sono triangolari superiori e quindi $\det(N_j) = a_{11}^{(1)} \cdots a_{jj}^{(j)} \neq 0$, per $j = 1, \dots, n-1$. Ma $\det(N_j) = \det(M_j)$ in quanto N_j è ottenuto da M_j mediante operazioni sulle righe che non cambiano il determinante e quindi $\det(M_k) \neq 0$ per ogni k. Infatti le operazioni sulle righe della matrice A_k, inducono operazioni sulle righe del minore dello stesso tipo (somma di una riga, con un multiplo di una riga sopra di essa).

Viceversa, ragioniamo per assurdo. Nell'ipotesi che $\det(M_k) \neq 0$ per ogni k, e che il metodo di Gauss abbia un breakdown al passo ℓ, cerchiamo una contraddizione. Siccome il metodo di Gauss ha breakdown al passo ℓ, è possibile costruire la matrice A_ℓ, ma $a_{\ell\ell}^{(\ell)} = 0$. Sia S_ℓ il minore principale di testa di ordine ℓ di A_ℓ, esso è triangolare superiore e il suo determinante coincide con quello di M_ℓ perché è ottenuto da esso mediante operazioni sulle righe che non cambiano il determinante, quindi $\det(M_\ell) = \det(S_\ell) = a_{11}^{(1)} \cdots a_{\ell\ell}^{(\ell)} = 0$. Assurdo, poiché avevamo supposto che $\det(M_\ell) \neq 0$. Ne concludiamo che il metodo di Gauss è applicabile.

□

Da questo teorema segue un'espressione esplicita per i pivot.

Corollario 3.9. *Sia $A \in \mathbb{K}^{n\times n}$ una matrice per cui il metodo di Gauss sia applicabile e siano M_k i suoi minori principali di testa e $a_{kk}^{(k)}$ i suoi pivot per $k = 1, \dots, n-1$. Allora $a_{11}^{(1)} = \det M_1$ e $a_{kk}^{(k)} = \det M_k / \det M_{k-1}$ per $k = 1, \dots, n-1$.*

Dimostrazione. Il primo caso si verifica direttamente, infatti $a_{11}^{(1)} = a_{11} = \det[a_{11}] = \det M_1$. Per $k > 1$ nel Teorema **??** si ha $\det M_k = a_{11}^{(1)} \cdots a_{kk}^{(k)}$ e $\det M_{k-1} = a_{11}^{(1)} \cdots a_{k-1,k-1}^{(k-1)}$, da cui segue che $a_{kk}^{(k)} = \det M_k / \det M_{k-1}$. □

Il legame tra la fattorizzazione LU di una matrice e l'algoritmo di Gauss ci suggerisce il seguente teorema.

Teorema 3.10. *Una matrice $A \in \mathbb{K}^{n\times n}$ ammette fattorizzazione LU unica se e solo se i suoi minori principali di testa di ordine k, per $k = 1, \dots, n-1$ sono invertibili.*

Stabilità del metodo di Gauss

L'applicabilità del metodo di Gauss è legata al fatto che i pivot siano tutti diversi da zero, tuttavia può succedere che il metodo sia applicabile ma, in aritmetica finita, non fornisca la soluzione cercata, cioè sia numericamente instabile come mostra il seguente esempio.

Si consideri il sistema lineare $Ax = b$ con

$$A = \begin{bmatrix} \varepsilon & 1 \\ 1 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 1+\varepsilon \\ 1 \end{bmatrix},$$

la cui soluzione unica, per ogni $\varepsilon > 0$, è $x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Questo sistema lineare, per piccoli valori di ε, non ha grossi problemi di condizionamento, infatti il numero di condizionamento di A in norma 1 è $(1+\varepsilon)^2$.

Risolvendo il sistema lineare con il metodo di Gauss per $\varepsilon = 0.3 \cdot 10^{-1}$ e lavorando in aritmetica finita in base 10 e con 2 cifre significative si ha

$$\begin{cases} 0.3 \cdot 10^{-1}\widetilde{x}_1 + \widetilde{x}_2 = 1, \\ \quad -0.33 \cdot 10^{2}\widetilde{x}_2 = -0.33 \cdot 10^{2}, \end{cases}$$

la cui soluzione è $\widetilde{x} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, ben lontana dalla soluzione effettiva. Se ne deduce che il metodo di Gauss può essere instabile!

In generale, l'instabilità si ha quando i pivot sono piccoli e quindi gli elementi della matrice U possono diventare anche molto grandi. Per compredere come l'errore sia legato alla grandezza degli elementi di L e U, diamo il seguente risultato.

Teorema 3.11. *Sia A una matrice di numeri di macchina di ordine n e siano $\widetilde{L}$ e $\widetilde{U}$ le matrici della fattorizzazione LU di A effettivamente calcolate con il metodo di Gauss, allora $\widetilde{L}\widetilde{U} = A + H$, dove $|h_{ij}| \leqslant 2nu(|a_{ij}| + |\widetilde{\ell}_{ij}||\widetilde{u}_{ij}|) + O(u^2)$, dove u è la precisione di macchina.*

Ci si aspetta quindi un grosso errore se gli elementi di L e U diventano grandi relativamente ad A.

Per fortuna esiste una variante del metodo di Gauss che risolve molti dei problemi di instabilità; essa verrà descritta nella prossima sezione.

Esercizio 3.12. Si risolva in aritmetica finita in base 10 e con 2 cifre significative il sistema lineare ottenuto scambiando le righe della matrice A e del vettore b nell'esempio numerico precedente (con $\varepsilon = 0.3 \cdot 10^{-1}$).

Variante del massimo pivot parziale

La condizione di applicabilità vista sopra non è molto agevole da verificare, in quanto occorre dimostrare che i minori principali di testa di A sono invertibili. (Esistono tuttavia dei casi in cui la condizione di applicabilità si può dimostrare per via teorica.) Questo è un serio problema del metodo di Gauss.

Naturalmente si può pensare che il caso in cui la matrice abbia minori singolari è un caso raro. In aritmetica finita, tuttavia, ci sono seri problemi anche se il valore di $a_{kk}^{(k)}$ non è nullo, ma è molto piccolo, come mostra l'esempio nella precedente sezione.

Ne concludiamo che il metodo di Gauss è un algoritmo instabile. Fortunatamente, una piccola variante di esso si è rivelata estremamente stabile nei problemi pratici: si tratta della variante del massimo pivot parziale o *Gaussian elimination with partial pivoting (GEPP)*.

Siccome il problema è dato dal fatto che al passo k, l'elemento $a_{kk}^{(k)}$ può essere piccolo, la variante consiste nel cercare nella colonna k tra gli elementi $a_{kk}^{(k)}, a_{k+1,k}^{(k)}, \ldots, a_{nk}^{(k)}$, quello più grande in modulo, diciamo $a_{sk}^{(k)}$ e scambiare la riga k con la riga s. In questo modo si cerca di controllare l'aumento dei moduli degli elementi di L e U.

Vediamo l'implementazione in pseudocodice

```
for k=1:n-1
   % ciclo che scorre i passi del metodo di Gauss
   % prima si cerca l'indice dell'elemento di massimo modulo
   s=k; m=abs(a(k,k));
   for i=k+1:n
      t=abs(a(i,k));
      if (t>m)
         s=i; m=t;
      end
   end
   % se k non coincide con s si scambiano le righe k e s
   if (k!=s)
      for j=k:n
         swap=a(k,j); a(k,j)=a(s,j); a(s,j)=swap;
      end
      swap=b(k); b(k)=b(s); b(s)=swap;
   end
   for i=k+1:n
      % come nel metodo di Gauss standard
      ...
   end
end
```

La variante del massimo pivot parziale richiede $n-k$ confronti al passo k, che sommati danno un numero totale di confronti pari a $O(n^2)$. Se assumiamo che il costo di un confronto sia dell'ordine del costo di un'operazione aritmetica, l'aumento di costo risulta trascurabile rispetto al costo delle operazioni di eliminazione. Si può concludere che la variante è sostanzanzialmente gratuita.

Il metodo di Gauss con la variante del massimo pivot parziale, se applicato alla soluzione di un sistema lineare, costruisce la sequenza

$$[\widehat{A}_1|\widehat{b}_1] = [A|b] \to [A_1|b_1] \to [\widehat{A}_2|\widehat{b}_2] \to [A_2|b_2] \to \cdots \to [A_n|b_n] =: [U|\widetilde{b}],$$

dove $[\widehat{A}_k|\widehat{b}_k]$ è ottenuta eseguendo un eventuale scambio di righe su $[A_k|b_k]$.

Se il metodo di Gauss fallisce è perché ad un certo passo k, si ha che $a_{kk}^{(k)} = 0$; la variante del massimo pivot parziale cerca di porre un rimedio scambiando la riga k con una riga in cui l'elemento sulla colonna k ha massimo modulo. In linea di principio anche il metodo di Gauss con pivoting può fallire, se tutti gli elementi $a_{kk}^{(k)}, a_{k+1,k}^{(k)}, \ldots, a_{nk}^{(k)}$, sono nulli, tuttavia questo non può accadere se la matrice A è invertibile, come mostra il seguente risultato.

Teorema 3.13. *Sia $A \in \mathbb{K}^{n\times n}$ invertibile allora l'algoritmo di Gauss con la variante del massimo pivot parziale è applicabile alla matrice A.*

Dimostrazione. Supponiamo che il metodo di Gauss con pivoting non sia applicabile alla matrice A e mostriamo che essa non è essere invertibile, in questo modo avremo dimostrato il teorema.

Supponiamo che il metodo di Gauss con pivoting si arresti al passo $1 \leqslant k \leqslant n-1$, allora si avrà che $a_{kk}^{(k)} = 0$ (e anche $a_{jj}^{(j)} \neq 0$ per $1 \leqslant j < k$, si noti che questa condizione è vuota per $k = 1$).

Avendo supposto che il metodo fallisca al passo k, si ha che gli elementi $a_{ik}^{(k)}$ della colonna k della matrice A_k, per $i = k, \dots, n$, sono nulli. Concentriamoci sulla sottomatrice ottenuta selezionando le prime k colonne di A_k, che chiamiamo $C_k \in \mathbb{K}^{n\times k}$. La matrice C_k ha righe nulle dalla k-esima in poi, quindi ha al più $k-1$ righe non nulle, da cui il suo rango è al più $k-1$ (non può essere k perché tutti i minori di ordine k hanno una riga nulla e quindi hanno determinante 0). Ma siccome C_k ha rango minore di k, una sua colonna è combinazione lineare delle altre.

Ma questa colonna di C_k è anche una colonna di A_k, che è combinazione lineare delle altre (tutte le colonne di C_k sono colonne di A_k) e quindi A_k non è invertibile e $\det(A_k) = 0$.

Le operazioni eseguite dal metodo di Gauss con pivoting possono cambiare solo il segno del determinante, allora si ha che $\det(A) = \det(A_k) = 0$ e quindi anche la matrice A non è invertibile. □

Come abbiamo suggerito nella precedente sezione l'instabilità è dovuta al fatto che nel metodo di Gauss gli elementi di L e U possono diventare grandi a piacere. Ora vediamo come crescono gli elementi di L e U nel caso in cui si usi la variante del massimo pivot parziale.

Innanzitutto si osserva che, nel caso in cui si usi la variante del massimo pivot parziale, gli elementi di L sono tutti minori o uguali a uno in modulo. Infatti, $|\ell_{ik}| = \frac{|\widehat{a}_{ik}^{(k)}|}{|\widehat{a}_{kk}^{(k)}|} \leqslant 1$ poiché $|\widehat{a}_{kk}^{(k)}| \geqslant |\widehat{a}_{ik}^{(k)}|$.

Per quanto riguarda gli elementi di U, detto $M = \max_{i,j=1,\dots,n} |a_{ij}|$, si hanno i due casi $|a_{ij}^{(2)}| = |\widehat{a}_{ij}^{(1)}| \leqslant M$ o $|a_{ij}^{(2)}| = |\widehat{a}_{ij}^{(1)} - \ell_{i1}\widehat{a}_{ki}^{(1)}| \leqslant |\widehat{a}_{ij}^{(1)}| + |\ell_{i1}||\widehat{a}_{ki}^{(1)}| \leqslant 2M$ e quindi ogni ogni elemento di A_2 è in modulo minore o uguale a $2M$. In modo analogo si dimostra che ogni elemento di A_3 è in modulo minore o uguale a $4M$ e così via fino a mostrare che ogni elemento di $U = A_n$ è in modulo minore o uguale a $2^{n-1}M$.

In questo modo abbiamo mostrato che gli elementi di L e U non possono crescere a dismisura, ma sono controllati. La maggiorazione data sopra è ottimale (cfr. esercizio **??**) quindi la crescita degli elementi di A_k è esponenziale nel caso peggiore. Tuttavia, per un motivo non del tutto chiaro, il caso peggiore non capita se non lo si va a cercare e nella pratica l'algoritmo di Gauss con la variante del massimo pivot parziale è usato comunemente con risultati ottimi.

Nel caso dell'algoritmo di Gauss standard applicato alla matrice A, le matrici L e U che vengono costruite sono tali che $A = LU$, cioè viene calcolata la fattorizzazione LU di A. Nel caso della variante del massimo pivot parziale questa equivalenza sparisce, tuttavia si può dimostrare che le matrici L e U ottenute dalla GEPP sono tali che LU sia uguale alla matrice ottenuta dalla matrice A permutando opportunamente le righe (la permutazione è quella ottenuta componendo tutti gli scambi di righe).

Esercizio 3.14. Si consideri la matrice

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

Si mostri che l'elemento di massimo modulo della matrice $U = A_4$ ottenuta applicando l'algoritmo GEPP alla matrice A è 8 cioè realizza la maggiorazione vista sopra. Trovare una matrice $n \times n$ il cui elemento di massimo modulo è 1 e tale che il massimo modulo di U è 2^{n-1}.

Soluzione. Applicando il metodo di Gauss con pivoting, si osserva che non sono necessari scambi di righe e quindi

$$A_1 = \begin{bmatrix} 1 & 0 & 0 & 1 \\ -1 & 1 & 0 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 \end{bmatrix} \to A_2 = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & -1 & 1 & 2 \\ 0 & -1 & -1 & 2 \end{bmatrix} \to A_3 = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & -1 & 4 \end{bmatrix} \to A_4 = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 8 \end{bmatrix},$$

da cui l'elemento di massimo modulo di A_4 è 8. In generale una matrice $n \times n$ tale che ha 1 sulla diagonale e sull'ultima colonna, mentre vale -1 sotto la diagonale e zero altrove produce una matrice U tale che l'elemento di massimo modulo è 2^{n-1}. □

Scambi di righe come prodotto di matrici

Supponiamo di voler permutare le righe di una matrice $A \in \mathbb{K}^{m\times m}$, secondo una permutazione $\sigma \in \mathfrak{S}_m$ data. Questa operazione può essere eseguita moltiplicando la matrice A per un'opportuna matrice, detta di permutazione, che è costruita a partire da σ.

Si può costruire una matrice di permutazione nel seguente modo: detti $e_1, \dots, e_m$ i vettori della base canonica di $\mathbb{K}^n$ e data $\sigma \in \mathfrak{S}_n$, si ha

$$\sigma \to \Pi_\sigma = [e_{\sigma(1)} \,|\, e_{\sigma(2)} \,|\, \cdots \,|\, e_{\sigma(m)}].$$

Esempio 3.15. Se $\sigma = \mathrm{id} \in \mathfrak{S}_m$, cioè la permutazione identica, per cui $\sigma(i) = i$ per ogni i, allora si ha $\rho(e) = I$. Infatti

$$\rho(e) = \left[e_{\sigma(1)} | e_{\sigma(2)} | \cdots | e_{\sigma(m)}\right] = [e_1 | e_2 | \cdots | e_m] = I.$$

Esempio 3.16. Se $\sigma \in \mathfrak{S}_2$ è la permutazione che scambia 1 e 2, cioè tale che $\sigma(1) = 2$ e $\sigma(2) = 1$, allora

$$\rho(\sigma) = \left[e_{\sigma(1)}|e_{\sigma(2)}\right] = [e_2|e_1] = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Esempio 3.17. Se $\sigma \in \mathfrak{S}_3$ è la permutazione tale che $\sigma(1) = 2$, $\sigma(2) = 3$ e $\sigma(3) = 1$, allora

$$\rho(\sigma) = \left[e_{\sigma(1)}|e_{\sigma(2)}|e_{\sigma(3)}\right] = [e_2|e_1|e_3] = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Applicando una matrice di permutazione Π_σ a un vettore $v \in \mathbb{K}^m$, gli elementi di v saranno permutati secondo σ, infatti

$$\Pi_\sigma v = [e_{\sigma(1)} \mid \cdots \mid e_{\sigma(m)}] \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix} = v_1 e_{\sigma(1)} + \cdots + v_m e_{\sigma(m)} =: \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix},$$

dove $v_i = w_{\sigma(i)}$, per $i = 1, \ldots, m$. Infatti v_1, moltiplicato per $e_{\sigma(1)}$, va a finire nella posizione $\sigma(1)$ del vettore w, da cui $v_1 = w_{\sigma(1)}$ e questo si ripete per ognuna delle componenti.

In questo modo siamo riusciti a mandare l'elemento v_1 nella posizione $v_{\sigma(1)}$, l'elemento v_2 nella posizione $v_{\sigma(2)}$ e così via, riuscendo a permutare il vettore a piacere tramite una moltiplicazione.

Se moltiplichiamo la matrice Π_σ per un'altra matrice A, le sue righe saranno permutate secondo σ.

Esempio 3.18. Sia $A = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$, e supponiamo di volerne scambiare le righe, allora la permutazione da usare è quella tale che $\sigma(1) = 2$ e $\sigma(2) = 1$, che corrisponde alla matrice $\left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$. Verifichiamo che lo scambio viene effettuato

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} c & d \\ a & b \end{bmatrix}.$$

Esempio 3.19. Sia

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ -1 & 0 & 2 & 1 \\ 0 & 0 & 1 & 0 \\ 2 & 2 & 2 & 2 \end{bmatrix}$$

e supponiamo di voler scambiare la riga di indice 2 con la riga di indice 4. Allora la permutazione che cerchiamo è tale che $\sigma(1) = 1, \sigma(2) = 4, \sigma(3) = 3$ e $\sigma(4) = 2$. La matrice da creare è

$$\Pi = [e_1|e_4|e_3|e_2] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Si può verificare che ΠA ha le 2 righe scambiate.

Esempio 3.20. Supponiamo di avere una matrice 4×4 e di voler mandare la riga 2 nella riga 4, la 4 nella 3 e la 3 nella 2. La permutazione che cerchiamo sarà

$$\sigma(1) = 1, \qquad \sigma(2) = 4, \qquad \sigma(3) = 2, \qquad \sigma(4) = 3,$$

e quindi la matrice che effettua lo scambio voluto è

$$\Pi_\sigma = [e_{\sigma(1)} \mid e_{\sigma(2)} \mid e_{\sigma(3)} \mid e_{\sigma(4)}] = [e_1 \mid e_4 \mid e_2 \mid e_3] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Si verifica che $\Pi_\sigma A$ ha le righe scambiate nel modo voluto.

Apparentemente questo non porta un vantaggio, visto che la stessa permutazione si può operare molto più semplicemente. Tuttavia, vedere la permutazione come un prodotto per matrice è utile per ottenere una fattorizzazione di tipo LU anche nella variante del metodo di Gauss con pivoting. Infatti nella variante del massimo pivot parziale si ottinee una fattorizzazione, non più della matrice A, ma di una sua permutazione.

Teorema 3.21. *Sia U la matrice ottenuta all'ultimo passo del metodo di Gauss con la variante del massimo pivot parziale applicata ad A e L la matrice triangolare inferiore con 1 sulla diagonale ottenuta raccogliendo i rapporti ℓ_{ik} usati per eliminare le righe, allora esiste una matrice di permutazione Π tale che*

$$\Pi A = LU.$$

La matrice Π è associata alla permutazione ottenuta raccogliendo tutti gli scambi di riga effettuati durante l'algoritmo.

Si può dimostrare che le matrici di permutazione sono invertibili. Se disponiamo della fattorizzazione $\Pi A = LU$, e vogliamo risolvere il sistema lineare $Ax = b$, le sue soluzioni coincideranno con quelle del sistema lineare $\Pi Ax = \Pi b$ (cfr. Esercizio **??**) che può essere riscritto come $LU = \Pi b$ e quindi il problema si risolve permutando il vettore b e risolvendo due sistemi triangolari (come nel caso della fattorizzazione LU standard).

Esercizio 3.22. Siano $A, M \in \mathbb{K}^{n \times n}$ con M invertibile e sia $b \in \mathbb{K}^n$, mostrare che le soluzioni dei sistemi lineari $Ax = b$ e $MAx = Mb$ coincidono. Mostrare inoltre che se M non è invertibile i due sistemi possono avere soluzioni diverse, per opportuni valori di A e b.

Soluzione. Sia x tale che $Ax = b$, allora i due vettori Ax e b coincidono. Se ad essi si applica una matrice M il risultato sarà lo stesso e quindi $MAx = Mb$. Viceversa, se $MAx = Mb$, utilizzando il ragionamento precedente e moltiplicando ambo i membri per M^{-1} si ottiene $M^{-1}MAx = M^{-1}Mb$ cioè $Ax = b$.

Per mostrare che la proprietà non vale se M non è invertibile basta pensare al caso $n = 1$ e $a \neq 0$, $m = 0$ e b qualsiasi, allora l'equazione $ax = b$ ha l'unica soluzione $x = b/a$, mentre l'equazione $max = mb$, che è $0x = 0$ ha infinite soluzioni. □

Calcolo del determinante

Oltre alla soluzione di sistemi lineari, l'algoritmo di Gauss può essere usato per altri problemi dell'algebra lineare numerica, come il calcolo del determinante e il calcolo dell'inversa di una matrice. Descriveremo prima come si può calcolare il determinante di una matrice $A \in \mathbb{K}^{n\times n}$.

L'osservazione chiave è che il passo del metodo di Gauss non cambia il determinante, cioè per ogni $k = 1, \dots, n-1$, si ha che $\det(A_k) = \det(A_{k+1})$. Questo segue dal fatto che la matrice A_{k+1} è ottenuta dalla matrice A_k con operazioni sulle righe che non cambiano il determinante (è noto che se in una matrice $n \times n$ si sostituisce la riga i-esima con la somma della riga i-esima e un multiplo di un'altra riga, il determinante non cambia).

Siccome $\det(A_k) = \det(A_{k+1})$ si ha che $\det(A) = \det(A_1) = \det(A_n) = \det(U)$ e quindi il determinante di A coincide con il determinante di U, ma U è una matrice triangolare e quindi $\det(A) = \det(U) = u_{11} \cdots u_{nn} = \prod_{i=1}^{n} u_{ii}$.

Nel caso del metodo di Gauss con la variante del massimo pivot parziale, oltre alle combinazioni di righe vengono effettuati degli scambi di righe. Siccome ogni scambio di righe cambia il segno del determinante, si avrà che $\det(A) = \det(U)$ se il numero di scambi di righe è pari, e $\det(A) = -\det(U)$ se il numero di scambi di righe è dispari. Più sinteticamente $\det(A) = (-1)^s \det(U)$ dove s è il numero di scambi di righe.

In conclusione l'algoritmo per il calcolo del determinante è il seguente:

- si calcola la matrice triangolare superiore U dalla matrice A con il metodo di Gauss con la variante del massimo pivot parziale contando il numero di scambi di righe s;
- si calcola $\det(A) = (-1)^s \prod_{i=1}^{n} u_{ii}$.

Il costo computazionale di questo algoritmo è dato dal costo del metodo di Gauss, in quanto il prodotto $\prod_i u_{ii}$ si calcola con $n-1$ ops e quindi è trascurabile.

Sistema lineare con termine noto multiplo e inversa di una matrice

Un problema che capita molto di frequente è quello di risolvere più di un sistema lineare in cui la matrice dei coefficienti è la stessa, ma cambia il termine noto: cioè occorre risolvere i sistemi

$$Ax_1 = b_1, \dots, Ax_s = b_s,$$

dove $b_1, \dots, b_s$ sono opportuni termini noti. Il modo più immediato ma inefficiente di risolvere tale problema è di risolvere separatamente i sistemi lineari. In questo modo si ha un costo computazionale pari a $\frac{2}{3}sn^3$, dato dal prodotto del costo della soluzione di ciascun sistema lineare tramite il metodo di Gauss, moltiplicato per il numero di sistemi lineari.

Si può considerare un algoritmo migliore. Si parte dalla matrice $[A|b_1|b_2|\cdots|b_s]$ ottenuta incollando alla matrice A tutti i termini noti e si applica l'eliminazione gaussiana (con pivoting) come nel caso in cui ci sia un solo termine noto. Dopo $n-1$ passi si ottiene la matrice $[U|\widetilde{b}_1|\widetilde{b}_2|\cdots|\widetilde{b}_s]$, tale che, per $i = 1, \dots, s$, il sistema $Ux_i = \widetilde{b}_i$ ha la stessa soluzione del sistema $Ax_i = b_i$.

L'implementazione del metodo è simile a quella del metodo di Gauss, con la differenza che ora non bisogna aggiornare un solo vettore b ma s di essi che metteremo per comodità in una matrice $n \times s$. Nell'implementazione è sufficiente sostituire la riga

```
b(i)=b(i)-ell(i,k)*b(k);
```

con le righe

```
for t=1:s
   b(i,t)=b(i,t)-ell(i,k)*b(k,t);
end
```

Il costo di questo algoritmo è dato dal costo dall'eliminazione gaussiana sulla matrice A, che richiede $\frac{2}{3}n^3$ ops, a cui vanno sommate le operazioni necessarie per modificare i vettori b_i e per risolvere gli s sistemi triangolari. Le operazioni sui vettori b_i al passo k sono $2s$ per ogni $i = k+1, \dots, n$, per un totale di $\sum_{k=1}^{n-1} 2s(n-k) \approx sn^2$ ops. Il costo della soluzione dei sistemi lineari triangolari è di sn^2 ops, se si usa la sostituzione all'indietro.

In conclusione si possono risolvere s sistemi lineari con la stessa matrice dei coefficienti e con membro destro diverso con un algoritmo che richiede $\frac{2}{3}n^3 + 2sn^2$ ops.

Questo procedimento richiede che i vettori b_i siano tutti noti quando iniziamo l'eliminazione gaussiana, ma invece può succedere di dover risolvere i sistemi in tempi diversi o che addirittura il valore del vettore b_{i+1} dipenda dalla soluzione del sistema $Ax_i = b_i$ (come nel caso del metodo delle potenze inverse per il calcolo

degli autovalori di una matrice). Per questo motivo è preferibile usare la fattorizzazione LU di A (o di una sua permutazione, nel caso della variante del pivoting).

I sistemi $Ax_i = b_1, \ldots, Ax_i = b_s$, diventano $LUx_i = \Pi b_1, \ldots, LUx_i = \Pi b_s$, dove Π è la permutazione ottenuta componendo gli scambi del metodo di Gauss con pivoting. Ciascuno di questi ultimi sistemi può essere risolto in due passi, usando la sostituzione $Ux_i = y_i$. In totale si eseguono, come nel caso precedente $\frac{2}{3}n^3 + 2sn^2$ ops.

Per finire, consideriamo il problema del calcolo dell'inversa di una matrice. Ricordiamo che l'inversa di una matrice $A \in \mathbb{K}^{m\times m}$ è una matrice $X \in \mathbb{K}^{m\times m}$ tale che $AX = I$. Se denotiamo con $x_1, \ldots, x_m$, le colonne della matrice X e con $e_1, \ldots, e_m$ le colonne della matrice I, si ottiene

$$AX = A[x_1|x_2|\cdots|x_m] = [Ax_1|Ax_2|\cdots|Ax_n] = [e_1|e_2|\cdots|e_m] = I,$$

che equivale a risolvere gli m sistemi lineari $Ax_1 = e_1, \ldots, Ax_m = e_m$, ciascuno dei quali fornisce una colonna della matrice inversa.

Questo è un problema di termine noto multiplo pertanto il costo totale di questo algoritmo è $\frac{2}{3}m^2 + 2m^3 = \frac{8}{3}m^3$. In realtà si può fare di meglio, poiché i sistemi lineari da risolvere hanno una struttura speciale, infatti i vettori e_i, per $i = 1, \ldots, m$, sono tutti nulli escluso l'elemento i che è 1. Utilizzando questa struttura si riesce a trovare un algoritmo che calcola l'inversa di una matrice e che richiede $2m^3$ ops (cfr. esercizio **??**). Si noti che è lo stesso costo del prodotto tra due matrici, anche se all'apparenza si tratta di un problema molto più difficile.

Esercizio 3.23. Trovare un algoritmo che calcoli l'inversa di una matrice $m \times m$ che ammette fattorizzazione LU e che abbia un costo computazionale asintotico di $2m^3$ ops.

Esercizio 3.24. Si risolva l'esercizio precedente nel caso di una qualsiasi matrice invertibile, utilizzando il metodo di Gauss con la variante del massimo pivot parziale.

Riferimenti bibliografici

L'eliminazione gaussiana (con la variante del massimo pivot) è l'algoritmo più usato per risolvere sistemi lineari e la sua descrizione può essere trovata in qualsiasi testo di Analisi Numerica. Per una trattazione incentrata maggiormente sulle fattorizzazioni, si consulti il capitolo 4 di [1].

Un riferimento in inglese è [2], sezione 4.1.

[1] D. Bini, M. Capovani, O. Menchi. *Metodi Numerici per l'Algebra Lineare.* Zanichelli, Bologna, 1988.

[2] J. Stoer, R. Bulirsch. *Introduction to Numerical Analysis.* Third Edition. Springer, 2002.

4 Interpolazione

Premessa applicativa

Le immagini digitali si distinguono in due grosse categorie

- immagini rasterizzate, rappresentate nel caso bidimensionale da griglie rettangolari (pixmap) di punti (pixel) a ciascuno dei quali viene assegnato un colore, questi punti spesso sono piccoli e indistinguibili dando l'impressione della continuità. Nel caso tridimensionale si usano griglie cubiche (voxelmap). Questo tipo di immagine è usata nelle fotografie digitali, nella diagnostica medica per immagini, ecc.
- immagini vettoriali, descritte attraverso le loro proprietà geometriche (per esempio le coordinate di alcuni loro punti, la forma, ecc.) o il loro contenuto informativo (per esempio un testo) e visualizzate come immagini rasterizzate da un opportuno programma. Questo tipo di immagine è usata nella grafica pubblicitaria ed editoriale e in gran parte della grafica 3D.

Una differenza sostanziale tra le immagini rasterizzate e le immagini vettoriali è il fatto che queste ultime possono essere *scalate* senza perdere dettagli. D'altro canto, non tutto è adatto ad essere descritto in modo vettoriale se non in modo estremamente costoso.

Un problema fondamentale della grafica vettoriale può essere formulato nel modo seguente:

Come unire dei punti in modo da ottenere qualcosa di gradevole e al tempo stesso facile da calcolare?

La soluzione più facile (ma non banale nel caso di punti nello spazio) è di unire i punti formando nel caso del piano una spezzata, nel caso dello spazio una superficie poligonare (mesh). Tuttavia questa soluzione non è affatto gradevole alla vista, a causa della presenza di spigoli e dei vertici.

Un possibile rimedio è quello di aumentare i punti in modo che l'occhio non percepisca questi spigoli, tuttavia è una soluzione costosa e non del tutto soddisfacente. Oggi ci sono soluzioni migliori.

La presenza di spigoli in un grafico di funzione, si traduce nel fatto che la funzione non è derivabile in alcuni punti. Quindi una prima richiesta può essere quella di trovare una curva che raccordi i punti e sia derivabile con continuità.

Appare subito chiaro che con questo solo vincolo le soluzioni sono infinite. Ci si concentra perciò sulla "facilità di calcolo" e per questo si cerca di lavorare con classi di funzioni che hanno buona regolarità e che sono adatte al calcolo: i polinomi, le funzioni razionali, i polinomi trigonometrici, e così via.

Interpolazione polinomiale

Concentriamoci su un caso semplice, quello in cui si hanno $n+1$ punti del piano $\begin{bmatrix} x_i \\ y_i \end{bmatrix}$, per $i = 0, \dots, n$, tali che $x_i \neq x_j$ per $i \neq j$. Si cercano polinomi tali che $p(x_i) = y_i$.

Le condizioni da verificare sono $n+1$ quindi è ragionevole cercare tra i polinomi di grado al più n (che dipendono da $n+1$ parametri). Nel caso di due punti, dai tempi di Euclide, è ben noto che esiste un polinomio di grado al più uno che passa per essi (ma può avere anche grado 0 o indefinito).

Per più di due punti punti la risposta non è ovvia. Cerchiamo quindi di risolvere il *problema dell'interpolazione polinomiale*: dati $n+1$ numeri reali nell'intervallo $[a, b]$, $x_0, \dots, x_n$ e dati $n+1$ numeri reali, $y_0, \dots, y_n$, trovare un polinomio di grado al più n che verifichi le condizioni $p(x_i) = y_i$. I punti $x_0, \dots, x_n$ sono detti *nodi* dell'interpolazione.

Risolvere il problema equivale a risolvere il sistema $p(x_i) = y_i$, per $i = 0, \dots, n$, nelle incognite $a_0, a_1, \dots, a_n$, dove $p(x) = \sum_{i=0}^{n} a_i x^i$. Provando ad espandere le equazioni

$$\begin{cases} a_0 + a_1 x_0 + a_2 x_0^2 + \cdots + a_n x_0^n = y_0, \\ a_0 + a_1 x_1 + a_2 x_1^2 + \cdots + a_n x_1^n = y_1, \\ \quad \vdots \\ a_0 + a_1 x_n + a_2 x_n^2 + \cdots + a_n x_n^n = y_n, \end{cases}$$

ci si accorge che il sistema è lineare nelle incognite $a_0, a_1, \dots, a_n$ e si può riscrivere come $Va = y$, dove

$$V = \begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix}, \quad a = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}, \quad y = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

La matrice V è detta *matrice di Vandermonde* relativa ai nodi $x_0, \dots, x_n$.

A questo punto l'algebra lineare ci assicura che il sistema ha un'unica soluzione se e solo se la matrice V è invertibile, che equivale a dire $\det(V) = 0$. Ci si accorge immediatamente che se $x_i = x_j$ per $i \neq j$, la matrice ha due righe uguali e quindi non è invertibile ($\det(V) = 0$), e non si ha soluzione unica.

In realtà si può dimostrare che la condizione che i punti $x_0, \dots, x_n$ siano tutti distinti è anche sufficiente per l'esistenza e unicità della soluzione, infatti si può mostrare che

$$\det V = \prod_{\substack{i,j=0 \\ i>j}}^{n} (x_i - x_j) = \prod_{j=0}^{n} \prod_{i=j+1}^{n} (x_i - x_j),$$

ottenendo il seguente teorema, che dimostreremo in modo diverso.

Teorema 4.1. *Siano $x_0, x_1, \dots, x_n \in [a, b]$ distinti e $y_0, y_1, \dots, y_n \in \mathbb{R}$, allora esiste un unico polinomio di grado al più n tale che $p(x_i) = y_i$ per $i = 0, \dots, n$. Esso è detto polinomio di interpolazione.*

In questo modo abbiamo risolto il problema dell'interpolazione polinomiale. Si noti che la condizione che i nodi siano distinti è necessaria e sufficiente per l'unicità, in quanto, se due nodi coincidono, la matrice V non è invertibile.

In particolare, l'equivalenza con i sistemi lineari ci permette di enunciare il seguente risultato.

Corollario 4.2. *Il problema dell'interpolazione polinomiale è ben posto se e solo se i nodi sono distinti.*

I polinomi di Lagrange e il calcolo del polinomio di interpolazione

Ora cerchiamo di capire come si calcola in modo efficiente il polinomio di interpolazione. Un primo approccio consiste nel risolvere il sistema lineare $Va = y$ con il metodo di Gauss. L'algoritmo corrispondente ha costo computazionale pari a $O(n^3)$ ops, ma purtroppo la matrice dei coefficienti non ha un buon condizionamento (con dovute eccezioni) quindi per n grande la soluzione non viene calcolata in modo accurato.

Di fatto, nella pratica, quello che interessa non è di trovare i coefficienti del polinomio di interpolazione, ma di valutarlo in uno o più punti (diversi dai nodi), per questo motivo può essere equivalentemente utile rappresentarlo in una base di $\mathbb{R}_n[x]$ diversa dalla base canonica $\{1, x, \dots, x^n\}$, che sarà la base dei polinomi di Lagrange.

L'idea è di costruire polinomi di grado n che valgano 1 in un nodo e 0 negli altri. In questo, anche se la base è più complicata di quella canonica, le coordinate del polinomio di interpolazione rispetto a tale base, saranno i valori $y_0, \dots, y_n$.

Siano $x_0, \dots, x_n \in [a, b]$ punti distinti, si definiscono i polinomi di Lagrange

$$L_0(x) = \prod_{\substack{k=0 \\ k\neq 0}}^{n} \frac{x - x_k}{x_0 - x_k}, \quad L_1(x) = \prod_{\substack{k=0 \\ k\neq 1}}^{n} \frac{x - x_k}{x_1 - x_k}, \quad \dots, \quad L_n(x) = \prod_{\substack{k=0 \\ k\neq n}}^{n} \frac{x - x_k}{x_n - x_k},$$

o, che è lo stesso,

$$L_i(x) = \frac{(x - x_0)(x - x_1) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)}{(x_i - x_0)(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)}, \qquad i = 0, \dots, n. \tag{3}$$

Esempio 4.3. I polinomi di Lagrange relativi a due nodi x_0 e x_1, sono

$$L_0(x) = \frac{x - x_1}{x_0 - x_1}, \qquad L_1(x) = \frac{x - x_0}{x_1 - x_0}.$$

Se i nodi sono $x_0 = 1$, $x_1 = 2$, si ha

$$L_0(x) = 2 - x, \qquad L_1(x) = x - 1.$$

Esempio 4.4. I polinomi di Lagrange relativi a tre nodi x_0, x_1 e x_2 sono

$$L_0(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}, \qquad L_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)}, \qquad L_2(x) = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}.$$

Se i nodi sono $x_0 = -1$, $x_1 = 0$, $x_2 = 1$, si ha

$$L_0(x) = \frac{x(x-1)}{(-1)(-2)} = \frac{1}{2}x^2 - \frac{1}{2}x, \quad L_1(x) = \frac{(x+1)(x-1)}{-1} = -x^2 + 1, \quad L_2(x) = \frac{(x+1)x}{2} = \frac{1}{2}x^2 + \frac{1}{2}x.$$

Si osserva subito che sono polinomi di grado n perché sono prodotti di n fattori lineari il cui coefficiente principale non è nullo, inoltre è possibile dimostrare alcune interessanti proprietà.

Lemma 4.5. *Siano $L_0(x), \dots, L_n(x)$ i polinomi di Lagrange relativi ai nodi $x_0, \dots, x_n$ distinti in $[a, b]$ allora*

(a) $L_i(x_j) = \delta_{ij}$ per ogni $i, j = 0, \dots, n$ (cioè $L_i(x_i) = 1$ e $L_i(x_j) = 0$ se $i \neq j$).

(b) I polinomi formano una base di $\mathbb{R}_n[x]$ (lo spazio dei polinomi di grado al più n).

Dimostrazione. (a) Quando si valuta $L_i(x_i)$, sostituendo x_i a x nell'equazione (**??**), il numeratore diventerà identico al denominatore

$$L_i(x_i) = \frac{(x_i - x_0)(x_i - x_1)\cdots(x_i - x_{i-1})(x_i - x_{i+1})\cdots(x_i - x_n)}{(x_i - x_0)(x_i - x_1)\cdots(x_i - x_{i-1})(x_i - x_{i+1})\cdots(x_i - x_n)} = 1.$$

Mentre se si valuta $L_i(x_j)$ con $j \neq i$, si ha, se per esempio $j < i$,

$$L_i(x_j) = \frac{(x_j - x_0)(x_j - x_1)\cdots(x_j - x_j)\cdots(x_j - x_{i-1})(x_j - x_{i+1})\cdots(x_j - x_n)}{(x_i - x_0)(x_i - x_1)\cdots(x_i - x_j)\cdots(x_i - x_{i-1})(x_i - x_{i+1})\cdots(x_i - x_n)} = 0,$$

perché al numeratore compare la differenza $x_j - x_j = 0$ che annulla il prodotto. Stessa cosa accade quando $j > i$.
(b) Siccome lo spazio $\mathbb{R}_n[x]$ ha dimensione $n + 1$ e i polinomi di Lagrange sono $n + 1$ è sufficiente dimostrare che sono linearmente indipendenti, cioè che se $\sum_{i=0}^{n} \alpha_i L_i(x) \equiv 0$, per $\alpha_i \in \mathbb{R}$, allora $\alpha_0 = \alpha_1 = \cdots = \alpha_n = 0$.

Dire che

$$\alpha_0 L_0(x) + \cdots + \alpha_n L_n(x) \equiv 0,$$

vuol dire che l'uguaglianza vale in ogni punto, quindi vale in x_i, per $i = 0, \dots, n$,

$$\alpha_0 L_0(x_i) + \cdots + \alpha_{i-1} L_{i-1}(x_i) + \alpha_i L_i(x_i) + \alpha_{i+1} L_{i+1}(x_i) + \cdots + \alpha_n L_n(x_i) = 0,$$

ma dal punto (a) sappiamo che $L_0(x_i) = \cdots = L_{i-1}(x_i) = L_{i+1}(x_i) = \cdots = L_n(x_i) = 0$ e $L_i(x_i) = 1$ e quindi la precedente uguaglianza diventa $\alpha_i = 0$. Data l'arbitrarietà di i, abbiamo dimostrato che i polinomi sono linearmente indipendenti e quindi formano una base di $\mathbb{R}_n[x]$. □

Il precedente corollario ci fornisce un modo per dimostrare l'esistenza e unicità del polinomio di interpolazione nel caso in cui i nodi siano distinti. Inoltre, ci fornisce anche un'espressione esplicita del polinomio da cui ricavare un algoritmo per la valutazione.

Teorema 4.6. *Siano $L_i(x)$, per $i = 0, \dots, n$ i polinomi di Lagrange relativi ai nodi $x_0, \dots, x_n$ distinti in $[a, b]$ e siano, inoltre, $y_0, \dots, y_n \in \mathbb{R}$. Esiste un unico polinomio $p(x)$ di grado al più n tale che $p(x_i) = y_i$ per ogni i e si può scrivere come*

$$p(x) = y_0 L_0(x) + y_1 L_1(x) + \cdots + y_n L_n(x) = \sum_{i=0}^{n} y_i L_i(x).$$

Dimostrazione. Sia $p(x)$ un polinomio di grado al più n tale che $p(x_i) = y_i$, per $i = 0, \dots, n$. Poiché $L_0(x), \dots, L_n(x)$ formano una base, si può scrivere

$$p(x) = z_0 L_0(x) + z_1 L_1(x) + \cdots + z_n L_n(x) = \sum_{i=0}^{n} z_i L_i(x).$$

Imponendo l'uguaglianza $p(x_i) = y_i$, si ha che

$$y_i = p(x_i) = z_0 L_0(x_i) + \cdots + z_{i-1} L_{i-1}(x_i) + z_i L_i(x_i) + z_{i+1} L_{i+1}(x_i) + \cdots + z_n L_n(x_i) \overset{L_i(x_j)=0}{=} z_i L_i(x_i) \overset{L_i(x_i)=1}{=} z_i,$$

da cui deduciamo che $y_i = z_i$ per ogni i. Quindi le coordinate del polinomio sono determinate dalle condizioni di interpolazione e si ha

$$p(x) = y_0 L_0(x) + \cdots + y_n L_n(x),$$

ed il polinomio è unico perché le coordinate rispetto a una base sono uniche. □

Ora occupiamoci della valutazione del polinomio di interpolazione. Possiamo scrivere in modo compatto

$$p(x) = \sum_{i=0}^{n} y_i L_i(x) = \sum_{i=0}^{n} y_i \frac{\prod_{k\neq i}(x - x_k)}{\prod_{k\neq i}(x_i - x_k)}$$

e ci accorgiamo che alcuni elementi contenuti nella sommatoria vengono ripetuti in ogni addendo. In particolare, ponendo

$$z_i = \frac{y_i}{\prod_{k\neq i}(x_i - x_k)}, \qquad \pi_n(x) = \prod_{i=0}^{n}(x - x_i) = (x - x_0)(x - x_1)\cdots(x - x_n),$$

si osserva che

$$\prod_{k\neq i}(x - x_k) = (x - x_0)\cdots(x - x_{i-1})(x - x_{i+1})\cdots(x - x_n)$$
$$= \frac{(x - x_0)\cdots(x - x_{i-1})(x - x_i)(x - x_{i+1})\cdots(x - x_n)}{x - x_i} = \frac{\pi_n(x)}{x - x_i}$$

e quindi si ottiene

$$p(x) = \sum_{i=0}^{n} \frac{z_i \pi_n(x)}{x - x_i}.$$

L'ultima espressione ci consente di sviluppare un algoritmo per la valutazione di $p(x)$ in più punti.

L'idea è di calcolare prima le quantità che non dipendono da x:

1. tutte le differenze $x_i - x_j$ per $i = 0, \ldots, n$; si osserva che $x_i - x_i = 0$ e $x_i - x_j = -(x_j - x_i)$ e quindi è sufficiente calcolare solo quelle per cui $i > j$ che sono $n(n+1)/2$ che è il numero di operazioni necessarie;
2. per ogni i si calcola il prodotto $\prod_{k\neq i}(x_i - x_k) = (x - x_0)\cdots(x - x_{i-1})(x - x_{i+1})\cdots(x - x_n)$ e questo richiede $(n-1)$ prodotti, per un totale di $(n+1)(n-1)$ ops;
3. per ogni i il valore z_i, che richiede una divisione, per un totale di $(n+1)$ ops.

In seguito, nella valutazione di p in un punto $\widetilde{x}$ si calcola

4. $\pi_n(\widetilde{x}) = (\widetilde{x} - x_0)\cdots(\widetilde{x} - x_n)$ con $n+1$ sottrazioni ed n prodotti e si salvano le differenze, per un totale di $2n+1$ ops;
5. $p(\widetilde{x}) = z_0 \frac{\pi_n(\widetilde{x})}{(\widetilde{x} - x_0)} + \cdots + z_n \frac{\pi_n(\widetilde{x})}{(\widetilde{x} - x_n)}$ in cui si usano le quantità $\widetilde{x} - x_i$ calcolate al punto 4, ma occorre eseguire $n+1$ quozienti e prodotti e n somme, per un totale di $3n+2$ operazioni elementari.

In definitiva, riusciamo a valutare il polinomio in un punto con un costo asintotico pari a $\frac{3}{2}n^2$ ops, che è il costo dei passi 1-3. Per valutarlo in un ulteriore punto, sono richiesti solo i passi 4-5 per un totale di $5n$ ops.

In totale per valutare il polinomio di interpolazione in k punti con questo algoritmo sono necessarie $\frac{3}{2}n^2 + 5nk$ ops.

L'interpolazione per approssimare funzioni

L'interpolazione polinomiale può essere usata anche per l'approssimazione di funzioni continue.

Sia $f \in C[a, b]$ e siano dati $x_0, \ldots, x_n \in [a, b]$ distinti. Il polinomio di interpolazione di f nei nodi $x_0, \ldots, x_n$ è l'unico polinomio di grado al più n tale che $p(x_i) = f(x_i)$.

Si noti che il problema è lo stesso di quello studiato nelle sezioni precedenti, dove anziché dare $n+1$ valori $y_0, \ldots, y_n$ si dà una funzione. Ponendo $y_i = f(x_i)$ il problema è identico.

Una volta calcolato il polinomio di interpolazione $p_n(x)$ di una funzione $f(x)$, si può pensare di approssimare la funzione con tale polinomio. In tal caso si commette un errore

$$r(x) = f(x) - p_n(x),$$

detto resto dell'interpolazione.

Se la funzione f è sufficientemente regolare, cioè se è derivabile $n+1$ volte con derivata $(n+1)$-esima continua su $[a, b]$, allora è possibile dare un'espressione per il resto dell'interpolazione in ogni punto, descritta nel seguente risultato per la cui dimostrazione si veda [1, pp. 358-359].

Teorema 4.7. *Sia $f \in C^{k+1}[a,b]$ e sia $p_n(x)$ il polinomio di interpolazione relativo ai nodi $x_0, \dots, x_n \in [a,b]$ distinti. Allora per ogni $x \in [a,b]$, esiste $\xi(x) \in [a,b]$, tale che*

$$r(x) = f(x) - p_n(x) = \frac{f^{(k+1)}(\xi(x))}{(n+1)!}(x-x_0)(x-x_1)\cdots(x-x_n).$$

In generale, non è detto che aumentando il numero dei nodi, l'approssimazione migliori, anche se la funzione da approssimare è molto regolare.

Un esempio classico è la funzione di Runge $f(x) = 1/(1+x^2)$ che è derivabile infinite volte (è addirittura analitica) nell'intervallo $[-5,5]$, ma se p_n è il polinomio di interpolazione su n nodi equispaziati

$$x_\ell = a + \frac{b-a}{n}\ell, \qquad \ell = 0, 1, \dots, n,$$

si può dimostrare che p_n non converge a $f(x)$ per alcuni $x \in [-5,5]$.

Tuttavia, potendo scegliere i nodi liberamente, è possibile trovare successioni di polinomi di interpolazione che convergono alla funzione puntualmente. Una scelta migliore consiste nel prendere come nodi i coseni delle soluzioni dell'equazione $\cos(nx) = 0$ sull'intervallo $[-1,1]$ che sono

$$x_\ell = \cos\Big(\frac{\pi(2k+1)}{2(n+1)}\Big), \qquad k = 0, \dots, n.$$

Essi possono essere definiti su ogni intervallo $[a,b]$ usando la funzione affine $\varphi : [-1,1] \to [a,b]$, tale che $\varphi(x) = \frac{b-a}{2}x + \frac{b+a}{2}$, da cui

$$x_\ell = \frac{b+a}{2} + \frac{b-a}{2}\Big(\frac{\pi(2k+1)}{2(n+1)}\Big), \qquad k = 0, \dots, n.$$

Si può dimostrare che se $\{q_n\}_n$ è la successione dei polinomi di interpolazione sui nodi di Chebyshev su $[a,b]$ e $f \in C^1[a,b]$ allora la successione $\{q_n\}$ converge puntualmente a f in $[a,b]$. Questo vale in particolare per la funzione di Runge.

Purtroppo non è sempre possibile scegliere i nodi liberamente. Per esempio, in computer grafica, i nodi sono spesso dati da punti assegnati.

Funzioni spline

Un secondo modo di interpolare è considerando una funzione che sia semplice per ogni intervallo in cui viene diviso l'intervallo $[a,b]$ dai nodi, come nel caso della funzione lineare a tratti.

Si consideri una partizione Δ dell'intervallo $[a,b]$, cioè un insieme di punti $x_0, \dots, x_n$ con la proprietà che

$$x_0 = a < x_1 < \ldots < x_{n-1} < x_n = b.$$

La differenza, rispetto al caso precedente è che ora i nodi sono ordinati e il primo deve essere a e l'ultimo b, mentre nell'interpolazione polinomiale queste condizioni non sono richieste.

L'insieme delle spline di grado k subordinate alla partizione Δ sono le funzioni che, sull'intervallo $[a,b]$, sono derivabili con continuità fino all'ordine $k-1$ e, ristrette a ogni intervallo $[x_i, x_{i+1}]$ della partizione, sono polinomi di grado al più k. Più precisamente,

$$\mathcal{S}_{\Delta,k} = \{s \in C^{k-1}[a,b] \,:\, s|_{[x_i,x_{i+1}]}(x) \equiv q_i(x),\, q(x) \in \mathbb{R}_k[x], i = 0, \dots, n-1\}.$$

Per $k = 1$ si hanno le funzioni lineari a tratti e continue. Per $k = 3$ si parla di spline cubiche, che sono funzioni $C^2[a,b]$ e che su ogni intervallo coincidono con polinomi di grado al più 3. Esse sono molto usate nella pratica perché sono versatili e permettono di ottenere risultati esteticamente gradevoli, inoltre sono computazionalmente interessanti.

Si può porre il problema dell'interpolazione tramite spline: data una partizione $\Delta = \{x_0, \dots, x_n\}$ dell'intervallo $[a,b]$ e dati $y_0, \dots, y_n \in \mathbb{R}$, trovare una spline $s \in \mathcal{S}_{\Delta,k}$ tale che $s(x_i) = y_i$.

Si può dimostrare che l'insieme $\mathcal{S}_{\Delta,k}$ è uno spazio vettoriale di dimensione $n+k$. Questo ci fa capire che il problema dell'interpolazione tramite spline di grado 1 può avere soluzione unica e si può costruire abbastanza facilmente. La spline di grado 1 interpolante è quella che sull'intervallo $[x_i, x_{i+1}]$ vale

$$s_i := s_{[x_i,x_{i+1}]} = \frac{y_{i+1}-y_i}{x_{i+1}-x_i}x + \frac{y_i x_{i+1} - y_{i+1}x_i}{x_{i+1}-x_i},$$

che può essere valutata in N punti con $7 + 2(N-1)$ ops.

Invece è più complicato risolvere il problema dell'interpolazione tramite spline cubiche, infatti il numero di condizioni è minore del numero di parametri delle spline cubiche. Il problema si risolve imponendo delle condizioni al bordo, precisate dal seguente risultato (per la dimostrazione si veda il libro: Metodi Numerici, di Bevilacqua, Bini, Capovani, Menchi, Editrice Zanichelli, Bologna, Capitolo 5, Sezione 14).

Teorema 4.8. *Esiste un'unica spline cubica interpolante $s(x)$ tale che $s''(a) = s''(b) = 0$ e i parametri che la definiscono possono essere calcolati con $O(n)$ operazioni.*

Tramite l'algoritmo dato dal precedente teorema, è possibile avere le formule esplicite per la spline cubica nell'intervallo

$$s_i := s_{[x_i, x_{i+1}]} = a_i x^3 + b_i x^2 + c_i x + d_i,$$

che si può valutare in un punto con 6 operazioni. Tuttavia per valutare $s(x)$ è necessario preliminarmente scoprire in quale intervallo si trova x e poi valutare il polinomio di grado 3 corrispondente.

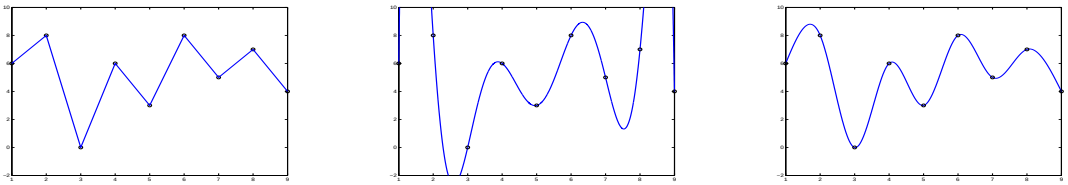


Figura 1: Confronto tra la spezzata (a sinistra), il polinomio di interpolazione (al centro) e la spline cubica interpolante (a destra): dal punto di vista estetico la spline è la soluzione migliore al problema di unire i punti.

Curve di Bézier

Discutiamo ora delle curve di Bézier molto usate in grafica vettoriale. L'idea è quella di unire due punti utilizzando punti ulteriori che determinano un insieme convesso in cui giace la curva.

Dati $n+1$ vettori $v_0, v_1, \ldots, v_n \in \mathbb{R}^N$ che, per aiutare l'intuizione, possono essere immaginati come punti nel piano o nello spazio, la curva di Bézier con punti di controllo $v_0, \ldots, v_n$ è

$$B_{012\ldots n}(t) = \sum_{i=0}^{n} \binom{n}{i} t^i (1-t)^{n-i} v_i, \qquad t \in [0,1].$$

Alcuni casi particolari sono i seguenti

$$\begin{aligned} B_0(t) &= v_0, \\ B_{01}(t) &= (1-t)v_0 + tv_1, \\ B_{012}(t) &= (1-t)^2 v_0 + 2t(1-t)v_1 + t^2 v_2, \\ B_{0123}(t) &= (1-t)^3 v_0 + 3t(1-t)^2 v_1 + 3t^2(1-t)v_2 + t^3 v_3, \end{aligned}$$

con $t \in [0,1]$. La curva di Bézier $B_{012\ldots k}(t)$ è detta di grado k, il cui punto iniziale è v_0 e quello finale v_k, e i punti $v_0, \ldots, v_k$ sono detti punti di controllo. Si osserva che la curva di Bézier di grado 1 è il segmento che unisce i due punti che la definiscono. Inoltre, si osserva che ciascuna coordinata, $B_{012\ldots k}(t)$ rappresenta un polinomio di grado al più k.

Nota 4.9. Nella nostra definizione tutti i punti $v_0, \ldots, v_n$ necessari a costruire una curva sono chiamati punti di controllo. In alcuni testi, invece, i punti di controllo sono solo i punti $v_1, \ldots, v_{n-1}$.

Esempio 4.10. Si consideri la curva di Bézier relativa a $v_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ e $v_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Essa è data da

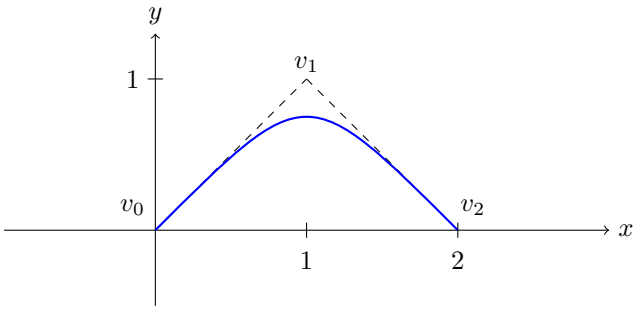
$$B_{01}(t) = (1-t)v_0 + tv_1 = (1-t)\begin{bmatrix} 1 \\ 0 \end{bmatrix} + t\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1-t \\ t \end{bmatrix},$$

ed è il segmento che unisce v_0 e v_1.

Esempio 4.11. Si consideri la curva di Bézier relativa ai punti del piano $v_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $v_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $v_2 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$. Essa è data da

$$\begin{aligned} B_{012}(t) = (1-t)^2 v_0 + 2t(1-t)v_1 + t^2 v_2 &= (1-t)^2\begin{bmatrix} 0 \\ 0 \end{bmatrix} + 2t(1-t)\begin{bmatrix} 1 \\ 1 \end{bmatrix} + t^2\begin{bmatrix} 2 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} 2t(1-t)+2t^2 \\ 2t(1-t) \end{bmatrix} = \begin{bmatrix} 2t \\ 2t(1-t) \end{bmatrix}. \end{aligned}$$

Ponendo $x = 2t$ e $y = 2t(1-t)$ ci si accorge che la curva verifica l'equazione $y = x - \frac{x^2}{2}$ e quindi è un arco di parabola. Da $x = 2t$, si scopre che deve essere $0 \leqslant x \leqslant 2$.



Per le curve di Bézier, vale il seguente risultato.

Teorema 4.12. *Sia $B_{01\ldots n}(t)$ la curva di Bézier relativa ai punti $v_0, \ldots, v_n \in \mathbb{R}^N$ allora:*

1. *la curva parte da v_0 e termina in v_n, cioè $B_{01\ldots n}(0) = v_0$, $B_{01\ldots n}(1) = v_n$;*
2. *la curva è contenuta nell'inviluppo convesso di $v_0, \ldots, v_n$;*
3. *vale la formula di ricorrenza*

$$B_{01\ldots n}(t) = (1-t)B_{01\ldots n-1}(t) + tB_{1\ldots n}(t);$$

4. *il vettore tangente nel punto iniziale è $B'_{01\ldots n}(0) = n(v_1 - v_0)$ e nel punto finale è $B'_{01\ldots n}(1) = n(v_n - v_{n-1})$.*

Dimostrazione. Per dimostrare la prima proprietà basta osservare che nel valutare $B_{01\ldots n}(0)$ tutti i termini che contengono t^i con $i > 0$ sono nulli, mentre nel valutare $B_{01\ldots n}(1)$ sono nulli tutti i termini che contengono $(1-t)^{n-i}$ per $i < n$ e quindi (si noti la convenzione $0^0 = 1$, comune nei coefficienti binomiali)

$$B_{01\ldots n}(0) = \binom{n}{0} 0^0 (1-0)^{n-0} v_0 = v_0, \qquad B_{01\ldots n}(1) = \binom{n}{n} 1^n (1-1)^0 v_n = v_n.$$

Per dimostrare la seconda proprietà è sufficiente osservare che $B_{01\ldots n}(t)$ è una combinazione lineare dei punti $v_0, \ldots, v_n$ i cui coefficienti sono non negativi e sommati danno 1, infatti,

$$\binom{n}{i} t^i (1-t)^{n-i} \geqslant 0, \qquad \sum_{i=0}^{n} \binom{n}{i} t^i (1-t)^{n-i} = (t + (1-t))^n = 1^n = 1,$$

dove nell'ultima parte si è usata la formula del binomio di Newton.

La terza proprietà segue dalle proprietà dei coefficienti binomiali

$$\begin{aligned} B_{01\ldots n}(t) &= \sum_{i=0}^{n} \binom{n}{i} t^i (1-t)^{n-i} v_i = \sum_{i=0}^{n-1} \binom{n-1}{i} t^i (1-t)^{n-i} v_i + \sum_{i=1}^{n} \binom{n-1}{i-1} t^i (1-t)^{n-i} v_i \\ &= \sum_{i=0}^{n-1} \binom{n-1}{i} t^i (1-t)^{n-i-1} (1-t) v_i + \sum_{j=0}^{n-1} \binom{n-1}{j} t^{j+1} (1-t)^{n-j-1} v_{j+1} = (1-t)B_{01\ldots(n-1)}(t) + tB_{12\ldots n}(t). \end{aligned}$$

Si è usato

$$\binom{n}{0} = \binom{n-1}{0}, \qquad \binom{n}{i} = \binom{n-1}{i-1} + \binom{n-1}{i}, \quad i = 1, \dots, n-1, \qquad \binom{n}{n} = \binom{n-1}{n-1}.$$

La quarta proprietà infine, segue da un calcolo diretto,

$$\frac{d}{dt}\binom{n}{0}t^0(1-t)^{n-0} = \frac{d}{dt}(1-t)^n = -n(1-t)^{n-1}, \qquad \frac{d}{dt}\binom{n}{n}t^n(1-t)^{n-n} = \frac{d}{dt}t^n = nt^{n-1},$$

e per ogni $i = 1, 2, \dots, n-1$,

$$\frac{d}{dt}\binom{n}{i}t^i(1-t)^{n-i} = \binom{n}{i}\frac{d}{dt}t^i(1-t)^{n-i} = \binom{n}{i}\big(it^{i-1}(1-t)^{n-i} - (n-i)t^i(1-t)^{n-i-1}\big).$$

da cui

$$B'_{01\dots n}(0) = -nv_0 + nv_1 = n(v_1 - v_0), \qquad B'_{01\dots n}(1) = nv_n - nv_{n-1} = n(v_n - v_{n-1}).$$

□

Dal precedente teorema si ha che ciascun punto della curva di Bézier $B_{012\dots k}(t)$ è combinazione convessa di $v_0, \dots, v_k$, quindi tutta la curva giace nell'inviluppo convesso di $v_0, \dots, v_k$.

Inoltre si osserva che

$$B_{012}(t) = (1-t)B_{01}(t) + tB_{12}(t),$$

e più in generale

$$B_{0123}(t) = (1-t)B_{012}(t) + tB_{123}(t).$$

Questa relazione per ricorrenza permette di ottenere un algoritmo per il calcolo efficiente di $B_{012\dots k}(t)$, detto algoritmo di de Casteljau. L'idea, nel caso $k = 3$ è di seguire lo schema

$$\begin{array}{llll} B_0 = v_0 & & & \\ B_1 = v_1 & B_{01}(t) = (1-t)B_0 + tB_1 & & \\ B_2 = v_2 & B_{12}(t) & B_{012}(t) = (1-t)B_{01} + tB_{12} & \\ B_3 = v_3 & B_{23}(t) & B_{123}(t) & B_{0123}(t) \end{array}$$

ricavando al primo passo $B_{01}(t), B_{12}(t)$ e $B_{23}(t)$, al secondo passo $B_{012}(t)$ e $B_{123}(t)$ e al terzo passo $B_{0123}(t)$.

Algorithm 4.13 (Algoritmo di de Casteljau)**.** Input: $v_0, \dots, v_n, t$.
Output: $B_{012\dots n}(t)$
for $k = 1, 2, \dots, n+1$ $b_{k1} = v_{k-1}$, end
for $h = 2, 3, \dots, n+1$
  for $k = 1, 2, \dots, n+2-h$
    $b_{kh} = (1-t)b_{k,h-1} + tb_{k+1,h-1}$
  end
end
$B_{012\dots k}(t) = b_{1,k+1}$

L'algoritmo per la valutazione di una curva di Bézier, usando un solo vettore B che conterrà, al passo i la colonna i, si può scrivere nel seguente modo. I vettori $v_0, \dots, v_n$ sono le colonne della matrice `v`.

```
1 s=1-t;
2 % si assegnano i punti iniziali
3 B=v;
4 % si scorre la matrice
5 for j=1:N
6   for i=1:N-j+1
7     B(:,i)=s*B(:,i)+t*B(:,i+1);
8   end
9 end
```

Il costo computazionale è dell'ordine di $O(n^2)$.

Riferimenti bibliografici

Per approfondimenti sull'interpolazione polinomiale e tramite spline si faccia riferimento a [1] (le sezioni 1-2 e Teorema 5.5 per l'interpolazione; la sezione 5.14 per le spline).

Per approfondimenti sulle curve di Bézier e il loro uso in computer grafica si faccia riferimento a [2].

Un riferimento in inglese per questo capitolo è il libro [3], sezioni 2.1.1, 2.1.4, 2.4.1.

[1] R. Bevilacqua, D. Bini, M. Capovani, O. Menchi. *Metodi Numerici.* Zanichelli, Bologna, 1992.

[2] G. Farin. *Curves and surfaces for computer-aided geometric design: a practical guide.* Elsevier. 2014.

[3] J. Stoer, R. Bulirsch. *Introduction to Numerical Analysis.* Third Edition. Springer, 2002.

Richiami

Numeri reali e complessi

Indichiamo con $\mathbb{R}^n$ lo spazio dei vettori di n componenti reali, con le operazioni di somma e prodotto per scalare usuale. Indicheremo comunemente i vettori incolonnando le loro coordinate rispetto alla base canonica, ottenendo

$$v = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}.$$

Per esplicitare questo fatto, parleremo di *vettori colonna* che saranno identificati con matrici con n righe e 1 colonna. Applicando l'operatore trasposta a una matrice $n \times 1$ si ottiene una matrice $1 \times n$, cioè una colonna viene trasformata in una riga, avremo quindi la matrice

$$v^T = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix},$$

che verrà detta *vettore riga*.

Nel caso $n = 0, 1, 2, 3$ lo spazio $\mathbb{R}^n$ si può rappresentare con un punto, una retta, un piano e lo spazio tridimensionale, rispettivamente.

L'insieme dei numeri complessi $\mathbb{C}$ è costituito da oggetti del tipo $a + \mathbf{i}b$, dove $a, b \in \mathbb{R}$, sono detti rispettivamente parte reale e parte immaginaria di $a + \mathbf{i}b$ e $\mathbf{i}$ è detta unità immaginaria, che verifica l'identità $i^2 = -1$. Sui numeri complessi sono definite le operazioni

$$(a + \mathbf{i}b) + (c + \mathbf{i}d) = (a + c) + \mathbf{i}(b + d), \qquad \alpha(a + \mathbf{i}b) = \alpha a + \mathbf{i}(\alpha b), \qquad a, b, c, d, \alpha \in \mathbb{R}.$$

In questo modo l'insieme dei numeri complessi è uno spazio vettoriale reale di dimensione 2 e quindi è isomorfo a $\mathbb{R}^2$. È possibile rappresentare i numeri complessi come punti di un piano (detto piano di Gauss).

I numeri complessi sono molto di più di uno spazio vettoriale reale, infatti grazie alla condizione $\mathbf{i}^2 = -1$ è possibile definire un prodotto tra numeri complessi

$$(a + \mathbf{i}b)(c + \mathbf{i}d) = ac + \mathbf{i}ad + \mathbf{i}bc + \mathbf{i}^2bd = (ac - bd) + \mathbf{i}(ad + bc),$$

che lo rende, insieme all'addizione, un campo (come $\mathbb{R}$).

Indichiamo con $\mathbb{C}^n$ lo spazio dei vettori con n componenti complesse, con le operazioni di somma e prodotto per scalare (ricordando che lo scalare questa volta è un numero complesso). Gli elementi di $\mathbb{C}^n$ saranno rappresentati con vettori colonna le cui componenti sono complesse.

Si noti che ogni somma di numeri complessi richiede due somme di numeri reali, mentre ogni prodotto richiede quattro prodotti e due somme di numeri reali.

Nel seguito, nello studio di alcuni algoritmi noi useremo comunemente $\mathbb{K}$ per indicare indifferentemente $\mathbb{R}$ o $\mathbb{C}$.

Strutture algebriche

Una struttura algebrica di base molto usata è il gruppo.

Definizione 4.14. Un insieme G dotato di un operazione $*$ è detto gruppo se l'operazione gode delle seguenti proprietà:

1. esiste $e \in G$ tale che $e * a = a * e = a$ per ogni $a \in G$ (esistenza dell'elemento neutro);
2. per ogni $a, b, c \in G$ vale $(a * b) * c = a * (b * c)$ (proprietà associativa);
3. per ogni $a \in G$ esiste $b \in G$ tale che $a * b = b * a = e$, in tal caso si scrive $b = a^{-1}$ (esistenza dell'inverso).

In un gruppo non è detto che $ab = ba$ per ogni a e b, i gruppi per cui questo vale sono detti abeliani.

Definizione 4.15. Un insieme G dotato di un operazione $*$ è detto gruppo abeliano se verifica le tre proprietà del gruppo e inoltre

4. per ogni $a, b \in G$ si ha $ab = ba$ (proprietà commutativa).

Se l'operazione è la somma, l'elemento neutro si indica con 0, mentre l'inverso dell'elemento a si indica con $-a$ (opposto).

Un insieme può essere dotato di più operazioni.

Definizione 4.16. Un insieme $\mathbb{F}$ dotato di due operazioni $+, \times$ è detto campo se rispetto alla prima operazione è un gruppo abeliano e, detto 0 l'elemento neutro della prima operazione, valgono le seguenti proprietà:

1. esiste $1 \in \mathbb{F}$ tale che $1 \times a = a \times 1 = a$ per ogni $a \in \mathbb{F}$ (esistenza dell'unità);
2. per ogni $a, b, c \in \mathbb{F}$ vale $(a \times b) \times c = a \times (b \times c)$ (proprietà associativa);
3. per ogni $a, b, c \in \mathbb{F}$ vale $a \times (b+c) = (a \times b) + (a \times c)$ e $a + (b \times c) = (a+b) \times (a+c)$ (proprietà distributiva);
4. per ogni $a \in \mathbb{F} \setminus \{0\}$, esiste $b \in G$ tale che $a \times b = b \times a = e$, in tal caso si scrive $b = a^{-1}$ (esistenza dell'inverso degli elementi non nulli);
5. per ogni $a, b \in \mathbb{F}$ vale $a \times b = b \times a$ (proprietà commutativa).

Esempi di campi di uso comune sono, l'insieme dei numeri razionali $\mathbb{Q}$, dei reali $\mathbb{R}$, dei complessi $\mathbb{C}$ con le usuali operazioni di somma e prodotto. Anche $\mathbb{Z}_p$ è un campo se p è primo.

Applicazioni lineari e matrici

Data una matrice quadrata $A \in \mathbb{K}^{n \times n}$, la coppia (λ, v) con $\lambda \in \mathbb{C}$ e $v \in \mathbb{C}^n \setminus \{0\}$ è detta *autocoppia* per A se $Av = \lambda v$. In tal caso v è detto autovettore e λ è detto autovalore.

Il polinomio carattaristico di A è dato da $\det(A - \lambda I)$ ed ha grado n. Gli autovalori sono le soluzioni dell'equazione $\det(A - \lambda I) = 0$ e sono n numeri complessi, se contati nella loro molteplicità (per il teorema fondamentale dell'algebra).

Se la matrice è reale, allora il suo polinomio caratteristico è reale e quindi i suoi autovalori possono essere o reali o coppie di complessi coniugati.

Un'interpretazione geometrica delle autocoppie risiede nel fatto che un autovettore identifica una direzione invariante e cioè una retta che non viene modifica dall'applicazione lineare che definisce la matrice. Nel caso reale, sono solo le autocoppie reali che identificano direzioni invarianti in $\mathbb{R}^n$. L'autovalore, invece, indica l'omotetia che viene realizzata lungo la direzione invariante, ad esempio, se $\lambda > 1$ allora in quella direzione ci sarà una dilatazione, se $\lambda < 0$ i punti verranno spostati in direzioni opposte rispetto all'origine e così via.

Curve

Una curva in $\mathbb{R}^n$ è una funzione continua $\gamma : [a, b] \to \mathbb{R}^n$, dove $[a, b]$ è un intervallo reale. Se $n = 2$ si parla di curva nel piano, se $n = 3$ si parla di curva nello spazio. L'immagine della curva, che è un sottoinsieme di $\mathbb{R}^n$, è detto *supporto della curva* ed è quello che nel senso comune viene identificato con il concetto di curva. (Per evitare situazioni patologiche, si assume che la curva sia C^1 a tratti e cioè che sia possibile dividere l'intervallo $[a, b]$ in un numero finito di intervalli più piccoli in cui la curva è derivabile con continuità.)

Una curva quindi avrà valori in $\mathbb{R}^n$ e potrà essere scomposta in n funzioni $\gamma_i : [a, b] \to \mathbb{R}$ che sono le sue componenti. Ad esempio

$$\gamma : [0, 2\pi] \to \mathbb{R}^2, \ \vartheta \to \begin{bmatrix} \cos\vartheta \\ \sin\vartheta \end{bmatrix}$$

descrive una circonferenza e le sue due componenti sono $\gamma_1(\vartheta) = \cos\vartheta$ e $\gamma_2(\vartheta) = \sin\vartheta$. A volte si scrive $x(t)$, $y(t)$, $z(t)$ anziché $\gamma_1(t)$, $\gamma_2(t)$ e $\gamma_3(t)$ utilizzando la notazione standard per le coordinate cartesiane.

Si dice che la funzione $f : (a, b) \to \mathbb{R}^n$ è derivabile se per ogni $t_0 \in (a, b)$ esiste il limite del rapporto incrementale, cioè

$$f'(t) = \lim_{t \to t_0} \frac{f(t_0) - f(t)}{t - t_0}.$$

Il vettore $f'(t)$ è detto vettore tangente.

Con un piccolo tecnicismo è possibile definire la differenziabilità anche per una curva $\gamma : [a, b] \to \mathbb{R}^n$ che è derivabile in $[a, b]$ se esiste un numero positivo ε e una funzione $f : (a - \varepsilon, b + \varepsilon) \to \mathbb{R}^n$ derivabile e tale che $f(t) = \gamma(t)$ per ogni $t \in [a, b]$. In alternativa si può definire la derivata in a tramite la derivata destra e la derivata in b tramite la derivata sinistra.

Se $\gamma(t) = (x(t), y(t), z(t))$ allora il suo vettore tangente in $t \in (a, b)$ è $(x'(t), y'(t), z'(t))$, dove le derivate delle componenti sono le classiche derivate di funzioni di una variabile.

Inviluppi convessi

Dati due punti $v_0, v_1 \in \mathbb{R}^n$, il segmento che li unisce è il supporto della curva $(1-t)v_0 + tv_1$ per $t \in [0,1]$.

Un insieme $C \in \mathbb{R}^n$ si dice convesso se per ogni coppia di punti $x, y \in C$, il segmento che li unisce è interamente contenuto in C.

Dato un insieme $K \in \mathbb{R}^n$ (per esempio un insieme di punti), il suo inviluppo convesso, indicato con $\mathrm{conv}(K)$ è l'intersezione di tutti gli insiemi convessi che contengono K.

Si può dimostrare che l'inviluppo convesso esiste e se $K = \{v_0, \ldots, v_n\}$ è un insieme di punti, esso può essere descritto analiticamente come

$$\mathrm{conv}(K) = \Big\{ \sum_{i=0}^{n} t_i v_i \,:\, t_i \geqslant 0, \sum_{i=0}^{n} t_i = 1 \Big\}.$$

L'inviluppo convesso è l'insieme delle combinazioni lineari del tipo

$$t_0 v_0 + t_1 v_1 + \cdots + t_n v_n,$$

dove i coefficienti t_i sono tutti non negativi e hanno somma 1. Tali combinazioni sono dette combinazioni convesse.

Si può verificare che, dati due punti $v_0, v_1 \in \mathbb{R}^n$, il loro inviluppo convesso è il segmento che li unisce, infatti se $P = t_0 v_0 + t_1 v_1$, allora siccome $t_0 + t_1 = 1$, si ha $t_0 = 1 - t_1$ e si può porre $t_1 = t$, $t_0 = 1 - t$ e quindi $P = (1-t)v_0 + tv_1$. Si osserva inoltre che $t_1 \geqslant 0$ per definizione e $t_1 \leqslant 1$ perché $t_1 + t_0 = 1$, quindi $t \in [0,1]$, e questo vale anche per $1 - t$.

Alfabeto greco

alfa	α	A
beta	β	B
gamma	γ	Γ
delta	δ	Δ
epsilon	ε	E
zeta	ζ	Z
eta	η	H
theta	ϑ	Θ
iota	ι	I
kappa	κ	K
lambda	λ	Λ
mi	μ	M
ni	ν	N
xi	ξ	Ξ
omicron	o	O
pi	π	Π
rho	ρ, ϱ	P
sigma	σ, ς	Σ
tau	τ	T
ypsilon	υ	Y,Υ
phi	ϕ, φ	Φ
chi	χ	X
psi	ψ	Ψ
omega	ω	Ω

Simboli vari

infinito	∞
aleph	$\aleph$
S gotica	$\mathfrak{S}$
fine della dimostrazione	$\square$